Appunti di Probabilità e Statistica

Riccardo Ricci

Università di Firenze, Facoltà di Scienze M.F.N.
Corso di Laurea in
Scienze Biologiche

Anno Accademico 2003-2004

29 ottobre 2004

Indice

Capitolo 1

Statistica descrittiva

Lo scopo della statistica descrittiva è quello condensare una grande quantità di dati in modo da conciliare al meglio due richieste antitetiche: da un lato la completezza di una descrizione dettagliata, dall'altro la semplicità di una descrizione sintetica.

A questo scopo sono state sviluppate molte tecniche sia per quanto riguarda l'acquisizione dei dati, sia per quanto riguarda la presentazione dei dati stessi. Una corretta acquisizione dei dati è fondamentale per la correttezza delle deduzioni che da tali dati vengono fatte. Il processo di acquisizione dei dati dipende fortemente dal tipo di dati e dal contesto generale dell'indagine. Per uscire dal vago, un conto è fare una statistica sulle preferenze politiche dei diciottenni, un altro è fare una statistica sulle caratteristiche fisiche delle marmotte delle Alpi. E' quindi chiaro che ogni disciplina, ovvero ogni tipo di dati da raccogliere, avrà bisogno di tecniche particolari per l'aquisizione.

Esistono però alcune somiglianze "strutturali" che accomunano i differenti tipi di statistiche.

Si può quindi tentare una classificazione dei dati, indipendentemente dal loro significato "concreto" facendo attenzione a certe caratteristiche, che potremo definire "grammaticali", comuni a vari tipi di dati, che riassumiamo nelle seguenti parole chiave:

1. Popolazione;
2. Individuo,
3. Variabile;
4. Frequenza;
5. Distribuzione.

Cerchiamo di chiarire tramite un esempio questi concetti chiave della statistica. Consideriamo, nello schedario di un ospedale, le schede cliniche dei ricoverati: ogni scheda contiene un certo numero di informazioni sul paziente: nome e cognome, età, data del ricovero, patologia principale, pressione sanguigna e temperatura corporea misurate con una data periodicità (p.e. due misure di ognuna al giorno), varie terapie somministrate, etc.

A partire da queste schede è possibile ricavare delle informazioni, non solo sul singolo paziente, ma anche sull'insieme dei ricoverati. Questo tipo di informazioni non sono destinate a curare un malato particolare ma a fare, per esempio, scelte generali

sulla "politica ospedaliera" (tipo e quantità di medicinali da acquistare, ...) o a studiare una forma di patologia e le strategie di cura (successo di una terapia, tempi di guarigione o di sopravvivenza, etc.)

Vediamo chi sono, nel nostro esempio, popolazione, individuo, etc.

1) Per *popolazione* si intende l'insieme di tutti i possibili oggetti dell'indagine statistica. Lo stesso nome *statistica* deriva da *Stato* e ha origine dai censimenti romani. Per estensione quindi si dà il nome di popolazione a tutto ciò che, in un indagine statistica, ha lo stesso ruolo della popolazione propriamente detta in un censimento. Nel nostro esempio la popolazione è lo schedario dell'ospedale, o per meglio dire la raccolta di tutte le schede cliniche.

2) Un *individuo* è un qualsiasi "oggetto" della popolazione. Nel nostro esempio un "individuo" è una singola scheda (in "rappresentanza" del paziente).

3) Una *variabile* è una qualsiasi caratteristica di ogni individuo della popolazione (p.e. l'età del paziente come risulta dalla scheda) soggetta a possibili variazioni da individuo a individuo. Dal punto di vista matematico una *variabile* è una *funzione* definita sulla popolazione che associa a ogni individuo uno tra tutti i possibili valori della caratteristica in esame. Per chiarire, a ogni scheda clinica posso associare il peso del paziente così come riportato dalla scheda. In questo caso, la *variabile* "peso" associerà a ogni scheda (p.e. "Mario Rossi, n.23423") un numero (p.e. 75 Kg, che sarà il"peso del degente Mario Rossi, n.23423").

Qui è bene introdurre una distinzione fondamentale tra i tipi di variabili che si possano incontrare in statistica. Sempre con riferimento al nostro esempio posso considerare, per esempio, le seguenti variabili:

i) il peso;

ii) il numero di ricoveri precedenti;

iii) gruppo sanguigno;

iv) il titolo di studio del paziente.

Nel caso del *peso* siamo di fronte a quella che viene detta una *variabile numerica continua* cioè una variabile il cui valore è un numero che può assumere un qualsiasi valore in un certo intervallo (questo significa che se p.e Mario Rossi pesa 73.4 Kg e Giovanni Bianchi 73.8 Kg, può esistere un sign. Giuseppe Verdi che pesa 73.6 Kg).

Nel caso **ii)** siamo di fronte a una *variabile numerica discreta*, cioè una variabile che può assumere solo valori che sono numeri naturali (0, 1, 2, 3, ...).

Nei casi **iii)** e **iv)** parleremo di *variabile nominale* o *categoriale* in quanto i valori assunti dalle due variabili sono dei *nomi* (A, B, AB, 0 nel primo caso licenza elementare, licenza media, diploma, laurea nel nel secondo). Anche in questo caso però esiste una differenza.

Nel secondo caso i valori possibili si presentano naturalmente ordinati: non posso essere laureato se non ho una licenza di scuola media. In questo caso si parla di variabili *ordinali*.

Nel caso dei gruppi sanguigni non ci sono ordinamenti "oggettivi" naturali (anche se, a volte, è possibile ordinarli secondo un qualche criterio, p.e. un donatore di gruppo 0 è "più utile" di un donatore del gruppo A)

Nel seguito avremo a che fare soprattutto con variabili numeriche. Ciò non perché esse siano più "interessanti" di quelle nominali, ma in quanto le variabili numeriche permettono (e richiedono) una più ricca analisi dei dati.

Per il momento abbiamo solo introdotto alcune distinzioni tra possibili dati, ma non abbiamo detto ancora niente su come condensare le informazioni.

Il primo e fondamentale metodo è quello che consiste nel *contare gli individui a seconda del valore assunto dalla variabile in esame*. Per esempio nel caso dei "ricoveri precedenti", conterò quanti hanno già subito 0 ricoveri, quanti 1 ricovero, quanti 2 ricoveri, etc.

Supponiamo che dalla popolazione esaminata ottenga, p.e i seguenti dati

Numero di individui nella popolazione (ovvero numero di schede ospedaliere) = 50
Numero di individui con nessun ricovero = 5
Numero di individui con un ricovero = 19
Numero di individui con due ricoveri = 19
Numero di individui con tre ricoveri = 5
Numero di individui con quattro ricoveri = 5
Numero di individui con più di quattro ricoveri = 0

Posso riassumere questi dati in una tabella

Numero di ricoveri	freq. assoluta	freq. %	freq. cumul.	cum. %
0	5	10	5	10
1	19	38	24	48
2	19	38	43	86
3	5	10	48	96
4	2	4	50	100
$\geq$5	0	0	50	100

La tabella è stata costruita associando a ogni possibile valore (o a insiemi di possibili valori come nell'ultima riga, ≥ 5) il numero di individui della popolazione sui quali la variabile in esame assume quel valore. Questi numeri sono detti *frequenze* del valore, distinguendo tra frequenze assolute e frequenze percentuali (ovvero normalizzate a un totale di 100).

Nelle due restanti colonne sono riportate le frequenze *cumulative* ovvero, per ogni valore x della variabile, la somma delle frequenze corrispondenti ai valori minori o uguali a x. Questo ragruppamento ha senso solo se i valori della variabile possono essere ordinati: se un paziente è gìstato ricoverato 2 volte, è necessariamente stato ricoverato *almeno* una volta.

La funzione che a ogni valore di una variabile associa la sua frequenza (in genere normalizzata a 1, ovvero la frequenza divisa per la numerosità della popolazione) prende il nome di *distribuzione* della variabile. La funzione che associa a ogni valore la frequenza cumulativa viene detta *distribuzione cumulativa*.

Nota: si osservi che la distribuzione cumulativa è una funzione sempre non-decrescente e che si può ricostruire la distribuzione a partire dalla distribuzione cumulativa le differenza tra la frequanze cumulativa associata a una categoria e quella associata alla categoria immediatamente precedente.

Un particolare trattamento va riservato alle variabili continue. In questo caso non è possibile raggruppare i dati secondo le frequenze dei possibili valori della variabile, in quanto questi valori sono infiniti. Inoltre, se la variabile è veramente continua, ovvero può essere misurata con infinita precisione, ci aspettiamo che nessuno dei possibili valori sia assunto più di una volta (p.e. nessuno peserà "esattamente" quanto un altra persona).

In questo caso quindi si ricorre a una "discretizzazione" della variabile prima di effettuare il conteggio dei dati. Ovvero si considerano al posto di singoli valori della

Figura 1.1: Istogramma

variabile, degli *intervalli di valori*. Nel caso del peso, per esempio, possiamo contare gli individui della nostra popolazione secondo un criterio del tipo di quello che segue:

Numero di individui che pesano meno di 40 Kg
Numero di individui con peso maggiore uguale a 40 Kg e minore di 50 Kg
Numero di individui con peso maggiore uguale a 50 Kg e minore di 60 Kg
Numero di individui con peso maggiore uguale a 60 Kg e minore di 70 Kg
Numero di individui con peso maggiore uguale a 70 Kg e minore di 80 Kg
Numero di individui con peso maggiore uguale a 80 Kg e minore di 90 Kg
Numero di individui con peso maggiore uguale a 90 Kg

Ovviamente si potevano scegliere anche altri intervalli (p.e. con un'ampiezza di 5 Kg invece che 10 Kg). La scelta "ottimale" dipende in genere dalla numerosità del campione e dalla "dispersione" dei dati, nel senso che maggiore è il numero dei dati tanto maggiori possono essere gli intervalli; tanto più i dati di accumulano presso certi valori, tanto più gli intervalli devono essere piccoli per poter distinguere i dati, etc.

Una volta decisi quali siano gli intervalli si può procedere come nel caso delle variabili discrete.

1.0.1 Medie

Limitiamo ora la nostra attenzione alle variabili numeriche. Il primo, e più noto, indicatore sintetico di una distribuzione è la *media* della variabile.

Un esempio certamente familiare è quello della media dei voti (p.e. nella frase "negli esami fin qui sostenuti Rossi ha la media del 25"). Vediamo di ritrovare le nostre parole chiave in questo esempio: la popolazione è l'insieme degli esami che lo studente Rossi ha attualmente superato; la variabile è il voto V che Rossi ha ricevuto a ogni singolo esame (quindi V è una variabile numerica discreta che può assumere per valori i numeri interi tra 18 e 30, lasciando perdere le lodi). Al momento dell'esame di laurea, una prima e brutale valutazione della "bontà" dello studente Rossi è affidata proprio alla sua "media". Come questa si calcoli è noto a tutti: si sommano i voti e si divide per il numero degli esami sostenuti.

Cominciamo a introdurre un po' di notazioni: , indichiamo con $\mathbf{x}$ la variabile e con x_i (leggi "x i") i valori che la variabile ha assunto su l'individuo i-esimo della popolazione, cioè quell'individuo che in una possibile enumerazione degli individui della popolazione occupa il posto numero i (nota che qui i è a sua volta una variabile:

p.e. nell'esempio dei voti, se ordiniamo in ordine di data gli esami sostenuti x_5 sarà il voto ottenuto nell'esame dato per quinto)

Meglio essere un po' prolissi finché le cose sono semplici: è bene capire subito la distinzione tra la variabile $\mathbf{x}$ e i valori che essa ha assunto sui vari individui. $\mathbf{x}$ è, matematicamente parlando, una funzione ovvero qualcosa che associa a ogni individuo della popolazione un valore tra quelli che possono essere assunti.

x_5 sta a indicare è il valore effettivamente associato all'individuo numero cinque (nel caso dei voti di Rossi, l"'individuo" è il quinto esame, e non Rossi; p.e. potremmo avere $x_5 = 27$): diremo che x_5 è una *realizzazione* della variabile $\mathbf{x}$.

Se la numerosità della popolazione è N ovvero la popolazione è composta da N individui, la media è definita dalla formula

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.1}$$

il cui significato è: si sommano i valori x_i per i che va da 1 fino a N e poi si divide il tutto per N. Una notazione alternativa, forse più semplice ma meno precisa, è data da

$$\frac{1}{N}(x_1 + x_2 + ... + x_N)\,.$$

Questa media è chiamata più precisamente *media aritmetica*, ed è di gran lunga la più usata nelle applicazioni, soprattutto per le sue buone caratteristiche "matematiche".

Essa non è però la sola media possibile. Accanto alla media aritmetica possimo costruire la *media geometrica*

$$G = [x_1 x_2 ... x_N]^{\frac{1}{N}}$$

ottenuta facendo la radice $N-$esima del prodotto degli N valori x_i, e la *media armonica* è definita come *l'inverso della media algebrica degli inversi dei valori* x_i ovvero

$$\frac{1}{H} = \frac{1}{N}\left[\frac{1}{x_1} + \frac{1}{x_2} + ...\frac{1}{x_N}\right]\,.$$

Vediamo un esempio in cui la media aritmetica non dà una "giusta" indicazione della realtà. Supponiamo di avere due automobili che percorrono (in condizioni "normali") rispettivamente 10Km e 20Km con un litro di carburante. Posso quindi dire che la loro "percorrenza" è di 10Km/L e 20Km/L rispettivamente. Sarei tentato quindi di dire che la "percorrenza media" è di 15Km/L, ma ha senso? Supponiamo di dover percorre 100Km con entrambe le vetture. Se mi baso sulla percorrenza media come l'ho definita sono tentato di dire che ho bisogno di 2*(numero di Kilometri)/(percorrenza media)=2*100/15 Litri = 13.333 Litri, risposta ovviamente sbagliata poiché ho bisogno di 10L per la prima auto e 5L per la seconda, ovvero di un totale di 15L. Quindi la sola conoscenza della "percorrenza media" definita dalla media algebrica, mi porta a una conclusione sbagliata. Vediamo ora cosa succede se definiamo la percorrenza media mediante la media armonica. Avremo quindi

$$\left(\frac{\frac{1}{10} + \frac{1}{20}}{2}\right)^{-1} = 13.333 Km/L$$

e quindi per percorrere 100 Km con le due vetture ho bisogno di 2*(numero di Kilometri)/(percorrenza media)=2*100/13.333 Litri = 15 Litri, che la giusta risposta[1].

L'esempio tipico in cui la "giusta" media è la media geometrica è quello delle percentuali. Supponiamo di misurare l'inflazione su scala annua, e che in tre anni successivi si abbiano rispettivamente tassi di inflazione del 2.5%, 2%, 1.5% rispettivamente. E' corretto dire che l'inflazione media su questi tre anni è stata del 3% (la media aritmetica dei tre dati)? La risposta è no. Infatti, se così fosse il prezzo di un bene "medio" (ovvero di un bene ideale il cui prezzo sia cresciuto esattamente come l'inflazione) il cui prezzo iniziale era p sarebbe, dopo tre anni, $p * (1.02) * (1.02) * (1.02) = p * 1.061208$. Ma quello che avviene è che dopo un anno il prezzo è diventato $p_1 = p * (1.025)$; alla fine del secondo anno il prezzo è aumentato del 2%, quindi 'è passato da p_1 (il prezzo all'inizio del secondo anno) a $p_2 = p_1 * (1.02)$. Analogamente alla fine del terzo anno il prezzo $p_3 = p_2 * (1.015)$. Quindi alla fine dei tre anni, il prezzo sale da p a $p * (1.025) * (1.02) * (1.015) = p * 1.0611825$.

Questo risultato si ottiene utilizzando la media geometrica $MG = ((1.025) * (1.02) * (1.015))^{1/3} = 1.01999$ (circa), avendo ovviamente $p * MG * MG * MG = p * 1.0611825$. La differenza è minima ma non trascurabile quando si tratti di grandi cifre come i bilanci statali. Si noti anche che questo "errore" è analogo a quello, più grave, che consiste nel sommare i tassi di inflazione annui, dicendo quindi nel nostro esempio che nel complesso dei tre anni l'inflazione è stata del (2.5+2+1.5)%=6% contro un'inflazione vera del 6.11825%; tuttavia in questo caso si dà una "sottostima" dell'inflazione, mentre la media aritmetica dà sempre una "sovrastima" dell'inflazione vera[2]).

Una generalizzazione della definizione di media aritmetica è quella che si ottiene "pesando" gli individui in modo diverso. Per restare nel nostro esempio, alcuni corsi di laurea gli esami dei primi anni vengono pesati meno di quelli successivi (era il casi di alcuni vecchi corsi di laurea in ingegneria dove, al momento della media i 10 esami del primo biennio venivano considerati, nel fare la media aritmetica con i restanti 22 esami, come un solo esame in cui lo studente avesse preso un voto pari alla media aritmetica dei voti ottenuti nel biennio).

Da un punto di vista matematico questo significa scegliere N numeri maggiori o uguali a zero, f_i, $i = 1, ..., N$, e modificare la definizione di (1.1) nel modo seguente:

$$\overline{x} = \frac{1}{\sum_{i=1}^{N} f_i} \sum_{i=0}^{N} f_i x_i \tag{1.2}$$

detta *media pesata*. Questa volta si sono sommati non i valori x_i ma i prodotti di questi valori per i loro "pesi" f_i, inoltre non si è diviso per il numero degli individui N ma per il "peso totale" dato dalla somma degli f_i; così facendo la definizione di media pesata non varia se tutti i pesi vengono moltiplicati per uno stesso numero, o, in altre parole, è indipendente dall'unità di misura scelta per i pesi (si noti la somiglianza di questa definizione con la definizione di baricentro di N punti pesanti).

La media aritmetica è un caso particolare di media pesata, quando tutti i pesi siano uguali (ovvero f_i non dipenda da i).

[1] Non a caso il consumo di un automobile nelle specifiche tecniche è indicato con la misura *Litri per 100 Km* ovvero di quanti litri ho bisogno per percorrere una distanza di 100 Km; in questo caso il consumo medio è effettivamente dato dalla media aritmetica dei consumi.

[2] Questo è dovuto a un "principio di massimo": il prodotto di N numeri che variano mantenando fissa la loro somma, è massimo quando i numeri sono uguali tra loro

Esercizio: Determinare i pesi f_i nel caso sopra descritto della media finale adottata nel corso di laurea in ingegneria.

La media pesata si usa in particolare per determinare la media aritmetica a partire da una distribuzione assegnata. Supponiamo di conoscere la distribuzione di una variabile , ovvero il numero di individui della popolazione sui quali la variabile assume un determinato valore. Abbiamo quindi l'insieme dei valori possibili, $x_i,\ k = 1, ..., n$, che supponiamo in numero finito, e le frequenze rispettive f_i: definiamo allora la media della variabile (diremo anche la media della distribuzione) tramite la formula

$$\overline{x} = \frac{1}{\sum_{i=0}^{n} f_i} \sum_{i=0}^{n} f_i x_i \tag{1.3}$$

Il risultato coincide con la media aritmetica fatta direttamente su tutta la popolazione. Infatti calcolando direttamente quest'ultima avremo

$$\underbrace{(x_1 + ... + x_1}_{f_1 \text{volte}} + \underbrace{x_2 + ... + x_2}_{f_2 \text{volte}} + ... + \underbrace{x_n + ... + x_n)}_{f_n \text{volte}} / N_P$$

dove $N_P = f_1 + f_2 + ... + f_n$ è il numero di individui della popolazione.

Si noti che si ottene lo stesso valore sia che le f_i siano le frequenze assolute sia che siano le frequenze relative. Questo consente, nel caso siano note le frequenze relative, di calcolare la media anche senza conoscere la numerosità della popolazione.

1.0.2 Moda e Mediana

Altri importanti indicatori che si utilizzano per sintetizzare la distribuzione di una variabile sono la *mediana* e la *moda*.

La moda può essere definita per la distribuzione di una variabile categoriale, ed è data dal *valore con maggiore frequenza*. Nel caso che il profilo della distribuzione presenti due (o più) massimi (di frequenze paragonabili) si parla allora di distribuzione bimodale (o tri-modale, etc.).

La mediana è definita invece per una variabile numerica $\mathbf{x}$ come quel valore M tale che $\mathbf{x} < \mathbf{M}$ per metà degli individui della popolazione (e ovviamente $\mathbf{x} > \mathbf{M}$ per l'altra metà).

Al contrario della media aritmetica, questi due indicatori sono in genere affetti da una certa indeterminatezza nella definizione e non si prestano a elaboarazioni matematiche. Però, in alcuni casi, danno un informazione più significativa della media aritmetica.

Una generalizzazione della mediana è il *percentile*. Si dice n-esimo percentile il valore che lascia alla sua sinistra una percentuale dell'n% degli individui della (analogamente si parla di quantili se invece delle percentuali si usano frazioni normalizzate a uno, i.e. 25-esimo percentile=quantile di ordine 1/4). La specificazione di un ragionevole numero di percentili (p.e. per intervalli del 20%) dà una buona idea della forma della distribuzione.

1.0.3 Varianza

Una caratteristica importante di una distribuzione è, oltre alla sua "tendenza cedntrale" che abbiamo rappresentato tramite la media aritmetica (o in alternativa, tramite la mediana) è la sua più o meno grande "dispersione". Per dispersione si intende lo sparpagliamento dei dati su valori distanti dal valore centrale di riferimento (la media).

Una misura della "dispersione" di una distribuzione deve essere quindi costruita a partire dalle quantità $x_i - \overline{x}$. Ovviamente se sommiamo su i le quantità $x_i - \overline{x}$ otteniamo una quantità nulla poiché le quantità positive sono cancellate da quelle negative, proprio per la definizione di media aritmetica (fare il calcolo!). Potremmo sommare i valori assoluti $|x_i - \overline{x}|$, e così facendo otteremmo un indicatore sensato di dispersione, ma assai poco maneggevole a causa delle sue poche proprietà algebriche.

Un indicatore con i requisiti richiesti e con buone caratteristiche algebriche è dato invece dalla *varianza*. Essa si ottiene sommando i quadrati degli scarti dalla media, ovvero essa è definita dalla formula

$$V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{1.4}$$

ovvero la si ottiene sommando per tutti gli elementi della popolazione i quadrati delle differenze tra il valore della variabile e la sua media aritmetica.

Nel caso si parta dalla distribuzione f_i allora la varianza è data da

$$V = \frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=0}^{n} f_i(x_i - \overline{x})^2 \tag{1.5}$$

La varianza è quindi sempre positiva e si annulla *se e solo se tutti i valori della varabile* $\mathbf{x}$ *coincidono con la sua media*, in altri termini, se la "variabile" è una costante su tutti gli individui della popolazione.

A posto della varianza si usa spesso indicare il valore della sua radice quadrata, $\mathrm{s.d.} = \sqrt{V}$, detta *deviazione standard*.

Il calcolo della varianza è facilitato dal cosiddetto teorema di König che dice che

$$V = \frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=0}^{n} f_i x_i^2 - \overline{x}^2 \tag{1.6}$$

che si può leggere dicendo che la varianza è la media (aritmetica) dei quadrati meno il quadrato della media. Ovviamente partendo da (1.4) otteniamo

$$V = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2 \tag{1.7}$$

La verifica della (1.7) è elementare ma richiede un po' di abitudine nella manipolazione dei simboli di somma: il primo passo consiste nello sviluppare il quadrato nella (1.4), ottenendo

$$V = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\overline{x}\frac{1}{N}\sum_{i=0}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N}\overline{x}^2$$

Le ultime due somme sono rispettivamente la definizione della media aritmetica (moltiplicata per N) e la somma di N termini tutti uguali al quadrato della media, e quindi

$$V = \frac{1}{N}\sum_{i=0}^{N} x_i^2 - 2\overline{x}\,\overline{x} + \overline{x}^2$$

ovvero la (1.7).

Nota: C'è un osservazione importante da fare in merito alla (1.7). Supponiamo di aver già calcolato media e varianza di un insieme di N dati, che indichiamo con $\overline{x}_N$ e V_N. Aggiungiamo ora un nuovo dato x_{N+1}: il calcolo della media del nuovo insieme di dati si fa semplicemente tramite la formula

$$\overline{x}_{N+1} = \frac{N\overline{x}_N + x_{N+1}}{N+1} \,.$$

Anche la varianza può essere ricalcolata immediatamente grazie alla (1.7): abbiamo infatti

$$V_{N+1} = \frac{N(V_N + \overline{x}_N^2) + x_{N+1}^2}{N+1} - \overline{x}_{N+1}^2 \,.$$

Capitolo 2

Calcolo Combinatorio

2.1 I principi del conteggio

Il calcolo combinatorio è l'insieme delle tecniche che permettono di contare efficientemente il numero di possibili scelte, combinazioni allineamenti etc. di oggetti scelti da insiemi con un numero finito di elementi.

I principi base hanno l'aria di banalità, ma presto le cose si fanno piuttosto difficili, quindi è bene prestare attenzione alla progressione delle tecniche che verranno introdotte.

Principio dell'addizione *Se un oggetto può essere scelto in p modi da un insieme A e in q modi da un insieme B, disgiunto da A, allora esso può essere scelto in $p + q$ modi diversi.*

Questo è equivalente a dire che se due insiemi disgiunti hanno *cardinalità finita* (cioè contengono un numero finito di elementi) allora la loro unione contiene un numero di elementi pari alla somma degli elementi dei due insiemi. Questo principio si generalizza nel modo ovvio a un numero finito qualsiasi di insiemi (a due a due disgiunti, e con un numero finito di elementi ciascuno[1]), ed è il principio base di cio che comunemente si intende per *contare*.

Un po' meno ovvio è il seguente

Principio della moltiplicazione *Se A è un insieme di q oggetti e B un insieme di p oggetti, allora l'insieme delle coppie ordinate (a, b) con $a \in A$ e $b \in B$ contiene $p \times q$ elementi*

Cosa abbia a che fare questo principio con la moltiplicazione è spiegato dalla sua formulazione equivalente: *Se si può scegliere in p modi diversi un primo oggetto, e per ognuna di queste scelte, si può scegliere in q modi diversi il secondo oggetto, allora il numero di tutte le scelte possibili di questa coppia di oggetti è $p \times q$*

Anche in questo caso la generalizzazione a un numero arbitrario (finito) di insiemi è immediata ma val la pena enunciarla espicitamente:

Principio della moltiplicazione (seconda formulazione) *Se un oggetto si forma facendo una successione di k scelte tali che ci siano n_1 possibilità per la prima scelta, n_2*

[1] Val la pena di notare che questo principio, come pure il successivo, è *falso* nel caso di insiemi infiniti: p.e. l'unione dei numeri pari e dei numeri dispari ha lo "stesso numero di elementi" sia dei pari che dei dispari; e per quanto è del principio seguente, i razionali sono "tanti quanto" gli interi

per la seconda, ... , n_k possibilità per la k-esima scelta, allora il numero complessivo di oggetti che si possono formare è dato dal prodotto

$$n_1 n_2 ... n_k$$

Questo principio ci consente di calcolare tutte le situazioni di interesse nelle applicazioni. Il problema che ci si pone nella pratica del calcolo è che non sempre è chiaro quale sia la successione di scelte e quale, per ogni scelta, l'insieme da cui scegliere.

Per esemplificare vediamo di fare il conto di quanti elementi contenga l'*insieme delle parti* di un insieme A (da ora in poi non ripeteremo più l'aggettivo finito, ma esso sarà sempre sottinteso). Ricordiamo che l'insieme delle parti è l'insieme i cui elementi sono tutti i sottoinsiemi di A (compreso A medesimo e l'insieme vuoto).

Per fare il nostro conteggio dobbiamo riformulare nei termini del principio della moltiplicazione "come si costruisce un sottoinsieme" di A. Supponiamo di aver ordinato gli elementi di A: ora per ogni elemento nella successione degli elementi di A così ordinati, decidiamo se questo elemento appartiene al sottoinsieme oppure no. Quindi effettuiamo una serie di n scelte, dove $n = |A|$ è la cardinalità di A, e ogni volta possiamo scegliere in due modi, ovvero uno dei due elementi dell'insieme $\{SI,\ NO\}$ (se scegliamo sempre NO otteniamo l'insieme vuoto, sempre SI l'insieme A).

Quindi il totale delle delle nostre possibili scelte, ovvero la cardinalità dell'insieme delle parti di A, è dato dal prodotto di 2 n volte, ovvero 2^n. E' per questo che spesso si indica l'insieme delle parti di A con il simbolo 2^A.

Esercizio: Calcolare il numero di colonne differenti che si possono giocare al totocalcio.

Per ognuna delle tredici partite dobbiamo scegliere un risultato tra i tre possibili $\{1, 2, X\}$, quindi le colonne sono $3^{13} = 1.594.320$.

Esercizio: Calcolare in quanti modi diversi di possono mettere 3 palline distinguibili (p.e. una blu, una bianca e una rossa) in tre scatole distinguibili (p.e. U_1, U_2 e U_3).

L'insistenza sull'aggettivo "distinguibili" significa che consideriamo diverse p.e. il caso in cui la pallina blu è in U_1, quella bianca in U_2 e quella rossa in U_3 dal caso in cui la pallina bianca è in U_1, quella blu in U_2 e quella rossa in U_3.

In questo caso il conteggio è identico al caso delle colonne del totocalcio anche se forse è meno intuitivo. Infatti il ruolo delle partite è ora tenuto dalle palline mentre il ruole dei risultati $\{1, 2, X\}$ dalle urne. Il conteggio avviene in questo modo: per ogni pallina dobbiamo scegliere in quale urna vogliamo metterela. Quindi otteniamo $3^3 = 27$ possibili casi. I numeri scelti sono volutamente ingannevoli, in quanto abbiamo lo stesso numero di palline e di scatole: ma in quanti modi possibili si possono mettere k palline in n scatole?

Definizione: Quando da una scelta all'altra non cambia l'insieme delle possibili scelte ed è quindi possibile ripetere anche una scelta già fatta si dice che abbiamo a che fare con *disposizioni con ripetizione.*

Esempi di applicazione dello stesso principio della moltiplicazione, ma con numero delle possibili scelte che varia da scelta a scelta è dato dai seguenti esempi.

Esercizio: In un ristorante c'è un menu a prezzo fisso composto da antipasto, primo, secondo, dolce. Il menu propone al cliente la scelta tra 2 antipasti, 3 primi, 2 secondi e 4 dolci. Quanti pranzi diversi si possono scelgliere con questo menu?

Esercizio: Quanti sono i numeri dispari di quattro cifre? In questo caso si deve far attenzione perché le cifre della decina e della centinaia possono essere numeri qualsiasi

tra 0 e 9, quindi 10 possibilità, mentre per le migliaia non si può scegliere lo 0 e per le unità la scelta è ristretta ai numeri dispari $1, 3, 5, 7, 9$, si ha quindi $9\times10\times10\times5 = 4500$ numeri dispari.

E quanti sono i numeri dispari con quattro cifre diverse tra loro? (attenzione qui il problema è complicato dal fatto che la scelta di una cifra condiziona la scelta delle altre.)

Definizione: Per *disposizioni senza ripetizione* di k tra n oggetti, si intende i possibili esiti del processo di scelta di k oggetti in un insieme di n, $n \geq k$, senza poter scegliere più di una volta uno stesso elemento.

La distinzione tra queste due forme di disposizione diventa più chiara se la rifrasiamo in termini di estrazioni.

Se estraiamo un numero da un urna, come nel gioco della tombola, e dopo ogni estrazione il numero estratto viene rimesso nell'urna, il numero delle cinquine si calcola come nel caso della schedina del totocalcio, e si hanno 90^5 risultati possibili (anche qui si tiene conto dell'ordine in cui i numeri sono estratti, per cui la cinquina $\{3, 34, 21, 18, 76\}$ deve considerarsi diversa, p.e., dalla cinquina $\{34, 3, 21, 18, 76\}$) .

Se invece effettuiamo una serie di estrazioni senza reinserire i numeri estratti, il numero estratto alla k-esima estrazione non può ovviamente essere uno di quelli estratti nelle estrazioni precedenti. Se quindi vogliamo contare quante sono le possibili cinquine (ordinate!) che si possono ottenere su una ruota nell'estrazioni del lotto, dobbiamo tener conto che il primo numero può essere estratto tra 90 numeri diversi, il secondo tra 89 e così via. Abbiamo quindi $90 \times 89 \times 88 \times 87 \times 86 = 5.273.912.160$ possibili cinquine ordinate.

In generale una successione di k estrazioni da un insieme di n oggetti senza reinserimento abbiamo $n \times (n-1) \times .... \times (n-k+1)$ esiti possibili.

2.1.1 Permutazioni e combinazioni

Un modo molto efficiente per effettuare questo tipo di conteggi è quello di che utilizza il concetto di *permutazione*.

Consideriamo l'insieme I_n dei numeri interi compresi tra 1 e n.

Definizione *Una permutazione è una qualsiasi applicazione invertibile di I_n in sé.*

In modo meno astratto, una permutazione è un qualsiasi ordinamento di n oggetti, in rapporto a un ordinamento "base" scelto arbitrariamente.

Il conteggio di tutte le possibili permutazioni è analogo a quello delle delle disposizioni senza ripetizione, di cui la permutazione è il caso particolare di disposizione senza ripetizione di "n tra n oggetti".

Indichiamo con α la permuazione: abbiamo n possibili valori per $\alpha(1)$ (in altre parole n possibili scelte del nuovo primo elemento dell'ordinamento), poi $n-1$ per $\alpha(2)$ in quanto $\alpha(2) \neq \alpha(1)$, e così via fino ad arrivare ad $\alpha(n)$ che è determinato dalle scelte precedenti (quindi una sola possibile scelta). Applicando il principio della moltiplicazione abbiamo quindi $n \times (n-1) \times .... \times 2 \times 1$ possibili permutazioni. Il numero $n \times (n-1) \times .... \times 2 \times 1$ si indica con il simbolo $n!$ (leggi n fattoriale).

E' facile rendersi conto che tutte le permutazioni sono in corrispondenza biunivoca con le matrici $n \times n$ i cui elementi sono o 1 o 0 e in cui le somme per righe e per colonne sono sempre uguale a 1 (il che implica che c'è un solo 1 in ogni riga e in ogni colonna). Basta infatti fare il prodotto righe per colonne di una tale matrice con il

vettore colonna $(1, 2, 3, 4, ..., n)^T$ e constare che il risultato è una permutazione degli elementi del vettore.

Esercizio: Dimostrare che l'insieme delle permutazioni forma un gruppo rispetto alla composizione di applicazioni (o il prodotto di matrici).

Esercizio: Quale matrice corrisponde alla permutazione generata dalla "traslazione" $\alpha(i) = i + 1 (\mathrm{mod}\ n)$?

Se scegliamo un numero r minore di n definiamo $r-$permutazione una qualsiasi disposizione ordinata di r degli n oggetti. Il numero delle $r-$permutazioni è dato da

$$P(n, r) = n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}$$

dato che il primo elmento si può scegliere in n modi distinti, il secondo in $n-1$ modi e così via, calando di una unità il numero delle possibili scelte fino ad arrivare all'ultimo, r-esimo, elemento che si può quindi scegliere in $n - r + 1$ modi distinti.

Il fattoriale è una funzione rapidamente crescente di n, p.e. 1!=1, 2!=2, 3!=6, 4!=24, e già 10!=3628800. Per valori elevati di n, il fattoriale è approssimato dalla famosa *formula di Stirling*[2]

$$n! \sim \sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n}$$

che fa bizzarramente apparire, accanto al prodotto di numeri naturali, i più famosi numeri irrazionali $\sqrt{2}$, π ed e.

Una volta contate le permutazioni possiamo introdurre le *combinazioni*.

Definizione: si dice *combinazione* di k oggetti scelti tra n un qualsiasi sottoinsieme di cardinalità k di oggetti di un insieme di cardinalità n.

La novità rispetto alle disposizioni consiste nel fatto che ora non si distinguono scelte che differiscono solo nell'ordine in cui viene fatta la scelta, in altre parole due sottoinsiemi differiscono solo se uno contiene almeno un elemento che non è contenuto nell'altro. Il numero delle possibili combinazioni quindi conta quanti sottoinsiemi distinti di k elementi si possono fare da un insieme contenente n elementi.

Per completezza si considerano anche combinazioni di 0 (zero) oggetti su n e di n oggetti su n. In questo caso i sottoinsiemi ottenuti sono rispettivamente l'insieme vuoto e l'intero insieme. Quindi c'è una sola combinazione di 0 oggetti su n e una sola di n oggetti su n (qualsiasi sia n).

Per contare quante siano le combinazioni di k oggetti su n basta osservare che un sottoinsieme si può formare nel modo seguente: ordiniamo in un modo qualsiasi gli n oggetti e prendiamone i primi k. L'ordinamento si può fare, come sappiamo, in $n!$ modi diversi. Tuttavia tutte le permutazioni che scambiano tra loro solo i primi k elementi o solo gli ultimi $n - k$ presentano gli stessi elementi tra i primi k. Quindi delle $n!$ permutazioni possibili, solo

$$C(n, k) = \frac{n!}{k!(n-k)!}$$

sono tali che i sottoinsiemi formati dai primi k elemnti differiscono tra loro.

[2]Il simbolo $\sim$ utilizzato nella formula significa che i dui termini sono *asintotici*, ovvero il limite per $n \to \infty$ del loro rapporto fa 1.

Il numero $C(n,k)$ si indica anche con il simbolo

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)...(n-k+1)}{k!} = \frac{n(n-1)(n-2)...(k+1)}{(n-k)!}$$

e prende il nome di *coefficiente binomiale*.

Esercizio: Vericare che

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right)$$

e l'identità

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n-1 \\ k \end{array} \right) + \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right)$$

I coefficienti binomiali si possono calcolare tramite il triangolo di Pascal (o di Tartaglia)

Triangolo di Pascal

k	0	1	2	3	4	5	6	7	8	9
n										
0	1									
1	1	1								
2	1	2	1							
3	1	3	3	1						
4	1	4	6	4	1					
5	1	5	10	10	5	1				
6	1	6	15	20	15	6	1			
7	1	7	21	35	35	21	7	1		
8	1	8	28	56	70	56	28	8	1	
9	1	9	36	84	136	136	84	36	9	1

che è costruito tramite la regola

$$\begin{array}{cccc} ... & ... & ... & ... \\ ... & c_{n,k} & c_{n,k+1} & ... \\ ... & ... & c_{n+1,k+1} & ... \\ ... & ... & ... & ... \end{array}$$

con

$$c_{n+1,k+1} = c_{n,k} + c_{n,k+1}$$

Il nome "coefficiente binomiale" proviene dal fatto che essi forniscono i coefficienti dello sviluppo della potenza $n-$esima di un binomio secondo la *formula di Newton*

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}$$

Come immediata conseguenza della formula di Newton abbiamo

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) = 2^n$$

che possiamo reinterpretare dicendo che *il numero di tutti i sottoinsiemi di un insieme con N elementi è la somma su k del numero dei suoi sottoinsiemi di numerosità k.*

Capitolo 3

Probabilità

3.1 Introduzione

Il calcolo delle probabilità è uno strumento essenziale per la statistica. Esso dà una risposta a quello che possiamo considerare come il problema inverso di quello della statistica inferenziale. Mentre la statistica cerca di determinare tramite la conoscenza dei risultati di un esperimento (o più esperimenti) quali siano le caratteristiche della popolazione su cui l'esperimento è stato eseguito, nel calcolo delle probabilità si assume che tutte le caratteristiche della popolazione siano note (senza preoccuparsi del come ciò sia possibile) e si vuole calcolare *a priori* la "probabilità" che un esperimento abbia un determinato risultato.

Come per tutti i concetti fondamentali è impossibile definire cosa si intenda per probabilità senza far ricorso a un idea intuitiva del concetto stesso. Cercheremo qui di illustrare con alcuni esempi cosa si intende per probabilità e, soprattuttto, estrarre da questi esempi le "regole del gioco" (una volta chiarite queste regole potremo enunciare la definizione assiomatica di probabilità che sarà utile per lo sviluppo del calcolo).

Il calcolo della probabilità trae le sue origini dal gioco dei dadi e quindi possiamo iniziare dal problema del "lancio di un dado". Ho usato le virgolette perché la prima cosa da fare è definire bene, quando si abbia a che fare con un problema di probabilità, il contesto. Dando per noto cosa sia un dado (con facce numerate da 1 a 6), per "lancio di un dado" si intende che il dado venga lanciato in aria con sufficiente impulso e rotazione, si attenda che cada a terra su una superficie orizzontale, e che si fermi con una sua faccia adagiata al suolo. Il risultato del lancio sarà il numero che si legge sulla faccia opposta a quella al suolo. Perché una descrizione tanto prolissa di una cosa che tutti sanno? Perché prima di procedere a calcolare delle probabilità è necessario chiarire alcune cose:

- l'*esperimento* deve essere *casuale*, o *aleatorio*, nel senso che non si possa prevederne con certezza il risultato in anticipo (chi sarebbe disposto a scommettere su un "lancio" del dado che avvenga prendendo un dado, ponendolo a 3 mm dal suolo con la faccia numero 6 rivolta verso l'alto e facendolo cadere da fermo?);

- deve essere chiaro quale sia lo *spazio campionario* $\mathcal{S}$ soggiacente, ovvero l'insieme di tutti i possibili esiti dell'esperimento (nel nostro caso abbiamo $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$). Chiameremo *evento elementare* ogni singolo elemento di $\mathcal{S}$.

Chiameremo poi *evento* ogni sottoinsieme E dello spazio campionario. Diremo che un evento si è verificato, o realizzato, quando il risultato dell'esperimento (che è un evento elementare) è un elemento di E. Per esempio, nel lancio del dado ho l'evento $E = \{\text{il numero uscito è pari}\}$, ovvero $E = \{2, 4, 6\}$, che si verifica se il risultato del lancio è 2 oppure 4 oppure 6.

Gli eventi, in quanto sottoinsiemi, ereditano tutte le operazioni definite dalla teoria elementare degli insiemi. Avremo quindi, dati due eventi E_1 e E_2, l'evento *unione* $E_1 \cup E_2$, che è formato da tutti gli eventi elementari che appartengono ad almeno uno dei due sottoinsiemi, e l'evento *intersezione* $E_1 \cap E_2$, formato dagli eventi che appartengono a entrambi i sottoinsiemi.

Diremo che due eventi E_1 e E_2 sono *mutuamente esclusivi*, o *incompatibili*, se $E_1 \cap E_2 = \emptyset$, ovvero se non hanno eventi elementari in comune (nota quindi che il realizzarsi di E_1 esclude che si verifichi, contemporaneamente, E_2 e viceversa, in particolare due eventi elementari (distinti) sono sempre incompatibili). Infine, dato un evento E, chiameremo evento *complementare*, che indicheremo con $\overline{E}$, l'insieme degli elementi di $\mathcal{S}$ che non appartengono a E. Ovviamente $E \cap \overline{E} = \emptyset$ e $E \cup \overline{E} = \mathcal{S}$.

Possiamo ora dare una definizione formale (assiomatica) di che cosa si intende per probabilità matematica. Dato uno spazio campionario $\mathcal{S}$, sia P una funzione definita sugli eventi di $\mathcal{S}$ a valori reali, ovvero una legge che a ogni evento E associa un numero $P(E)$, con le seguenti proprietà:

(i) $0 \leq P(E) \leq 1$

(ii) $P(\mathcal{S}) = 1$

(iii) per ogni coppia di eventi E_1 e E_2 incompatibili, si ha

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Il numero $P(E)$ si dice la *probabilità* dell'evento E.

Commentiamo un po' questa definizione interpretando la probabilità come il grado di fiducia che noi assegnamo al verificarsi o meno di un certo evento.

La proprietà (i) dice semplicemente che la probabilità è un numero non negativo che va da zero (nessuna fiducia sul verificarsi dell'evento) a 1 (completa fiducia che l'evento si realizzi). Nel linguaggio comune è più usuale esprimersi in termini di percentuali per cui il nostro valore 1 diviene il 100%.

La seconda proprietà ci dice che l'evento formato da tutti i possibili eventi elementari ha probabilità 1. Ma $\mathcal{S}$ deve per forza verificarsi poiché è l'insieme di *tutti i possibili risultati*, ovvero è un evento *certo* (e il solo evento certo). In altri termini la (ii) ci dice che se siamo sicuri che un evento si realizzi la sua probabilità deve essere 1 (come vedremo più avanti, il viceversa non è necessariamente vero se $\mathcal{S}$ ha infiniti elementi).

Infine la (iii) ci dice che se E_1 ed E_2 non hanno eventi elementari in comune, allora la probabilità che si verifichi almeno uno dei due eventi è la somma delle singole probabilità.

Nell'esempio del lancio del dado, se p.e. $E_1 = \{1, 2\}$ e $E_2 = \{3, 4\}$, allora la probabilità che si verifichi $E_1 \cup E_2 = \{1, 2, 3, 4\}$ è la somma delle probabilità $P(E_1)$ e $P(E_2)$.

Nel caso che $\mathcal{S}$ sia formato da infiniti elementi, questa proprietà è sostituita da

(iii') per successione di eventi $E_1, E_2, ...$ a due a due incompatibili, cioè tali che $E_i \cap E_j = \emptyset$ se $i \neq j$ si ha

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) \ ,$$

Queste proprietà, assieme con le operazioni di unione e intersezione permettono di definire P per tutti gli eventi E. Per esempio, se $p = P(E)$ è la probabilità di un evento E, allora la (ii) e la (iii) ci dicono che $P(\overline{E}) = 1 - p$. A partire da queste regole è ora possibile sviluppare tutto il meccanismo del *Calcolo delle Probablità*.

Esse tuttavia non ci dicono quanto deve valere $P(E)$ in un determinato caso concreto. Abbiamo, per il momento, fissato soltanto le proprietà formali della probabilità, ma non abbiamo detto come assegnare i valori di P ai singoli eventi. Questo problema esula dal calcolo delle probabilità nella sua accezione puramente matematica e ha piuttosto a che fare con la "filosofia" della probabilità.

Vediamo come possiamo comportarci nel nostro esempio del dado. Se non abbiamo alcun sospetto che il dado sia "truccato", cioè se pensiamo che non ci siano ragioni che ci facciano ritenere che un numero abbia più possibilità di uscire di un altro, allora ci possiamo accordare per assegnare a ogni numero una uguale probabilità, ovvero $P(e) = 1/6$ dove e indica un qualunque evento elementare di $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$. In generale se $\mathcal{S}$ è uno spazio campionario formato da n elementi che riteniamo *equiprobabili*, allora la probabilità di un singolo evento elementare sarà data da $P(e) = 1/n$. Sulla base di questa definizione di P è possibile verificare che ogni evento E ha una probabilità $P(E)$ che possiamo calcolare tramite il rapporto dei *casi favorevoli* su i *casi possibili*, ovvero il numero di eventi elementari contenuti in E diviso n (p.e. la probabilità di ottenere un numero pari nel lancio di un dado non truccato è $3/6 = 1/2$). Questa definizione di probabilità fu introdotta da B. Pascal e P. de Fermat alla metà del secolo XVII.

In questo approccio il problema di determinare la probabilità di un singolo evento si riduce al conteggio dei casi favorevoli (e dei casi possibili), cioè diventa un problema di *Calcolo Combinatorio* e a questo rimandiamo per le tecniche specifiche.

Questa definizione classica tuttavia non dice niente in due casi fondamentali.

Primo, cosa fare se abbiamo il sospetto che il dado sia truccato, o più in generale se sospettiamo che gli n eventi elementari di $\mathcal{S}$ non siano tutti equiprobabili?.

Secondo, questo approccio è banalmente inutilizzabile nel caso che $\mathcal{S}$ sia formato da un numero infinito di eventi elementari.

Seguendo un altro approccio della probabilità, detto *frequentista*, possiamo ragionare come segue (almeno nel caso di $\mathcal{S}$ finito). Supponiamo di fare un numero molto grande (ma, ovviamente, finito) di esperimenti, diciamo N. Contiamo quante volte un singolo evento elementare e_i compare in questa successione di prove, N_i e poniamo infine $P(e_i) = N_i/N$ (per esercizio verificare che questa definizione è coerente con gli assiomi (i)-(iii)).

Questa "definizione" di probabilità si presta a diverse critiche (p.e. anch'essa non può darci alcuna risposta coerente per il caso di $\mathcal{S}$ infinito) e in alternativa si è sviluppato il cosidetto approccio *soggettivista* o *bayesiano*, che grosso modo dice che l'assegnazione della probabilità è comunque frutto di una scelta personale e l'unico strumento di cui disponiamo è un meccanismo per rivedere a posteriori la scelta iniziale alla luce dei risultati degli esperimenti (questo strumento è il teorema di Bayes che vedremo in seguito). Le discussioni su questi due approcci sono tuttora accanite e esulano non solo dagli scopi di questo note, ma direi dallo stesso uso della probablilità nella corrente pratica delle scienze applicate.

3.2 Relazioni elementari

Ritorniamo quindi al calcolo delle probabilità supponendo di aver fissato lo spazio campionario $\mathcal{S}$ e la funzione P.

Abbiamo già detto che $P(\overline{E}) = 1 - P(E)$ come possiamo verificare con il seguente calcolo (ricordando che E e $\overline{E}$ sono eventi incompatibili)

$$1 = P(\mathcal{S}) = P(E \cup \overline{E}) = P(E) + P(\overline{E}) \Longrightarrow P(\overline{E}) = 1 - P(E)$$

Se E_2 è un evento contenuto nell'evento E_1, in simboli $E_2 \subset E_1$, cioè se tutti gli eventi elementari di E_2 appartengono anche a E_1, allora si ha (proprietà di monotonia)

$$P(E_2) \leq P(E_1)$$

che si deduce facilmente osservando che $E_1 = E_2 \cup (E_1 - E_2)$ ($A - B = A \cap \overline{B}$ indica l'insieme degli elementi di A che non appartengono a B) e che, poiché E_2 e $E_1 - E_2$ hanno intersezione vuota,

$$P(E_1) = P(E_2) + P(E_1 - E_2).$$

Di conseguenza, per l'unione di due insiemi qualsiasi vale

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2).$$

Questa diseguaglianza diventa "ovvia" se si osserva che nell'evento $E_1 \cup E_2$ gli, eventuali, eventi elementari che appartengono all'intersezione $E_1 \cap E_2$ vengono contati una sola volta, mentre nella somma $P(E_1) + P(E_2)$ essi vengono contati due volte essendo somma delle singole probabilità di tutti gli eventi elementari in E_1 più le probabilità di tutti gli eventi elementari in E_2.

Si può anche essere più precisi osservando che

$$\begin{aligned} E_1 &= (E_1 \cap E_2) \cup (E_1 - E_2) \\ E_2 &= (E_1 \cap E_2) \cup (E_2 - E_1) \\ E_1 \cup E_2 &= (E_1 - E_2) \cup (E_2 - E_1) \cup (E_1 \cap E_2) \end{aligned}$$

e che gli eventi a destra del segno di uguale sono eventi incompatibili e quindi possiamo sommarne le probabilità

$$\begin{aligned} P(E_1) &= P(E_1 \cap E_2) + P(E_1 - E_2) \\ P(E_2) &= P(E_1 \cap E_2) + P(E_2 - E_1) \\ P(E_1 \cup E_2) &= P(E_1 - E_2) + P(E_2 - E_1) + P(E_1 \cap E_2) \end{aligned}$$

sommando infine le prime due uguaglianze e sottraendo la terza otteniamo la formula

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2), \tag{3.1}$$

in accordo con quanto osservato prima sul doppio conteggio. Questa formula può essere generalizzata al caso di n eventi $E_1, E_2, ..., E_n$

$$\begin{aligned} P(E_1 \cup ..... \cup E_n) = \sum_i P(E_i) - \sum_{i \neq j} P(E_i \cap E_j)) \\ + \sum_{i \neq j \neq k} P(E_i \cap E_j \cap E_k) - ..... + (-1)^{n-1} P(E_1 \cap .... \cap E_n). \end{aligned} \tag{3.2}$$

Questa formula si "legge" così : prima si sommano tutte le probabità degli insiemi $E_1, ..., E_n$. In questo modo, come nel caso di due soli insiemi, abbiamo contato due volte gli eventi elementari che appartengano all'intersezione di due eventi diversi; dobbiamo quindi togliere queste probabilità. Così facendo però abbiamo tolto due volte (almeno) la probabilità di un evento elemtare che appartiene all'intersezione di tre eventi diversi e quindi dobbiamo sommare le probabilità delle intersezione a tre a tre di eventi diversi. E così via. In definitiva bisogna sommare le probabilità di tutte le intersezioni di un numero dispari di eventi diversi (anche le "intersezioni" di un solo evento) e togliere quelle di tutte le intersezioni id un numero pari di eventi diversi.

3.3 Probabilità condizionata

Supponiamo di essere interessati al realizzarsi di un evento A. Qualcuno compie un esperimento e ci dice che si è realizzato l'evento B. Cosa possiamo dire ora sul fatto che A si sia realizzato, sapendo che B si è realizzato?

Così enunciata, al questione può sembrare piuttosto cervellotica. Si tratta tuttavia di un caso comunissimo nelle applicazioni del "ragionamento probabilistico". Si pensi a quanto avviene in ambito giudiziario: si vuol sapere se "Caino ha ucciso Abele" (evento A); nelle nostre indagini scopriamo che "sotto le unghie di Abele ci sono capelli biondi" (evento B, supponiamo che Caino sia biondo e Abele bruno). Come cambia il nostro pregiudizio sulla innocenza (o colpevolezza) di Caino alla luce dell'evento B ?

Ovviamente se A e B sono eventi incompatibili, A non può essersi realizzato in contemporanea con B, quindi la probabilità che assegneremo al realizzarsi di A *condizionatamente* alla realizzazione di B sarà 0. Se invece $A = B$ (o, più in generale, $B \subset A$) siamo certi che A si è realizzato, quindi la sua probabilità *condizionata* a B sarà 1.

Attenzione a non commettere un errore grossolano: il realizzarsi di B implica che A si è realizzato solo se *tutti gli elementi di B sono contenuti in A* e non il viceversa. (Questo è un punto importante e non si deve proseguire se non è chiaro! quando diciamo che "si è realizzato l'evento B" intendiemo che il risultato dell'esperimento è *un evento elementare e contenuto in B*: se B è un sottoinsieme di A allora e appartiene anche ad A e quindi "A si è realizzato.

Viceversa se $A \subset B$ allora B può realizzarsi anche senza che si realizzi A: basta che l'evento elementare e sia un elemento di B che non appartiene ad A.)

Quello che segue può essere omesso in un prime lettura e si può passare direttamente alla definione di probabilità condizionata.

Formalizziamo quanto abbiamo detto finora: vogliamo definire una nuova funzione di probabilità, che indicheremo con $P(A|B)$ e chiameremo *probabilità condizionata dell'evento A rispetto all'evento B* (quando non possa insorgere confusione diremo semplicemente probabilità condizionata di A). Ovviamente $P(A|B)$ deve soddisfare agli assiomi **(i)**, **(ii)**, **(iii)** di una funzione di probabilità.

Inoltre abbiamo visto che deve valere:

$$P(A|B) = 0, \text{ se } A \cap B = \emptyset,$$

e anche

$$P(A|B) = 1, \ se \ B \subset A.$$

Ma cosa succede se A e B non sono incompatibili e A non contiene B?

Osserviamo che basta decidere cosa succede per i sottoinsiemi di B. Infatti, dato un evento qualsiasi A, lo possiamo scomporre nei due eventi incompatibili $A_1 = A \cap B$ e $A_2 = A \cap \overline{B}$.

Poiché A_2 è incompatibile con B, $P(A_2|B) = 0$, e avremo

$$P(A|B) = P(A_1|B) + P(A_2|B) = P(A_1|B).$$

Supponiamo quindi che C e D siano entrambi sottoinsiemi di B. Vogliamo legare le "nuove" probabilità $P(C|B)$ e $P(D|B)$ alle "vecchie" probabilità $P(C)$ e $P(D)$ (dette in questo caso probabilità a priori) che gli eventi hanno *prima di sapere che si è verificato l'evento* B. Questo legame non discende dagli assiomi e dalle richieste finora fatte sulla probabilità condizionata e quindi è frutto di una scelta "arbitraria", che deve solo essere coerente (cioè deve soddisfare alle richieste degli assiomi di una funzione di probabilità). D'altra parte questa scelta dovrà essere legittimata dalla sua capacità di "funzionare" nelle applicazioni.

C'è comunque un argomento che ci guida nella scelta: se sappiamo solo che B si è realizzato, non abbiamo nessuna indicazione su quale evento elementare di B si sia realizzato (a meno che B non sia fatto di un solo elemento). Quindi, se C e D sono entrambi sottoinsiemi di B, non abbiamo nessun motivo per "preferire" C a D rispetto alla nostra valutazione a priori. Questo equivale a dire che il rapporto tra le nuove probabilità (quelle condizionate a B) e quelle a priori non è cambiato, ovvero

$$\frac{P(C|B)}{P(D|B)} = \frac{P(C)}{P(D)}.$$

Possiamo riscrivere questa relazione come

$$\frac{P(C|B)}{P(C)} = \frac{P(D|B)}{P(D)},$$

ovvero, per ogni sottoinsieme E di B vogliamo che sia costante il rapporto $P(E|B)/P(E)$. In particolare, poiché B è un sottoinsieme di B stesso, dobbiamo avere

$$\frac{P(E|B)}{P(E)} = \frac{P(B|B)}{P(B)}, \forall E \subset B.$$

Possiamo ora concludere ricordando che $P(B|B) = 1$, da cui otteniamo $P(E|B)/P(E) = 1/P(B)$, per ogni $E \subset B$.

Definiamo quindi la probabilità condizionata in accordo con quanto detto.

Definizione *Sia B tale che $P(B) > 0$, si dice probabilità condizionata di un evento A rispetto all'evento B il numero*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.3}$$

(Ovviamente le considerazioni fatte sopra perdono di senso qualora $P(B) = 0$. Questo va d'accordo con l'intuizione: se si verifica un evento che aveva probabilità nulla, allora "può succedere di tutto".)

Resta da verificare che questa definizione soddisfa effettivamente agli assiomi di probabilità: la verifica è lasciata al lettore.

Per esemplificare questa definizione torniamo al caso del dado. Scommettiamo che il risultato del lancio sia un numero pari, evento $A = \{2, 4, 6\}$. Nell'ipotesi di dado non truccato, quindi la probabilià di vincere la scommessa è di $1/2$ (o 50%). Ora qualcuno lancia il dado (senza che noi lo possiamo vedere) e ci dice che il numero uscito è maggiore o uguale a 4, evento $B = \{4, 5, 6\}$. Cosa posso dire sulla probabilità di aver vinto la scommessa?

Ora il realizzarsi dell'evento A è più probabile che a priori.

Infatti nell'insieme B ci sono due numeri pari su tre, contro i tre su sei dello spazio campionario originale. In accordo la probabilità dell'evento "è uscito un numero pari"

deve passare da $1/2$ a $2/3$. D'altra parte, l'intersezione tra A e B è formata da due numeri e quindi ha una probabilità a priori di $2/6 = 1/3$. Applicando la formula per la probabilità condizionata otteniamo $P(A|B) = \dfrac{1/3}{1/2} = 2/3$.

Nota: *Nel caso di uno spazio campionario finito e di eventi tutti equiprobabili, è facile verificare che la (3.3) discende dalla regola* $\dfrac{\text{casi favorevoli}}{\text{casi possibili}}$ *applicata agli eventi contenuti nel "nuovo" spazio campionario* B

In alcuni casi può convenire leggere al contrario la formula precedente e utilizzarla per calcolare la probabilità dell'intersezione di due eventi, una volta nota la probabilità condizionata di uno rispetto all'altro. Per esempio se conosciamo la probabilità dell'evento B e la probabilità condizionata di A su B, otteniamo

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

mentre se conosciamo la probabilità dell'evento A e la probabilità condizionata di B su A, otteniamo

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

3.4 Eventi indipendenti

Un concetto molto importante per le applicazioni statistiche della probabilità è quello di *eventi indipendenti*. Due eventi A e B si dicono indipendenti quando la conoscenza del verificarsi di uno dei due non ci dà alcuna informazione sul verificarsi dell'altro. Vediamo ancora il caso del lancio del dado: l'evento A è "il numero uscito è divisibile per tre" e l'evento B è "il numero uscito è pari", i.e. $A = \{3, 6\}$, $B = \{2, 4, 6\}$, sono due eventi indipendenti. Abbiamo, a priori, $P(A) = 1/3$ e $P(B) = 1/2$. Supponiamo ora di sapere che l'evento A si è verificato. Qual'è la probabilità di B condizionata al verificarsi di A? Affinché anche B sia verificato deve essersi verificato l'evento elementare $e = 6$ che è $A \cap B$. Questo evento ha probabilità $1/2$ di verificarsi se sappiamo che A si è verificato (controllare tramite la formula della probabilità condizionata). Ma $1/2$ era la probabilità "a priori" di B, quindi non ho avuto alcuna variazione della mia "fiducia" sul verificarsi di B. Viceversa se sappiamo che B si è verificato, abbiamo $P(A|B) = P(A \cap B)/P(B) = (1/6)/(1/2) = 1/3 = P(A)$.

Possiamo quindi definire due eventi *indipendenti* se si verifica una delle due relazioni seguenti

$$P(A|B) = P(A) \text{ o } P(B|A) = P(B).$$

(nota che la congiunzione "o" è *non-esclusiva*, infatti in questo caso, lo si dimostri per esercizio, queste due condizioni solo equivalenti e quindi una della due è vera se e solo se è vera anche l'altra)

Alternativamente possiamo prendere come definizioni di eventi indipendenti la seguente:

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

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

ovvero se la probabilità che siano entrambi realizzati è il prodotto delle singole probabilità.

Nota: Attenzione a non confondere eventi indipendenti e eventi incompatibili. Se due eventi sono incompatibili allora **non** sono indipendenti (il verificarsi di uno ci dà la certezza che l'altro non può verificarsi).

3.5 Teorema di Bayes

Supponiamo che lo spazio campionario $\mathcal{S}$ sia diviso in due sottoinsiemi (eventi) A e B tali che $A \cup B = \mathcal{S}$ e $A \cap B = \emptyset$, ovvero che gli eventi siano mutuamente esclusivi (diremo che $\{A, B\}$ è una *partizione* di $\mathcal{S}$). Supponiamo anche che sia stata definita una probabilità P sullo spazio campionario e che $P(A) > 0$ e $P(B) > 0$ siano le probabilità di A e B rispettivamente.

Consideriamo ora un terzo evento C di cui si conoscano le probabilità condizionate rispetto ad A e a B, $P(C|A)$ e $P(C|B)$. Supponiamo di effettuare un esperimento e constatare che in questo esperimento C si è verificato. Ci domandiamo: Qual'è la probabilità che sia sia verificato A piuttosto che B? (Questa situazione è assai comune nella pratica sperimentale, come vedremo nel prossimo paragrafo, quando parleremo di test diagnostici).

Se la probabilità di C è nota, la risposta è data dalla formula della probabilità condizionata,

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

(attenzione all'ordine delle probabilità condizionate!)

Ma quello che sappiamo su A, B e C può essere utilizzato per sostituire $P(C)$ nella formula precedente. Infatti abbiamo che $C = (C \cap A) \cup (C \cap B)$ e questi due eventi sono mutuamente esclusivi poiché $A \cap B = \emptyset$. Ne segue che

$$P(C) = P(C \cap A) + P(C \cap B) = P(C|A)P(A) + P(C|B)P(B) \tag{3.4}$$

che, una volta sostituita nella formula per $P(A|C)$ ci dà

$$P(A|C) = \frac{P(C|A)P(A)}{P(C|A)P(A) + P(C|B)P(B)}. \tag{3.5}$$

La formula (3.5) si può generalizzare al caso di n eventi mutuamente esclusivi $E_1, E_2, ..., E_n$, con $P(E_i) > 0$ e tali che $\bigcup_{i=1}^{n} E_i = \mathcal{S}$. Se E è un evento con $P(E) > 0$, abbiamo

$$P(E_j|E) = \frac{P(E|E_j)P(E_j)}{\sum_{i=1}^{n} P(E|E_i)P(E_i)}. \tag{3.6}$$

La formula (3.6) prende il nome di *Teorema di Bayes*.

Essa gioca un ruolo fondamentale nella teoria soggestivistica della probabilità, in quanto ci fornisce un *meccanismo per modificare la nostra opinione in funzione degli esiti di un esperimento*.

3.6 Test diagnostici

Vediamo ora una delle applicazioni più utili del calcolo delle probabilità in campo medico, quella ai *test disgnostici*. Quando un test sembra indicare la presenza della malattia, si dice che il risultato del test è *positivo*, quando il test sembra escluderla, si

dice che il risultato del test è *negativo*. Un test ideale dovrebbe individuare sempre con esattezza una malattia in un soggetto malato ed escluderne la presenza in un soggetto sano, ovvero i malati dovrebbero essere tutti e soli quelli per cui il test è positivo e i sani tutti e soli quelli per cui il test è negativo. In realtà questo non avviene e si possono presentare quattro situazioni distinte:

1. La malattia è presente e il test ne rileva la presenza. Diremo in questo caso che il soggetto è un *vero-positivo*.
2. La malattia è presente ma il test non la rileva, ovvero risulta negativo. In questo caso diciamo che il soggetto è un *falso-negativo*.
3. Il soggetto non è malato ma il test indica la presenza della malattia. Si dice che il soggetto è un *falso-positivo*.
4. Il soggetto non è malato e il test non indica la presenza della malattia. Il soggetto si dice un *vero-negativo*.

Nei casi 1 e 4 il test è corretto mentre nei casi 2 e 3 si è commesso un errore. Un buon test deve avere una probabilità di errore la più piccola possibile. Precisiamo quindi le misure di questi errori, detti *tassi di errore*, in termini di probabilità condizionate.

def. 1 Il tasso di falso-positivo in un test si indica con α ed è dato da

$$\alpha = P(\text{test positivo}|\text{soggetto sano})$$

def. 2 Il tasso di falso-negativo in un test si indica con β ed è dato da

$$\beta = P(\text{test negativo}|\text{soggetto malato})$$

La quantità $1 - \alpha$ si dice *specificità* del test e indica la probabilità che il test sia negativo per un soggetto sano (detto anche *vero-negativo*).

La quantità $1 - \beta$ si dice *sensibilità* del test e indica la probabilità che il test sia positivo per un soggetto malato (detto anche *vero-positivo*.

Si dice infine *accuratezza* di un test la probabilità che fornisca una risposta esatta quale che sia lo stato del paziente (quindi i casi 1 e 4). Nota che mentre specificità e sensibilità sono caratteristiche proprie del test, indipendenti dalla popolazione in esame, l'accuratezza del test dipende dalla popolazione a cui esso è applicato. Infatti se $P(S)$ è la probablilià che un individuo della popolazione sia sano (e $P(M) = 1 - P(S)$ quella che sia malato), l'accuratezza del test è data da $(1-\alpha)P(S)+(1-\beta)P(M) = 1 - \alpha P(S) - \beta P(M)$.

I test diagnostici sono un'importante campo di applicazione per il teorema di Bayes. Infatti un test viene fatto per "sapere" se un soggetto è malato oppure no. Supponiamo che ci sia nota (almeno con una buona approssimazione) la percentuale di malati nella popolazione, ovvero sia nota $P(M)$. Di conseguenza $P(S) = 1 - P(M)$ sarà la probabilità che un individuo sia sano.

Scegliamo ora un individuo a caso e sottoponiamolo al test. Supponiamo che il risultato sia *positivo*; ci domandiamo: qual'è la probabilità che il soggetto sia effettivamente malato?

Per rispondere basta osservare che questa è esattamente la situazione descritta nel paragrafo precedente. Abbiamo partizionato la popolazione tra malati e sani e conosciamo le due probabilità $P(M)$ e $P(S)$. Inoltre conosciamo le probabilità condizionate

di esito positivo sia rispetto alla condizione di essere malato, $(1-\beta)$, sia a quella di essere sano, α. Di conseguenza, applicando la (3.5), otteniamo

$$P(\text{essere malato} \,|\, \text{risultare positivo}) = \frac{(1-\beta)P(M)}{(1-\beta)P(M) + \alpha P(S)}.$$

La (3.6) ha conseguenze importanti. Se l'incidenza della malattia è bassa, i.e. $P(M)$ è piccola, anche un test molto efficiente, cioè con piccole α e β, dà una risposta apparentemente (ma solo apparentemente) paradossale: dopo un test positivo può essere più probabile essere sani che ammalati! (provare con $\alpha = \beta = 0.05$ e una percentuale di malati dell'uno per mille; la probabilità di essere malato per un individuo scelto a caso che risulti positivo al test è minore del 2%).

Questa conclusione può apparire paradossale e desta sempre una certa perplessa diffidenza da parte dei medici nei "trucchi della matematica". In effetti la situazione che abbiamo presentato è quella che si presenterebbe in un ipotetico "screening" casuale della popolazione su base, p.e., nazionale. Nella pratica medica corrente, il medico decide di sottoporre a un test un suo paziente solo se ha un qualche sospetto che questi si trovi nelle condizioni "favorevoli" alla malattia (le cosidette popolazioni a rischio). In questo caso la probabilità *a priori* che il paziente sia malato, ovvero $P(M)$, è ben superiore all'incidenza della malattia su tutta la popolazione nazionale come riportata dalle statistiche. E' bene osservare che, qualsiasi sia $P(M)$, se il test risulta positivo, la probabilità a posteriori $P(\text{essere malato} \,|\, \text{risultare positivo})$ è più grande della probabilità a priori $P(M)$.

3.7 Appendice

Per "visualizzare" le leggi del calcolo delle probabilità ci si può aiutare con uno schema simile a quello dei "diagrammi di Venn" nella teoria degli insiemi.

Disegnamo una regione R su di un foglio e scegliamo l'unità di misura in modo che l'area della regione sia uguale a una unità di superficie, Area$(R) = 1$.

Suddividiamo R in tante regioni $e_k,\ k = 1, ..., N$ e pensiamo a queste regioni del piano come i nostri eventi elementari. Gli eventi E saranno quindi quelle sotto-regioni di R formate da unioni di sotto-regioni e_k. Assegnamo come probabilità di un evento elemenetare e_i l'area della regione e_i stessa.

Si verifica facilmente che questa definizione soddisfa agli assioni di probabilità.

E' opportuno rivedere tutti i concetti e le regole presentate in questo capitolo alla luce di questo esempio, usando le comuni proprietà dell'area delle figure piane.

Capitolo 4

Variabili aleatorie

4.1 Variabili aleatorie discrete

Una *variabile aleatoria* è una funzione a valori reali X definita su uno spazio campionario $\mathcal{S}$, $X : \mathcal{S} \to \mathbf{R}$. A ogni esperimento otteniamo quindi un numero, $X(e)$, che è il valore che la variabile aleatoria assume sul risultato dell'esperimento, l'evento elementare e. Possiamo quindi considerare l'insieme di tutti i valori possibili (detto il *range* della variabile aleatoria) come un nuovo spazio campionario e assegnare una probabilità ai possibili valori della variabile aleatoria: a ogni valore x nel range della variabile aleatoria X, assegnamo la probabilità che X assuma il valore x. Questo valore è dato dalla probabilità $P(E)$ dell'evento $E = \{e \in \mathcal{S} | X(e) = x\}$ ovvero la retroimmagine di x tramite X.

Otteniamo così , al posto dello spazio campionario $\mathcal{S}$, che in genere è assai complesso, un semplice spazio campionario formato da un insieme di numeri. Il maggiore vantaggio di questa sostituzione è che molte variabili aleatorie, definite su spazi campionari anche molto diversi tra loro, danno luogo a una stessa "distribuzione" di probabilità sull'asse reale.

Denoteremo con lettere romane maiuscole le variabili aleatorie e con lettere romane minuscole i valori assunti da una variabile aleatoria. Con $P(X = x)$ indicheremo la *probabilità che la variabile aleatoria X assuma il valore x*.

Una variabile aleatoria si dirà *discreta* se essa può assumere solo un numero finito o numerabile di valori. In questo caso a ogni valore x sarà associato il numero $f(x) := P(X = x)$. La funzione f si dice funzione di probabilità o *funzione di densità di probabilità*. Essa si estende a tutti i valori reali, ponendo il suo valore uguale a 0 al di fuori dei valori che può assumere X. La funzione f soddisfa la condizione di normalizzazione $\sum_x f(x) = 1$ dove la somma è estesa a tutti i possibili valori assunti da X, che ci dice che la probabilità che X assuma almeno uno dei valori possibili è 1 .

Si definisce funzione di distribuzione cumulativa, o semplicemente *funzione di distribuzione* della variabile aleatoria X, la funzione definita da

$$F(x) := P(X \leq x). \tag{4.1}$$

F quindi rappresenta la probabilità che la variabile aleatoria X assuma un qualunque valore minore o uguale a x. La funzione di distribuzione gode della seguenti proprietà :

(i) $F(x)$ è una funzione non decrescente di x;

(ii) $\lim_{x\to+\infty} F(x) = 1$;

(iii) $\lim_{x\to-\infty} F(x) = 0$;

(iv) F è continua a destra, ovvero $\lim_{x\to x_0^+} F(x) = F(x_0)$, per ogni $x_0 \in \mathbf{R}$;

Le proprietà (ii) e (iii) sono "ovvie": esse ci dicono semplicemente che la probabilità di assumere un qualsiasi valore è 1 e quella di non assumere alcun valore è 0. Anche la (i) è semplice da spiegare: poiché se $y > x$ si ha $\{e \in \mathcal{S}|X(e) \leq x\} \subset \{e \in \mathcal{S}|X(e) \leq y\}$, ne segue che $P(X \leq x) \leq P(X \leq y)$. La proprietà (iv) ci dice che la F può ammettere delle discontinuità solo del tipo $\lim x \to x_0^- F(x) < F(x_0)$; in questo caso la differenza tra il limite da sinistra e $F(x_0)$ è il valore di $f(x_0)$.

Tramite la funzione di distribuzione è possibile esprimere tutte la quantità riguardanti la probabilità di X. Per esempio

$$P(a < X \leq b) = F(b) - F(a)\,, \text{ per ogni } a < b, \;\; P(X < b) = \lim_{x\to b^-} F(x)\,,$$

(si noti che il limite è fatto da sinistra). In particolare

$$P(X = b) = F(b) - \lim_{x\to b^-} F(x).$$

Infine la probabilità

$$P(x > a) = 1 - F(a)$$

è detta *probabilità di coda*.

Uno dei maggiori vantaggi dell'uso della funzione di distribuzione è che permette un trattamento unificato del caso delle variabili aleatorie discrete e di quelle continue, come vedremo tra poco.

Vediamo un esempio di variabile aleatoria discreta. Consideriamo una sequenza di 5 lanci di una moneta non truccata. Lo spazio campionario è ora formato da tutte le stringhe di lunghezza 5 di simboli T o C (testa, croce), che contiene $2^5 = 32$ elementi. Se siamo interessati a sapere quante teste escono in una successione di 5 lanci (indipendentemente dall'ordine di uscita) possiamo definire una variabile aleatoria X che conta le teste in ogni singola sequenza. Dunque la X ha per range l'insieme formato dai numeri $\{0, 1, 2, 3, 4, 5\}$, rispettivamente nessuna testa, una testa, etc. A questi valori corrispondono uno o più elementi dello spazio campionario, retroimmagine del valore tramite la X. Per esempio, $X = 0$ ha per retroimmagine la sola sequenza C, C, C, C, C mentre $X = 1$ ha per retroimmagine le cinque sequenze

$$T, C, C, C, C \quad C, T, C, C, C \quad C, C, T, C, C \quad C, C, C, T, C \quad C, C, C, C, T$$

Le probabilità da assegnare a ogni singolo valore della variabile aleatoria si contano dividendo i casi favorevoli per i casi possibili, quindi $P(X = 0) = \frac{1}{32}$, $P(X = 1) = \frac{5}{32}$, $P(X = 2) = \frac{10}{32}$, $P(X = 3) = \frac{10}{32}$, $P(X = 4) = \frac{5}{32}$, $P(X = 5) = \frac{1}{32}$.

La funzione di distribuzione è quindi data da $F(x) = 0,\ x < 0,\ F(x) = \frac{1}{32},\ 0 \leq x < 1,\ F(x) = \frac{6}{32},\ 2 \leq x < 3,\ F(x) = \frac{16}{32},\ 3 \leq x < 4,\ F(x) = \frac{26}{32},\ 4 \leq x < 5,\ F(x) = \frac{31}{32},\ 4 \leq x < 5$ e infine $F(x) = \frac{32}{32} = 1,\ x \geq 5$. Si noti che in corrispondenza dei valori del range la F è discontinua da sinistra.

4.2 Variabili aleatorie continue

Una variabile aleatoria X si dice continua quando il suo range è tutta la retta reale o unione di intervalli sulla retta reale. Per esempio se può assumere tutti i valori compresi tra 0 e 20, o tutti i valori positivi, etc.

Nel seguito ci limiteremo a variabili aleatorie continue per cui esiste una funzione $f(x)$, detta *densità di probabilità*, per cui valga

$$P(X \in \mathcal{B}) = \int_{\mathcal{B}} f(x)\, dx \,, \tag{4.2}$$

dove $\mathcal{B}$ è un qualsiasi sottoinsieme di $\mathbf{R}$, ovvero la probabilità che il valore della variabile aleatoria X cada nell'insieme $\mathcal{B}$ è data dall'integrale esteso a $\mathcal{B}$ della funzione f. Ponendo $f = 0$ al di fuori del range di X, possiamo quindi fare a meno di specificare il range di una variabile aleatoria continua. La funzione f deve essere tale da soddisfare la condizione di normalizzazione

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

che significa semplicemente che la probabilità di assumere un qualsiasi valore reale è 1.

(Nota: così enunciata la definizione di variabile aleatoria continua manca assai di rigore matematico; sviluppare il contesto matematico per precisare questi concetti necessiterebbe un corso a parte, forse anche due!)

Come già fatto per le varibili discrete, possiamo definire la funzione di distribuzione

$$F(x) := \int_{-\infty}^{x} f(s)\, ds = P(X \leq x) \,. \tag{4.3}$$

La funzione di distribuzione così definita gode delle stesse proprietà (i),(ii),(iii) della funzione di distribuzione per le variabili aleatorie discrete. In più la funzione di distribuzione di una varibile aleatoria continua risulta essere una funzione continua (e non solo continua a destra): nella nostra definizione non esiste più distinzione tra $P(X \leq x)$ e $P(X < x)$, la probabilità che una variabile aleatoria continua assuma esattamente il valore x essendo nulla. Osserviamo che ciò non significa che X non possa assumere il valore x: possiamo però assegnare un valore positivo solo alla probabilità di assumere un valore qualsiasi compreso tra x meno qualcosa e x più qualcosa (ovvero a un intervallo, piccolo quanto vogliamo ma di lunghezza positiva, che contenga il numero x).

Come esempio consideriamo il caso di una variabile *uniformemente distribuita* nell'intervallo (a, b), cioè di una variabile con funzione di densità uguale a $\frac{1}{b-a}$ per $x \in (a, b)$ e $f(x) = 0$ per tutti gli altri x. In questo caso la funzione di distribuzione f sarà data da

$$F(x) = \begin{cases} 0 & \text{per } x \leq a \\ \frac{x-a}{b-a} & \text{per } a < x < b \\ 1 & \text{per } b \leq x \end{cases} .$$

4.3 Valor medio di una variabile aleatoria

Introduciamo qui un concetto fondamentale per la probabilità e per la statistica e per illustrarlo ci serviremo ancora del gioco dei dadi.

Consideriamo ancora il caso semplice del lancio di un dado (non truccato), e supponiamo di giocare con le seguenti regole: un giocatore (il banco) lancia il dado e, a seguito del risultato del lancio, paga a un altro giocatore (lo scommettitore) certe cifre fissate in anticipo. Per fissare le idee diciamo che non paga niente se il risultato del lancio è minore o uguale a 3, che paga 40 se esce il 4, 30 se esce il 5 e 50 se esce il 6. Ovviamente il banco chiede allo scommettitore di pagare, in anticipo, una certa somma per poter giocare, posta. Come si calcola questa somma in modo che il gioco sia "equo"? (prima di tutto bisogna definire cosa si intende per gioco equo! diremo che il gioco è equo se i due giocatori sono, razionalmente, disposti a invertire le parti nel gioco).

Ovviamente la posta sarà un numero maggiore di zero (la minima vincita possibile) perché il banco rischia di dover pagare qualcosa allo scommettitore, d'altra parte essa sarà minore di 50 altrimenti lo scommettitore rifiuterebbe di giocare.

Consideriamo la variabile aleatoria X che ci dà la vincita dello scommettitore. Quindi X può assumere i valori 0, 30, 40 e 50. La probabilità che la variabile aleatoria assuma il valore 0 è $3/6 = 1/2$, mentre $X = 20$ ha probabilità $1/6$ come pure $X = 30$ e $X = 50$. Ricordiamo che (p.e. alla luce dell'interpretazione frequentista) ciò significa che in una "lunga" serie di lanci, ci aspettiamo che X assuma circa metà delle volte il valore 0 e circa un sesto delle volte ciascuno degli altri tre valori, $40,\ 30,\ 50$. Supponiamo che questa ripartizione degli esiti sia esatta, cioè su diciamo 600 prove ci siano esattamente 300 prove in cui $X = 0$, 100 con $X = 40$, 100 con $X = 30$ e 100 con $X = 50$. Con questa situazione, la "vincita" totale sarebbe di $300 \times 0 + 100 \times 40 + 100 \times 30 + 100 \times 50 = 12'000$. Questa sarà la posta equa da pagare. Suddividendo la posta nelle 600 prove otteniamo una posta di 20 per ogni partita. Questo valore corrisponde alla somma delle vincite X per le rispettive probabilità: $0 \times \frac{1}{2} + 40 \times \frac{1}{6} + 30 \times \frac{1}{6} + 50 \times \frac{1}{6} = 20$.

Questo è il cosiddetto *valore atteso* della variabile aletatoria (o *medio* o ancora *speranza matematica*, in inglese *mathematical expectation* dal termine *expectatio* introdotto da Christian Huygens in *De ratiociniis in alea ludo*, 1657, il primo trattato di probabilità che sia stato pubblicato) che indicheremo con $E(X)$ o anche con μ quando sia chiaro a quale variabile aleatoria sia riferisca. Quindi questa quantità rappresenta il guadagno medio che si attende dal gioco, nel senso che a lungo andare (reiterando più volte il gioco) lo scommettitore si attende un guadagno se la posta da giocare è minore di 20 e una perdita se la posta è maggiore di 20.

La definizione generale del valor medio, indicato con $E(X)$ o con μ, è data da

$$E(X) = \begin{cases} \sum_i x_i P(X = x_i) & \text{se } X \text{ è una variabile discreta} \\ \int_{-\infty}^{+\infty} x f(x)\, dx & \text{se } X \text{ è una variabile continua} \end{cases} \tag{4.4}$$

nell'ipotesi che la somma o l'integrale convergano assolutamente.

4.4 Funzioni di variabili aleatorie

Sia X una variabile aleatoria con funzione di distribuzione F_X e funzione di densità f_X.

Sia $g : \mathbf{R} \to \mathbf{R}$ una funzione. Possiamo definire una nuova variabile aleatoria $Y = g(X)$ come quella variabile aleatoria che assume il valore $y = g(x)$ ogni qualvolta la variabile aleatoria X assume il valore x.

Perché la variabile aleatoria Y sia ben definita occorre ovviamente che il dominio della funzione g contenga il "range" della variabile aleatoria X, ovvero che sia pos-

sibile calcolare $g(x)$ per ogni valore x che può essere assunto da X. Per esempio, qualsiasi sia la variabile aleatoria X possiamo sempre definire la varibile aleatoria $Y = \exp(X)$, ma possiamo definire $Y = \ln(X)$ solo se la variabile aleatoria X assume soltanto valori positivi.

Una volta che ci si sia accertati della possibilità di definire Y, si pone il problema di calcolarne la funzione di distribuzione F_Y e la funzione di densità f_Y a partire dalla conoscenza di g, F_X e f_X.

Ricordiamo che, per definizione, $F_Y(x) = \mathrm{P}(Y \leq x)$ ovvero $F(x)$ è la probabilità che la variabile aleatoria Y assuma valori minori o uguali a x. Per come è definita la Y abbiamo quindi

$$F_Y(x) = \mathrm{P}(Y \leq x) = \mathrm{P}(g(X) \leq x) = \int_{G_x} f_X(s)\, ds\ ,$$

dove l'integrale è esteso a tutto l'insieme $G_x = \{s \in \mathbf{R} : g(s) \leq x\}$. Per esempio, prendiamo $g(x) = x^2$. Allora avremo che

$$F_Y(x) = 0,\ \text{per ogni}\ x < 0$$

in quanto, qualunque sia il valore assunto dalla variabile aleatoria X, il suo quadrato sarà un numero maggiore o uguale a 0. Se invece vogliamo calcolare $F_Y(2)$, dovremo tener conto che la Y assume un valore compreso nell'intervallo $(0, 2)$ ogni qualvolta X assume un valore compreso tra $-\sqrt{2}$ e $\sqrt{2}$. Inoltre, poiché Y non può assumere valori negativi, $F_Y(2) = \mathrm{P}(Y \leq 2) = \mathrm{P}(0 \leq Y \leq 2)$. Di conseguenza abbiamo

$$F_Y(2) = \int_{-\sqrt{2}}^{\sqrt{2}} f_X(s)\, ds = F_X(\sqrt{2}) - F_X(-\sqrt{2}).$$

Nota che non è detto che l'insieme G_x sia un intervallo. Prendiamo come esempio $g(x) = -x^2$. In questo caso avremo "ovviamente" $F_Y(x) = 1$ per ogni $x > 0$ (perché ?) mentre se vogliamo calcolare $F_Y(-2)$ dobbiamo calcolare l'integrale della funzione f_X su tutto l'insieme in cui $-x^2 < 2$, che in questo caso è fatto dall'unione degli intervalli $(-\infty, -\sqrt{2})$ e $(\sqrt{2}, \infty)$. Avremo quindi

$$F_Y(-2) = \int_{-\infty}^{-\sqrt{2}} f_X(s)\, ds + \int_{\sqrt{2}}^{\infty} f_X(s)\, ds = F_X(-\sqrt{2}) + 1 - F_X(\sqrt{2}).$$

Inoltre è possibile che G_x abbia intersezione non vuota con l'insieme in cui f_X si annulla. Per esempio, cosa succede del conto precedente se la X è una variabile aleatoria uniformemente distribuita nell'intervallo $(-4, 7)$?

Nel caso che la funzione g sia strettamente monotona (assumeremo nel calcolo che segue che g sia anche derivabile e che $g' > 0$) si può scrivere una formula generale per la funzione densità della variabile aleatoria Y. Infatti abbiamo

$$\begin{aligned} F_Y(x) &= \mathrm{P}(g(X) \leq x) = \mathrm{P}(X \leq g^{-1}(x)) \\ &= \int_{-\infty}^{g^{-1}(x)} f_X(s)\, ds = \int_{g(-\infty)}^{x} \frac{f_X(g^{-1}(z))}{g'(g^{-1}(z))}\, dz\ , \end{aligned} \tag{4.5}$$

dove abbiamo effettuato il cambiamento di variabili $z = g(s)$ e abbiamo indicato con $g(-\infty)$ il limite di g per $x \to -\infty$. La funzione

$$f_Y(x) = \frac{f_X(g^{-1}(x))}{g'(g^{-1}(x))} \tag{4.6}$$

è quindi la funzione di densità della variabile Y (se $g(-\infty > -\infty$ la f_Y si pone uguale a zero in $(-\infty, g(-\infty))$).

Esercizio: cosa cambia se $g' < 0$?

Esercizio: usando la (4.6) trovare le funzioni di densità delle variabili aleatorie $Y = \alpha X + \beta$ e $Y = \mathrm{arctg}\,(X)$.

4.5 Valor medio di funzione di var. aleat.

Se è piuttosto laborioso ricavare la funzione di di densità di una variabile aleatoria $Y = g(X)$ in termini della funzione di densità della X, è invece molto semplice calcolarne il valor medio. Infatti vale la seguente formula

$$E(Y) = E(g(X)) = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx. \tag{4.7}$$

La dimostrazione si ottiene con un po' di calcolo dalla (4.6)

$$E(Y) = \int_{-\infty}^{+\infty} y f_Y(y)\, dy = \int_{-\infty}^{+\infty} y \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}\, dy\ , \tag{4.8}$$

Operando ora il cambiamento di variabile $y = g(x)$ nella (4.8) otteniamo immediatamente la (4.7).

Nel caso di una variabile discreta abbiamo invece

$$E(g(y)) = \sum_{x|f(x)>0} g(x) f_X(x), \tag{4.9}$$

dove la somma è estesa a tutti i valori x del range di X (dove quindi la $f(x) > 0$).

La (4.9) può comunque essere dimostrata direttamente come segue: fissato y abbiamo che $P(Y = y)$ è uguale alla probabilità che la X assuma un qualsiasi valore x tale $g(x) = y$ ovvero $x \in g^{-1}(y)$, quindi $P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x)$. Abbiamo quindi

$$E(Y) = \sum_y y P(Y = y) = \sum_y \sum_{x \in g^{-1}(y)} y P(X = x) = \sum_y \sum_{x \in g^{-1}(y)} g(x) f_X(x)$$

ma la doppia somma non è nient'altro che la somma su tutti gli x tali che $P(X = x) > 0$, e quindi otteniamo la (4.9).

4.6 Varianza di una variabile aleatoria

Per ogni intero ≥ 1, la quantità $E(X^n)$ è detta momento di ordine n-esimo della variabile aleatoria X. Essa può calacolarsi con la formula (4.7). Per $n =$, il momento coincide con il valor medio.

Di più frequente sono i *momenti centrali*: per ogni $m \geq 2$ definiamo il momento centrale di ordine m la quantità $E[(X - \mu)^m]$, dove $\mu = E(X)$ è il valor medio.

Di particolare importanza è il momento centrale del second'ordine, detto anche *varianza* e indicato generalmente con $\mathrm{var}(X)$ oppure con σ^2:

$$\begin{aligned} \mathrm{var}(X) = \sigma^2 = E[(X - E(X))^2] \\ = \begin{cases} \sum_i (x_i - \mu)^2 P(X = x_i) & X \text{ variabile discreta} \\ \int_{-\infty}^{+\infty} (x - \mu)^f(x)\, dx & X \text{ variabile continua} \end{cases} \end{aligned} \tag{4.10}$$

La radice quadrata della varianza $\sigma = \sqrt{\mathrm{var}(X)}$ viene detta *deviazione standard* della variabile aleatoria. Queste due quantità danno un'informazione sulla dispersione della varaibile aleatoria attorno al valor medio.

Per il calcolo della varianza, in alternativa alla (4.10), si utilizza la seguente formula

$$\mathrm{var}(X) = E(X^2) - (E(X))^2, \tag{4.11}$$

che ci dice che il calcolo della varianza si effettua calcolando il valor medio della variabile aletoria X^2 (il quadrato della X) e poi sottraendo il quadrato del valor medio di X. La dimostrazione della (4.11) è lasciata per esercizio.

Diseguaglianza di Čebišev *Se X è una variabile aleatoria di media μ e varianza σ^2, allora per ogni $\varepsilon > 0$ si ha*

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \tag{4.12}$$

La dimostrazione della (4.12) segue dalla *diseguaglianza di Markov*: *se X è una varibile aleatoria che assume solo valori non negativi, allora per ogni $a > 0$ si ha*

$$P(X \geq a) \leq \frac{E(X)}{a}. \tag{4.13}$$

Diamo la dimostrazione della (4.13) nel caso di una variabile aleatoria continua di densità f. Abbiamo

$$\begin{aligned} E(x) = \int_0^{+\infty} x f(x)\, dx \;\geq \int_a^{+\infty} x f(x)\, dx \;\geq \\ \int_a^{+\infty} a f(x)\, dx \;= a P(X \geq a) \end{aligned}$$

La dimostrazione (che lasciamo per esercizio) della (4.12) si ottiene applicando la (4.13) alla variabile aleatoria non negativa $(X - \mu)^2$ con $a = \varepsilon^2$.

La diseguaglianza di Čebišev può anche essere scritta, scegliendo $\varepsilon = k\sigma$ con $k > 0$

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{4.14}$$

che ci dice che la probabilità che la variabile aleatoria assuma valori che si scostano dalla media per piu di k volte la deviazione standard è minore di $1/k^2$.

Osserviamo che la disuguaglianza di Čebišev permette di ottenere una maggiorazione della probabilità dalla sola conoscenza del valor medio e della varianza, che in molte applicazioni statistiche sono tutta la conoscenza (sotto forma di stime) che abbiamo della popolazione.

4.7 Variabili aleatorie vettoriali

Consideriamo una popolazione (nel senso usuale del termine) e prendiamo come variabile aleatoria X l'altezza di un individuo scelto a caso. Accanto consideriamo la variabile aleatoria Y data dal peso dello stesso individuo. La coppia (X, Y) è un esempio di vettore aleatorio, cioè di una funzione che associa a ogni evento elementare uan coppia (più in generale una n-pla) di numeri reali. Possiamo ovviamente immaginare vettori aleatori di dimesione n qualsiasi, per esempio otteniamo un vettore tridimensionale aggiungendo come Z l'età dell'individuo.

Per semplicità di notazione ci limiteremo al caso di due variabili.

Definiano la *funzione di probabilità cumulativa congiunta* del vettore aleatorio (X, Y) la funzione di due variabili

$$F(x, y) = P(X \leq x, Y \leq y), \quad (x, y) \in \mathbf{R}^2. \tag{4.15}$$

La distribuzione si dirà *discreta* se esiste un insieme numerabile di punti (x_i, y_j) tali che

$$p_{i,j} = P(X = x_i, Y = y_j) > 0, \quad \sum_i \sum_j p_{i,j} = 1$$

La distribuzione si dirà *continua* se esiste una funzione di due variabili $f(x, y) \geq 0$ tale che

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{\infty}^{y} f(u, v)\, dudv$$

La funzione f si dice *densità congiunta* di X e Y.

Se g è una funzione di due variabili, possiamo calcolare il valor medio della variabile aleatoria $g(X, Y)$ tramite la generalizzazione della formula (4.7),

$$E(g(X, Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y)\, dxdy, \tag{4.16}$$

e analogamente per le variabili discrete

$$E(g(X, Y)) = \sum_{i,j} p_{i,j} g(x, y)\, dxdy, \tag{4.17}$$

Distribuzioni marginali Se conosciamo la distribuzione congiunta di due variabili aleatorie possiamo ricavare le distribuzione delle singole variabili X e Y, dette anche *distribuzioni marginale*.

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \leq +\infty) = F(x, \infty)$$
$$F_Y(y) = P(Y \leq y) = P(X \leq +\infty, Y \leq y) = F(\infty, y)$$

e le densità

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy; \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx. \tag{4.18}$$

Nel caso discreto abbiamo

$$p_i = P(X = x_i) = \sum_j p_{ij} \ , \quad q_j = P(Y = y_i) = \sum_i p_{ij}$$

Utilizzando le densità condizionate possiamo dimostrare nel caso di g lineare che la (4.16) diventa

$$E(aX + bY) = aE(X) + bE(Y), \tag{4.19}$$

dove $E(X)$ e $E(Y)$ sono i valor medi delle due variabili aleatorie, che si possono calcolare usando le densità marginali. Questo risultato si generalizza a una combinazione lineare di un numero n qualsiasi di variabili, nella seguente

Proposizione *Se $X_1, X_2, ..., X_N$ sono N variabili aleatorie, allora il valor medio della media aritmetica delle X_i è la media aritmetica dei valor medi:*

$$E\left(\frac{X_1 + X_2 + ... + X_N}{N}\right) = \frac{E(X_1) + E(X_2) + ... + E(X_N)}{N}. \tag{4.20}$$

Nota: Se X e Y sono due variabili aleatorie di cui si conosce la distribuzione congiunta, allora le distribuzioni delle singole variabili aleatorie coincidono con le distribuzioni marginali. Infatti si ha

$$P(X = x) = P(\{e \in \mathcal{S} | X(e) = x\}), \ P(Y = y) = P(\{e \in \mathcal{S} | Y(e) = y\}),$$
$$P(X = x, Y = y) = P(\{e \in \mathcal{S} | X(e) = x, Y(e) = y\}),$$

sommando $P(X = x, Y = y)$ rispetto a tutti i valori di y si ottiene

$$\sum_y P(X = x, Y = y) = P(\bigcup_y \{e \in \mathcal{S} | X(e) = x, Y(e) = y\}).$$

Basta infine verificare che l'evento (l'insieme) $\bigcup_y \{e \in \mathcal{S} | X(e) = x, Y(e) = y\}$ coincide con l'evento $\{e \in \mathcal{S} | X(e) = x\}$.

Distribuzioni condizionate Nel caso di variabili discrete, accanto alle formule per la distribuzione marginale che sono simili a quelle del caso continuo, possiamo definire la *funzione di probabilità condizionata* di X nell'ipotesi che Y assuma un valore definito, $Y = y_j$.

$$p_{X|Y}(x_i|y_j) = P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{i,j}}{q_j}$$

e analogamante per le funzioni di distribuzione.

Per le variabili continue, possiamo define concetti analoghi per le densità di probabilità.

Esempio Sia Y una variabile aleatoria funzione di una variabile aleatoria X, i.e. $Y = g(X)$. Allora se conosciamo la distribuzione di probabilità p_i di X, la distribuzione congiunta di (X, Y) si ricava facilmente osservando che Y assume il valore y se e solo se esiste un valore x assunto da X tale che $y = g(x)$. Quindi la coppia (x_h, y_k) ha probabilità $p_{h,k}$ che vale p_h se $y_k = g(x_h)$ o 0 se $y_k \neq g(x_h)$. Ponendo $y_i = g(x_i)$ si ha

$$p_{i,j} = P(X = x_i, Y = y_j) = \delta_{i,j} p_i, \tag{4.21}$$

dove $\delta_{i,j} = 1$ se $i = j$ e $\delta_{i,j} = 0$ se $i \neq j$.

Variabili indipendenti Due variabili aleatorie X e Y si dicono indipendenti se la funzione di densità congiunta $f(x, y)$ si può esprimere come il prodotto di una funzione $f_X(x)$ della sola variabile x e una funzione $f_Y(y)$ della sola y, $f(x, y) = f_X(x)f_Y(y)$. Confrontando con formula (4.18) per le densità marginali, le funzioni f_X e f_Y sono le densità (marginali) delle variabili X e Y rispettivamente.

E' importante osservare che nel caso di variabili aleatorie indipendenti, oltre alla (4.19) vale la seguente formula per la varianza

$$\mathrm{var}(aX + bY) = a^2\mathrm{var}(X) + b^2\mathrm{var}(Y), \tag{4.22}$$

che in generale non è vera per una coppia di variabili aleatorie qualsiasi (in effetti la (4.22) è vera se e solo se $\mathrm{cov}(X, Y) = 0$). Anche la (4.22) si generalizza a un numero qualsiasi di variabili aleatorie indipendenti.

Proposizione *Se $X_1, X_2, ..., X_N$ sono N variabili aleatorie indipendenti con la stessa varianza σ^2 abbiamo*

$$\mathrm{var}\left(\frac{X_1 + X_2 + ... + X_N}{N}\right) = \frac{\sigma^2}{N}, \quad s.d.\left(\frac{X_1 + X_2 + ... + X_N}{N}\right) = \frac{\sigma}{\sqrt{N}}. \tag{4.23}$$

Dalla (4.23), abbiamo che la deviazione standard della media di N osservazioni indipendenti decresce come $\sqrt{N}$ al crescere del numero di osservazioni.

4.8 Teoremi sul limite

Possiamo infine illustrare due teoremi fondamentali sia per l'interpretazione della probabilità sia per le applicazioni ai problemi di inferenza statistica.

Il primo è la cosiddetta *legge (debole) dei grandi numeri*

Teorema 4.1 (Legge dei grandi numeri) *Sia $X_1, X_2, ..., X_N$ una successione di variabili aleatorie indipendenti con la stessa media μ e la stessa varianza σ^2. Allora per ogni $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + ... + X_N}{N} - \mu\right| < \varepsilon\right) = 1. \tag{4.24}$$

Questo teorema fu stabilito per la prima volta, nel caso di una successione di variabili aleatorie che obbedivano tutte alla stessa distribuzione binomiale, da Jacob Bernoulli (1654-1705) e pubblicato nel suo trattato postumo *Ars conjectandi* (1713).

Il teorema stabilisce che la media aritmetica di una successione di variabili aleatorie avente la stessa distribuzione, di qualunque tipo essa sia, converge, con probabilità 1, al valor medio della distribuzione.

La dimostrazione segue facilmente dalla disuguaglianza di Čebišev (4.12) e dalla formula per la varianza (4.23). Abbiamo infatti

$$P\left(\left|\frac{X_1 + X_2 + ... + X_N}{N} - \mu\right| \geq \varepsilon\right) = \frac{\sigma^2}{\varepsilon^2 N},$$

che possiamo riscrivere

$$P\left(\left|\frac{X_1 + X_2 + ... + X_N}{N} - \mu\right| < \varepsilon\right) = 1 - \frac{\sigma^2}{\varepsilon^2 N}, \tag{4.25}$$

da cui il teorema segue facilmente passando al limite $N \to \infty$. Osserviamo che la (4.25) ci permette anche di valutare la velocità di convergenza a 1 della probabilità.

Il secondo teorema, fondamentale per le applicazioni alla statistica, è il *teorema centrale del limite*

Teorema 4.2 (Teorema centrale del limite) *Sia $X_1, X_2, ..., X_N$ una successione di variabili aleatorie indipendenti con la stessa media μ e la stessa varianza σ^2. Allora la distribuzione della variabile aleatoria*

$$\frac{X_1 + X_2 + ... + X_N - N\mu}{\sigma\sqrt{N}}$$

tende alla distribuzione normale standard per $N \to \infty$. Ovvero, per ogni $x \in \mathbf{R}$ sia ha

$$\lim_{N\to\infty} P\left(\frac{X_1 + X_2 + ... + X_N - N\mu}{\sigma\sqrt{N}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt. \qquad (4.26)$$

Torneremo su questo teorema nel capitolo dedicato al campionamento.

4.9 Covarianza

Il numero

$$\sigma_{XY} = \mathrm{cov}(X, Y) := E\left[(X - E(X))(Y - E(Y))\right]$$

si dice *covarianza* tra X e Y. Vale inoltre una formula per il calcolo della covarianza, analoga alla (4.11) per il calcolo della varianza,

$$\mathrm{cov}(X, Y) = E(XY) - E(X)E(Y), \qquad (4.27)$$

in accordo con il fatto che $\mathrm{cov}(X, X) = \mathrm{var}(X)$.

La covarianza è considerata un indice della tendenza delle variabili X e Y a "variare assieme", p.e. Y cresce se X cresce (in questo caso $\mathrm{cov}(X, Y) > 0$) o decresce ($\mathrm{cov}(X, Y) < 0$).

Se le variabili X e Y sono indipendenti allora $\mathrm{cov}(X, Y) = 0$, ma l'inverso non è vero. Infatti bisogna essere un po' prudenti nell'interpretare la covarianza come indice della dipendenza di una variabile aleatoria dall'altra. Vediamo con un esempio, che possiamo avere $Y = g(X)$ ma $\mathrm{cov}(X, Y) = 0$. Basta prendere una variabile aleatoria X che assume, con ugual probabilità, i valori $\{-3, -2, -1, 0, 1, 2, 3\}$ e $Y = X^2$. Questo esempio si generalizza a tutte quelle variabili aleatorie X che abbiano una funzione (o densità) di probabilità simmetrica rispetto al valor medio e con funzioni g pari rispetto a $X - E(X)$.

Il concetto di covarianza è in effetti legato alla variazione lineare della variabile aleatoria Y in funzione della variabile aleatoria X. Calcoliamo $\mathrm{cov}(X, Y)$ nel caso in cui $Y = \alpha X + \beta$. Ricordiamo che $E(Y) = \alpha E(X) + \beta$ e $\mathrm{var}(Y) = \alpha^2 \mathrm{var}(X)$. E' inoltre immediato verificare che, in questo caso, $E(XY) = \alpha E(X^2) + \beta E(X)$ (fare il calcolo per esercizio). Introducendo queste relazioni in (4.27), e ricordando la (4.11), otteniamo

$$\mathrm{cov}(X, \alpha X + \beta) = \alpha \mathrm{var}(X), . \qquad (4.28)$$

Accanto alla covarianza si introduce anche il numero

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(x)\mathrm{var}(Y)}}, \tag{4.29}$$

detto *coefficiente di correlazione*. Questo coefficiente, al contrario della covarianza, è indipendente dalla scala utilizzata per misurare i valori delle variabili aleatorie, e quindi offre una migliore misura del "legame" tra X e Y.

Osserviamo che la covarianza soddisfa alla disequazione

$$\left(\mathrm{cov}(X,Y)\right)^2 \leq \mathrm{var}(X)\mathrm{var}(Y), \tag{4.30}$$

nota come *disequaglianza di Cauchy-Schwarz*. Il segno di uguale nella (4.30) vale se, e solo se, esistono tre costanti a, b, c, non tutte nulle, tali che $P(aX + bY = c) = 1$.

In virtù della (4.30) si ha

$$|\rho| \leq 1.$$

Confrontando con la (4.28), abbiamo che $\rho = 1$ o $\rho = -1$ se $Y = \alpha X + \beta$ con $\alpha > 0$ o $\alpha < 0$ rispettivamente. Un valore di $|\rho|$ vicino a uno è quindi indice di una relazione *lineare* tra Y e X.

Torneremo su questi concetti nel capitolo dedicato alla regressione lineare.

Capitolo 5

Distribuzioni discrete

5.1 Distribuzione binomiale

Molti tipi di esperimenti hanno in comune la caratteristica che i loro risultati possono essere raggruppati in due classi, generalmente indicate con i nomi convenzionali di *successo* e *insuccesso*. L'esempio paradigmatico è quello del lancio di una moneta, dove si può considerare, p.e. successo l'uscita di una "testa" e insuccesso l'uscita di una "croce".

La variabile aleatoria rilevante in questo tipo di esperimenti è quella che conta il numero di successi su un dato numero di ripetizioni *indipendenti* dello stesso esperimento.

Un'altra importartante caratteristica di molte serie di esperimenti è che i singoli eperimenti della successione sono indipendenti, ovvero l'esito di un esperimento **non** influenza gli esperimenti precedenti.

Questo è quanto avviene in una serie di lanci di una moneta. Il risultato del lancio $n+1$-esimo non è influenzato dai precedenti n lanci, nel senso che la probabilità di ottenere una testa o una croce non dipende da quante teste e quante croci si sono ottenute nei lanci precedenti. Inoltre, come è lecito assumere se si lancia sempre la stessa moneta, la probabilità di successo rimane invariata per ogni esperimento della succesione.

Questo tipo di esperimenti ripetuti è comunemente indicato con il nome di *prova di Bernoulli* (Bernoulli trial in inglese) dalla famiglia di matematici svizzeri Bernoulli che annovera tra i sui membri alcuni dei fondatori della teoria della probabilità.

Ripetiamo che le ipotesi fondamentali che stanno dietro alla assunzione che una serie di prove sia di Bernoulli sono che

- ogni esperimento della serie ha solo due possibili risultati (successo-insuccesso);
- i singoli esperimenti della serie sono *eventi indipendenti*;
- la probabilità di successo resta invariata da un esperimento all'altro. In genere si indica con $p \in (0,1)$ la probabilità di successo e con $q = 1-p$ quella di insuccesso.

A partire da queste assunzioni possiamo ricavare la distribuzione della variabile aleatoria X={numero di successi in n prove}.

Fissiamo il numero di prove, n, ed esaminiamo una singola sequenza di prove, che è un evento elementare del nostro spazio campionario delle successioni di n prove. Supponiamo, per esempio, di esaminare un evento E tale da far assumere alla variabile aleatoria X il valore k, ovvero una successione con k successi e $n-k$ insuccessi. Poiché i risultati delle singole prove sono indipendenti, la probablità di E è data dal prodotto delle probabilità degli esiti delle singole prove che lo compongono, ovvero

$$P(E) = \underbrace{pp...p}_{k\,\text{volte}} \ \underbrace{qq...q}_{n-k\,\text{volte}} = p^k(1-p)^{n-k}$$

Poiché ogni serie di lanci è un evento elementare, ci resta soltanto da contare quante sono le successioni di lanci con k successi. La risposta è il numero di combinazioni di k oggetti scelti da n oggetti, il cui numero è dato da

$$\left(\begin{array}{c} n \\ k \end{array} \right)$$

. Quindi la probabilità di ottenere k successi in n prove è

$$P(k \text{ successi}) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \tag{5.1}$$

Come controllo, verifichiamo che la probabilità di ottenere un numero qualsiasi di successi è 1. Poiché gli eventi $(k \text{ successi})$ e $(h \text{ successi})$ con $k \neq h$ hanno intersezione nulla, la probablità di ottenere un numero qualsiasi di successi tra 0 e n si ottiene sommando le probablità di ottenere k successi per $k = 0, 1, ..., n$, ottenendo

$$P(\#\text{ qualsiasi di successi}) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}.$$

Ricordando lo sviluppo di Newton per la potenza n-esima di un binomio, otteniamo

$$P(\#\text{ qualsiasi di successi}) = (p + (1-p))^n = 1.$$

Media e varianza della distribuzione binomiale Come ci si può aspettare ricordando quanto detto nel capitolo precedente, il valor medio di X sarà dato dal prodotto np, ovvero si ottengono "mediamente" np successi su n prove. Per dimostrarlo dobbiamo calcolare

$$\mu = \sum_{k=0}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}. \tag{5.2}$$

Ricordiamo ora la formula per il coefficiente binomiale

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

e osserviamo che il primo termine della somma in (5.2) può essere omesso in quanto moltiplicato per 0. Abbiamo quindi

$$k \left(\begin{array}{c} n \\ k \end{array} \right) = n \frac{((n-1)!}{(k-1)!(n-k)!}$$

e, tornando alla somma

$$\mu = \sum_{k=1}^{n} n \frac{((n-1)!}{(k-1)!(n-k)!} p^k (1-p)^{n-k}. \tag{5.3}$$

Ponendo $h = k - 1$ in (5.3) otteniamo

$$\mu = n\,p \sum_{h=0}^{n-1} \frac{((n-1)!}{h!(n-1-h)!} p^h (1-p)^{n-1-h}. \tag{5.4}$$

dove abbiamo messo in evidenza il fattor comune np nella somma. Infine possiamo riconoscere nella somma lo sviluppo del binomio $(p+(1-p))^{n-1} = 1$, il che completa la dimostrazione.

Per la varianza si può dimostrare che $\mathrm{var}(X) = npq = np(1-p)$.

Per completezza riportiamo la dimostrazione. Sfuttando la (4.11) dobbiamo mostrare che

$$\sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} = (np)^2 + np(1-p). \tag{5.5}$$

Riscriviamo il primo membro della (5.5) come

$$np \sum_{k=1}^{n} k \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) p^{k-1} (1-p)^{n-1-(k-1)}. \tag{5.6}$$

e poniamo $h = k - 1$ nella (5.6)

$$np \sum_{h=0}^{n-1} (h+1) \left(\begin{array}{c} n-1 \\ h \end{array} \right) p^h (1-p)^{n-1-h} = np[p(n-1)+1]. \tag{5.7}$$

dove abbiamo applichiamo nella (5.7) la proprietà distributiva del prodotto rispetto alla la somma $(h+1)$, ottenendo così la somma di due sommatorie: la prima (quella con fattore h) che è il valor medio della varaibile aletoria Y che conta i successi su una serie di $n-1$ lanci, l'altra (quella col fattore 1) che è il solito sviluppo del binomio di Newton per $(p+(1-p))^{n-1}$.

Nel seguito, per indicare che una variabile aleatoria X che obbedisce a una distribuzione binomiale su n prove con probabilità di successo p, scriveremo $X \sim \mathcal{B}(n,p)$. Ricapitoliamo i risultati ottenuti nella seguente tabella

$X \sim \mathcal{B}(n,p)$	
$P(X=k)$	$\left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$
$\mu = E(X)$	np
$\sigma^2 = \mathrm{var}(X)$	$np(1-p)$

Nota: E' bene chiarire un punto: la distribuzione binomiale, così come l'abbiamo descritta, assume che la struttura dell'esperimento sia soggiacente sia **nota**, e, in particolare, sia nota la probabilità di successo p. A partire da questa conoscenza possiamo calcolare "a priori" qual'è la probabilità di k sucessi su n prove.

Una serie di prove può essere usata invece per determinare "sperimentalmente" il valore della probabilità di successo p (p.e. nel caso si sospetti che una moneta sia truccata). Anche in questo caso dobbiamo comunque assumere alcune ipotesi: che le prove della serie siano indipendenti; che la probabilità di successo p (inognita) non cambi da una prova all'altra. Una volta assunte queste ipotesi, la variabile aleatoria che conta il numero dei sucessi segue una distribuzione binomiale con p incognita. La stima di p si può effettuare "a posteriori" tramite la stima del valor medio np ottenuta a partire dal valor medio campionario, come vedremo in seguito.

Distribuzione multinomiale Conderiamo ora il caso in cui i risultati possibili di una prova in una successione di prove siano più di due. Effettuaiamo una successione di n prove indipendenti, dove le probabilità dei singoli esiti si mantengano costanti come nel caso delle prove di Bernoulli viste prime. Supponiamo che gli esiti possibili siano almeno tre, e concentriamo la nostra attenzione sui due primi possibili esiti della prova, e_1 ed e_2, con probabilità p_1 e p_2 rispettivamente. Indichiamo con X_1 e con X_2 il numero degli esiti uguali a e_1 ed e_2 rispettivamente. Consideriamo la variabile aleatoria vettoriale $\mathbf{X} = (X_1, X_2)$ che ha per range l'insieme delle coppie di numeri interi non negativi (x_1, x_2) tali che $x_1 + x_2 \leq n$.

La distribuzione congiunta della variabile vettoriale $\mathbf{X}$ è la distribuzione *trinomiale* di parametri n e (p_1, p_2) data da

$$P(X_1 = x_1, X_2 = x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}$$

che soddisfa $E(X_i) = np_i,\ \mathrm{var}(X_i) = np_i(1-p_i),\ i = 1, 2$ (queste uguaglianze sono ovvie alla luce della distribuzione binomiale) e $\mathrm{cov}(X_1, X_2) = -np_1p_2$. Il fatto che la covarianza sia negativa è giustificato dal fatto che, se si ottengono "molti" risultati e_1, di conseguenza ci si aspettano "pochi" risultati e_2.

Distribuzione geometrica Una distribuzione legata alla binomiale, è la distribuzione geometrica che è quella a cui obbedisce la variabile aleatoria X che conta, in una successione di prove di Bernoulli indipendenti di probabilità p, il numero di fallimenti prima di ottenere il primo successo, p.e. se l'uscita di testa è il *successo*, la serie $\{C, C, T, C, T\}$ ci dà $X = 2$.

Si ha $P(X = r) = p(1 - p)^r,\ r = 0, 1, 2, ...$ e $E(X) = \frac{1-p}{p}$ e $\mathrm{var}(X) = \frac{1-p}{p^2}$.

Distribuzione ipergeometrica Consideriamo un'urna con N_1 palline bianche e N_2 palline nere e poniamo $X = \{\text{numero di palline bianche estratte}\}$. Allora, in una successione di n estrazioni, la variabile X sarà distribuita secondo $\mathcal{B}(N, p),\ N = N_1 + N_2,\ p = N_1/N$, se, dopo ogni estrazione, *la pallina estratta viene reintrodotta nell'urna*; diremo in questo caso che abbiamo effettuato una serie di estrazioni con reintroduzione.

Se, invece, effettuiamo le estrazioni *senza reintroduzione*, la distribuzione binomiale non può essere usata poiché la probabilità non è costante da una prova all'altra (e le prove non sono indipendenti tra loro).

In questo caso la variabile X deve soddisfare la limitazione $0 \leq X \leq N_1$, in quanto non si possono avere più di N_1 successi e $n - N_2 \leq X$, in quanto si possono avere al più N_2 fallimenti (questa disequaglianza diventa significativa se ci sono più di N_2 estrazioni).

Si può dimostrare che la variabile X obbedisce alla distribuzione ipergeometrica

$$P(X=x)=\frac{\left(\begin{array}{c}Np\\x\end{array}\right)\left(\begin{array}{c}N(1-p)\\n-x\end{array}\right)}{\left(\begin{array}{c}N\\n\end{array}\right)}$$

per ogni x tale che $\max(0,n-N(1-p))\leq x\leq\min(n,Np)$. Si ha

$$E(X)=np,\quad \mathrm{var}(X)=\frac{np(1-p)(N-n)}{N-1}.$$

Se il numero N è molto grande rispetto a n allora la distribuzione ipergeometrica si confonde alla distribuzione binomiale (e ci tende nel senso delle distribuzioni se $N\to+\infty$).

Ciò significa che nei campionamenti da popolazioni molto numerose, e con campioni poco numerosi rispetto alla numerosità della popolazione, si può utilizzare, al posto della distribuzione ipergeometrica, la distribuzione binomiale per la variabile che conta il numero di successi nel campione (questa approssimazione, o meglio, come vedremo, l'approssimazione tramite la distribuzione normale, è comunemente usata nelle applicazioni statistiche). Questo fatto non deve sorprendere: tornando al caso delle palline, se nell'urna ci sono, p.e., 20000 palline bianche e 30000 palline nere, l'estrazione di 10 palline non altera sensibilmente, qualunque sia l'esito delle estrazioni, la probabilità di estrarre una pallina bianca alla undicesima estrazione.

5.2 Distribuzione di Poisson

Se il numero di prove n in una prova di Bernoulli è molto grande, ma la probabilità di successo p molto piccola (paragonabile a $1/n$) allora la distribuzione binomiale è approssimata dalla *distribuzione di Poisson*.

Questa distribuzione, oggi largamente usata per la descrizione dei problemi di arrivo (p.e. le code a un casello autostradale o a uno sportello postale), fu introdotta dal matematico francese S.D. Poisson (1781-1840).

Una variabile aleatoria X si dice distribuita secondo Poisson con parametro $m>0$ se la sua funzione di probabilità è data da

$$P(X=x)=e^{-m}\frac{m^x}{x!},\ x=0,1,2,....,\quad P(X=0)\ \text{altrimenti.}\tag{5.8}$$

Scriveremo in questo caso $X\sim\mathcal{P}(m)$.

Osserviamo che la distribuzione di Poisson può essere ottenuta dalla seguente formula di ricorrenza

$$P(X=0)=e^{-m},\quad P(X=x+1)=P(X=x)\frac{m}{x+1}.$$

Il valore atteso $E(X)$ di una distribuzione di Poisson è dato dal valore del parametro stesso e così pure la sua varianza

$$\mu=E(X)=m,\quad \sigma^2=\mathrm{var}(X)=m.\tag{5.9}$$

Che la distribuzione di Poisson soddisfi la condizione di normalizzazione $\sum_{i=1}^{\infty} P(X = i) = 1$, non che le (5.9), può essere verificato abbastanza agevolmente ricordando che $\sum_{i=1}^{\infty} \frac{m^i}{i!} = e^m$. I calcoli sono lasciati per esercizio.

Processi di Poisson

Nelle moderne applicazioni ai problemi di arrivo (auto che arrivano in un'ora a una stazione di servizio, numero di clienti che entrano in un negozio in un periodo di tempo fissato, tasks inviati a una unità di stampa di un calcolatore per minuto, numero di impulsi ricevuti da una cellula nervosa per unità di tempo, ma anche numero di errori di stampa per pagina o rigo in un libro, numero di pezzi difettosi prodotti al giorno da un impianto, etc.) il quadro di applicabilità della distribuzione di Poisson può essere descritto come segue.

Consideriamo un fenomeno che si svolge nel tempo (o nello spazio). Uno o più eventi possono prodursi in un intervallo di ampiezza T con una data probabilità; indichiamo con $X(\Delta)$ la variabile aleatoria che conta le occorrenze (il numero degli eventi verificatesi) in un sottointervallo di ampiezza Δ, e assumiamo che:

1. $X(\Delta)$ dipende solo dall'ampiezza dell'intervallo e non dal suo istante iniziale e questa propabilità resta costante per tutto il processo;

2. Le occorrenze in ogni sottointervallo sono indipendenti, ovvero se $\Delta_1, \Delta_2, ...$ sono sottointervalli disgiunti, le variabili $X(\Delta_1), X(\Delta_2), ...$ sono variabili aleatorie indipendenti;

3. La probabilità di una singola occorrenza in un sottointervallo "piccolo" è proporzionale all'ampiezza δ del sottointervallo, ovvero $P(X(\delta) = 1) = \lambda\delta + o(\delta)$ dove $\lim_{\delta \to 0} o(\delta)/\delta = 0$;

4. La probabilità che in un sottointervallo "piccolo" accadano due o più eventi è sufficientemente piccola da poter essere trascutata, ovvero $P(X(\delta) > 1) = o(\delta)$.

Un processo che soddisfa (entro un ragionevole margine di approssimazione) questo modello prende il nome di *processo di Poisson*

La probabilità di un'occorrenza in un sottointervallo "piccolo" di ampiezza δ essendo uguale a $\lambda\delta$, ci dice che λ rappresenta il "limite" del numero medio di occorrenze quando l'ampiezza δ tende a zero: la chiameremo quindi *velocità* del processo di Poisson. Di conseguenza, poiché questa velocità è costante durente il processo, per ogni unità di tempo (o spazio) usata nel nostro processo avvengono, in media, λ eventi. (Nota che la nozione di intervallo "piccolo" dipende dal fenomeno in esame: un intervallo temporale di un secondo può essere considerato piccolo se prendiamo in esame l'arrivo di un cliente in un negozio (non in supermercato) ma non se prendiamo in esame i tasks inviati alla CPU di un computer che svolge un milione di operazioni al secondo).

λ è l'unico parametro necessario a caratterizzare questa distribuzione. Nelle applicazioni esso deve essere determinato sperimentalmente.

Una volta noto λ la funzione di probabilità per la variabile aleatoria $X(t)$, numero di occorrenze in un intervallo di tempo (o spazio) di ampiezza t, è data da una distribuzione di Poisson con parametro $m = \lambda t$, i.e. $X(t) \sim \mathcal{P}(\lambda t)$.

$$P(X(t) = x) = e^{-\lambda t} \frac{(\lambda t)^x}{x!}, \quad x = 0, 1, 2, ... \tag{5.10}$$

L'introduzione della distribuzione di Poisson può apparire vagamante magica, quindi è bene dare una spiegazione del perché essa sia in grado di dar conto di quelli che abbiamo chiamato processi di Poisson.

Come abbiamo osservato, si tratta di processi che involvono eventi *rari* ma che vengano contati su un gran numero di osservazioni del fenomeno. Come abbiamo visto nella sezione precedente, il numero di eventi in una successione di esperimenti lunghezza n, quando ogni singolo evento può accadere con probabilità p (e si abbia indipendenza dei singoli esperimenti) obbedisce alla distribuzione binomiale con parametri n e p. Il caso che ci interessa è quello di "lunghe" successioni e di "piccola" probabilità, ovvero n grande e p piccolo. In effetti, preso un intervallo temporale finito di ampiezza t, lo si divida in n intervalli "piccoli" in cui si possa assumere che la probabilità di una singola occorrenza sia $p = \lambda t/n$. Poiché assumiamo anche che non possano verificarsi due o più occorrenze in un intervallino, e che le occorrenze in intervallini distinti siano indipendenti tra loro, questo processo diventa una prova di Bernoulli con probabilità p e n prove (abbiamo un successo se nell'intervallino l'evento si verifica, un insuccesso se no).

La grandezza di n e la piccolezza di p sono legate tra loro dal fatto che il prodotto $np = \lambda t = m$ è fissato dalla costanza della velocità del processo.

A questo punto guardiamo cosa succede alla distribuzione binomiale $\mathcal{B}(n, p)$ quando $n \to \infty$ con $p = m/n$ e k (numero delle occorenze) resta fissato.

$$\begin{aligned} &\lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{m}{n}\right)^k \left(1 - \frac{m}{n}\right)^{n-k} \\ &= \frac{m^k}{k!} \lim_{n\to\infty} \frac{(n-k+1)(n-k+2)...n}{n^k} \lim_{n\to\infty} \left(1 - \frac{m}{n}\right)^{n-k} \\ &= \frac{m^k e^{-m}}{k!} \end{aligned}$$

ovvero la distribuzione binomiale ammette come limite la distribuzione di Poisson quando $n \to \infty$ e il prodotto $np = m$ si mantiene costante.

Tabella ricapitolativa della distribuzione di Poisson

$X \sim \mathcal{P}(m)$	
$P(X = k)$	$\frac{e^{-m} m^k}{k!}$
$\mu = E(X)$	m
$\sigma^2 = \mathrm{var}(X)$	m

Figura 5.1:

Figura 5.2:

Capitolo 6

Distribuzioni continue

6.1 Distribuzione normale

La distribuzione di gran lunga più importante nelle applicazioni è la cosiddetta *distribuzione normale* o di Gauss in onore di K.F. Gauss (1777-1855). Essa compare in una molteplicità di occasioni. Da un lato essa descrive la distribuzione degli errori in un processo di misurazione sperimentale, ovvero come le misure effettivamente osservate si scostino valore "vero" della quantità che si vuol misurare. Dall'altro essa fornisce un'utilissima approssiamazione sia della distribuzione binomiale sia di quella di Poisson. Infine, tramite il *Teorema centrale del limite*, essa compare come distribuzione asintotica della media campionaria estratta da una popolazione di cui siano note valor medio e varianza, *qualunque sia la distribuzione originaria* da cui è estratto il campione.

Una variabile aleatoria X è distribuita *normalmente* con valor medio μ e varianza $\sigma^2 > 0$ quando X ha densità

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbf{R},\ \mu \in \mathbf{R},\ \sigma > 0. \tag{6.1}$$

Scriveremo $X \sim \mathcal{N}(\mu, \sigma^2)$. Quando $\mu = 0$ e $\sigma = 1$, la distribuzione si dirà *normale standard*. Indicheremo spesso con Z una variabile distribuita secondo la normale standard, $Z \sim \mathcal{N}(0, 1)$.

Il grafico della funzione f è simmetrico rispetto alla retta $x = \mu$ (il valor medio) che è anche l'unico punto di massimo della f. La curva ha due flessi in $\mu - \sigma$ e $\mu + \sigma$.

6.1.1 Standardizzazione

La distribuzione normale ha due parametri μ e σ. I valori di una variabile X distribuita normalmente sono usualmente descritti in termini di *quante volte la deviazione standard* essi sono lontani dal valor medio. Si ha infatti che la probabilità che X sia contenuta in un intervallo centrato in μ e di ampiezza t volte la deviazione standard, ovvero

$$P(\mu - t\sigma \leq X \leq \mu + t\sigma),$$

non dipende da μ e σ ma solo da t.

Per verificare questa proprietà basta ricorre al seguente cambiamento di variabili $z = (x - \mu)/\sigma$. Se $X \sim \mathcal{N}(\mu, \sigma^2)$ allora la variabile aleatoria

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \tag{6.2}$$

è distribuita secondo la normale standard e si ha

$$\int_{\mu+a\sigma}^{\mu+b\sigma} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

La standardizzazione è fondamentale nelle applicazioni in quanto consente di calcolare quale sia la probabilità che una variabile normalmente distribuita assuma valori in un certo intervallo riccorrendo alle tavole per la distribuzione normale standard.

Ricordiamo che la funzione di distribuzione per la normale standard è definta da

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds. \tag{6.3}$$

Questo funzione integrale non è esprimibile tramite combinazioni finite di funzioni elementari, quindi non è possibile eseguire il calcolo di $F(z)$ con i metodi abituali del calcolo integrale. Quindi per sapere quanto valga $F(z)$ per un dato z si ricorre a delle tavole che riportano il valore di $F(z)$ (oppure di $F(z) - 0.5 = P(0 \leq Z \leq z)$)

Queste stesse tabelle possono essere utilizzate per calcolare anche $P(X \leq x)$ quando $X \sim \mathcal{N}(\mu, \sigma^2)$ attraverso la standardizzazione.

Per chiarire il procedimento vediamo un semplice esempio. Supponiamo che X sia una variabile aleatoria distribuita normalmente, con valor medio $\mu = 3$ e varianza $\sigma^2 = 4$. Vogliamo sapere quanto vale la probabilità che X assuma valori nell'intervallo $[1, 4]$ ovvero vogliamo calcolare $P(1 \leq X \leq 4)$.

Per prima cosa *standardizziamo* il problema: poiché $X \sim \mathcal{N}(3, 4)$ la (6.2) ci dice che la variabile distribuita secondo la normale standard sarà

$$Z = (X - 3)/2. \tag{6.4}$$

A questo punto dobbiamo trasformare la diseguaglianza

$$1 \leq X \leq 4$$

in una disguaglianza per Z.

$$1 - 3 \leq X - 3 \leq 4 - 3, \Rightarrow -1 \leq \frac{X - 3}{2} = Z \leq 0.5,$$

che ci dice che la X è compresa nell'intervallo $[1, 4]$ se e solo se la Z è compresa nell'intervallo $[-1, 0.5]$. Quindi

$$P(1 \leq X \leq 4) = P(-1 \leq Z \leq 0.5) = p$$

Per valutare $p = P(-1 \leq Z \leq 0.5)$ possiamo ora ricorre alle tavole. In un testo di statistica si può trovere una delle seguenti tre tavole

1. valori della funzione di distribuzione $F(z)$ per $z \geq 0$, ovvero i valori di $P(Z \leq z)$ per $z \geq 0$

2. valori della funzione $F(z) - 0.5$ per $z \geq 0$, ovvero i valori di $P(0 \leq Z \leq z)$ per $z \geq 0$

3. valori della *coda* della distribuzione di Z per $z \geq 0$, ovvero i valori di $P(Z \geq z)$ per $z \geq 0$

Ovviamente i valori ricavati da una qualunque di queste tavole possono essere trasformati facilmente nei valori riportati da un altra: p.e. se la tavola è del tipo (3) e abbiamo $z = 1$ otteniamo un valore per $P(Z \geq 1) \simeq 0.159$ (tutti i valori su queste tavole sono approssimati con un numero di decimali esatti che dipende dalla accuaratezza delle tavole), mentre se la tavola è del tipo (2) otteniamo, per $z = 1$, il valore per $P(0 \leq z \leq 1) \simeq 0.3413$. Il legame tra i due valori è dato da $0.5 + 0.3413 = 0.8413 \simeq 0.841 = 1.0 - 0.159$ (non ottengo esattamente lo stesso valore a causa degli errori di arrotondamento) come ci dice l'uguaglianza $P(Z \leq 0) + P(0 \leq Z \leq 1) = P(Z \leq +\infty) - P(Z \geq 1)$.

Supponiamo ora di avere una tavola del tipo (1) e calcoliamo la probabilità $p = P(-1 \leq Z \leq 0.5)$. Abbiamo

$$p = P(Z \leq 0.5) - P(Z \leq -1). \tag{6.5}$$

Il primo addendo nella (6.5) si ricava direttamente dalle tavole: $P(Z \leq 0.5) \simeq 0.6915$. Il secondo valore non si può ricavare direttamente poiché $-1 < 0$. Ma la distribuzione di Z è simmetrica rispetto al suo valor medio $\mu = 0$, quindi $P(Z \leq -1) = P(Z \geq 1)$ e, a sua volta, $P(Z \geq 1) = 1 - P(Z \leq 1)$. Quest'ultimo valore è riportato nella tavola $P(Z \leq 1) \simeq 0.8413$. Infine il valore cercato è $p \simeq 0.6915 - (1 - 0.8413) = 0.5328$.

Riassumiamo tutto quanto in uno schema

Problema: Calcolare $P(a \leq X \leq b)$ con $X \sim \mathcal{N}(\mu, \sigma^2)$

↓

1: standardizzare la variabile $Z = \dfrac{x - \mu}{\sigma} \sim \mathcal{N}(0, 1)$

↓

2: trasformare $P(a \leq X \leq b)$ in $P(\alpha = \sigma a + \mu \leq Z \leq \sigma b + \mu = \beta)$

↓

3: trasformare $P(\alpha \leq Z \leq \beta)$ in disuguaglianze adatte alle proprie tavole

ovvero

3.1: tavola tipo (1) $P(\alpha \leq Z \leq \beta) = P(Z \leq \beta) - P(Z \leq \alpha)$

3.2: tavola tipo (2)
$P(\alpha \leq Z \leq \beta) = P(Z \leq \beta) - P(Z \leq \alpha)$, se $0 < a < b$
$P(\alpha \leq Z \leq \beta) = P(Z \leq \beta) + P(Z \leq -\alpha)$, se $a < 0 < b$
$P(\alpha \leq Z \leq \beta) = P(Z \leq -\alpha) + P(Z \leq -\beta)$, se $a < b < 0$

3.3: tavola tipo (3) $P(\alpha \leq Z \leq \beta) = P(Z \geq \alpha) - P(Z \geq \beta)$

6.1.2 Approssiamazione tramite la distr. normale

La distribuzione normale può essere utilizzata per il calcolo della distribuzione binomiale. Si ha infatti

Teorema 6.1 (DeMoivre-Laplace) *Sia $X_n \sim \mathcal{B}(n,p)$. Allora*

$$\frac{X_n - np}{(np(1-p))^{1/2}} \to Z \sim \mathcal{N}(0,1) \quad \text{per } n \to \infty$$

nel senso che $\lim_{n\to\infty} P\left(a \leq \frac{X_n - np}{(np(1-p))^{1/2}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{z^2/2} dz$.

Il significato del teorema 6.1 è che, se la variabile X_n è distribuita secondo la distribuzione binomiale, allora per n grande, la variabile aleatoria X_n è distribuita in modo approssimabile con la distribuzione normale standard, ovvero

$$\frac{X_n - np}{(np(1-p))^{1/2}} \approx \mathcal{N}(0,1). \tag{6.6}$$

Quanto grande deve essere n perché questa approssimazione sia accettabile nella applicazioni? Comunemente si accetta questa approssimazione se si ha $np \geq 5$ e $n(1-p) \geq 5$ (in alternativa a questa "regola" si ha $np(1-p) \geq 3$). Se n è grande ma $np < 5$ ricordiamo che la distribuzione binomiale viene approssimata dalla distribuzione di Poisson. Nel calcolo bisogna fare attenzione al fatto che la distribuzione binomiale è una distribuzione discreta mentre la normale è una distribuzione continua. Per usare la (6.6) occorre introdurre un corretivo che consiste nel sostiuire alla probalilità che $X_n = k$ con l'evento $k - 0.5 \leq X_n \leq k + 0.5$ prima di effettuare il calcolo.

Per illustrare questa procedura supponiamo di voler calcolare la probabilità che il numero 3 esca almeno 6 volte in una successione di 30 lanci di un dado non truccato. Usando la distribuzione binomiale dovremmo calcolare la somma

$$\sum_{k=6}^{30} P(X=k) = \sum_{k=6}^{30} \left(\begin{array}{c} 20 \\ k \end{array} \right) (1/6)^k (5/6)^{30-k}$$

che è assai laboriosa da calcolare (esistono anche tavole per la distribuzione binomiale ma sono di faticosissima lettura!). Poiché $n = 30$ e $p = 1/6$ abbiamo $np = 5$ e $n(1-p) = 25$ (e anche $np(1-p) = 25/6 > 3$) possiamo applicare l'approssimazione (6.6). In accordo con quanto detto si tratta quindi di calcolare $P(5.5 \leq X \leq 30.5)$ e quindi $P(0.245 \leq Z \leq 12.492)$ dove abbiamo sostituito la variabile Z distribuita secondo la normale standard per $\frac{X-np}{(np(1-p))^{1/2}} = \sqrt{6}\frac{X-5}{5}$. Inoltre, poiché la coda della distribuzione normale a destra di 12.492 ha area trascurabile, possiamo limitarci a calcolare $P(0.245 \leq Z) \simeq 0.403$, ovvero una probabilità di circa il 40%.

6.1.3 Altre proprietà della distr. normale

Teorema 6.1 *Siano $X_1, X_2, ..., X_N$ variabili aleatorie indipendenti distribuite normalmente con medie $\mu_1, \mu_2, ..., \mu_N$ e varianze $\sigma_1^2, \sigma_2^2, ..., \sigma_N^2$ rispettivamente. Allora la variabile aleatoria $X = a_1X_1 + a_2X_2 + ... + a_NX_N + b$ è ancora una variabile distribuita normalmente con media $\mu = \sum_{i=1}^N a_i\mu_i + b$ e varianza $\sigma^2 = \sum_{i=1}^N a_i^2\sigma_i^2$.*

In particolare questa proprietà può essere applicata alla media aritmetica

$$X = \frac{X_1 + X_2 + ... + X_N}{N}$$

di N variabili aleatorie indipendenti aventi la stessa media μ e la stessa varianza σ, tutte distribuite normalmente. In questo caso otteniamo che $X \sim \mathcal{N}(\mu, \sigma/N)$, ovvero

la media aritmetica delle variabili è distribuita normalmente e ha la stessa media delle variabili X_i e varianza uguale a un N-esimo della varianza delle variabili X_i.

Questa proprietà ha una generalizzazione di fondamentale importanza nelle applicazioni statistiche.

Teorema 6.2 (Teorema Centrale del Limite) *Sia $X_1, X_2, ..., X_N$ una successione di variabili aleatorie indipendenti con la stessa media μ e la stessa varianza σ. Allora la distribuzione della variabile aleatoria*

$$X = \frac{X_1 + X_2 + ... + X_N}{N}$$

tende alla distribuzione normale con media μ e varianza σ^2/N per $N \to \infty$.

In altre parole si ha, per ogni $x \in \mathbf{R}$

$$\lim_{N\to\infty} P\left(\frac{X_1 + X_2 + ... + X_N - N\mu}{\sigma\sqrt{N}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{6.7}$$

E' fondamentale osservare che il teorema non richiede altro alle singole distribuzioni delle variabili aleatorie, a parte di avere la stessa media e la stessa varianza. La "bontà" dell'approssimazione, ovvero quanto grande debba essere N affinché si possa trascurare l'errore commesso sustituendo la distribuzione normale a quella della media atritmetica, dipenderà tuttavia dalla forma delle distribuzioni (abbiamo visto che nel caso di distribuzioni normali non si commette alcun errore).

Vedremo come si utilizza questo teorema nel capitolo dedicato ai compionamenti.

6.2 Distribuzione esponenziale

La distribuzione esponenziale è una distribuzione continua legata ai processi di Poisson. Quando abbiamo introdotto i processi di Poisson, abbiamo definito la distribuazione di Poisson che "conta" quanti eventi accadono in un dato intervallo temporale (o spaziale) di ampiezza t. Ricordiamo che la probabilità che in un processo di Poisson di velocità λ si verifichino k eventi nell'intervallo $(T, T + t)$ è data da

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

da cui otteniamo che la probabilità che si verifichi almeno un evento in un intervallo $(T, T + t)$ è

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda t}. \tag{6.8}$$

L'ultimo termine nell'uguaglianza (6.8) è una funzione di t che possiamo considerare la *probabilità di dover aspettare al più un tempo t prima che si verifici un nuovo evento*. La funzione

$$F(t) = \begin{cases} 1 - e^{-\lambda t} & \text{per } t \geq 0 \\ 0 & \text{per } t < 0 \end{cases} \tag{6.9}$$

soddisfa le proprietà di una funzione di distribuzione con densità

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{per } t \geq 0 \\ 0 & \text{per } t < 0 \end{cases} \tag{6.10}$$

e prende il nome di distribuzione esponenziale.

La media e la varianza della distribuzione esponenziale sono date da

$$\boxed{\mu = \frac{1}{\lambda}, \quad \sigma^2 = \frac{1}{\lambda^2}}$$

6.3 La distribuzione χ^2

Una distribuzione di frequente impiego, come vedremo, nei test statistici è la distribuzione χ^2 (chi-quadro). Essa è strettamente legata alla distribuzione normale. Siano $X_i,\ i = 1, ..., n$ variabili aleatorie indipendenti distribuite normalmente, $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ e siano $Z_i = \frac{X_i - \mu_i}{\sigma_i}$ le corrispondenti variabili standardizzate. Consideriamo ora la nuova variabile aleatoria

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2$$

ottenuta sommando i quadrati delle variabili aletorie Z_i. Questa variabile, che ovviamente può assumere solo valori positivi, dà una misura dello scarto totale delle variabili aleatorie dalla loro media. Il pedice n serve per ricordare che abbiamo preso la somma di n variabili. Esso ha la funzione di parametro per la distribuzione della variabile aleatoria χ_n^2 e viene detto *grado di libertà* della distribuzione.

La funzione di densità per la distribuzione della χ_n^2 si ricava abbastanza agevolmente. Partiamo costruendo la funzione di distribuzione della χ_n^2: per $x < 0$ ovviamente avremo $P(\chi_n^2 \leq x) = 0$. Vogliamo ora calcolare $P(\chi_n^2 \leq x)$, per ogni $x > 0$ ovvero

$$P(\chi_n^2 \leq x) = P(\sum_{i=1}^{n} Z_i^2 \leq x).$$

Possiamo pensare alla $\sum_{i=1}^{n} Z_i^2$ come una funzione della variabile aleatoria vettoriale $(Z_1, ..., Z_n)$ che ha distribuzione nota poiché le variabili aleatorie Z_i sono indipendenti. Quindi per calcolare $P(\chi_n^2 \leq x)$ basta integrare la funzione di densità congiunta della variabile $(Z_1, ..., Z_n)$ sulla ipersfera n dimensionale definita da $\sum_{i=1}^{n} Z_i^2 \leq x$.

Il risultato che si ottiene è che

$$P(\chi_n^2 \leq x) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^x t^{n/2-1} e^{-t/2} dt, \quad x \geq 0,$$

dove la Γ è la funzione gamma di Eulero.

In ogni caso la conoscenza della funzione di densità ha un scarsa importanza. I valori di $P(\chi_n^2 \leq x)$ si ricavano da opportune tabelle (una per ogni grado di libertà).

La media e varianza della distribuzione χ_n^2 sono date da

$$\boxed{\mu = n, \quad \sigma^2 = 2n}$$

Capitolo 7

Campionamenti

Lo scopo principale della statistica induttiva è quello di stimare i parametri di una popolazione, o di sottoporre a esame delle ipotesi su di una popolazione, tramite l'osservazione di un numero ridotto di elementi della popolazione stessa: un *campione*.

Un campione consiste in una collezione finita di osservazioni, ognuna delle quali rappresenta la realizzazione di una variabile aleatoria $\boldsymbol{x}_i$, $i = 1, .., n$.

Per esempio, supponiamo di voler determinare l'altezza media dei giovani maschi italiani. Possiamo usare come campione le altezze misurate durante una visita di leva in qualche distretto militare italiano: l'altezza misurata per ogni recluta fornisce così la realizzazione di una variabile aletoria (l'altezza di un elemento scelto a caso dalla popolazione). L'insieme di queste osservazioni è il nostro campione.

A partire da questi dati possiamo formare una quantità che pensiamo ci possa fornire qualche indicazione sul parametro che ci interessa stimare (nel nostro caso l'altezza media). Questa quantità si dice una *statistica*. Si tratta, ancora una volta, di una variabile aletoria, funzione delle n variabili aleatorie $\boldsymbol{x}_i$ del nostro campione.

Il problema che ci poniamo per primo è quello di definire, per i vari parametri della popolazione, quali siano le statistiche giuste per stimare questi parametri, e come queste statistiche siano distribuite (in quanto variabili aleatorie). La conoscenza della distribuzione delle statistiche ci permetterà di valutare la nostra *fiducia* sulla bontà della stima.

Tornando all'esempio dell'altezza media, è abbastanza ovvio pensare che una buona stima dell'altezza media della popolazione sia fornita dall'altezza media del campione. Questa si misura come la media di una popolazione, facendo attenzione al modo in cui sono stati raggruppati i dati (se lo si è fatto). Questa statistica prende il nome di *media campionaria* e si indica con $\overline{\boldsymbol{x}}$. La sua realizzazione su un campione è data da:

$$\text{Media Campionaria}: \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{c} x_j f_j, \tag{7.1}$$

dove la prima somma si estende a tutti gli elementi del campione, sommando tutte le singole realizzazioni x_i delle variabili aleatorie $\boldsymbol{x}_i$, mentre la seconda somma si utilizza nel caso in cui i dati del campione siano stati raggruppati in c classi corrispondenti ai valori x_j e con frequenze f_j (nota che $\sum_{j=1}^{c} f_j = n$).

Accanto alla media campionaria, consideriamo un'altra statistica utile nelle appli-

cazioni, la varianza campionaria $\boldsymbol{s}^2$, la cui realizzazione è:

$$\text{Varianza Campionaria}: \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{j=1}^{c}(x_j - \overline{x})^2 f_j, \quad (7.2)$$

dove le due somme somme sono definite come sopra.

Notiamo due differenze nella (7.2) rispetto alla definizione della varianza di una popolazione. La prima è che in (7.2) gli scarti sono calcolati rispetto alla media campionaria e non alla vera media della popolazione μ, che in generale non è nota. La seconda differenza è che si divide per il fattore $(n-1)$ anziché per n. Questo è dovuto al fatto che in questo modo $\boldsymbol{s}^2$ diventa un buon stimatore (nel senso che verrà spiegato poi) della vera varianza della popolazione.

Accanto a (7.2) si ha la seguente formula di calcolo

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{n-1}\left(\sum_{j=1}^{c} x_j^2 f_j - n\overline{x}^2\right), \quad (7.3)$$

Definiamo infine *deviazione standard campionaria* s la radice quadrata della varianza campionaria.

Ricordiamo che sia $\overline{\boldsymbol{x}}$ che $\boldsymbol{s}^2$ sono due variabili aleatorie, il loro valore cioè cambia a ogni campionamento in modo aleatorio. Per fare delle previsioni è quindi necessario conoscere come queste variabili aleatorie sono distribuite. Come vedremo, possiamo rispondere a questa questione in modo "esatto" in alcuni casi significativi, e in modo "approssimato" nella maggioranza dei casi di interesse pratico.

Cominciamo da due definizioni di carattere generale:

Definizione *Diremo che una statistica y è uno stimatore* **non distorto** *di un parametro p di una popolazione se il suo valor medio $E(y)$ coincide con p.*

Definizione *Diremo che una statistica y è uno stimatore* **consistente** *di un parametro p di una popolazione se converge in probabilità a p quando la dimensione del campione tende all'infinito.*

Un primo risultato è il seguente:

Teorema *Siano $\overline{x}_i,\ i = 1, ..n$ n variabili aleatorie indipendenti aventi la stessa media μ. Allora la media campionaria $\overline{\boldsymbol{x}}$ è uno stimatore non distorto della media μ.*

La dimostrazione di questo teorema si riduce ad osservare che il valor medio della media di n variabili aleatorie indipendenti è uguale alla media dei valor medii.

Questo teorema ci dice che se effetuiamo un campionamento da una popolazione tramite n osservazioni indipendenti, allora la media campionaria $\overline{\boldsymbol{x}}$ sarà uno stimatore non distorto della media μ della popolazione soggiacente, ovvero: il valor medio della media campionaria è uguale alla media della popolazione.

E' bene stare attenti al possibile pasticcio linguistico che può generare l'affermazione precedente. La media campionaria $\overline{\boldsymbol{x}}$ è una variabile aleatoria, e come tale ha un suo valor medio $E(\overline{\boldsymbol{x}})$, che spesso è indicato con $\mu_{\overline{\boldsymbol{x}}}$. Il teorema ci dice che questo numero coincide con il valor medio μ della popolazione da cui si effettua il campionamento,

$$\mu_{\overline{\boldsymbol{x}}} = \mu. \quad (7.4)$$

Ancora rifacendoci alla teoria delle variabili aleatorie, possiamo calcolare la varianza della variabile aleatoria $\overline{\boldsymbol{x}}$, nell'ipotesi che il campionamento sia fatto di n osservazioni indipendenti estratte da una popolazione di varianza σ. In questo caso la

varianza di $\overline{\boldsymbol{x}}$, indicata con $\mathrm{Var}(\overline{\boldsymbol{x}})$ o con $\sigma^2_{\overline{\boldsymbol{x}}}$ è uguale a $1/n$ per la varianza della popolazione soggiacente

$$\sigma^2_{\overline{\boldsymbol{x}}} = \frac{1}{n}\sigma^2. \tag{7.5}$$

In particolare (7.5) ci dà una dimostrazione di un altro risultato sulla media campionaria:*la media campionaria è uno stimatore consistente della media della popolazione.* Infatti quando $n \to \infty$ la varianza di $\overline{\boldsymbol{x}}$ tende a zero, e quindi la $\overline{\boldsymbol{x}}$ tende, in probabilità, al suo valor medio.

Anche qui è bene fare attenzione ai pasticci linguistici: la varianza della media campionaria, $\sigma^2_{\overline{\boldsymbol{x}}}$, è una cosa totalmente diversa dalla varianza campionaria, $\boldsymbol{s}^2$.

Questi risultati sono di carattere generale e non fanno alcuna ipotesi sulla distribuzione della popolazione soggiacente al campionamento. D'altra parte essi non ci danno alcuna informazione su come sia distribuita la variabile aleatoria $\overline{\boldsymbol{x}}$. In questo modo non è possibile valutare con precisione quale sia la "bontà" della stima di μ fatta a partire da un valore di $\overline{x}$ ottenuto da un campionamento (la sola stima possibile, se si conoscono solo il valor medio e la varianza di una distribuzione, è quella che si ottiene dalla disuguaglianza di Čebišev, che non è molto precisa).

7.1 Popolazione normale

Nel caso che la popolazione soggicente sia distribuita normalmante allora possiamo migliorare sensibilmente la situazione. Infatti in questo caso la variabile aleatoria media campionaria ha una distribuzione nota.

Teorema *Se la popolazione da cui fa il campionamento è distribuita secondo una distribuzione normale di media μ e di varianza σ^2, e il campionamento consiste di n osservazioni indipendenti, allora la variabile aleatoria $\overline{\boldsymbol{x}}$ è distribuita normalmente con media μ e varianza σ^2/n*

(la novità è che ora sappiamo che $\overline{\boldsymbol{x}}$ è distribuita normalmente) La dimostrazione di questo teorema è un "esercizio" sulle variabili aleatorie (anche se un po' complicato). Quello che è importante è l'uso che di esso si può fare.

7.1.1 Popolazione normale, σ^2 nota

Supponiamo di sapere che la popolazione da cui si fa il campionamento sia una popolazione distribuita normalmente, di cui ci sia ignoto il valor medio μ, ma di cui si conosca la varianza σ^2.

Osserviamo che l'assunzione che la popolazione soggiacente sia normale è soddisfatta in molti casi reali, ovvero in molti casi si sa (o almeno è lecito assumere) che la variabile aleatoria che stiamo campionando è distribuita secondo la normale, anche se non ne conosciamo i parametri, valor medio e varianza. L'assunzione di conoscere σ^2, ma non μ, può apparire più bislacca. Vi sono casi, tuttavia in cui anche questa ipotesi può considerarsi ragionevolmente soddisfatta. Per esempio, se stiamo misurando una qualche caratteristica di una sostanza tramite uno strumento, (p.e. una concentrazione in una soluzione) ogni misura sarà affetta da un errore che dipende, in buona parte almeno, dall'apparecchiatura di misura e non dal valore misurato. E' abbastanza comune, in questo caso, assumere che i valori misurati oscillino attorno al valore vero in modo da essere distribuiti secondo una legge normale con media il valore da misurare e con una varianza che dipende dallo strumento (e quindi può esserci nota da precedenti esperienze).

In questo caso il teorema precedente ci offre un immediata regola di condotta. Infatti abbiamo che la variabile

$$\boldsymbol{z} = \frac{\overline{\boldsymbol{x}} - \mu}{\sigma/\sqrt{n}} \tag{7.6}$$

è distribuita secondo la normale standard (ovvero ha media nulla e varianza 1). A partire da questa osservazione è possibile fare delle previsioni su μ che sarà espresso da

$$\mu = \overline{\boldsymbol{x}} - \frac{\sigma}{\sqrt{n}}\boldsymbol{z}, \tag{7.7}$$

ovvero il valore di μ è dato dal valore medio dei valori osservati più un errore aleatorio, di cui conosciamo la distribuzione di probabilità.

7.1.2 Popolazione normale, σ^2 sconosciuta

Cosa succede invece nel caso in cui non si conosca la varianza della popolazione soggiacente, pur sapendo che essa è distribuita normalmente? In questo caso, ricordando che σ^2 è il valore atteso della varianza campionaria, possiamo tentare di sostituire la varianza della popolazione con la varianza campionaria $\boldsymbol{s}^2$, e fare la stessa stessa trasformazione

$$\boldsymbol{t} = \frac{\overline{\boldsymbol{x}} - \mu}{\boldsymbol{s}/\sqrt{n}} \tag{7.8}$$

ottenendo una nuova variabile aleatoria, che abbiamo indicato con $\boldsymbol{t}$ (nota che ora sia $\overline{\boldsymbol{x}}$ che $\boldsymbol{s}$ sono variabili aleatorie, cioè il loro valore varia da campione a campione.)

Questa nuova variabile aleatoria *non* è più distribuita secondo la normale, ma secondo una legge di probabilità la cui funzione di distribuzione ha un grafico assai simile alla normale, ma più disperso. Per essere più precisi $\boldsymbol{t}$ segue una legge di probabilità che dipende da n, ovvero abbiamo una famiglia di distribuzioni, parametrizzata da n. La distribuzione a cui obbedisce la $\boldsymbol{t}$ prende il nome di $\boldsymbol{t}$-di-Student con $\nu = (n-1)$ **gradi di libertà**, indicata spesso con $\boldsymbol{t}_\nu$. Essa ha media nulla e varianza $\nu/(\nu-2)$ se $\nu \geq 3$. Per questa distribuzione esistono delle tavole (una per ogni grado di libertà).

All'aumentare dei gradi di libertà, la distribuzione $\boldsymbol{t}_\nu$ converge alla distribuzione normale standard e viene generalmete confusa con essa per $\nu > 30$ (ovvero, per $\nu > 30$ si utilizzano le tavole della normale standard al posto di quelle per la $\boldsymbol{t}$ per effettuare i calcoli.)

7.2 Popolazione non normale

Cosa dobbiamo fare invece se l'ipotesi di normalità per la popolazione soggiacente non può essere considerata valida?

Anche in questo caso dobbiamo distinguere il caso in cui sia nota la varianza della popolazione σ^2 da quello in cui essa ci sia ignota.

7.2.1 Popolazione non normale, σ^2 nota

In questo caso possiamo far appello al Teorema centrale del limite che ci dice che la distribuzione della variabile aleatoria

$$\boldsymbol{y} = \frac{\overline{\boldsymbol{x}} - \mu}{\sigma/\sqrt{n}}$$

tende, in probabilità, alla normale standard per $n \to \infty$, qualunque sia la distribuzione della popolazione soggiacente.

La regola di comportamento che ne traiamo è che se "n è grande", allora possiamo considerare $\boldsymbol{y}$ come se fosse realmente distribuita secondo la normale stantard, e fare i conti utilizzando le tabelle per $\boldsymbol{z}$.

Resta il problema pratico di decidere cosa significhi che "n è grande". Questo dipende molto dalla forma della distribuzione soggiacente. Se la distribuzione da cui si fa il campionamento è simmetrica e unimodale oppure è una distribuzione uniforme su un intervallo finito, allora si ha un buon accordo per valori di n già relativamente piccoli (p.e. $n = 10$); in generale per $n \geq 30$ si ha un buon accordo per qualsiasi tipo di distribuzione soggiacente, e si accetta nella pratica di considerare la variabile $\boldsymbol{y}$ come se fosse distribuita normalmente.

7.2.2 Popolazione non normale, σ^2 sconosciuta

Anche in questo caso vale una regola pratica simile alla precedente. Ci si comparta come nel caso della popolazione soggiacente normale assumendo che la variabile

$$\boldsymbol{t} = \frac{\overline{\boldsymbol{x}} - \mu}{\boldsymbol{s}/\sqrt{n}}$$

sia distribuita secondo la $\boldsymbol{t}$ di Student con $\nu = n - 1$ gradi di libertà. Ciò è tanto più lecito qunto più n è grande. Al crescere di n ($n > 30$), si può approssimare a sua volta la $\boldsymbol{t}$ di Student con la normale standard e compartarsi, in pratica, come nel caso di σ^2 nota, con la sola differenza che la standardizzazione si calcola usando la media campionaria.

7.3 Popolazioni finite

Infine è bene ricordare che tutto quello che abbiamo detto fino ad ora presupponeva che le osservazioni fossero indipendenti. Questo accade nel caso di popolazioni (potenzialmente) infinite, p.e. nel caso di misurazioni tramite uno strumento, purché un processo di osservazione non influenzi i successivi.

Nel caso di una popolazione finita, l'ipotesi di indipendenza presuppone che dopo ogni osservazione la situazione sia riportato allo stato che la precedeva. Per esempio, se facciamo un sondaggio di mercato, l'ipotesi di indipendenza della osservazioni implica che una stessa persona possa essere intervistata più di una volta (in linea di principio anche n volte!) Questo ovviamente non corrisponde alla pratica reale, dove il sondaggio viene condotto "senza rimbussolamento". Da un punto di vista pratico la differenza è inapprezzabile se la dimensione N della popolazione soggiacente è abbastanza grande (essendo il campione aleatorio, la probabilità di intervistare più di una volta anche solo una persona è piccola).

In ogni caso, a questo problema si pone rimedio "correggendo" la varianza della media campionaria $\overline{\boldsymbol{x}}$ moltiplicando σ^2/n per il fattore correttivo $(N - n)/(N - 1)$. In particolare la standardizzazione della variabile media campionaria diventa

$$\boldsymbol{z} = \frac{\overline{\boldsymbol{x}} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}. \tag{7.9}$$

Notiamo che il fattore corretivo $(N - n)/(N - 1)$ tende 1 quando $N \to \infty$, e che è "praticamente" uguale a 1 se N è grande rispetto a n.

7.4 Distribuzione della varianza campionaria

Anche per la varianza campionaria è possibile dare la distribuzione nel caso che la popolazione soggiacente sia una popolazione distribuita secondo la normale. In questo caso si può dimostrare che la variabile

$$\boldsymbol{\chi}^2_{n-1} = \frac{(n-1)\boldsymbol{s}^2}{\sigma^2} \tag{7.10}$$

è distribuita secondo al distribuzione $\boldsymbol{\chi}^2$ (chi-quadro) con $n-1$ gradi di libertà.

Un altro caso in cui è nota la distribuzione è quello del rapporto tra le varianza campionarie di due campioni aleatori indipendenti di numerosità n_1 e n_2 tratti da due popolazioni distribuite normalmente. In questo caso la variabile $\boldsymbol{s}_1^2/\boldsymbol{s}_2^2$ segue una distribuzione detta F_{ν_1,ν_2} di Fischer con due parametri (detti ancora gradi di libertà) $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$.

7.5 Intervalli di confidenza

Una volta note le distribuzioni di probabilità degli stimatori puntuali dei parametri di una popolazione, è possibile precisare la "bontà" della stima che un campionamento ci dà di un parametro incognito.

Iniziamo con un esempio: supponiamo di voler stimare il valor medio μ di una popolazione che sappiamo già essere distribuita normalmente con varianza $\sigma^2 = 4$.

A tale scopo, effettuiamo un campionamento mediante $n = 36$ osservazioni indipendenti. Sappiamo che lo stimatore per la media μ della popolazione è la media campionaria delle nostre osservazioni, $\overline{\boldsymbol{x}}$.

Avendo assunto che la popolazine sia distribuita normalmente e con varianza nota, la $\overline{\boldsymbol{x}}$ è una variabile aleatoria distribuita normalmente, con valor medio μ e varianza $\sigma^2/n = 1/9$.

Supponiamo infine che dal nostro campionamento noi abbiamo ottenuto un valore $\overline{x} = 13.8$ per la media campionaria.

Come si usa questa informazione?

A partire dai dati in nostro possesso possiamo costruire una "stima" di tipo probabilistico del parametro μ. Questo significa che possiamo determinare un intervallo (μ_1, μ_2) in modo che la media μ abbia una probabilità p (scelta a nostro piacimento) di essere compresa tra i valori μ_1 e μ_2, ovvero $P(\mu \in (\mu_1, \mu_2)) = p$.

Qualche osservazione:

Per prima cosa osserviamo che l'intervallo non è univocamente determinato, in generale avremo infiniti intervalli che godono di questa proprietà. Per rendere univoca la scelta si adotta il criterio di scegliere l'intervallo in modo che le probablità che μ appartenga a uno due intervalli $(-\infty, \mu_1)$ o $(\mu_2, +\infty)$ siano entrambe $(1-p)/2$.

La seconda e più importante osservazione è che la stima *non garantisce* che il valore del parametro sia compreso tra i due valori μ_1 e μ_2, ma solo che c'è una certa probabilità che questo sia vero. Inoltre *più grande scegliamo p, più grande risulta essere l'intevallo* (μ_1, μ_2); in altri termini, la stima deve bilanciare due richieste "negativamente correlate": la precisione (ovvero un intervallo (μ_1, μ_2) "piccolo") e l"'affidabilità" della stima (cioè un valore elevato di p).

Torniamo ora al nostro esempio e vediamo come si costruisce l'intervallo per la media.

Una variabile aleatoria $\mathbf{X}$ distribuita secondo la normale $\mathcal{N}(\mu, 1/9)$ ha una probabilità nota, p, di trovarsi in un intervallo centrato attorno al suo valor medio e di ampiezza data $2a$, $p = P(\mu - a < \mathbf{X} < \mu - a)$. p è una funzione, invertibile, di a, ovvero per ogni $a > 0$ esiste uno e un solo p che ci dà la probabilià che $\mathbf{X}$ appartenga all'intervallo $(\mu - a, \mu + a)$

$$p = P(\mu - a < \mathbf{X} < \mu + a) = F(\mu + a) - F(\mu - a). \tag{7.11}$$

Viceversa, assegnata p, possiamo determinare univocamente a in modo che la (7.11) sia soddisfatta.

L'approccio classico del problema è dunque il seguente. Fissiamo un valore di p, p.e. $p = 0.95$ ovvero una probabilità del 95%, e determiniamo $a_{.95}$ modo da soddisfare la (7.11).

Possiamo affermare che il vero valor medio μ si trova, con probabilità $p = .95$, nell'intervallo $(\overline{x} - a_{.95}, \overline{x} + a_{.95})$.

Infatti il valore $\overline{x} = 13.8$ da noi determinato nel campionamento aveva una probabilità, a priori, del 95% di trovarsi nell'intervallo $(\mu - a_{.95}, \mu + a_{.95})$. Ma la disuguaglianza $\mu - a_{.95} < \overline{x} < \mu + a_{.95}$ è equivalente a $\overline{x} - a_{.95} < \mu < \overline{x} + a_{.95}$, cioè i due eventi $\overline{x} \in (\mu - a_{.95}, \mu + a_{.95})$ e $\mu \in (\overline{x} - a_{.95}, \overline{x} + a_{.95})$ concidono e quindi hanno la stessa probablità. Ciò significa che, una volta effettuato il campionamento e calcolato $\overline{x} = 13.8$, il vero valore μ del valor medio della popolazione ha una probabilità del 95% di trovarsi nell'intervallo $(13.8 - a_{.95}, 13.8 + a_{.95})$.

Resta da calcolare il valore di $a_{.95}$. Per far questo basta ricordare quanto detto a proposito della standardizzazione di una variabile normale. La variabile $\boldsymbol{z} = (\overline{\boldsymbol{x}} - \mu)/(\sigma/\sqrt{n})$ è distribuita secondo la normale standard e la disuguaglianza $\overline{x} - a_{.95} < \mu < \overline{x} + a_{.95}$ è equivalente a $-a_{.95}\sqrt{n}/\sigma < \boldsymbol{z} < a_{.95}\sqrt{n}/\sigma$. Quindi si ha $a_{.95} = \sigma z_c/\sqrt{n}$ dove z_c è il valore per cui $P(-z_c < \boldsymbol{z} < z_c) = 0.95$ (questi valori sono detti *valori critici*), ovvero il valore z_c tale che le due code rispettivamente a destra di z_c e a sinistra di $-z_c$ abbiamo entrambe probabilià uguale a $(1 - 0.95)/2 = 0.025$ ossia del 2.5%. In una tabella per la distribuzione cumulativa bisogna quindi cercare il valore di z per cui $F(z) = 0.975$ e lo indicheremo con $z_c = z_{.975} \approx 1.96$.

Possiamo ora concudere con i dati del nostro esempio: il vero valore del valor medio μ si trova, con probabilità del 95% nell'intervallo $(13.8 - 1.96/9, 13.8 + 1.96/9)$.

Un tale intervallo si dice *un intervallo di confidenza al 95%* per il valor medio.

Notiamo che per ogni campionamento si otterranno intervalli di confidenza diversi. Ognuno di essi è un intervallo di confidenza lecito.

C'è un modo un po' diverso di interpretare lo stesso calcolo. Secondo questo punto di vista diremo che il valor medio μ appartiene all'intervallo appena costruito *al livello di errore del 5%*. Questo significa che se si assume che il valor medio appartiene all'intervallo si può commettere un errore (cioè μ può anche non appartenere all'intervallo), ma la probabilità di sbagliare è solo del 5%.

Se effettuiamo il campionamento a partire da una popolazione distribuita normalmente, ma di cui ignoriamo sia il valor medio che la varianza, possiamo procedere come sopra semplicemente sostituendo la varianza campionaria alla varianza della popolazione e la distribuzione t-Student (con gli opportuni gradi di libertà) alla distribuzione normale standard. Di conseguenza, al posto dei valori critici z_c, avremo dei valori critici $t_c(n-1)$, dipendenti questa volta non solo dalla probabilità che l'intervallo deve avere di contenere il valor medio della popolazione, ma anche dalla dimensione n del campione.

Capitolo 8

Test di ipotesi

Una tecnica importante della statistica deduttiva è quella che va sotto il nome di *Test di Ipotesi*. Essa consiste nel porre a confronto un'ipotesi su una caratteristica di una popolazione con un insieme di dati sperimentali.

L'ipotesi che si sceglie di mettere a confronto con l'esperimento prende tradizionalmente il nome di *ipotesi nulla* e viene indicata con la "sigla" H_0. Essa si presenta in genere nella forma di una assegnazione del valore di un parametro di una distribuzione parametrica di forma nota. Questa ipotesi è messa a confronto con una possibile alternativa (che prende in nome di *ipotesi alternativa* e si indica tradizionalmente con la sigla H_1).

Vediamo un esempio: supponiamo di aver sviluppato una nuova tecnica di inseminazione artificiale. Vogliamo vedere se essa è più efficiente della tecnica attualmente in uso. L'enunciazione di questo problema sembra già chiara, tuttavia se vogliamo "fare dei calcoli" dobbiamo dare una struttura statistico-matematica più precisa al problema.

Intanto come misuriamo l'efficienza della tecnica usata? Dobbiamo associare un numero a l'idea di efficienza: è abbastanza ovvio, in questo caso, che questo numero sia il rapporto tra il numero delle inseminazioni che hanno avuto successo e il numero totale delle inseminazioni effettuate. La vecchia tecnica ha quindi un tasso di successo $p_0 = \frac{\text{successi}}{\text{tentativi}}$ che ci è noto dai dati storici in nostro possesso. Ovviamente il numero p_0 è un dato sperimentale, soggetto a modificarsi se si fanno nuove inseminazioni con la vecchia tecnica, tuttavia a questo punto dobbiamo corsiderarlo come un dato "oggettivo" e interpretarlo come la probabilità che una ulteriore inseminazione (effettuata con la vecchia tecnica) abbia successo. In altre parole, adottiamo come modello probabilistico per i nostri esperimenti quello delle *prove di Bernoulli* con probabilità di successo p_0. Tanto per fissare le idee supponiamo $p_0 = 0.3$, ovvero una percentuale di successi del 30%.

L'affermazione che la nuova tecnica è più efficiente si traduce nel dire che essa ha un tasso di successi $p > p_0$.

A questo punto si procede scegliendo come *ipotesi nulla* H_0 che *la nuova tecnica ha la stessa efficienza della vecchia* e la raffrontiamo con l'*ipotesi alternativa* H_1 che la nuova tecnica è più efficente (ovvero ha un tasso di successi $p \geq 0.3$).

Ora dobbiamo analizzare un campione di n inseminazioni condotte con la nuova tecnica. Dobbiamo supporre che questo campione sia un campione aleatorio: questo ci garantisce che il rapporto $P = \frac{s}{n}$ tra la variabile aleatoria $\mathbf{s}$ che conta il numero di successi ottenuti il numero n di tentativi effettuati è uno stimatore per il parametro p (nota che nella pratica questo è un punto delicatissimo specie nella pratica clinica; infatti i

soggetti disposti alla sperimentazione con la nuova tecnica sono spesso quelli con i quali la vecchia tecnica ha fallito!).

Lo stimatore che abbiamo scelto ha una distribuzione nota se assumiano vera l'ipotesi nulla. Infatti in questo caso la variabile aleatoria $\mathbf{s}$ che conta il numero di successi sulle n prove, è distribuita secondo una distribuzione binomiale con probabilità di successo p_0.

Il test di ipotesi si basa su questo fatto: se l'ipotesi nulla è vera, allora sarà "poco probabile" che s sia "molto più grande" di un dato valore s_0. Se gli esperimenti mi danno un risultato $\mathbf{s} > s_0$, allora propendo a credere che l'ipotesi di partenza sia sbagliata, e la *respingo* a favore dell'ipotesi alternativa che $p > p_0$.

Dobbiamo decidere come fissare la soglia s_0 che discrimina la decisione. Per far ciò è necessario stabilire quanto vogliamo "rischiare" prendendo la decisione: in altri termini dobbiamo stabilire un livello che *noi* riteniamo adeguato per la probabilità $P(\mathbf{s} > s_0)$. Fissando un valore relativamente alto, p.e. $P(\mathbf{s} > s_0) = 0.2$ ovvero il venti percento, ci esponiamo al rischio di respingere l'ipotesi nulla (e quindi accettare, nel nostro esempio, il fatto che la nuova tecnica sia più efficace) con una probabilità del venti percento di sbagliare. Una scelta "conservatrice" sarà, quindi, quella di scegliere $P(\ s > s_0)$ molto piccolo, p.e. l'uno percento. In questo caso ci si espone a un rischio sensibile di non ritenere più efficente la nuova tecnica anche se lo è veramente. Per poter quantificare questo rischio, ovvero per assegnare una probabilità anche a questo tipo di errore, occorre però specificare quanto "più efficiente" sia la nuova tecnica (ovvero bisogna fissare un probabilità di successo p_1, p.e. $p_1 = 0.4$, da confrontare con p_0).

Riassumiano quindi la "struttura" del test di ipotesi:

1. Si formula un ipotesi riguardo a una certa caratteristica di una variabile aleatoria definita sulla nostra popolazione, l'*ipotesi nulla* indicata con H_0. Si individua inoltre l'*ipotesi alternativa*, indicata con H_1, ovvero l'insieme dei casi che si possono verificare se non si verifica l'ipotesi nulla;

2. Si sceglie uno stimatatore $\mathbf{x}$ per valutare, tramite campionamento, il valore della caratteristica sotto esame;

3. Si determina quale distribuzione abbia la variabile aleatoria scelta come stimatore, *nel caso che l'ipotesi nulla sia vera*;

4. Si decide un livello di errore α e si costruisce l'intervallo di confidenza A_0 per lo stimatore, corrispondente al livello di errore scelto (si sceglie A_0 in modo che $P(\mathbf{x} \in A_0) = 1 - \alpha$);

5. Si effettua il campionamento;

6. Si confronta il valore ottenuto dal campione con l'intervallo di confidenza calcolato;

infine *si accetta o si respinge l'ipotesi a seconda che il valore campionario appartenga o meno all'intervallo di confidenza* A_0.

Da un punto di vista "interpretativo", si privilegia piuttosto il *respingere* l'ipotesi rispetto all'*accettarla*.

In accordo a questo modo di pensare si pone come ipotesi alternativa proprio quello che "vorremmo si realizzasse" (p.e. nel caso dell'inseminazione, un maggior tasso di sucessi), mentre si pone come ipotesi nulla la situazione che si "vorrebbe negare".

Inoltre si tende a fissare un livello di errore "piccolo" (una scelta tipica è in livello di errore α del 5%). Questo significa che, *se l'ipotesi nulla è vera*, allora è piccola la probabilità di ottenere valori al di fuori dell'intervallo di confidenza a causa di fluttazioni aleatorie del campionamento.

Di conseguenza, se il valore calcolato nell'esperimento cade fuori dall'intervallo, allora è "più ragionevole" pensare che sia falsa l'ipotesi di partenza.

Da questo punto di vista possiamo fissare un qualsiasi valore del livello di errore $\alpha \in (0, 1)$ e dire che l'ipotesi viene respinta *al livello di errore* α se il risultato del test cada fuori dell'intervallo di confidenza del $(1 - \alpha) \times 100\%$, e dire invece che i dati *non ci permettono di respingere l'ipotesi nulla* (sempre *al livello di errore* α) se il risultato cade nell'intervallo di confidenza.

Vediamo un altro esempio di test di ipotesi:

Analizziamo la seguente situazione: *il sindaco di una grande città è stato eletto con il 70% dei voti e vuole sapere come è composto il suo elettorato dal punto di vista del sesso.*

Si presuppone che non ci sia differenza in percentuale tra gli uomini e le donne che hanno votato per il sindaco, ovvero che il 70% degli elettori maschi e il 70% delle elettrici, abbiano votato per lui: Ipotesi H_0.

L'ipotesi alternativa in questo caso è che ci sia differenziazione sessuale del voto, quindi che le percentuali degli uomini che hanno votato per il sindaco sia diversa da quella delle donne: Ipotesi H_1.

La variabile aleatoria che sottoponiamo al test è la differenza $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$ delle frequenze relative, nei campioni sottoposti al sondaggio, degli elettori del sindaco rispettivamente tra gli uomini ($\mathbf{r}_1$) e tra le donne ($\mathbf{r}_2$). L'ipotesi nulla corrisponde ad affermare che $\mathbf{r} = 0$. L'ipotesi alternativa è semplicemente $\mathbf{r} \neq 0$.

Supponiamo di effettuare un sondaggio su un campione di 150 uomini e 90 donne che hanno votato. A ognuno viene chiesto se ha votato per il sindaco oppure no.

Vediamo ora qual'è la distribuzione della variabile aleatoria $\mathbf{r}$.

Dobbiamo fare un'ipotesi "strutturale" sul campione: assumiamo, come al solito, che sia gli uomini che le donne intervistate siano stati scelti aleatoriamente.

Se l'ipotesi H_0 è vera, la probabilità che un singolo elettore, scelto a caso, sia un elettore che abbia votato per il sindaco è quindi 0.7 sia per gli uomini che per le donne.

Il sondaggio è quindi una serie di estrazioni, senza reintroduzione in quanto non si intervista due volte una stessa persona. Poiché la popolazione soggiacente è "grande", possiamo trascurare questa sottigliezza e considerare il processo di campionamento come una serie di Bernoulli con probabilità di successo $p = 0.7$ (per "successo" assumiamo che la persona intervistata abbia votato per il sindaco). Abbiamo così la distribuzione di probabilità delle due variabili aleatorie X_1 e X_2 che conta il numero di successi tra gli intervistati uomini e donne rispettivamente: entrambe sono distribuite secondo la distribuzione binomiale, con valor medio $\mu_1 = 0.7 \times 150$ e $\mu_2 = 0.7 \times 90$, e con varianza $\sigma_1^2 = 150(0.7)(0.3)$ e $\sigma_2^2 = 90(0.7)(0.3)$ rispettivamente.

Possiamo ancora semplificare osservando che il campione è sufficientemente numeroso da poter sostituire la distribuzione binomiale con la normale (di stessa media e varianza).

Quindi approssimiamo sia X_1 che X_2 con due variabili aleatorie, che indichiamo ancora con X_1 e X_2, distribuite normalmente. Infine, ponendo $\mathbf{r}_1 = X_1/150$ e $\mathbf{r}_2 = X_2/90$ abbiamo che la variabile aleatoria $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$ può essere considerata come la differenza di due variabili aleatorie distribuite normalmente, entrambe con valore atteso 0.7 e con varianze, rispettivamente $0.7 \times 0.3/150$ e $0.7 \times 0.3/90$. Quindi anche

r è distribuita normalmente e la sua media è la differenza delle medie, mentre la sua varianza è data dalla somma delle varianze: $\mu = 0,\ \sigma^2 = (0.7)(0.3)\left[\frac{1}{150} + \frac{1}{90}\right]$.

A questo punto standardizziamo la variabile r ponendo

$$Z = \frac{\mathbf{r}}{\sigma}$$

di modo che la Z sia distribuita secondo la normale standard.

Supponiamo ora che il risultato del sondaggio sia che 115 uomini e 55 donne hanno votato il sindaco. Vogliamo vedere se, in base a questo risultato, si può respingere, *al livello di errore del 5%*, l'ipotesi di voto in percentuali uguali.

Calcolando il valore di Z realizzato nel nostro nel sondaggio otteniamo

$$z = \frac{\frac{115}{150} - \frac{55}{90}}{\sqrt{(.7)(.3)\left[\frac{1}{150} + \frac{1}{90}\right]}} = 2.545875..$$

Questo valore è fuori dall'intervallo $(-1.96, 1.96)$ che è l'intervallo di confidenza del 95% per la variabile Z. Quindi la decisione da prendere in virtù di questi dati è di *respingere l'ipotesi*.

Nota che l'intervallo di accettazione per la variabile aleatoria $\mathbf{r}$ si ottiene tornando indietro con la standardizzazione, e si ha $A_0 = (-1.96\sigma, 1.96\sigma) = (-0.12, 0.12)$

Quanto abbiamo detto ha senso nel caso che l'ipotesi nulla abbia la forma: *il parametro γ è uguale al valore c.* Una tale ipotesi si dice *semplice*.

Possiamo anche immaginare delle ipotesi più elaborate, tipo: *il parametro γ appartiene all'intervallo (c_1, c_2)* (in questo caso parleremo di *ipotesi composita*)

In questo caso, fissato un livello di errore α, per ogni valore c dell'intervallo (c_1, c_2) otteniamo un corrispondente intervallo di confidenza (al livello di errore fissato), diciamo A_c. Se l'ipotesi nulla è vera abbiamo quindi una probabilità $1 - \alpha$ di trovare il valore dello stimatore di γ, *in uno qualsiasi degli intervalli* A_c; ovvero abbiamo una probabilità $1 - \alpha$ di trovare lo stimatore nell'unione $A = \bigcup_{c\in(c_1,c_2)} A_c$ di questi intervalli, che è l'insieme di accettabilità per questa ipotesi. Di conseguenza l'ipotesi sarà respinta se il valore fornito dall'esperimento non appartiene all'insieme A di accettabilità.

Un punto dove si tende a fare confusione, specialmente se si prende il test di ipotesi come una "ricetta" di calcolo, è la formulazione dell'ipotesi alternativa H_1. Da essa dipende anche la forma dell'intervallo di accettazione e quindi i valori dei livelli critici.

Riprendiamo per un attimo i due esempi che abbiamo fatto: nel primo caso mettevamo a confronto l'ipotesi nulla "tasso di successi del 30%" con l'ipotesi alternativa "tasso di successi *maggiore* del 30%"; nel secondo caso l'ipotesi nulla era "percentuale di voto uguale tra uomini e donne" contro l'ipotesi alternativa "percentuale di voto diversa tra uomini e donne".

Nel primo caso abbiamo scelto di confrontare i dati solo con un'alternativa "unilaterale" (nella letteratura in lingua inglese di dice *one-sided*), mentre nel secondo esempio abbiamo un alternativa "bilaterale" (*two-sided*), in quanto la percentuale del voto maschile può essere sia maggiore che minore di quella femminile.

In corrispondenza avremo una struttura "unilaterale" o "bilaterale" degli intevalli di accettazione. Nel primo esempio l'intevallo di accettazione dell'ipotesi nulla avrà la forma $\mathbf{s} \leq s_0$ dove s_0 è il valore critico scelto in modo che si abbia $P(\mathbf{s} > s_0) = \alpha$. Nel secondo caso avremo invece un intervallo di accettazione della forma $(-r_0, r_0)$ dove il valore critico r_0 è scelto in modo che $P(\mathbf{r} > r_0) = \alpha/2$ (in modo che ancora sia $P(-r_0 < \mathbf{r} < r_0) = 1 - \alpha$).

8.0.1 Tipi di errore di un test

Come nel caso dei test diagnostici, anche nei test di ipotesi possono presentarsi quattro possibili situazioni:

1. L'ipotesi nulla è vera e viene accettata;
2. L'ipotesi nulla è vera e viene respinta;
3. L'ipotesi nulla è falsa e viene accettata;
4. L'ipotesi nulla è falsa e viene respinta.

Nel secondo e terzo caso si commette un errore, detti rispettevamente errore errore del primo e del secondo tipo.

Per quanto riquarda l'errore del primo tipo, è chiaro che la probabilità di commettere un tale errore coincide con il livello di errore fissato per il test, ovvero la probabilità che si cada al di fuori dell'insieme di accetabilità, che abbiamo denotato con α.
Nota: è bene rendersi conto che si tratta di *probabilità condizionate: α è la probabilità di respingere H_0 se l'ipotesi nulla è vera.*

La probabilità di commettere un errore del secondo tipo si indica con β. In accordo a quanto detto è importante sapere quanto vale $1-\beta$, ovvero la probabilità di respingere (correttamente) un ipotesi falsa. Questo numero è detto *potenza* del test.

Un buon test dovrebbe avere contemporaneamente un α piccolo e un piccolo β (ovvero una grande potenza). Tuttavia non è possibile ridurre contemporaneamente queste due variabili, che sono tra loro correlate, con β che cresce al diminuire di α.

Inoltre bisogna fare attenzione alla stessa definizione di β.

Ricordiamo in cosa consiste un errore del secondo tipo: abbiamo detto che questo errore si commette se si accetta l'ipotesi nulla mentre è vera l'ipotesi alternativa. D'altra parte, accettiamo l'ipotesi nulla se il valore della stima del parametro cade dentro l'insieme di accettazione A_0 (che abbiamo determinato assumendo vera H_0). *La probabilità di compiere un errore del secondo tipo sarà quindi data dalla probabilità che ha la stima di appartenere all'insieme di accettazione* **se è vera l'ipotesi alternativa**.

Vediamo come si calcola β nel caso si possa supporre che l'ipotesi alternativa H_1 sia semplice (ovvero consista nell'assegnare un valore alternativo al parametro). In questo caso la distribuzione della variabile aleatoria che deve stimare il parametro è determinata univocamente dall'ipotesi alternativa.

Riprendiamo l'esempio del voto per il sindaco e supponiamo che l'ipotesi alternativa invece di essere data dall'ipotesi composita (percentuali di elettori uomini $\neq$ percentuale di elettrici donne) si data, p.e., da (perc. uomini $= 80\%$) e (perc. donne $= 60\%$) (in questo caso stiamo supponendo che ci siano tanti uomini quante donne nell'elettorato, cosa che non era necessaria per ipotesi nulla). Assumendo questa ipotesi, la variabile aletoria $\mathbf{r}$ non è più distribuita come nel caso in cui è vera l'ipotesi nulla, ma ora è una variabile (approssimativamente) normalmente distribuita con valore atteso $\mu_1 = 0.8 - 0.6 = 0.2$ e varianza $(0.8)(0.2)/150 + (0.6)(0.4)/90 = 0.004067$ ovvero $\sigma_1 = 0.06377$. La probabilità di errore di tipo 2 sarà quindi la probabilità dell'intervallo $A_0 = (-0.12, 0.12)$ per una variabile distribuita normalmente con valore atteso 0.2 e varianza 0.004067 (fare il calcolo per esercizio).

Veniamo ora al caso originale in cui H_1 era data da $E(\mathbf{r}) \neq 0$ (i.e. diversa proporzione tra gli elettori maschi e femmine). In questo caso **non è possibile calcolare** β senza fare qualche altra ipotesi. Infatti, dal conto che abbiamo appena fatto è

chiaro che il valore che abbiamo calcolato è la probabilità condizionata di A_0 all'ipotesi $\mathbf{r} \sim \mathcal{N}(\mu_1, \sigma_1^2)$, con $\mu_1 = 0.2$ e che diventa la probabilità di errore di secondo tipo quando la si moltiplica per la probabilità che il valor medio sia 0.2 quando è vera H_1. Poiché H_1 consisteva nell'affermazione che il valor medio **è** 0.2, quest'ultima probabilità è uguale a 1, e quindi il valore che abbiamo calcolato **è** la probabilità β.

Tutto ciò non è più vero se l'ipotesi H_1 contiene più di un caso possibile (è cioè un ipotesi composita). In questo caso per calcolare β bisognerebbe assegnare una probabilità a ogni caso possibile di H_1 (ovvero per ogni possibile valore del parametro μ_1), cosa che non sappiamo fare.

Le cose vanno ancora peggio se, come nel nostro esempio, l'ipotesi nulla è del tipo $\mu = \mu_0$ e l'ipotesi alternativa è $\mu \neq \mu_0$. In questo caso infatti potremmo assegnare una probabilità arbitrariamente vicina a 1 a un valore di μ_1 arbitrariamente vicino a μ_0 e, di conseguenza, ottenere un valore di β vicino quanto si vuole alla probabilità dell'intervallo di accettazione *sotto l'ipotesi nulla*, e quindi un β vicino quanto si vuole a $1 - \alpha$.

Questa difficoltà proviene dal fatto che l'assunzione che $\mu = \mu_0$ per l'ipotesi nulla, per quanto comoda per il calcolo, è "probabilisticamente" insensata se il parametro μ è una variabile continua, in quanto ha probabilità nulla di realizzarsi.

Possiamo rimediare a questo in due modi. O assumendo anche per l'ipotesi nulla la forma di un ipotesi composita (tipo $\mu \in (\mu_{\min}, \mu_{\max})$; oppure, conservando l'ipotesi nulla nella forma $\mu = \mu_0$, introdurre un livello di "errore significativo", $a > 0$, e sostituendo l'ipotesi alternativa $\mu \neq \mu_0$ con l'ipotesi $|\mu - \mu_0| > a$, ovvero che il parametro sia "sufficientemente diverso" da μ_0.

In questo caso si può calcolare la probabilità di errore del secondo tipo mettendosi nel "caso peggiore" ossia eseguendo il calcolo come nel caso di un ipotesi alternativa semplice (come abbiamo fatto nell'esempio) usando come valore quello (tra tutti i possibili μ di H_1) che rende massima la probabilità dell'insieme di accetazione A_0 (nel caso di una distribuzione simmetrica e unimodale come la normale, si tratta di $\mu_0 + a$).

8.0.2 Il test chi-quadro

Nella sezione precedente abbiamo visto le tecniche per effetuare test di ipotesi che riguardano il valore di un parametro sconosciuto per una distribuzione di forma nota.

Questi test si applicano quindi a casi in cui si abbia una variabile aleatoria di tipo numerico e si sia deciso a priori il tipo di distribuzione a cui questa variabile obbedisce.

In molti casi vogliamo mettere a confronto con i dati proprio la forma di una distribuzione: questo avviane in particolare quando si ha a che fare con dati di tipo categoriale, quindi non ci sia nessuna variabile aleatoria soggicente.

In questo caso il test che si usa va sotto il nome di test χ^2.

Vediamo come si arriva a questo test: supponiamo di avere una serie di dati sperimentali che possiamo suddividere in certo numero di classi $C_1, C_2, ..., C_k$. Per ogni classe abbiamo la frequenza dei dati sperimentali in quella classe, ovvero i numeri $N_1, N_2, ..., N_k$ di dati cadono nella classi $C_1, C_2, ..., C_k$ rispettivamente, e indichiamo con n la somma delle frequenze, $N = N_1 + ... + N_k$.

L'ipotesi che mettiamo a confronto con i dati è una *distribuzione teorica di probabilità sulle categorie* $C_1, C_2, ..., C_k$, ovvero una distribuzione dove $p_1, p_2, ..., p_k$ sono le probabilità per un dato di cadere nelle categorie $C_1, C_2, ..., C_k$ rispettivamente.

Dobbiamo assumere che le categorie siano esaustive per i nostri dati (i.e. un dato deve cadere in almeno una categoria) e mutualmente escusive (i.e. un dato può cadere in al più una categoria); in altre parole devono rapprentare una partizione dello spazio

campionario da cui provengono i dati. Di conseguenza le probabilità p_i, $i = 1, ..., k$ devono soddisfare la condizione di normalizzazione

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

L'ipotesi H_0 consiste quindi nell'assumere che i nostri dati vengano da un campionamento (aleatorio) di una popolazione divisa nelle nostre $C_1, ..., C_k$ categorie con probabilità $p_1, p_2, ..., p_k$. Se l'ipotesi nulla è vera, il valore più probabile (sui nostri N esperimenti) di risultati nella categoria C_i è dato da Np_i e il numero $(N_i - Np_i)^2/(Np_i)$ rappresenta una misura dello scarto della frequenza osservata nel nostro esperimento (relativamente alla categoria C_i) rispetto alla frequenza "attesa". La statistica data dalla somma di questi scarti per $i = 1, ..., k$, ovvero

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

prende il nome di *chi-quadro di Pearson*. Si assume che essa sia distribuita con una distribuzione χ^2_{k-1} (chi-quadro con $k-1$ gradi di libertà) il che è solo "approssimativamente vero", ma dà risultati ragionevoli nelle applicazioni per n sufficientemente grande e con valori p_i non troppo piccoli. Nella pratica si assume che per ogni i sia $np_i > 5$.

L'applicazione del test è semplice. Vediamo di esemplificarla con il più classico degli esempi, l'analisi degli esperimenti di G. Mendel sulle piante di piselli. Si tratta dell'osservazione di due coppie di caratteri che si escludono a vicenda: liscio-grinzoso e giallo-verde, i primi elemti della coppia essendo i caratteri dominanti. Usando le *Leggi* delle segregazione e dell'indipendenza (e l'ipotesi quantitativa che per entrambe le coppie il rapporto dominante:recessivo sia 3:1) si ottiene una distribuzione teorica delle coppie di caratteri sulla seconda generazione data da

C_1=liscio giallo p_1=9/16

C_2=liscio verde p_2=3/16

C_3=grinzoso giallo p_3=3/16

C_4=grinzoso verde p_4=1/16

Nelle osservazioni condotte il numero di grani osservati era $n = 556$. Si ha quindi una distribuzione teorica di $np_1 = 312.75$, $np_2 = np_3 = 104.25$ e $np_4 = 34.75$. Il numero delle categorie è ovviamente $k = 4$. Le frequenze osservate da Mendel furono $n_1 = 315,\ n_2 = 101,\ n_3 = 108,\ n_4 = 32$ per un totale di 556 osservazioni.

Abbiamo quindi

$$\chi^2 = \frac{(2.25)^2}{312.75} + \frac{(3.25)^2}{104.25} + \frac{(3.75)^2}{104.25} + \frac{(2.75)^2}{34.75} = 0.47$$

Cosa ne facciamo di questo valore? Poiché abbiamo $k = 4$, il numero calcolato deve essere messo in relazione con la distribuzione χ^2_3. Nella distribuzione χ^2_3 al valore 0.47 corrisponde un valore $P(\chi^2 > 0.47) = 1 - .07456892 = .92543108$.

Secondo lo "schema generale" del test di ipotesi si deve:

i) scegliere un ipotesi nulla;

ii) determinare l'intervallo di accettazione della nostra ipotesi nulla;

iii) confrontare il valore ottenuto dall'esperimento con l'intervallo di accettazione.

L'ipotesi nulla è che le frequenze attese nella varie categorie siano date dal prodotto del numero di esperimenti (n) per la probabilità p_i che il risultato dell'esperimento appartenga alla categoria C_i (questo è quanto abbiamo già fatto ne nostro esempio).

L'intervallo di accettazione viene scelto generalmente nella forma $(0, \chi^2_{crit})$ dove χ^2_{crit} è il valore per cui si ha $P(\chi^2_{k-1} > \chi^2_{crit}) = \alpha$, dove α è il livelo di errore che si è scelto. La scelta si fa utilizzando le tabelle dei valori critici. Nell'esempio, se scegliamo il solito livello di errore del 5%, otteniamo come valore critico $\chi^2_{crit} = 7.8147$. Il valore calcolato era $\chi^2 = 0.47$ che è più piccolo (e di molto) del valore critico, quindi "non possiamo respingere" l'ipotesi che Mendel avesse ragione.

Una possibile variante consiste nell'osservare che, se i dati sono veramente aleatori, anche un valore di χ^2 molto piccolo è assai improbabile. Questo osservazione diventa importante quando non siamo noi che abbiamo realizzato gli esperimenti, ma stiamo "controllando" i risultati riportati da altri. In questo caso si può sospettare che un valore molto basso di χ^2 sia il risultato non di esperimeti "reali" ma frutto di una (ingenua) falsificazione dei dati. Se si ha questo sospetto allora si può inglobare nell'intervallo di rifiuto dell'ipotesi nulla anche un intervallo della forma $(0, \chi^2_{crit_min})$ oltre all'intevallo $(0, \chi^2_{crit_Max})$. i valori critici $\chi^2_{crit_min}$ e $\chi^2_{crit_Max}$ si possono scegliere in modo che si abbia $P(\chi^2_{k-1} < \chi^2_{crit_min}) = \alpha/2$ e $P(\chi^2_{k-1} > \chi^2_{crit_Max}) = \alpha/2$, ovvero $P(\chi^2_{crit_min} < \chi^2_{k-1} < \chi^2_{crit}) = 1 - \alpha$. Nel nostro esempio, sempre prendendo un livello di errore del 5% otteniamo l'intervallo (0.2158, 9.3484).

Capitolo 9

Regressione lineare

9.1 La regressione lineare

La regressione è un *modello* che cerca di stabilire una relazione di "causalità" tra due variabili aleatorie.

Il nesso di causalità dipende dal contesto da cui le variabili aleatorie provengono, e la "spiegazione" di questo nesso è del tutto interna alla disciplina in esame. Tuttavia la struttura statistica è la stessa indipendentemente dal contesto interpretativo.

Selezioniamo due variabili $\boldsymbol{y}$ e $\boldsymbol{x}$ che vogliamo interpretare rispettivamente come *variabile dipendente* e *variabile indipendente*. Ciò significa che consideriamo la caratteristica della popolazione misurata dalla variabile $\boldsymbol{x}$ come una *causa* della caratteristica misurata dalla $\boldsymbol{y}$. Per esempio, possiamo misurare per un campione di N città, la quantità di polveri fini disperse nell'aria e il tasso di incidenza delle allergie più comuni: è ragionevole aspettarsi una "influenza" dei valori della prima variabile su quelli della seconda, e si tende a considerare la prima caratteristica come (una) causa della seconda.

Il modello più semplice per descivere quantitativamente un tale nesso causale è quello di *assumere che la variabile $\boldsymbol{y}$ sia una funzione lineare della variabile $\boldsymbol{x}$*, ovvero che sia possibile scrivere

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} \tag{9.1}$$

per due opportuni valori dei parametri α e β. Questo modello è ovviamente troppo rigido e non sarà mai validato dalla pratica. Un primo motivo è che, essendo le misurazioni comunque affette da errori, anche in presenza di un effettivo legame lineare tra le variabili, le misurazioni delle due variabili difficilmente soddisfanno la 9.1 (p.e. "sappiamo" che tra la lunghezza $\boldsymbol{x}$ del lato di un quadrato e la lunghezza $\boldsymbol{y}$ della sua diagonale esiste il legame "esatto" $\boldsymbol{y} = \sqrt{2}\boldsymbol{x}$: provate ora a disegnare un certo numero di quadrati, misurare rispettivamente i lati e le diagonali e vedere se i numeri si corrispondono secondo questa legge[1]). Un altro motivo è che altre concause potrebbero essere presenti nella determinazione di $\boldsymbol{y}$, a valori determinati di $\boldsymbol{x}$ (tornando all'esempio dell'allergia, la flora locale ha indubbiamente anch'essa un'influenza accanto al tasso di inquinamento).

[1] Anche "teoricamente" questo accordo è "impossibile" dato che $\sqrt{2}$ è un numero con infinite cifre decimali.

Il modello generale di regressione lineare si esprime quindi ponendo

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \mathbf{e} \tag{9.2}$$

dove la variabile aletaoria $\mathbf{e}$ rappresenta l'errore dovuto alle misurazioni o alle concause non considerate[2].

La stima dei parametri α e β si effettua con il metodo dei minimi quadrati, ovvero cercando i numeri a e b che rendono minimo l'*errore* definito da

$$E = \sum_{i=1}^{N} (y_i - a - bx_i)^2 \tag{9.3}$$

ovvero la somma degli scarti quadratici[3] tra i dati misurati y_i e i *dati previsti* $\hat{y}_i = a - bx_i$.

Da un insieme di dati $(x_i, y_i),\ i = 1, \ldots, N$ otteniamo la stima per b

$$b = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})} = \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{var}(\boldsymbol{x})} \tag{9.4}$$

(dove la covarianza e la varianza devono intendersi come varianza e covarianza delle variabili *stimate dai dati*, e l'intercetta a è determinata dalla condizione di "passaggio dal baricentro dei dati"

$$a = \overline{y} - b\overline{x} \tag{9.5}$$

dove $\overline{y}$ e $\overline{x}$ sono i valor medi (stimati dai dati) delle due variabili.

Il valore minimo dell'errore 9.3 è dato da

$$E_{min} = \sum_{i=1}^{N}(y_i - \overline{y})^2 - \frac{\sum_{i=1}^{N}\left((x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{9.6}$$

come si può verificare sviluppando il quadrato in 9.3 e sostituendo i valori di α e β dati da 9.5 e 9.4. La 9.6 si può riscrivere come

$$E_{min} = \mathrm{var}(\boldsymbol{y}) \left(1 - \frac{\left(\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})\right)^2}{\mathrm{var}(\boldsymbol{x})\mathrm{var}(\boldsymbol{y})}\right) \tag{9.7}$$

Ne segue che la regressione lineare dà una descrizione tanto più attendibile quanto le variabili (o meglio i dati) sono correlati, ovvero tanto più quanto il valore assoluto del coefficiente di correlazione

$$\rho = \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\mathrm{var}(\boldsymbol{x})\mathrm{var}(\boldsymbol{y})}}$$

[2] Scrivendo la 9.2 in realtà si sta già facendo un'ipotesi forte all'interno del modello: l'errore in 9.2 è una variabile aleatoria che *non dipende dal valore assunto da* $\boldsymbol{x}$. Nel gergo degli statitici questa situazione prende il nome regressione *omoscedastica*. Più in generale l'errore che si commette assumendo $y = \alpha + \beta x$ per un dato x (che è la distribuzione di $\boldsymbol{y}$ condizionata a $\boldsymbol{x} = x$) dipenderà dal valore x. Questa situazione è detta di regressione *eteroscedastica*. Nelle applicazioni l'ipotesi che la regressione sia omoscedastica è quasi sempre (implicitamente) assunta, e dà risultati ragionevoli.

[3] Questa è la ragione del perché il metodo prende il nome di *minimi quadrati*. Esso è dovuto a C.F. Gauss, ed è utilizzato come "ricetta universale" per trattare il caso di osservazioni "sovrabbondanti" (nel nostro caso quale il problema è quello di trovare la retta "più vicina" ai punti del piano rappresentati dalle coppie (x_i, y_i): questo è un tipico problema con dati sovrabbondanti e "contraddittori" poiché per ogni coppia di punti passa una sola retta e per tre punti, in generale, non ne passa alcuna!

si avvicina a 1 ($\rho = \pm 1$ quando i dati sono allineati lungo una retta, con $\rho = 1$ se i dati "crescono insieme", $\rho = -1$ se $\boldsymbol{y}$ descresce quando $\boldsymbol{x}$ cresce). Per valori di $|\rho|$ "discosti" da 1, la regressione lineare perde di senso, anche se ciò non esclude che tra i dati possa ancora sussistere una legame funzionale non lineare.

Capitolo 10

Generazione di numeri casuali

Un problema che si può porre quando si vogliano fare degli esperimenti "simulati", ovvero delle ricostruzioni al computer delle fenomenologie di esperimenti, è quello di generare dei numeri che costituiscano delle realizzazioni di una varibile aleatoria con una prescritta distribuzione.

Per esempio, se voglio "simulare" degli errori sperimentali dovuti a un apparecchio di misura, è ragionevole immaginare che l'errore sia distribuito normalmente, cioè la probabilità che l'errore sia compreso, p.e., tra $-a$ e a sia data da

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-a}^{a} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \tag{10.1}$$

In genere i computer (o meglio i programmi come fogli elettronici, linguaggi di programmazione, etc.) offrono un "generatore" di numeri *pseudo-casuali.* Questo significa che il risultato di una serie di chiamate di un'opportuna funzione (p.e. la funzione **RND()** nel Basic, o la funzione **CASUALE()** nella versione italiana di Excel) genera una successione di numeri dall'andamento "apparentemente casuale" (una precisa definizione di questo concetto è materia spinosissima!).

Possiamo quindi pensare a questa funzione come a una variabile aletoria X. Le implementazioni che generalmente si trovano nei software sono concepite in modo che questa variabile sia *uniformemente distribuita* nell'intervallo $[0, 1]$. Questo vuol dire che il risultato di una chiamata della funzione **RND()** è un numero compreso tra 0 e 1 e ha "ugual probabilità" di cadere in un punto qualsiasi dell'intervallo.

Come possiamo sfruttare questa funzione per generare dei numeri casuali che siano distribuiti in altro modo, p.e. secondo una distribuzione normale?

Per dare una risposta a questo problema, riformuliamolo matematicamente:

Problema: *Sia data una variabile aleatoria X con funzione di distribuzione cumulativa $F_X(x)$ e una funzione $F(y)$ monotona crescente e tale che*

$$\lim_{z\to-\infty} F(z) = 0 \leq F(y) \leq \lim_{z\to\infty} F(z) = 1$$

(queste sono le caratteristiche di un funzione di distribuzione cumulativa)

trovare una funzione g in modo che la variabile aleatoria $Y = g(X)$ abbia la funzione F come funzione di distribuzione cumulativa.

La risoluzione del problema è data dalla funzione $g(x) = F^{-1}(F_X(x))$.

Infatti, posto $Y = F^{-1}(F_X(X))$ si ha che $P(Y \leq y) = P(F^{-1}(F_X(X) \leq y) = P(X \leq F_X^{-1}(F(y)))$ che a sua volta è data da $F_X(F_X^{-1}(F(y))) = F(y)$.

Vediamo quindi com'è possibile costruire un generatore di numeri casuali distribuiti secondo la normale standard se disponiamo di un generatore di numeri casuali X uniformemente distribuiti nell'intervallo $[0, 1]$.

In questo caso la funzione F_X è data da:

$$F_X(x) = \begin{cases} 0 & \text{se } x \leq 0 \\ x & \text{se } 0 \leq x \leq 1 \\ 1 & \text{se } 1 \leq x \end{cases}$$

La funzione di distribuzione della normale standard è

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) \, ds.$$

per cui possiamo generare dei numeri casuali distribuiti secondo la normale semplicemente combinando la funzione che genera i numeri casuali uniformenente distribuiti con l'inversa della funzione F.

(Nota che la funzione inversa della distribuzione cumulativa della normale è presente nei maggiori programmi statistici e anche lei più diffusi fogli elettronici: p.e. il Excel, versione italiana, si chiama INV.NORM.ST(.) mentre la funzione che genera numeri casuali sui chiama CASUALE(); quindi per generare i numeri secondo la distribuzione normale standard si deve comporre le due funzioni e calcolare

INV.NORM.ST(CASUALE()),

senza argomento!)

Distribuzione Normale

Distribuzione cumulativa normale
$F(z) = P(Z \leq z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Valori critici t-Student

Valori critici della t-Student
$a(x) = P(\{t_n < -x\} \cup \{t_n > x\})$,
$Q(x) = P(-x < t_n < x)$, $F(x) = P(t_n < x)$
Esempio: x tale che $P(-x < t_8 < x) = 95\%$ è 2.306
n sono i gradi libertà

a	0.5	0.2	0.1	0.05	0.02	0.01	0.001
Q	0.5	0.8	0.9	0.95	0.98	0.99	0.999
F	0.75	0.9	0.95	0.975	0.99	0.995	0.9995
n							
1	1	3.0777	6.3137	12.706	31.821	63.656	636.58
2	0.8165	1.8856	2.92	4.3027	6.9645	9.925	31.6
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408	12.924
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041	8.6101
5	0.7267	1.4759	2.015	2.5706	3.3649	4.0321	6.8685
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074	5.9587
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995	5.4081
8	0.7064	1.3968	1.8595	2.306	2.8965	3.3554	5.0414
9	0.7027	1.383	1.8331	2.2622	2.8214	3.2498	4.7809
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693	4.5868
11	0.6974	1.3634	1.7959	2.201	2.7181	3.1058	4.4369
12	0.6955	1.3562	1.7823	2.1788	2.681	3.0545	4.3178
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123	4.2209
14	0.6924	1.345	1.7613	2.1448	2.6245	2.9768	4.1403
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467	4.0728
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208	4.0149
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784	3.9217
19	0.6876	1.3277	1.7291	2.093	2.5395	2.8609	3.8833
20	0.687	1.3253	1.7247	2.086	2.528	2.8453	3.8496
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188	3.7922
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.797	3.7454
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251
26	0.684	1.315	1.7056	2.0555	2.4786	2.7787	3.7067
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707	3.6895
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739
29	0.683	1.3114	1.6991	2.0452	2.462	2.7564	3.6595
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.75	3.646

Valori critici χ^2

Valori critici della distribuzione χ^2
$F(x) = P(\chi^2_n < x)$, n gradi di libertà

F	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
n										
1	4E-05	0.0002	0.001	0.0039	0.0158	2.7055	3.8415	5.0239	6.6349	7.8794
2	0.01	0.0201	0.0506	0.1026	0.2107	4.6052	5.9915	7.3778	9.2104	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.2514	7.8147	9.3484	11.345	12.838
4	0.207	0.2971	0.4844	0.7107	1.0636	7.7794	9.4877	11.143	13.277	14.86
5	0.4118	0.5543	0.8312	1.1455	1.6103	9.2363	11.07	12.832	15.086	16.75
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.239	1.6899	2.1673	2.8331	12.017	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.09	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1558	2.5582	3.247	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.92	24.725	26.757
12	3.0738	3.5706	4.4038	5.226	6.3038	18.549	21.026	23.337	26.217	28.3
13	3.565	4.1069	5.0087	5.8919	7.0415	19.812	22.362	24.736	27.688	29.819
14	4.0747	4.6604	5.6287	6.5706	7.7895	21.064	23.685	26.119	29.141	31.319
15	4.6009	5.2294	6.2621	7.2609	8.5468	22.307	24.996	27.488	30.578	32.801
16	5.1422	5.8122	6.9077	7.9616	9.3122	23.542	26.296	28.845	32	34.267
17	5.6973	6.4077	7.5642	8.6718	10.085	24.769	27.587	30.191	33.409	35.718
18	6.2648	7.0149	8.2307	9.3904	10.865	25.989	28.869	31.526	34.805	37.156
19	6.8439	7.6327	8.9065	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.4338	8.2604	9.5908	10.851	12.443	28.412	31.41	34.17	37.566	39.997
21	8.0336	8.8972	10.283	11.591	13.24	29.615	32.671	35.479	38.932	41.401
22	8.6427	9.5425	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.2604	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.8862	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.98	45.558
25	10.52	11.524	13.12	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.16	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.29
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672