Inferenza statistico-probabilistica

Luca Mari, versione 20.4.16

Contenuti

I principali concetti introdotti in questo capitolo

Introduzione

Come abbiamo visto, il problema generale della statistica descrittiva è: quali statistiche sintetizzano significativamente l'informazione portata da un campione? Grazie alla disponibilità di distribuzioni di probabilità note analiticamente, ci possiamo porre ora problemi complementari del tipo: il campione disponibile è stato ottenuto da una popolazione che segue una certa distribuzione di probabilità? oppure: dato un campione che si ipotizza ottenuto da una distribuzione, quali sono i parametri (per esempio media o deviazione standard) di tale distribuzione? oppure ancora: i due campioni disponibili sono relativi a variabili casuali statisticamente indipendenti?

Ciò che hanno in comune problemi di questo genere è il fatto che i dati disponibili sono solo parziali, essendo relativi a un campione, e non consentono perciò di giungere a una risposta certa, ma solo probabilistica, sulla popolazione: le conclusioni che si ottengono sono perciò inferenziali, appunto probabili e non certe, e le tecniche che si introducono si chiamano perciò di *inferenza statistico-probabilistica*.

Si può presentare la relazione tra tecniche statistico-probabilistiche descrittive e inferenziali con un semplice esempio. Qual è la probabilità di ottenere almeno una Testa in tre lanci di una moneta? Possiamo calcolare *la frequenza relativa* dell'evento E = 'almeno una Testa in tre lanci' in uno o più campioni dati, dunque eseguendo tre lanci per ogni campione, ma in generale dobbiamo aspettarci che le frequenze relative così ottenute potranno essere diverse. Al contrario, la domanda riguarda *la probabilità* $P(E)$, cosa che ci richiede di assumere un'ipotesi sulla struttura dell'esperimento, per esempio che la moneta possa produrre solo due eventi elementari, Testa e Croce, che eventi elementari ripetuti siano statisticamente indipendenti (cioè che la moneta non "abbia memoria", e quindi che la probabilità di ottenere T al secondo lancio non dipenda dall'esito del primo lancio), e che sia corretta e quindi tale che $P(\mathrm{T}) = P(\mathrm{C}) = 1/2$.

Ci sono almeno tre possibili strategie per trovare la risposta alla domanda su $P(E)$:

- l'evento E è equivalente a 'T al primo lancio (e T o C ai due successivi)' = T.., oppure 'C al primo lancio e T al secondo (e T o C al successivo)' = CT., oppure 'C al primo e al secondo lancio e T al terzo' = CCT; sono tre eventi disgiunti (nel senso che due di essi non possono accadere contemporaneamente), e dunque $P(E) = P(\text{T..}) + P(\text{CT.}) + P(\text{CCT}) = 1/2 + 1/4 + 1/8 = 7/8$;
- l'evento E è complementare a 'C nei tre lanci' = CCC; dunque $P(E) = 1 - P(\text{CCC}) = 1 - 1/2^3 = 7/8$;
- si può infine ricorrere a un esperimento, che oggi plausibilmente realizzeremmo mediante simulazione, generando un numero elevato di triple di T e C nelle condizioni specificate dalla struttura dell'esperimento e calcolando quindi la frequenza relativa dell'evento E, intesa come una stima di $P(E)$.

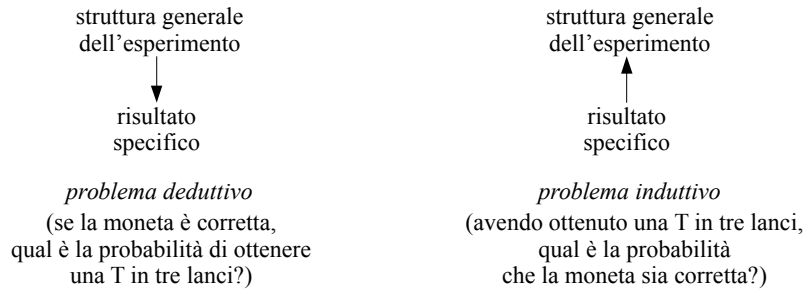
Un esempio della terza strategia, attuato per altro con un numero di ripetizioni (15) non sufficientemente elevato da garantire una buona approssimazione della frequenza relativa alla probabilità.

Esperimenti	Ripet1	Ripet2	Ripet3	Ripet4	Ripet5	Ripet6	Ripet7	Ripet8	Ripet9	Ripet10	Ripet11	Ripet12	Ripet13	Ripet14	Ripet15
	C	C	C	C	C	C	C	T	C	C	C	T	C	C	C
	T	T	C	T	T	C	C	T	T	C	T	T	C	C	T
	T	C	C	C	C	C	C	C	T	T	C	T	T	C	T
Num T	2	1	0	1	1	0	0	2	2	1	1	3	1	0	2
E verificato?	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE

Num *E* verif 11
Num *E* verif / Num esp **0.733** impiegato come stima di $P(E)$

In ogni caso, si tratta di un problema di tipo deduttivo, dal generale al particolare: si ipotizza la struttura dell'esperimento (la moneta è corretta e i lanci sono indipendenti), e se ne vuole ricavare l'informazione su un caso particolare.
Consideriamo ora quest'altro problema: qual è la probabilità che avendo ottenuto una sola T in tre lanci la moneta sia corretta? Si tratta dunque di un *problema inverso*:

- non: data la struttura dell'esperimento calcolare con quale probabilità / frequenza si ottiene un certo evento,
- ma, al contrario: dato un certo evento, qual è la struttura dell'esperimento che l'ha generato?



Per affrontare problemi induttivi con l'ausilio di strumenti probabilistici occorre dichiarare quali strutture dell'esperimento si ritengono possibili, in modo da fornire una risposta del tipo: "la struttura ipotizzata ... ha probabilità ...".
In questo caso, è necessario in particolare chiarire cosa si intende con "moneta corretta". A rigore, nessuna moneta è *perfettamente* simmetrica, e quindi è plausibile che nessuna moneta sia "perfettamente corretta", con la conseguenza che se ponessimo come ipotesi "moneta perfettamente corretta" e il suo complementare, "moneta non perfettamente corretta", potremmo ragionevolmente concludere che la prima ha probabilità 0, e quindi la seconda probabilità 1, anche senza compiere alcun esperimento. Per caratterizzare la struttura dell'esperimento potremmo allora distinguere i due casi di moneta "almeno sufficientemente corretta" e "non sufficientemente corretta". Al di là del fatto che, naturalmente, il concetto di 'sufficiente correttezza' dovrebbe essere precisato, si comprende la logica alla base del ragionamento inferenziale che stiamo introducendo:

- si specificano le possibili strutture dell'esperimento mediante due o più ipotesi alternative (nel caso potrebbero essere H = 'moneta almeno abbastanza corretta' e $\overline{H}$ = 'moneta non abbastanza corretta') (il simbolo H richiama il termine inglese *hypothesis*), e
- a partire dai dati D disponibili (per esempio, D = 'una T in tre lanci')
- si vuole calcolare la probabilità di una o più ipotesi specificate nelle condizioni date (nel caso, la probabilità che la moneta sia almeno abbastanza corretta avendo osservato una T in tre lanci, dunque la probabilità condizionata $P(H|D)$).

Ma come si calcolano le probabilità di ipotesi a partire da dati?

La regola di Bayes come strumento di inferenza: due ipotesi alternative

Per trovare una risposta a questo genere di domande, introduciamo un secondo problema, un poco più semplice di quello presentato sopra ma, come vedremo, con la stessa struttura di base, e basato sugli stessi elementi: le probabilità condizionate e, come vedremo, la regola di Bayes. Una volta impostato e risolto questo secondo problema, torneremo al primo.

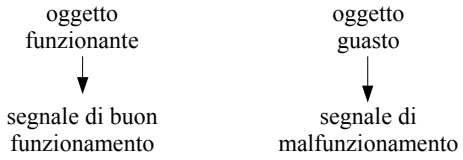
Un oggetto è scelto a caso da un lotto per essere sottoposto a un test finalizzato ad accertare suoi possibili malfunzionamenti. Il test, come accade spesso, non è perfetto: da analisi precedenti è noto che i casi di malfunzionamento sono sempre identificati, ma anche che in una piccola percentuale di casi, pari allo 0.2%, il test segnala malfunzionamenti anche quando l'oggetto è in effetti funzionante. Supponiamo inoltre che ancora da analisi precedenti si sappia che gli oggetti guasti sono lo 0.5% del totale. La domanda a cui siamo interessati è: *dato un risultato per il test, qual è la probabilità che l'oggetto in esame sia guasto?*
Il primo e più critico problema che si pone è quello di formalizzare questa descrizione, "traducendola dall'italiano" in cui è stata scritta nel linguaggio della matematica con cui potremo effettuare i calcoli.
Siamo in presenza di:

- due *ipotesi*, che qui operano come cause: H = oggetto funzionante e $\overline{H}$ = oggetto guasto,

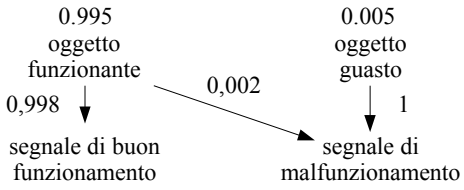
e corrispondentemente di:

- due *risultati osservabili*, e quindi effetti delle cause: D = segnalazione di buon funzionamento e $\overline{D}$ = segnalazione di malfunzionamento.

Si noti che il problema di ricostruire la causa che ha prodotto un certo risultato si pone perché il test non è perfettamente affidabile. Se il test avesse un comportamento ideale, la struttura dell'esperimento sarebbe infatti, semplicemente:



e l'informazione prodotta dal test sarebbe invertibile: osservando un segnale di malfunzionamento se ne potrebbe immediatamente concludere che l'oggetto è guasto. Ma abbiamo visto che non è così, dato che nello 0.2% dei casi il test segnala malfunzionamenti anche quando l'oggetto è funzionante. L'informazione disponibile sulla dipendenza dei risultati osservabili dalle cause ipotetiche, e quindi sulla struttura dell'esperimento, è più complessa, e può essere dunque sintetizzata così (i valori numerici si riferiscono alle probabilità indicate sopra):



In presenza di un segnale di buon funzionamento non ci sono ambiguità: l'oggetto è funzionante (nel diagramma: partendo dal risultato 'segnale di buon funzionamento' si può percorrere, inversamente, una sola freccia, che porta alla causa 'oggetto funzionante'). Ma se il test fornisce un segnale di mal funzionamento *siamo in una situazione di incertezza*, perché lo stesso dato, appunto il segnale di mal funzionamento, può derivare da due cause distinte: oggetto guasto e oggetto funzionante.

Consideriamo dunque la struttura del problema:

- date le probabilità per le ipotesi identificate:
  $P(\overline{H})=0.005$ [probabilità di oggetto non funzionante]
  $P(H)=1-P(\overline{H})=0.995$ [probabilità di oggetto funzionante]
- e date le probabilità di ottenere un certo risultato nel caso in cui una certa ipotesi sia vera:
  $P(D|H)=0.998$ [probabilità di segnale di buon funzionamento nel caso di oggetto funzionante: test corretto]
  $P(\overline{D}|H)=1-P(D|H)=0.002$ [probabilità di segnale di malfunzionamento nel caso di oggetto funzionante: test errato]
  $P(D|\overline{H})=0$ [probabilità di segnale di buon funzionamento nel caso di oggetto guasto: test errato]
  $P(\overline{D}|\overline{H})=1-P(D|\overline{H})=1$ [probabilità di segnale di malfunzionamento nel caso di oggetto guasto: test corretto]
- si vogliono calcolare le probabilità per le due ipotesi identificate a partire dal risultato del test, e quindi:
  $P(H|\overline{D})$ (qual è la probabilità che l'oggetto sia funzionante nonostante il test lo segnali malfunzionante?)
  e:
  $P(\overline{H}|\overline{D})$ (qual è la probabilità che l'oggetto sia effettivamente guasto se il test lo segnala malfunzionante?)

dove naturalmente $P(H|\overline{D})+P(\overline{H}|\overline{D})=1$.
Sfruttiamo a questo punto la regola di Bayes nella forma:

$$P(H\mid D)\propto P(D\mid H)\,P(H)$$

cioè:

$$\text{posterior}\propto\text{verosimiglianza}\times\text{prior}$$

Possiamo allora finalmente sostituire i valori numerici che avevamo ricavato sopra per i due problemi che vogliamo risolvere:

$$P(H\mid\neg D)\propto P(\neg D\mid H)\,P(H)=0.002\times0.995=0.002$$
$$P(\neg H\mid\neg D)\propto P(\neg D\mid\neg H)\,P(\neg H)=1\times0.005=0.005$$

Questi due valori sono corretti, ma sono solo proporzionali alle probabilità cercate, che si ottengono notando che le due posterior appena calcolate, $P(H|\neg D)$ e $P(\neg H|\neg D)$, sono esaustive e mutuamente esclusive, e quindi devono essere tali che:

$$P(H|\neg D) + P(\neg H|\neg D) = 1$$

Dobbiamo perciò normalizzare i due valori ottenuti, 0.002 e 0.005, in modo che la loro somma sia 1:

$$P(H\mid\neg D)=\frac{0.002}{0.002+0.005}\approx0.28$$

e:

$$P(\neg H\mid\neg D)=\frac{0.005}{0.002+0.005}\approx0.72$$

Si noti quindi che abbiamo trovato in questo modo una tecnica per calcolare il valore del denominatore nella regola di Bayes pur senza conoscere la probabilità del dato; in questo caso:

$$P(\neg D)=P(\neg D\mid H)\,P(H)+P(\neg D\mid\neg H)\,P(\neg H)$$

Dunque attraverso la regola di Bayes siamo passati:

- da una conoscenza generica e a priori, $P(H)$: il 99.5% dei prodotti è funzionante,
- alla conoscenza a posteriori, specifica sull'oggetto del test, dato che il test ha prodotto un segnale di malfunzionamento, la probabilità che l'oggetto sia comunque funzionante, $P(H|\neg D)$, è del 28% e la probabilità complementare, che l'oggetto sia invece guasto, $P(\neg H|\neg D)$, è del 72% .

La struttura dell'esperimento nel caso di due ipotesi e due possibili risultati

Un problema come quello appena presentato è caratterizzato da una struttura particolarmente semplice: occorre decidere tra due ipotesi alternative, H e $\neg H$, sulla base del fatto di aver ottenuto uno tra due possibili risultati, D e $\neg D$. Un esperimento di questo genere si presenta in modo particolarmente espressivo in una matrice a due colonne, per le ipotesi, e a due righe, per i risultati, i cui elementi rappresentano i quattro possibili casi:

	H	$\neg H$
D	risultato D nel caso in cui l'ipotesi sia H	risultato D nel caso in cui l'ipotesi sia $\neg H$
$\neg D$	risultato $\neg D$ nel caso in cui l'ipotesi sia H	risultato $\neg D$ nel caso in cui l'ipotesi sia $\neg H$

Mantenendo l'interpretazione data sopra per il test, i due casi in cui il test stesso ha un comportamento non corretto sono quelli descritti negli elementi della diagonale secondaria della matrice:

- quando il test fornisce il risultato $\neg D$ pur essendo H l'ipotesi corretta: si tratta di un caso di *falso negativo* ("il test indica di rifiutare l'ipotesi che invece avrebbe dovuto essere accettata": si chiama anche *errore di prima specie*);
- quando il test fornisce il risultato D pur essendo $\neg H$ l'ipotesi corretta: si tratta di un caso di *falso positivo* ("il test indica di accettare l'ipotesi che invece avrebbe dovuto essere rifiutata": si chiama anche *errore di seconda specie*).

Dunque:

	H	$\neg H$
D	*vero positivo*	*falso positivo*
$\neg D$	*falso negativo*	*vero negativo*

Come abbiamo visto, l'esperimento può essere dunque caratterizzato in termini probabilistici mediante le verosimiglianze, per esempio stimate su base statistica mediante le frequenze relative dei falsi positivi e dei falsi negativi:

	H	$\neg H$
D	$P(D\|H) = 1 - \mathrm{P}(\neg D\|H)$	$P(D\|\neg H)$, stimata mediante la frequenza relativa dei falsi positivi
$\neg D$	$\mathrm{P}(\neg D\|H)$, stimata mediante la frequenza relativa dei falsi negativi	$P(\neg D\|\neg H) = 1 - \mathrm{P}(D\|\neg H)$

Un esempio (adattato da I. Hacking, *An introduction to probability and inductive logic*, Cambridge University Press, 2001).
Un giudice deve decidere a proposito del seguente problema. In città ci sono due aziende di taxi, TaxiVerdi e TaxiBlu, che usano veicoli verdi e blu rispettivamente. L'85% dei taxi sono di TaxiVerdi. In una sera nebbiosa un taxi provoca un incidente ma non si ferma per gli accertamenti del caso. Un testimone dichiara che esso era blu. Per accertare l'affidabilità del testimone, il giudice lo sottopone a dei test, da cui emerge che in condizioni analoghe a quelle dell'incidente, il testimone riconosce correttamente il colore del taxi nell'80% delle volte. Cosa può concludere il giudice su questa base?

A partire da quanto riportato dal testimone, cioè il dato che il taxi era blu, D_B, mettiamoci nella condizione di calcolare attraverso la regola di Bayes le probabilità che il taxi fosse effettivamente blu oppure fosse verde.
Le ipotesi che si ritengono possibili sono dunque due: H_V = il responsabile dell'incidente è un taxi verde; H_B = il responsabile dell'incidente è un taxi blu. Le loro probabilità a priori sono $P(H_V) = 0.85$ e $P(H_B) = 0.15$.
La struttura della situazione, per come ottenuta dai risultati dei test, è descritta dalle verosimiglianze $P(D_B|H_B) = 0.8$ (e perciò $P(D_V|H_B) = 0.2$), $P(D_V|H_V) = 0.8$ (e perciò $P(D_B|H_V) = 0.2$).
Ciò è sufficiente per calcolare le probabilità a posteriori:

$$P(H_B|D_B) \propto P(D_B|H_B)\, P(H_B) = 0.8 \times 0.15$$
$$P(H_V|D_B) \propto P(D_B|H_V)\, P(H_V) = 0.2 \times 0.85$$

Il fattore di normalizzazione per cui dividere entrambi, cioè il denominatore nella regola di Bayes, è:

$$P(D_B|H_B)\, P(H_B) + P(D_B|H_V)\, P(H_V) = 0.8 \times 0.15 + 0.2 \times 0.85$$

Dunque:

$P(H_B|D_B) = 0.41$
$P(H_V|D_B) = 0.59$

Nonostante la testimonianza, la non completa affidabilità del testimone stesso è più che compensata dalla maggiore diffusione di taxi verdi: rimane ancora più probabile che il responsabile dell'incidente sia un veicolo dell'azienda TaxiVerdi.

E' interessante modificare i dati quantitativi del problema, per studiare alcuni casi particolari:

- se le prior fossero $P(H_V) = 0.5$ e $P(H_B) = 0.5$, ci si dovrebbe affidare al testimone, pur che fosse affidabile più che nel 50% dei casi;
- al limite, con prior uniformi se il testimone fosse completamente inaffidabile, cioè $P(D_B|H_B) = 0.5$ e $P(D_V|H_V) = 0.5$, la situazione sarebbe di completa indecisione;
- all'estremo opposto, con un testimone perfettamente affidabile, cioè $P(D_B|H_B) = 1$ e $P(D_V|H_V) = 1$, ci si affida al testimone a prescindere dalla prior (naturalmente purché $P(H_B) > 0$; ma un testimone affidabile non può aver visto un taxi blu se in città non ci sono taxi blu...).

E se arrivasse un secondo testimone, affidabile come il primo e indipendente nel suo giudizio da questo, e che pure indica che il taxi era blu? Usando la posterior come nuova prior, si ottiene:

$$P(H_B|D_{1B},D_{2B}) = P(D_B|H_B)\,P(H_B|D_{1B}) / [P(D_B|H_B)\,P(H_B|D_{1B}) + P(D_B|H_V)\,P(H_V|D_{1B})] =$$
$$= 0.8 \times 0.41 / [0.8 \times 0.41 + 0.2 \times 0.59] = 0.73$$

Due testimoni indipendenti ma concordi riescono a rendere più probabile la loro testimonianza nonostante la prior sfavorevole.

La regola di Bayes come strumento di inferenza: *n* ipotesi con distribuzione discreta

Siamo pronti, a questo punto, per tornare al primo problema introdotto (qual è la probabilità che avendo ottenuto una sola T in tre lanci la moneta sia corretta?), che è un poco più complesso di quello appena risolto. Dobbiamo partire da un'ipotesi generale sulla struttura dell'esperimento:

- supponiamo che la probabilità (ignota) di ottenere T in un lancio sia q e che essa sia costante (e dunque in particolare che i lanci siano eventi indipendenti); attraverso q esprimiamo quindi le possibili ipotesi H: in particolare, se la moneta dà sempre C, $q = 0$, se la moneta dà sempre T, $q = 1$, e la correttezza perfetta della moneta corrisponde all'ipotesi $q = 1/2$;
- indichiamo con D l'evento da cui partiamo, 'una T in tre lanci';
- allora, in accordo a questa ipotesi la probabilità di ottenere D segue una distribuzione binomiale: $\binom{3}{1} q^1 (1-q)^{3-1} = 3q(1-q)^2$

Questa è dunque la probabilità dell'evento D a partire dall'ipotesi H: è la probabilità condizionata $P(D|H)$:

$$P(D|H) = 3q(1-q)^2$$

Occorre ora "ragionare al contrario", trattando $P(D|H)$ come funzione dell'ipotesi H, e quindi, appunto, della probabilità q: al variare di q, e quindi dell'ipotesi che descrive parametricamente la struttura dell'esperimento, $P(D|H)$ fornisce la probabilità che dall'esperimento si ottenga D, cioè quello che effettivamente supponiamo sia stato ottenuto. Come abbiamo visto, quando usata in questo modo, una probabilità condizionata del tipo P(dato|ipotesi) è una verosimiglianza, ed è considerata funzione dell'ipotesi. Per meglio mettere in evidenza questo uso, potremo scrivere esplicitamente $P(D|H_i)$ per indicare la verosimiglianza dell'ipotesi i-esima nel fornire il dato D.
Avendo dunque formalizzato la nostra conoscenza sull'esperimento mediante la verosimiglianza, torniamo a "ragionare in modo diretto": che informazione probabilistica ci fornisce il dato D sul parametro H_i che descrive la struttura dell'esperimento? Cioè che valore ha la probabilità $P(H_i|D)$ per le diverse ipotesi H_i che possiamo formulare?
Torniamo a usare la regola di Bayes, che in questo caso, cioè in presenza di un insieme di ipotesi H_i, si scrive:

$$P(H_i|D) = \frac{P(D|H_i)\,P(H_i)}{\sum_j P(D|H_j)\,P(H_j)}$$

Ricordando nuovamente che $P(H_i|D)$ può essere intesa come una distribuzione di probabilità sull'insieme delle ipotesi $\{H_i\}$, e quindi che:

$$\sum_i P(H_i|D) = 1$$

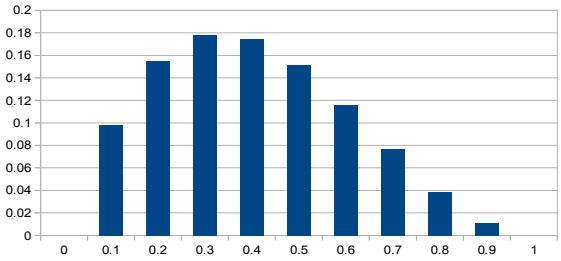
il denominatore nel termine a destra della formula può essere trattato come un fattore di normalizzazione e quindi tralasciato:

$$P(H_i|D) \propto P(D|H_i)P(H_i)$$

In accordo alla regola di Bayes, la posterior dell'ipotesi H_i a partire dal dato D, $P(H_i|D)$, è dunque proporzionale al prodotto della verosimiglianza di quell'ipotesi, $P(D|H_i)$, per la prior dell'ipotesi stessa, $P(H_i)$.

Definiamo ora la struttura parametrica del problema, per esempio, assumendo che le ipotesi possibili siano H_0: $P(\mathrm{T}) = 0$; H_1: $P(\mathrm{T}) = 0.1$; H_2: $P(\mathrm{T}) = 0.2$; ...; H_{10}: $P(\mathrm{T}) = 1$. Supponendo poi di non avere alcuna ulteriore informazione sulla moneta, potremo scegliere come prior la distribuzione uniforme, $P(H_i) = 1/11$ (le ipotesi scelte sono appunto 11). Allora:

ipotesi	q	$P(D\|H_i)$	$P(H_i)$	$P(D\|H_i)P(H_i)$	$P(H_i\|D)$
H_0	0	0.000	0.091	0.000	0.000
H_1	0.1	0.243	0.091	0.022	0.098
H_2	0.2	0.384	0.091	0.035	0.155
H_3	0.3	0.441	0.091	0.040	0.178
H_4	0.4	0.432	0.091	0.039	0.175
H_5	0.5	0.375	0.091	0.034	0.152
H_6	0.6	0.288	0.091	0.026	0.116
H_7	0.7	0.189	0.091	0.017	0.076
H_8	0.8	0.096	0.091	0.009	0.039
H_9	0.9	0.027	0.091	0.002	0.011
H_{10}	1	0.000	0.091	0.000	0.000



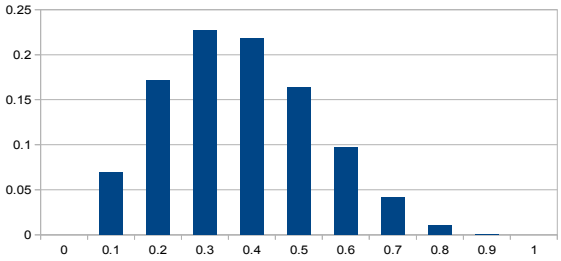
(nella scelta della prior avremmo potuto fare di meglio: se il dato è che in tre lanci si è ottenuta una T, e quindi due C, si sa già che le ipotesi H_0 e H_{10}, cioè che $P(\mathrm{T}) = 0$ e $P(\mathrm{T}) = 1$ rispettivamente, sono false, e quindi devono avere probabilità nulla).

Supponiamo a questo punto di ripetere l'esperimento e di ottenere come nuovo dato D_2 (essendo dunque il precedente D_1) ancora una T in tre lanci. Possiamo allora sfruttare questa informazione per migliorare la nostra stima delle probabilità delle diverse ipotesi H_i, usando la posterior $P(H_i|D_1)$ come nuova prior per calcolare la nuova posterior, $P(H_i|D_1,D_2)$, dunque mediante la regola di Bayes nella forma:

$$P(H_i|D_1,D_2) \propto P(D_2|H_i)P(H_i|D_1)$$

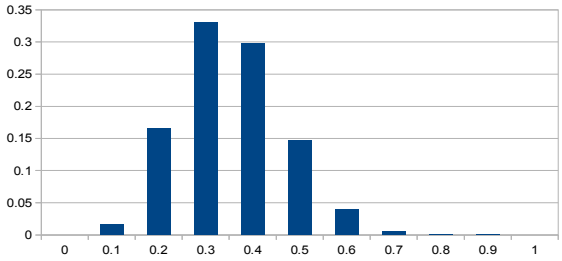
Allora:

ipotesi	q	verosim	prior	prodotto	posterior
H_0	0	0.000	0.000	0.000	0.000
H_1	0.1	0.243	0.098	0.024	0.069
H_2	0.2	0.384	0.155	0.060	0.172
H_3	0.3	0.441	0.178	0.079	0.227
H_4	0.4	0.432	0.175	0.075	0.218
H_5	0.5	0.375	0.152	0.057	0.164
H_6	0.6	0.288	0.116	0.034	0.097
H_7	0.7	0.189	0.076	0.014	0.042
H_8	0.8	0.096	0.039	0.004	0.011
H_9	0.9	0.027	0.011	0.000	0.001
H_{10}	1	0.000	0.000	0.000	0.000



Come si vede, questa seconda distribuzione di probabilità sulle ipotesi è un po' concentrata intorno al valore 1/3 di quanto fosse la precedente: il fatto che per due volte in tre lanci si sia ottenuta una T ha aumentato la probabilità delle ipotesi H_3: $P(\mathrm{T}) = 0.3$ e H_4: $P(\mathrm{T}) = 0.4$.

Ci possiamo aspettare, correttamente, che se questa situazione si ripetesse, e per esempio per cinque volte il dato fosse una T in tre lanci, la posterior $P(H_i|D_1,...,D_5)$ sarebbe ancora più concentrata, come infatti il grafico mostra:



La regola di Bayes come strumento di inferenza: *n* ipotesi con distribuzione continua

Le ipotesi di cui inferire la probabilità potrebbero riferirsi a un parametro di una distribuzione a dominio continuo, come nell'importante caso in cui ci si chiede se un elemento di valore x sia stato ottenuto da una popolazione a distribuzione gaussiana, con deviazione standard σ data, al variare della media μ, dove quindi le ipotesi sono relative proprio a μ. D'altra parte, nel continuo la probabilità di un singolo valore è comunque nulla, cioè se scegliessimo $H = \mu$, per un qualsiasi μ, avremmo la verosimiglianza:

$$P(D=x \mid H=\mu)=\int_{\mu}^{\mu} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)dm=0$$

che non potrebbe che produrre una posterior a sua volta a valore zero.
Naturalmente, il problema si risolve scegliendo come ipotesi non un singolo valore della media, ma un intervallo di valori, $H_i = [\mu_i, \mu_{i+1}]$, così che:

$$P(D=x \mid H=[\mu_i,\mu_{i+1}])=\int_{\mu_i}^{\mu_{i+1}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)dm$$

Per una generica distribuzione p, il valore di questo integrale può essere calcolato come:

$$\int_{\mu_i}^{\mu_{i+1}} p(m)\,dm=\int_{-\infty}^{\mu_{i+1}} p(m)\,dm-\int_{-\infty}^{\mu_i} p(m)\,dm$$

e quindi attraverso la distribuzione di probabilità cumulata F:

$$\int_{\mu_i}^{\mu_{i+1}} p(m)\,dm=F(\mu_{i+1})-F(\mu_i)$$

Supponiamo, per esempio, di aver ottenuto il valore $x=0.5$ e di voler inferire la probabilità delle ipotesi che esso sia stato ottenuto da una popolazione a distribuzione gaussiana, con deviazione standard σ = 1.0, al variare della media μ negli intervalli $H_1 = (-\infty, -2.0)$, $H_2 = (-2.0, -1.5)$, ..., $H_{10} = (2.0, +\infty)$. Supponiamo inoltre che la prior sia uniforme sull'insieme delle 10 ipotesi, $P(H_i) = 1/10$.
Questo è il risultato:

ipotesi		verosim	prior	prodotto	posterior
$H_1 =$	($-\infty$, -2.0)	0.006	0.1	0.001	0.006
$H_2 =$	(-2.0 , -1.5)	0.017	0.1	0.002	0.017
$H_3 =$	(-1.5 , -1.0)	0.044	0.1	0.004	0.044
$H_4 =$	(-1.0 , -0.5)	0.092	0.1	0.009	0.092
$H_5 =$	(-0.5 , 0.0)	0.150	0.1	0.015	0.150
$H_6 =$	(0.0 , 0.5)	0.191	0.1	0.019	0.191
$H_7 =$	(0.5 , 1.0)	0.191	0.1	0.019	0.191
$H_8 =$	(1.0 , 1.5)	0.150	0.1	0.015	0.150
$H_9 =$	(1.5 , 2.0)	0.092	0.1	0.009	0.092
$H_{10} =$	(2.0 , $+\infty$)	0.067	0.1	0.007	0.067

Dall'inferenza bayesiana ai test di ipotesi

Consideriamo nuovamente la regola di Bayes in presenza di un insieme di ipotesi H_i:

$$P(H_i|D)=\frac{P(D|H_i)P(H_i)}{\sum_j P(D|H_j)P(H_j)}$$

Se la prior $P(H_i)$ è uniforme, cioè $P(H_i) = 1/n$ nel caso di n ipotesi:

$$P(H_i|D)=\frac{P(D|H_i)\frac{1}{n}}{\sum_j P(D|H_j)\frac{1}{n}}=\frac{\frac{1}{n}P(D|H_i)}{\frac{1}{n}\sum_j P(D|H_j)}=\frac{P(D|H_i)}{\sum_j P(D|H_j)}$$

e quindi, considerando nuovamente $\sum_j P(D|H_j)$ come fattore di normalizzazione:

$$P(H_i|D)\propto P(D|H_i)$$

In questa situazione, *la posterior è proporzionale alla verosimiglianza*, che può dunque essere usata per stimare la posterior stessa

(e in particolare nel caso, come quello presentato sopra, in cui il problema sia di inferire la probabilità che un elemento di valore x sia stato ottenuto da una popolazione a distribuzione gaussiana, con deviazione standard σ data, al variare della media μ, le cose diventano ancora più semplici. Infatti dato che le ipotesi H_j sono mutuamente esclusive (cioè si riferiscono a intervalli disgiunti) ed esaustive (cioè complessivamente coprono tutto l'asse reale), il fattore di normalizzazione è:

$$\sum_j P(D|H_j)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)d\mu=1$$

Perciò, proprio come avevamo ottenuto nell'esempio precedente:

$$P(H_i|D)=P(D|H_i)$$

un risultato che può essere presentato così: in queste condizioni, l'ipotesi più probabile è quella che massimizza la verosimiglianza).

Se la regola di Bayes è uno strumento generale di inferenza per stabilire la distribuzione di probabilità su un insieme di ipotesi, sulla base della proporzionalità tra verosimiglianza e posterior – che dunque vale nel caso di prior uniforme – anche il punto di vista frequentistico alla probabilità ha sviluppato numerose tecniche per mettere alla prova delle ipotesi: introduciamo dapprima la logica dei test di ipotesi e quindi sperimentiamola in alcuni casi.

I test di ipotesi

Un campione (che per semplicità supporremo univariato, con valori indipendenti e ottenuti tutti da una stessa popolazione) porta informazione statistica sulla popolazione da cui è stato ottenuto. Si possono dunque formulare ipotesi su tale popolazione, e in particolare sulla sua distribuzione, usando quindi i dati forniti dal campione per controllare (in inglese: *to test*) le ipotesi.

Il caso più semplice, che impieghiamo anche per introdurre la logica generale dei *test di ipotesi*, è quello in cui ci si chiede se un singolo elemento, di valore x, sia stato ottenuto da una certa popolazione, per esempio a distribuzione gaussiana con media μ e deviazione standard σ, dunque $N(\mu,\sigma)$. Si formulano allora due ipotesi in opposizione:

- l'ipotesi di *default*, chiamata *ipotesi nulla*, H_0, che l'elemento sia stato effettivamente ottenuto dalla popolazione data, e quindi che eventuali differenze tra il valore x dell'elemento e la media della popolazione, μ, siano solo effetti dovuti al modo con cui il campionamento è stato effettuato, cioè dunque solo "effetti casuali";
- l'*ipotesi alternativa*, H_1, che l'elemento sia stato ottenuto da una popolazione non specificata ma comunque diversa da quella data.

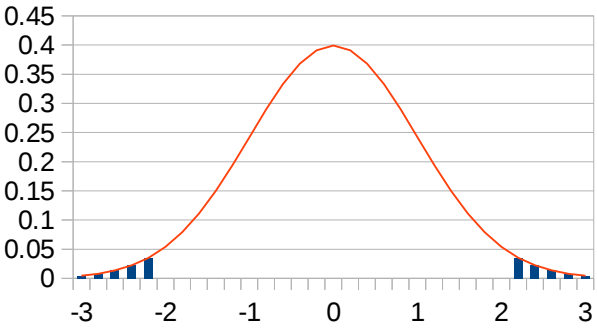
Se trattiamo il valore x come una realizzazione di una variabile casuale X, allora le due ipotesi si possono formulare così:

- ipotesi nulla H_0: $X\sim N(\mu,\sigma)$, e dunque se ripetessimo un numero sufficientemente elevato di volte il campionamento dovremmo aspettarci che la media dei valori degli elementi del campione coincida con la media della popolazione, e ciò perché il campione è ottenuto proprio da quella popolazione);
- ipotesi alternativa H_1: non $X\sim N(\mu,\sigma)$.

Per arrivare a decidere tra le due ipotesi, occorre dotarci di uno strumento di decisione, identificando una *statistica del test*, in questo caso:

$$Z=\frac{X-\mu}{\sigma}$$

Così come X, anche Z è una variabile casuale. La statistica del test è impiegata secondo una logica di tipo condizionale: *se* X fosse distribuita come la popolazione data, cioè se l'ipotesi nulla fosse vera, *allora* Z sarebbe distribuita secondo la gaussiana standard, che dunque in questo test rappresenta la distribuzione $p(Z|H_0)$ e descrive quanto sarebbe probabile ottenere i diversi valori di Z nel caso in cui H_0 sia vera (mentre se l'elemento fosse ottenuto da una popolazione diversa allora Z seguirebbe una distribuzione diversa, eventualmente ancora una gaussiana ma con altri parametri).
L'idea a questo punto è semplice: valori bassi di questa probabilità, corrispondenti a valori grandi di $|Z|$ (dovuti dunque al fatto che x è molto maggiore o molto minore, relativamente a σ, della media della popolazione), rendono poco verosimile l'ipotesi nulla H_0 e quindi suggeriscono di rifiutarla, accettando invece l'ipotesi alternativa H_1. La situazione è rappresentata nel grafico successivo: se per esempio l'elemento avesse un valore in una delle due code della gaussiana standard, evidenziate dalle barre, la probabilità che sia stato ottenuto appunto da questa distribuzione sarebbe piccola a sufficienza da poter supporre che più probabilmente è stato ottenuto da una qualche altra distribuzione.



Si noti l'asimmetria:

- l'ipotesi nulla è specifica: l'elemento è ottenuto da una ben determinata popolazione,
- l'ipotesi alternativa è generica: l'elemento è ottenuto da una qualsiasi altra popolazione.

Ciò implica che mentre un valore piccolo della verosimiglianza conduce a rifiutare H_0 e ad accettare H_1, da un valore grande non si dovrebbe concludere che H_0 è corretta, ma solo che i dati disponibili non consentono di rifiutare H_0: l'elemento potrebbe essere davvero ottenuto da una popolazione diversa da quella ipotizzata, e nondimeno il valore di $|Z|$ potrebbe essere piccolo.
Torniamo al valore x e quindi al corrispondente per la statistica del test:

$$z=\frac{x-\mu}{\sigma}$$

Supponiamo, per esempio, che $\mu=1.23$ e $\sigma=0.12$, e quindi:

$$Z=\frac{X-1.23}{0.12}$$

Supponiamo poi di aver ottenuto $x=1.45$ e quindi $z=(1.45-1.23)/0.12\approx 1.84$.
La probabilità che, sempre supponendo che H_0 sia vera, ripetendo il campionamento in condizioni analoghe si ottenga un valore non maggiore di $z=1.84$ sarebbe:

$$g_{\text{cum}}(1.84)=\int_{-\infty}^{1.84} g(y)\,dy\approx 0.97$$

dove g è la pdf gaussiana standard e g_{cum} è la sua cumulata, e quindi:

$$\int_{1.84}^{+\infty} g(y)\,dy=1-g_{\text{cum}}(1.84)\approx 0.03$$

è la probabilità di ottenere un valore maggiore di $z=1.84$.
Dunque se si campiona da una popolazione a distribuzione gaussiana con $\mu=1.23$ e $\sigma=0.12$, solamente nel 3% dei casi ci si può aspettare di ottenere un valore $x=1.45$ o maggiore. D'altra parte, in questo caso l'ipotesi nulla può essere rifiutata se x è sufficientemente distante da μ, e non solo nel caso in cui x sia maggiore di μ: stiamo dunque trattando un problema "a due code" (in inglese *two-tailed*; evidentemente si

pongono anche problemi "a una coda", *one-tailed*, in cui l'ipotesi nulla potrebbe essere, per esempio, di stabilire se x sia maggiore di μ). Anche il caso dei valori minori di $-z \approx -1.84$ deve essere dunque considerato:

$$g_{\text{cum}}(-1.84) = \int_{-\infty}^{-1.84} g(y)\,dy$$

che, per la simmetria della distribuzione, è anch'esso ≈ 0.03.
La somma delle probabilità nelle due code della distribuzione:

$$\int_{-\infty}^{-1.84} g(y)\,dy + \int_{1.84}^{+\infty} g(y)\,dy = g_{\text{cum}}(-1.84) + 1 - g_{\text{cum}}(1.84) \approx 0.06$$

è dunque la probabilità di ottenere l'elemento da una popolazione distribuita come specificato dall'ipotesi nulla, cioè la probabilità del dato data l'ipotesi nulla, $P(D|H_0)$. Ma data la proporzionalità tra verosimiglianza e posterior, che nei casi gaussiani come questo diventa addirittura identità, possiamo trattare il valore $P(D|H_0)$ come proporzionale, o appunto uguale, a $P(H_0|D)$, cioè alla probabilità dell'ipotesi nulla a partire dal dato. Per brevità, tale probabilità è spesso chiamata *p-value* (a volte tradotto in italiano "p-dei-dati").
Per trasformare il test di ipotesi in una vera e propria regola decisionale, occorre a questo punto stabilire come scegliere tra le due ipotesi, con la logica che se il p-value è abbastanza elevato l'ipotesi nulla è sufficientemente verosimile ma non poter essere rifiutata. Ma quanto elevato? La scelta della "soglia di rifiuto", chiamata *livello di significatività* del test, è evidentemente convenzionale, e si adotta spesso il livello $\alpha = 0.05$.
Nell'esempio, poiché il p-value è circa uguale a 0.06, se si sceglie $\alpha = 0.05$ la conclusione è che i dati a disposizione non sono sufficienti per rifiutare l'ipotesi nulla, cioè, più concretamente in questo caso, che x non è abbastanza lontano da μ per concludere che sia stato ottenuto da una popolazione diversa da quella di partenza. Con una regola decisionale di questo genere, il livello di significatività del test corrisponde dunque alla probabilità di commettere l'errore di rifiutare l'ipotesi nulla quando in effetti è corretta, cioè alla probabilità di quello che in precedenza abbiamo chiamato errore di prima specie.
In conclusione, si noti che si sarebbe potuto anche ragionare inversamente. Dopo aver fissato il livello di significatività, continuiamo a supporre $\alpha = 0.05$, ci si può chiedere quali sono i *valori critici* per la variabile X oltre i quali l'ipotesi nulla sarebbe rifiutata. Data la simmetria della distribuzione, ciò corrisponde a dividere a metà la probabilità limite α e trovare nella distribuzione cumulata il valore $\bar{x}_1$ entro cui sta il primo 2.5% e il valore $\bar{x}_2$ oltre il quale sta l'ultimo 2.5% della distribuzione stessa. Tali valori critici si ottengono calcolando la distribuzione inversa in $\alpha/2$ e in $1-\alpha/2$:

$\bar{z}_1 = g_{\text{inv}}(0.025) \approx -1.96$
$\bar{z}_2 = g_{\text{inv}}(0.975) \approx 1.96$

e quindi da questi:

$\bar{x}_1 = \bar{z}_1 \sigma + \mu \approx 0.99$
$\bar{x}_2 = \bar{z}_2 \sigma + \mu \approx 1.47$

Dunque nelle condizioni date i due intervalli $(-\infty, 0.99)$ e $(1.47, +\infty)$ costituiscono la *regione critica* (o "di rifiuto") per il test, cioè l'insieme dei valori che conducono a rifiutare l'ipotesi nulla e ad accettare l'ipotesi alternativa. In modo complementare, i valori nell'intervallo $[0.99, 1.47]$ non consentono di rifiutare l'ipotesi nulla. In questo modo si vede facilmente l'effetto di cambiare il livello di significatività del test, per esempio aumentandolo ad $\alpha = 0.2$. I nuovi valori:

$\bar{z}_1 = g_{\text{inv}}(0.1) \approx -1.28$, e quindi $\bar{x}_1 \approx 1.08$
$\bar{z}_2 = g_{\text{inv}}(0.9) \approx 1.28$, e quindi $\bar{x}_2 \approx 1.38$

corrispondono a un intervallo $[1.08, 1.38]$ più stretto intorno a μ, e quindi a un criterio di accettazione dell'ipotesi alternativa più debole.

In sintesi, la procedura di test di ipotesi può essere realizzata come segue:
1. scegliere l'ipotesi nulla H_0 e quindi l'ipotesi alternativa H_1, in termini della popolazione considerata e dei suoi parametri;
2. scegliere la statistica del test e stabilire la distribuzione della verosimiglianza della statistica data l'ipotesi nulla;
3. scegliere il livello di significatività α del test;
4. calcolare dai dati disponibili D il valore della statistica del test;

quindi alternativamente:
5.1. dal valore ottenuto per la statistica del test, calcolare il p-value $P(D|H_0)$;
5.2. confrontare p con α: se $p<\alpha$, H_0 è rifiutata e H_1 accettata; altrimenti H_0 non può essere rifiutata;
oppure:
6.1. dal livello di significatività α, calcolare i valori critici, e quindi la regione critica, del test;
6.2. se il valore ottenuto per la statistica del test è nella regione critica, H_0 è rifiutata e H_1 accettata; altrimenti H_0 non può essere rifiutata.

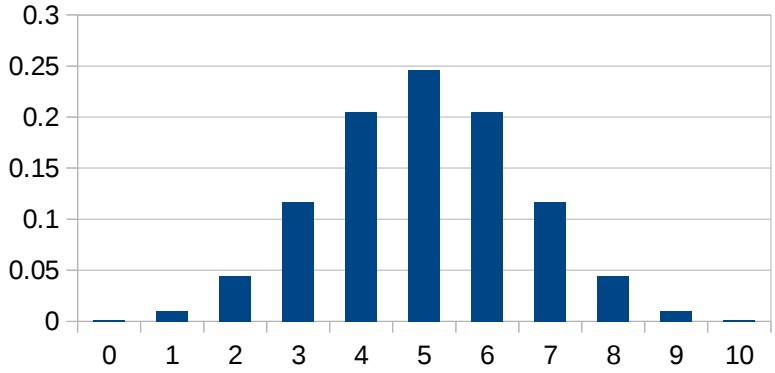
Ancora sulla logica dei test di ipotesi

Torniamo a considerare l'esempio precedente: ci siamo chiesti se un elemento di valore $x=1.45$ sia stato ottenuto da una popolazione a distribuzione gaussiana con $\mu=1.23$ e $\sigma=0.12$. La risposta è stata che in questo caso l'ipotesi nulla non può essere rifiutata: x è sufficientemente vicino a μ, relativamente a σ, e quindi l'ipotesi che x sia stato ottenuto dalla popolazione in questione non è abbastanza improbabile da essere scartata.
Supponiamo ora di disporre non più di un singolo elemento ma di un campione, per esempio di $n=4$ elementi e la cui media è anch'essa $m=1.45$. Il campione è stato ottenuto dalla popolazione? Il problema ha evidentemente la stessa struttura del precedente, e ci possiamo, correttamente, aspettare che la procedura di test di ipotesi appena sintetizzata sia identicamente applicabile anche in questo caso. La differenza con il problema precedente è solo per il fatto che ora il dato da confrontare con la media della popolazione si riferisce non a un singolo elemento ma alla media di un campione, che sappiamo essere più stabile dei suoi singoli valori. La statistica del test da impiegare è perciò una versione modificata di quella adottata sopra:

$$Z=\frac{m_X-\mu}{\frac{\sigma}{\sqrt{n}}}$$

dove m_X è la statistica media campionaria di un campione di n elementi e, come sappiamo, il nuovo denominatore, $\sigma/\sqrt{n}$, è la deviazione standard della media di un campione di n elementi ottenuto da una popolazione di deviazione standard σ. Perciò: $z = (1.45–1.23)/0.06 \approx 3.67$. Allora $g_{cum}(z) \approx 0.9999$, e quindi $1 – g_{cum}(z) \approx 0.0001$. Dato che il test è "a due code", anche in questo caso occorre raddoppiare questo valore, ma evidentemente il p-value di 0.0002, cioè lo 0.02%, è ora talmente minore del livello di significatività adottato, $\alpha=0.05$, da rendere perfettamente giustificata la scelta di rifiutare l'ipotesi nulla e di accettare invece l'ipotesi alternativa: il campione *non* è stato ottenuto dalla popolazione data.
La differenza nei due esempi è evidente: il valore di x nel primo caso e quello di m nel secondo sono identici, ma la minore variabilità della media fa sì che, a parità di distanza dalla media della popolazione, nel secondo caso H_0 è assai meno probabile, e quindi può essere appunto rifiutata. Questa differenza è stata ottenuta a causa dei dati diversi, che hanno anche richiesto di adottare statistiche di test diverse, ma la procedura seguita è stata esattamente la stessa.

Come altro esempio semplice di test di ipotesi, torniamo a considerare il problema di caratterizzare la probabilità che una moneta sia corretta a partire dall'esito di alcuni lanci ripetuti. Per rendere il caso più interessante, supponiamo che la moneta sia stata lanciata 10 volte, in condizioni tali da garantire l'indipendenza tra i lanci, e che si siano ottenute 2 T. Possiamo comunque considerare la moneta corretta? E' questa l'ipotesi nulla, H_0: $q=0.5$, che manterremo fino a che non ci sia evidenza sufficiente del contrario, cioè dell'ipotesi alternativa, H_1: $q\neq 0.5$. Anche in questo caso si presenta l'asimmetria notata sopra: mentre H_0 è un'ipotesi specifica, H_1, in quanto complemento di H_0, è generica, dato che non si specifica un valore di probabilità (la moneta potrebbe essere non corretta, per esempio, perché $q=0.2$ oppure $q=0.7$). Questa genericità giustifica il rigore del criterio di accettazione per H_1: H_0 deve essere molto poco probabile per essere rifiutata.
La statistica del test in questo caso descrive semplicemente i possibili esiti dei 10 lanci, cioè $X=$ 0 T, 1 T, ..., 10 T (questo esempio ben chiarisce il significato del termine "statistica del test": per esempio il valore 1 T sintetizza i 10 possibili esiti 'T al primo lancio e C negli altri 9', 'T al secondo lancio e C negli altri 9' ...), e la distribuzione di verosimiglianza $p(X|H_0)$ è binomiale, con probabilità pari a 0.5.



Dobbiamo ora scegliere come interpretare il dato ‘ 2 T su 10 lanci’ dal punto di vista della scelta tra le due ipotesi. A partire da questo dato, supponiamo che il dubbio sia non genericamente sulla possibile non correttezza della moneta, ma sull’ipotesi che essa favorisca la T (in fondo se la moneta è stabile non è credibile che sia truccata favorendo a volte la T e a volte la C...). Il p-value è allora da calcolare sugli esiti ‘≤2 T su 10 lanci’, dunque per un test “a una coda”, e quindi:

$$\binom{10}{0}0.5^0(1-0.5)^{10}+\binom{10}{1}0.5^1(1-0.5)^9+\binom{10}{2}0.5^2(1-0.5)^8\approx 0.055$$

Scegliendo ancora una volta il livello di significatività $\alpha=0.05$, il p-value è di poco superiore: non dovremmo rifiutare l’ipotesi che sia corretta una moneta che ha prodotto 2 T in 10 lanci.

Questa procedura può essere adottata anche in altri tipi di test di ipotesi, come nei casi che seguono.

Test di ipotesi: *goodness of fit*

Dato un campione (che supponiamo univariato e con valori indipendenti), si vuole mettere alla prova l’ipotesi che il campione sia stato ottenuto da una distribuzione di probabilità data, un test chiamato di *goodness of fit* (a volte tradotto in italiano “bontà di adattamento”).

Come abbiamo visto sopra, quando si fa test di ipotesi in un contesto probabilistico si ammette che i dati disponibili non consentiranno di confermare in modo definitivo l’ipotesi, ma solo, eventualmente, di mostrare che essa è sufficientemente improbabile da poter essere rifiutata. Nel caso del test di goodness of fit l’ipotesi nulla H_0 è che il campione sia stato ottenuto dalla distribuzione di probabilità data, e quindi l’ipotesi alternativa H_1 è che il campione sia stato ottenuto non da tale distribuzione ma da un’altra, una qualsiasi altra, non specificata.

La logica del test è analoga a quella seguita in precedenza: dai dati D disponibili si ottiene un valore per la probabilità di H_0, $P(H_0|D)$, cioè il p-value, e quindi lo si confronta con il livello di significatività del test, α: se $P(H_0|D)<\alpha$ l’ipotesi nulla è considerata appunto sufficientemente improbabile da poter essere rifiutata; nel caso contrario l’ipotesi nulla non può essere rifiutata e quindi la si potrà mantenere, almeno fino a che nuovi dati non consentiranno di effettuare un nuovo, e possibilmente più severo, test. Dunque il problema è tutto nella possibilità di calcolare il p-value $P(H_0|D)$: qual è la probabilità che il campione derivi dalla distribuzione di probabilità data?

La tecnica è la seguente.

Dal campione, $\langle x_1,\dots,x_n\rangle$, si costruisce la distribuzione a frequenze assolute, con m categorie scelte opportunamente, essendo O_j la frequenza osservata della j-esima categoria. Per le stesse categorie si ottiene dalla distribuzione di probabilità la frequenza attesa E_j, che nel caso di una distribuzione di probabilità discreta sulle m categorie è semplicemente:

$$E_j=n\,p_j$$

essendo dunque n il numero degli elementi del campione e p_j la probabilità della j-esima categoria nella distribuzione. Allora, quanto minori sono le differenze $|O_j-E_j|$ quanto più è credibile che il campione derivi dalla distribuzione data.

Per rendere più precisa questa idea, si considera che, per ogni categoria, O_j è una variabile casuale, e:

$$V_j=\frac{O_j-E_j}{\sqrt{E_j}}$$

è un'ulteriore variabile casuale distribuita come una gaussiana con media 0 e varianza $1-p_j$ (non dimostriamo questo risultato, che deriva dal teorema del limite centrale). Sommando su tutte le categorie i valori quadratici di queste variabili casuali si otterrà una nuova variabile casuale V^2:

$$V^2=\sum_{j=1}^{m} V_j^2=\sum_{j=1}^{m} \frac{(O_j-E_j)^2}{E_j}$$

distribuita appunto come una somma di gaussiane al quadrato.
E' nota una distribuzione, chiamata "chi quadro", proprio definita in questo modo, e cioè:

$$\chi_k^2=\sum_{j=1}^{k} Z_j^2$$

dove le Z_j sono variabili casuali a distribuzione gaussiana standard (cioè a media 0 e varianza 1) indipendenti. La conclusione è, in effetti (nuovamente tralasciamo i passaggi per giungere a questa conclusione), che se il campione di partenza, da cui sono state ottenute le frequenze O_j, segue la distribuzione corrispondente alle frequenze attese E_j, allora la variabile casuale V^2 segue una distribuzione che tende (in un senso tecnico che non è necessario chiarire qui) a χ_k^2, per un valore del parametro k che dipende dal numero m di categorie.
Il test per controllare l'ipotesi H_0 che il campione stesso sia stato ottenuto dalla distribuzione di probabilità data prevede allora di calcolare il p-value come integrale della coda destra della distribuzione χ_k^2 a partire dal valore calcolato per la variabile casuale V^2:

$$P(H_0|D)=1-\int_{-\infty}^{V^2} \chi_k^2(x)\,dx=1-\chi_{\text{cum}\,k}^2(V^2)$$

da trattare come al solito: se il p-value è minore del livello di significatività α, l'indicazione è di rifiutare l'ipotesi nulla.
Un ultimo problema da risolvere: per quale valore di k occorre calcolare la statistica di chi quadro? La variabile casuale V è una somma di m variabili a distribuzione gaussiana quadratica, e quindi parrebbe che si debba scegliere $k=m$. D'altra parte, date le probabilità su $m-1$ categorie la probabilità della m-esima categoria è fissata dalla condizione di normalizzazione:

$$\sum_{j=1}^{m} p_j=1$$

Se segue che la somma è calcolata su $m-1$ variabili indipendenti, e tale valore, chiamato *gradi di libertà* per il test, è quello da scegliere per k.

Un esempio.
Alla conclusione di un anno di corsi universitari, un campione di 120 studenti è stato intervistato a proposito della qualità percepita dei corsi, e i giudizi, in scala da 1 a 4, sono sintetizzati nella distribuzione a frequenze assolute:

$$\begin{bmatrix} C_j \\ n_j \end{bmatrix}=\begin{bmatrix} 1 & 2 & 3 & 4 \\ 18 & 27 & 40 & 35 \end{bmatrix}$$

Si formula l'ipotesi che la popolazione da cui il campione è stato ottenuto sia distribuita uniformemente rispetto alle quattro categorie, e si vuole mettere alla prova questa ipotesi con un livello di significatività $\alpha=0.05$. Elaborando i dati si ottiene:

C_j	O_j	E_j	V_j^2
1	18	30	4.800
2	27	30	0.300
3	40	30	3.333
4	35	30	0.833
	somma		V^2
	120		9.267
	α	k	p-value
	0.05	3	0.026

Il p-value, calcolato come $1-\chi^2_{\text{cum}\,k}(V^2)$, è pari a 0.026 e quindi è minore del valore scelto per α. L'ipotesi nulla è rifiutata al livello di significatività del 5%: i dati a disposizione mostrano che improbabile che la popolazione sia distribuita uniformemente.
Alla stessa conclusione si sarebbe potuto giungere anche operando inversamente, chiedendosi qual è il valore massimo per la statistica chi quadro al livello di significatività α scelto, sotto il quale l'ipotesi nulla dovrebbe essere rifiutata. Tale valore critico si calcola dalla distribuzione chi quadro (cumulata) inversa per la probabilità α, $\chi^2_{\text{inv}\,k}(\alpha)$. Il test è dunque:

se $V^2 > \chi^2_{\text{inv}\,k}(\alpha)$, allora rifiuta l'ipotesi nulla;
altrimenti non rifiutare l'ipotesi nulla.

Nel nostro caso, V^2 = 9.267 è maggiore del valore critico $\chi^2_{\text{inv}\,k}(0.05)=7.815$: si conferma perciò che al livello di significatività del 5% l'ipotesi nulla è rifiutata.

Un altro esempio.
Si suppone che un lotto contenga il 5% di prodotti difettosi e si vuole mettere alla prova questa ipotesi (nulla) estraendo per 100 volte 3 prodotti e contando quanti di questi sono difettosi. I dati che si ottengono, in cui le categorie C_j corrispondono a 0, 1, ... prodotti guasti in un gruppo di 3 prodotti, sono i seguenti:

$$\begin{bmatrix} C_j \\ n_j \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 89 & 9 & 2 & 0 \end{bmatrix}$$

Si confronta pertanto la distribuzione di frequenze osservate con la distribuzione ipotetica, che in questo caso è binomiale:

$$p_j = \binom{3}{j} p^j (1-p)^{3-j}$$

con p = 0.05, e si mette alla prova questa ipotesi con un livello di significatività α = 0.05. Elaborando i dati si ottiene:

Num. difettosi	O_j	p_j	E_j	V_j^2
0	89	0.857	85.74	0.124
1	9	0.135	13.54	1.521
2	2	0.007	0.71	2.327
3	0	0.000	0.01	0.013
	somma			V^2
	100			3.984
	α		k	p-value
	0.05		3	0.263

Il p-value, calcolato come $1-\chi^2_{\text{cum}\,k}(V^2)$, è pari a 0.263 e quindi è maggiore del valore scelto per α. L'ipotesi nulla non è rifiutata al livello di significatività del 5%: i dati a disposizione non confutano l'ipotesi che il lotto contenga il 5% di prodotti difettosi.
Adottando la tecnica inversa si giunge, naturalmente, allo stesso risultato: V^2 = 3.984 è minore del valore critico $\chi^2_{\text{inv}\,k}(0.05)=7.815$, e si conferma perciò che al livello di significatività del 5% l'ipotesi nulla non è rifiutata.

Un altro esempio.
Si dispone di un campione di 100 valori (nella tabella sotto sono visualizzati solo alcuni elementi) che si suppone ottenuto da una popolazione a distribuzione gaussiana con media 0 e deviazione standard 1, e si vuole mettere alla prova questa ipotesi. Si procede come nei casi precedenti, salvo che qui il supporto della distribuzione, e quindi del campione, è continuo, e occorre quindi prima di tutto discretizzarlo (come abbiamo già ricordato, si possono calcolare le probabilità p_j degli intervalli $[x_j, x_{j+1}]$ $p_{\text{cum}}(x_{j+1}) - p_{\text{cum}}(x_j)$, dove $p_{\text{cum}}(x)$ è il valore della distribuzione cumulata in x). Nuovamente, il parametro k della distribuzione chi quadro, cioè il numero di gradi di libertà, è calcolato come il numero delle categorie scelte meno 1.

X_i	C_j	O_j	p_j	E_j	V^2_j
1.3836	-3.4	0	0.000	0.034	0.034
-0.8805	-3	0	0.001	0.101	0.101
0.5614	-2.6	0	0.003	0.331	0.331
0.1864	-2.2	0	0.009	0.924	0.924
-0.2518	-1.8	1	0.022	2.203	0.657
0.9612	-1.4	5	0.045	4.483	0.060
0.8200	-1	5	0.078	7.790	0.999
0.8105	-0.6	7	0.116	11.560	1.799
-0.6927	-0.2	17	0.146	14.649	0.377
-1.5951	0.2	17	0.159	15.852	0.083
0.5814	0.6	13	0.146	14.649	0.186
1.2829	1	10	0.116	11.560	0.210
1.5817	1.4	14	0.078	7.790	4.951
-1.3738	1.8	7	0.045	4.483	1.414
-1.7622	2.2	1	0.022	2.203	0.657
-0.5222	2.6	3	0.009	0.924	4.662
0.9484	3	0	0.003	0.331	0.331
2.4510	3.4	0	0.001	0.101	0.101
-0.3255		somma			V^2
-0.5842		100			17.877
0.4689					
-0.1384		α		k	$\chi_k^2(V^2)$
-1.2112		0.05		17	0.397
1.2230					

Il p-value, calcolato come $1-\chi^2_{\text{cum}\,k}(V^2)$, è pari a 0.397 e quindi è maggiore del valore scelto per α. L'ipotesi nulla non è rifiutata al livello di significatività del 5%: i dati a disposizione non consentono di confutare l'ipotesi che la popolazione sia distribuita come una gaussiana.
Nuovamente, adottando la tecnica inversa si giunge allo stesso risultato: $V^2=17.877$ è minore del valore critico $\chi^2_{\text{inv}\,k}(0.05)=27.587$, e si conferma perciò che al livello di significatività del 5% l'ipotesi nulla non è rifiutata.

Test di ipotesi: indipendenza tra variabili

Dato un campione bivariato, $\langle\langle x_i, y_i\rangle\rangle$, si vuole controllare l'ipotesi che le due variabili siano statisticamente indipendenti. Una prima indicazione può essere ottenuta calcolando sul campione il coefficiente di correlazione campionaria, ma può accadere che le due variabili non siano indipendenti anche nel caso di correlazione bassa, e al limite nulla (si consideri, per esempio, la situazione in cui X è distribuito come una gaussiana a media 0 e $Y = X^2$: evidentemente le due variabili sono tutt'altro che indipendenti – conoscendo i valori di una si ricavano facilmente i valori dell'altra –, eppure la loro correlazione è nulla, o comunque molto bassa nel caso si considerino non le distribuzioni come tali ma dei campioni ottenuti da esse).
E' interessante che la tecnica introdotta per il test di goodness of fit può essere impiegata quasi identicamente anche in questo caso. L'idea di base è infatti la stessa. Dal campione si costruisce la distribuzione a frequenze assolute "osservate", O_j, che in questo caso è dunque una distribuzione congiunta, e nello stesso tempo si calcola la distribuzione a frequenze assolute "attese", E_j, così che, nell'ipotesi che le frequenze E_j siano state ottenute nel caso di indipendenza tra X e Y:

$$V^2=\sum_{j=1}^{m} V_j^2=\sum_{j=1}^{m}\frac{(O_j-E_j)^2}{E_j}$$

è distribuita come una somma di gaussiane al quadrato, da confrontare dunque con la distribuzione chi quadro.
Ma come si calcolano, in questo caso, le frequenze "attese"?
Partendo dal campione bivariato, si costruisce la distribuzione congiunta di XY e da essa le distribuzioni marginali:

		Y				
		Y_1	...	Y_k	...	Σ
X	X_1	$n_{1,1}$		$n_{1,k}$		$\Sigma_k n_{1,k}=r_1$
	...					...
	X_j	$n_{j,1}$		$n_{j,k}$		$\Sigma_k n_{j,k}=r_j$
	...					...
	Σ	$\Sigma_j n_{j,1}=c_1$	...	$\Sigma_j n_{j,k}=c_k$	...	$\Sigma_j\Sigma_k n_{j,k}=n$

Le frequenze "osservate" sono allora proprio i valori della distribuzione congiunta:

$$O_{j,k} = n_{j,k}$$

Ogni frequenza "attesa" è quel valore che si otterrebbe se l'ipotesi nulla fosse vera, cioè che le due variabili fossero indipendenti. Ricordiamo che per definizione, in caso di indipendenza:

$$p_{j,k} = p_j\, p_k$$

dove p_j e p_k sono le probabilità marginali. In termini di frequenze assolute, il valore j,k-esimo è dunque il prodotto delle corrispondenti frequenze marginali, r_j e c_k, normalizzato rispetto al numero complessivo di elementi del campione, n:

$$E_{j,k} = \frac{r_j c_k}{n}$$

Dato tutto ciò, il valore su cui calcolare la statistica di chi quadro è allora:

$$V^2 = \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} V_{j,k}^2 = \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} \frac{(O_{j,k} - E_{j,k})^2}{E_{j,k}}$$

dove m_1 e m_2 sono il numero di categorie per X e Y (cioè il numero di righe e di colonne della distribuzione congiunta) rispettivamente, e il numero di gradi di libertà è $(m_1–1)(m_2–1)$.

Un esempio.

Proviamo che un campione bivariato $\langle X, Y = X^2 \rangle$ di 100 elementi, in cui X è distribuito come una gaussiana a media 0 e deviazione standard 1, è tale che le due variabili hanno correlazione bassa ma non sono indipendenti (nella tabella sotto sono visualizzati solo alcuni elementi).

X	Y	Correlazione: -0.050
-1.311	1.718	
-0.478	0.229	
-0.028	0.001	
-0.222	0.049	

Per semplicità, categorizziamo le due variabili:

X	Y	X cat	Y cat
-1.311	1.718	-1.25	1.75
-0.478	0.229	-0.25	0.25
-0.028	0.001	-0.25	0.25
-0.222	0.049	-0.25	0.25

In questo modo è semplice generare la distribuzione congiunta e le due distribuzioni marginali:

Frequenze "osservate", $O_{j,k}$

Count -	Y cat													
X cat	0.25	0.75	1.25	1.75	2.25	2.75	3.25	3.75	4.25	4.75	5.25	5.75	6.25	**Total R**
-2.75													1	**1**
-2.25										1			1	**2**
-1.75						2	4		1					**7**
-1.25				6	1	1								**8**
-0.75	10	5												**15**
-0.25	14													**14**
0.25	21													**21**
0.75	3	9												**12**
1.25				5	3	1								**9**
1.75						1	4	2	1					**8**
2.25											1	1		**2**
2.75													1	**1**
Total R	**48**	**14**	**11**	**4**	**5**	**8**	**2**	**2**	**1**	**1**	**1**	**1**	**2**	**100**

Da qui si calcola la distribuzione per le frequenze "attese":

0.48	0.14	0.11	0.04	0.05	0.08	0.02	0.02	0.01	0.01	0.01	0.01	0.02
0.96	0.28	0.22	0.08	0.1	0.16	0.04	0.04	0.02	0.02	0.02	0.02	0.04
3.36	0.98	0.77	0.28	0.35	0.56	0.14	0.14	0.07	0.07	0.07	0.07	0.14
3.84	1.12	0.88	0.32	0.4	0.64	0.16	0.16	0.08	0.08	0.08	0.08	0.16
7.2	2.1	1.65	0.6	0.75	1.2	0.3	0.3	0.15	0.15	0.15	0.15	0.3
6.72	1.96	1.54	0.56	0.7	1.12	0.28	0.28	0.14	0.14	0.14	0.14	0.28
10.08	2.94	2.31	0.84	1.05	1.68	0.42	0.42	0.21	0.21	0.21	0.21	0.42
5.76	1.68	1.32	0.48	0.6	0.96	0.24	0.24	0.12	0.12	0.12	0.12	0.24
4.32	1.26	0.99	0.36	0.45	0.72	0.18	0.18	0.09	0.09	0.09	0.09	0.18
3.84	1.12	0.88	0.32	0.4	0.64	0.16	0.16	0.08	0.08	0.08	0.08	0.16
0.96	0.28	0.22	0.08	0.1	0.16	0.04	0.04	0.02	0.02	0.02	0.02	0.04
0.48	0.14	0.11	0.04	0.05	0.08	0.02	0.02	0.01	0.01	0.01	0.01	0.02

e da queste si ottengono i valori $V^2_{j,k}$:

0.48	0.14	0.11	0.04	0.05	0.08	0.02	0.02	0.01	0.01	0.01	0.01	48.02
0.96	0.28	0.22	0.08	0.10	0.16	0.04	0.04	48.02	0.02	0.02	48.02	0.04
3.36	0.98	0.77	0.28	7.78	21.13	0.14	5.28	0.07	0.07	0.07	0.07	0.14
3.84	1.12	29.79	1.45	0.90	0.64	0.16	0.16	0.08	0.08	0.08	0.08	0.16
1.09	4.00	1.65	0.60	0.75	1.20	0.30	0.30	0.15	0.15	0.15	0.15	0.30
7.89	1.96	1.54	0.56	0.70	1.12	0.28	0.28	0.14	0.14	0.14	0.14	0.28
11.83	2.94	2.31	0.84	1.05	1.68	0.42	0.42	0.21	0.21	0.21	0.21	0.42
1.32	31.89	1.32	0.48	0.60	0.96	0.24	0.24	0.12	0.12	0.12	0.12	0.24
4.32	1.26	16.24	19.36	0.67	0.72	0.18	0.18	0.09	0.09	0.09	0.09	0.18
3.84	1.12	0.88	0.32	0.90	17.64	21.16	4.41	0.08	0.08	0.08	0.08	0.16
0.96	0.28	0.22	0.08	0.10	0.16	0.04	0.04	0.02	48.02	48.02	0.02	0.04
0.48	0.14	0.11	0.04	0.05	0.08	0.02	0.02	0.01	0.01	0.01	0.01	48.02

La sintesi, infine, è:

m_1	m_2	V^2
12	13	553.389

α	gradi libertà	$\chi_k^2(V^2)$
0.05	132	0.000

Si conferma che la probabilità che le due variabili siano indipendenti è praticamente nulla.

Un esempio.
Dal database dell'International Monetary Fund - World Economic Outlook (WEO) (ottobre 2012, http://www.imf.org/external/pubs/ft/weo/2012/02) prendiamo i dati relativi al tasso di inflazione e al tasso di disoccupazione nel 2010 di un centinaio di nazioni, e ci chiediamo se le due variabili siano o meno indipendenti (nella tabella sotto sono visualizzati solo alcuni elementi).

Country	Inflation, average consumer prices, 2010	Unemployment rate, 2010
Albania	3.6	12.5
Algeria	3.9	9.961
Argentina	10.461	7.75
Armenia	7.274	19
Australia	2.845	5.225
Austria	1.69	4.4
Azerbaijan	5.666	6.048
Barbados	5.761	10.6

Operiamo esattamente come sopra:
– calcoliamo prima di tutto il coefficiente di correlazione campionaria: 0.02; è molto vicino a 0, ma questo non è ancora sufficiente per concludere che inflazione e disoccupazione siano variabili indipendenti;
– categorizziamo i valori delle due variabili, in modo da facilitare la costruzione della distribuzione;
– costruiamo la distribuzione congiunta e le distribuzioni marginali, trattate come distribuzioni di frequenze "osservate", $O_{j,k}$;
– da queste costruiamo la distribuzione delle frequenze "attese", $E_{j,k}$, e quindi i valori $V^2_{j,k}$ e la loro somma V^2;
– calcoliamo per V^2 la statistica di chi quadro, $\chi^2_k(V^2)$, in questo caso pari a 0.751; il fatto che sia ben superiore al livello di significatività del test, $\alpha = 0.05$, ci consente di concludere che non possiamo rifiutare l'ipotesi nulla, e possiamo dunque continuare a supporre che inflazione e disoccupazione siano variabili indipendenti.