Germano Rossi

Elementi di ragionamento statistico: per psicologia e scienze dell'educazione

versione 11

versione elettronica
ad uso degli studenti
http://psico.univr.it/germano/abcstat

Tabella delle modifiche

Siccome i primi capitoli di questa dispensa, verranno stampati dall'editore Carocci in un libro con il titolo di "Statistica descrittiva per psicologi", le parti relative sono state tolte.

Questo "libro elettronico" è un "work in progress" ovvero un libro che sto ancora scrivendo. Tuttavia, siccome la stesura delle varie parti dipende dalle necessità degli studenti, non verrà scritto in modo sequenziale, ma secondo le necessità.

La tabella che segue, indica, per ciascuno dei capitoli, la versione a cui è arrivato, in modo che possiate ri-stampare solo la parte che vi serve. Per il numero di versione ho usato la seguente regola:

- il primo numero indica la versione principale. Resterà a 0 finché non riterrò che il capitolo abbia assunto una forma abbastanza definita;
- il secondo numero aumenterà ad ogni modifica, anche piccola del contenuto, ma non della forma.
- una versione 0.0 significa, ovviamente, che non è ancora stato scritto nulla.

Le pagine sono numerate all'interno di ogni capitolo, in modo che non abbiate la necessità di ristampare tutto il testo solo perché ho aggiunto qualcosa all'inizio di un capitolo.

Infine, ho racchiuso fra parentesi quadre alcuni miei appunti personali (cose da fare o sui cui meditare, cose da aggiungere o da ri-scrivere…) oppure avvertimenti di lettura.

	Vers.
Introduzione	1.1
Teoria della probabilità	0.4
Cenni di teoria del campionamento	0.1
Introduzione all'inferenza statistica	0.0
La distribuzione binomiale	0.0
La distribuzione normale	0.0
Il test di chi-quadro	0.2
La correlazione	0.4
Inferenza sulla media	0.1
Appendice	0.3

Avvertenza per il lettore

In questo libro si è cercato di riportare le cognizioni essenziali per la comprensione della statistica. La semplicità del testo non deve portare il lettore a pensare che ciò che è stato scritto non sia importante. Al contrario, in ogni pagina si è cercato di mostrare come i ragionamenti della statistica siano passaggi logici, che hanno uno scopo ben preciso. Capire questo scopo è più importante che imparare a memoria le formule matematiche, perché significa capire la statistica.

Indice

Introduzione [1.1]

Da studente e poi durante gli anni di ricerca mi sono sempre stupito del fatto che, ogni qual volta una persona iniziava ad insegnare statistica, finiva spesso per scrivere un libro di statistica (un manuale) per il proprio corso di lezione. Ho sempre pensato fosse una sorta di "snobismo" nei confronti dei libri già scritti (e normalmente in commercio) sullo stesso argomento.

Trovandomi ora nella stessa situazione, invece, scopro che nessun testo dice le cose esattamente come tu vorresti spiegarle (ed in effetti le spieghi) e che dovresti adottare 5 o 6 testi diversi di cui poi utilizzare, di volta in volta, i pochi capitoli che sono in sintonia con te. E mi rendo anche conto (mi hanno costretto a farlo gli studenti) che la cosa è impossibile: ci sono studenti che frequentano le lezioni assiduamente, proprio per non (o perché non possono permettersi di) comperare i libri; altri che non frequentano affatto perché lontani o lavoratori e fanno fatica a capire perché ti servono così tanti libri diversi eppure uguali; altri ancora ti dicono che, sì, hai spiegato bene e che, sì, hanno capito, ma… sul libro di testo è spiegato in modo diverso…; e poi (sempre, con insistenza) ti chiedono di avere i lucidi che usi, vogliono avere il tempo di copiare tutto quello che scrivi alla lavagna…, ti chiedono perché non prepari una dispensa…

A questo punto, l'idea di fare un tuo libro di testo, tutto tuo, che sia in sintonia con te stesso… diventa quasi impellente, una necessità che soverchia la sensazione di "snobismo". Il problema diventa, allora, come scrivere un libro che non sia una copia dei tanti già esistenti, che, pur dicendo quanto "bisogna" dire (e quindi niente affatto 'originale'), riesca a non essere "banale"…

Non credo vi sia risposta a questo problema… bisogna correre il rischio della banalità, della ridondanza, dell'ovvietà, proprio per cercare di voler essere "comprensibili" da tutti, sia che abbiano una preparazione o una mentalità logica sia che non l'abbiano.

Ho scelto allora di fare un "*manuale in progress*" da rendere disponibile in formato Acrobat Reader© sulle mie pagine Web e di permetterne la fotocopia e la stampa da internet. In questo modo, il libro è sempre aperto alle modifiche, alle aggiunte, alle successive spiegazioni, ai nuovi esercizi o alla correzione degli errori. Ad ogni esame (purtroppo!), gli errori degli studenti (in questo caso nel ruolo sgradevole di "cavie") diventano lo spunto per migliorare una parte del testo, per chiarirlo o per scrivere una cosa del tutto nuova.

Per definizione, di questo libro non esiste (e non può esistere, logicamente) una versione definitiva. Qualunque versione "stampata" dal singolo utente o da un possibile editore (semmai vedesse la luce sotto questa forma) è, necessariamente, un'istantanea, che diventerà ben presto obsoleta.

Nel corso del testo, non farò nessun tentativo di essere esaustivo, completo, colto e di presentare e dettagliare tutti gli aspetti dell'argomento. Al contrario, tralascerò volutamente certi aspetti (seppur importanti) che, per gli intenti di questo lavoro, possono confondere il lettore alle prime armi. In particolare, questa 'approssimazione' accadrà nei capitoli che già nel titolo (*Cenni di…*, *Introduzione a…*) evidenziano un carattere introduttivo, necessario per comprendere altre parti del testo.

Su altri argomenti, invece, sarò addirittura ridondante, sempre volutamente, presentando riepiloghi o riprendendo lo stesso concetto più volte.

E prima di finire, iniziamo con le banalità.

La prima banalità è nel titolo che ho deciso di adottare (per capire la banalità, consultate la bibliografia):

- "elementi" sta a significare che non ho la pretesa di fare un vero e proprio libro di statistica, cioè un testo completo che diventi il fondamento per gli anni futuri, ma piuttosto uno strumento da usare, magari 'pasticciare' (come sembra sia "normale" fra gli studenti di oggigiorno), senza troppe pretese…
- "ragionamento statistico" è invece una pretesa (questa sì!), vale a dire il tentativo di *non scrivere un libro formale*, che faccia invidia a qualunque matematico per il suo rigore, per l'eleganza delle equazioni e/o dello sviluppo logico con cui sono presentate. Io mi pongo invece il problema di rendere il libro (e quindi la statistica) comprensibile ai miei studenti (e sta a loro, aiutarmi a scrivere un libro alla loro portata, dicendomi tutti i difetti che vi riscontrano);
- "psicologia e scienze dell'educazione" sono i riferimenti teorici in cui pongo gli esempi e anche l'ambito in cui verrà usato questo testo.

Non c'è libro che non termini con i ringraziamenti a qualcuno e un libro elettronico non fa eccezione.

Ecco i miei:

- Nebojša Jovan Živković, Dimitri Nicolau, Alessio Porceddu Cilione, James McMillan e il Kronos Quartet… sono alcuni tra le centinaia di compositori ed esecutori, che hanno costruito l'ambiente sonoro in cui ho lavorato e sto lavorando; la loro musica, seppur stimolante, non è invadente;
- gli studenti che hanno fatto l'esame con me e a cui ho dovuto correggere qualcosa: gli errori costanti sono state le prime cose che ho scritto in questo libro.

Germano

1 Teoria della probabilità [0.4]

Nella vita di tutti i giorni noi usiamo sia il concetto di probabilità sia termini o concetti che hanno a che fare con la probabilità: "E' impossibile", "sono sicurissimo", "ci metterei la mano sul fuoco", "non te lo garantisco, ma secondo me, ci sono buone possibilità di…", "non ti conviene". Anzi, spessissimo usiamo proprio il termine "probabilità": "è improbabile che riesca a…", "probabilmente, questa estate andrò in vacanza a…" e così via.

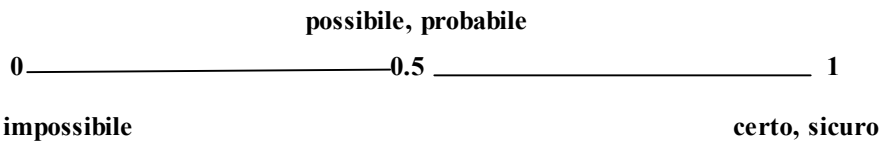
Ecco altri esempi di situazioni reali in cui, senza rendervene conto, usate il concetto di probabilità.

a) Siete in macchina e state guidando lungo una strada. Arrivate in prossimità di un semaforo che diventa giallo. Cosa fate? Cosa potete fare? Accelerare per passare prima che diventi rosso. Frenare subito. Cosa sceglieresteste? In base a cosa prendereste una decisione? La distanza che vi manca all'incrocio, la velocità a cui state procedendo, la fretta che avete di arrivare a destinazione, la quantità di traffico, la vostra propensione personale a correre rischi…

b) State giocando a "sette e mezzo" e con le prime due carte siete arrivati a 6. Che fate? Vi fermate o chiedete un'altra carta? Se siete all'inizio della mano, quasi certamente vi fermate; se siete alla fine del mazzo e sapete che i numeri alti sono già usciti quasi tutti e ci sono molte figure, allora, altrettanto certamente, continuerete. Se invece siete a metà mazzo e uno degli altri giocatori ha già fatto sei e mezzo, potrete decidere di non rischiare.

Anche se non ve ne siete resi conto, qualunque scelta facciate (anche quelle che io posso non aver elencato), dipende dal successo che pensate di ottenere: e questo è esprimibile come una probabilità.

Possiamo rappresentare graficamente la probabilità come un linea che parte da un evento impossibile e avanza verso un evento sicuro:



Associamo il valore 0 all'evento "impossibile" proprio perché è un evento che non potrà mai accadere e usiamo il valore 1 per l'evento sicuro perché, se accade, accade uno e un solo evento. Tutto quello che sta fra lo 0 e l'1 è il valore che associamo agli eventi che "possono" accadere e il valore che gli assoceremo corrisponderà alla probabilità della loro comparsa.

Indicheremo la probabilità con la lettera *p* e indicheremo tra parentesi l'evento a cui la probabilità si riferisce. In tal modo,

$P(x) = 1$

indicherà un evento "sicuro" e

$P(x) = 0$

un evento "impossibile". Se vi sono più eventi possibili e l'accadere di uno di questi esclude automaticamente l'accadere degli altri, diremo che:

- gli eventi sono "mutuamente esclusivi",
- la loro singola probabilità sarà minore di 1 e maggiore di 0: $0 < P(x) < 1$;
- la somma delle probabilità dei singoli eventi sarà 1, perché comunque uno e uno solo di questi eventi accadrà.

Consideriamo il lancio di una moneta. Gli eventi possibili sono le due diverse facce della moneta (testa e croce). Ciascuno dei due eventi avrà una probabilità compresa fra 0 e 1:

$0 < P(testa) < 1$, $0 < P(croce) < 1$

e la loro somma dovrà essere uguale a 1

$P(testa) + P(croce) = 1$

1.1 Principali teorie di probabilità

Nel corso del tempo sono state elaborate tre diverse teorie sulla probabilità che non si escludono a vicenda ma che, semplicemente, partono da presupposti diversi. Le tre teorie sono:

- la **teoria classica**: gli eventi possibili sono fra loro equiprobabili;
- la **teoria frequentista**: la probabilità di un evento dipende dalla frequenza con cui questo evento è comparso in passato;
- la **teoria soggettiva**: è una stima che ciascuno di noi fa sull'accadere di un determinato evento.

1.1.1 Probabilità classica

Viene chiamata teoria classica quella teoria che costituisce il primo approccio allo studio della probabilità e che è stata sviluppata sostanzialmente pensando ai giochi d'azzardo. E' la più semplice perché fa riferimento ad eventi che, di per sé, sono agglomerati di eventi semplici.

Alla base della concezione classica della probabilità vi sono alcuni presupposti:

- esistono diversi eventi possibili, tutti fra di loro mutuamente esclusivi:
- uno solo degli eventi possibili potrà accadere;
- alcuni eventi possono essere utili ai nostri scopi (sono quelli che ci interessano), altri no;
- tutti gli eventi possibili (se non diversamente indicato) hanno la stessa probabilità o, comunque, hanno probabilità teorica conoscibile a priori.

Sulla base di questi presupposti, la probabilità di un singolo evento si calcola come rapporto fra l'accadere di quell'evento e il numero totale degli eventi alternativi possibili.

Esempio:

p(faccia di un dado) = 1 / 6

p(faccia di una moneta) = 1 / 2

p(una carta da un mazzo di 40) = 1 / 40

Questa regola può essere ampliata pensando che possiamo considerare contemporaneamente tutti gli eventi che ci sono utili:

$$P(x) = \frac{casi\ favorevoli}{casi\ possibili} = \frac{f}{N} \qquad \text{Eq. 1-1}$$

Esempio:

La probabilità di ottenere un numero superiore o uguale a 5 lanciando un dado: poiché ci sono due eventi che ci sono utili (il 5 e il 6), la probabilità sarà P(5,6) = 2/6 = 1/3 = 0.33

La probabilità di estrarre un asso da un mazzo di 40 carte: poiché ci sono 4 assi all'interno del mazzo, P(asso) = 4/40.

La probabilità della teoria classica è conoscibile a priori, proprio usando la formula 4-1. Ma dobbiamo considerare che questo è vero solo all'infinito… vale a dire. Se io lancio una moneta una volta, potrò ottenere una sola delle due facce e se la lancio due volte posso ottenere sempre la stessa faccia. Se lancio la moneta dieci volte, potrò anche ottenere 5 teste e 5 croci, ma anche potrò ottenere 6 teste e 4 croci. Però lanciando la moneta un numero sempre maggiore di volte, la moneta tenderà a comportarsi sempre più come la teoria predice che debba comportarsi. La probabilità di 1/2 (ossia 0.5) è quindi un limite teorico, raggiungibile con un numero infinito di lanci.

1.1.2 Probabilità frequentista

La situazione di equiprobabilità è valida solo in certe particolari condizioni, spesso fittizie come quelle dei giochi d'azzardo, mentre nella realtà, gli eventi non presentano questa caratteristica.

Avere "capelli neri" è equiprobabile ad avere "capelli rossi"? Ragioniamo per assurdo. Se avere capelli neri o capelli rossi, fosse equiprobabile, dovrei incontrare giornalmente lo stesso numero di persone con capelli neri e con capelli rossi. Forse un giorno mi capiterà di incontrare un maggior numero di persone con capelli di un certo colore, ma se considero parecchi giorni consecutivi, dovrei trovarmi con una parità. L'esperienza ci dice invece che questo non è vero, perché non ci sono tante persone con capelli rossi quante sono quelle che hanno capelli neri. Quindi, la probabilità di "avere capelli rossi" è sicuramente inferiore (e comunque non uguale) a quella di "avere capelli neri".

Per calcolare esattamente la probabilità dei due eventi, dovremmo conoscere esattamente, in modo univoco e definitivo, quali sono i fattori che producono l'apparire del fenomeno "capelli rossi" o di quello dei "capelli neri". E questo non è possibile.

La teoria frequentista delle probabilità utilizza allora la conoscenza del passato come stima della probabilità di un evento. Vale a dire: facciamo un'indagine, raccogliamo le informazioni sul colore dei capelli, costruiamo una tabella di frequenze e quindi su questa base (il passato, il conosciuto, l'accaduto) costruiamo una stima della probabilità della caratteristica "avere capelli rossi".

Tabella 1.1 - Dati fittizi

	f	%
Neri	*352*	*35.2*
Castani	*282*	*28.2*
Marroni	*226*	*22.6*
Biondi	*83*	*8.3*
Rossi	*57*	*5.7*
Totale	*1000*	*100*

Applichiamo la stessa regola che abbiamo visto per la teoria classica: $\frac{f}{N}$. Siccome abbiamo 57 persone con i capelli rossi su un totale di 1000, la probabilità sarà:

$$P(rossi) = \frac{57}{1000} = 0.057$$

L'uso della formula generale della probabilità classica è giustificata dal fatto che possiamo pensare alle 57 persone con i capelli rossi come ad "eventi favorevoli" sul totale di tutti gli eventi possibili.

Notate anche come la stessa identica formula viene usata sia per calcolare la proporzione all'interno di una tabella delle frequenze, sia per la stima della probabilità frequentista.

1.1.3 Probabilità soggettiva

Questa concezione della probabilità è stata sviluppata agli inizi del 1900 da DeFinetti e da Ramsey e cerca di spiegare quel meccanismo per cui ciascuno di noi ha un certo grado di sicurezza in relazione a determinati eventi. Si capisce meglio questa teoria se teniamo presente il mondo anglosassone, più abituato rispetto agli italiani al mondo delle scommesse e soliti esprimere tali scommesse come rapporti.

E' la stima che ciascuno di noi fa, sull'accadere di un evento su cui non abbiamo informazioni sicure. Normalmente si usa nelle scommesse: "quanto scommetti che…" oppure nei giudizi: "mi consigli di fare…", "devo fare così… o così…"

1.2 Distribuzione di probabilità

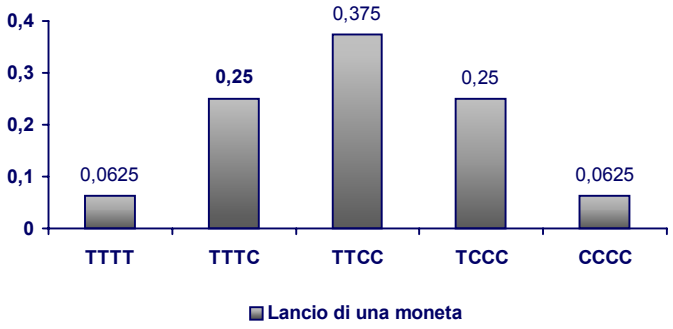
Poiché la somma delle probabilità di tutti gli eventi alternativi (e possibili) è (e **dev'essere**) pari a 1, perché uno solo è l'evento che può accadere, allora possiamo considerare queste probabilità come valori di una distribuzione di dati e costruire la distribuzione di probabilità di quel fenomeno.

Consideriamo il lancio di una moneta. Se la lanciamo 10 volte di seguito otteniamo i seguenti eventi (dove T indica una faccia e C l'altra): TTTT, TTTC, TTCT, TCTT, CTTT, TTCC, TCCT, CCTT, CTCT, TCTC, CTTC, TCCC, CTCC, CCTC, CCCT, CCCC. Ciascuno di questi eventi possibili ha la stessa probabilità di 1/16. Ma, alcuni di questi eventi sono simili, perché non abbiamo ragioni per discriminare fra TTTC e CTTT. Così se raggruppiamo questi eventi secondo categorie di equivalenza e costruiamo una distribuzione di frequenza, otteniamo:

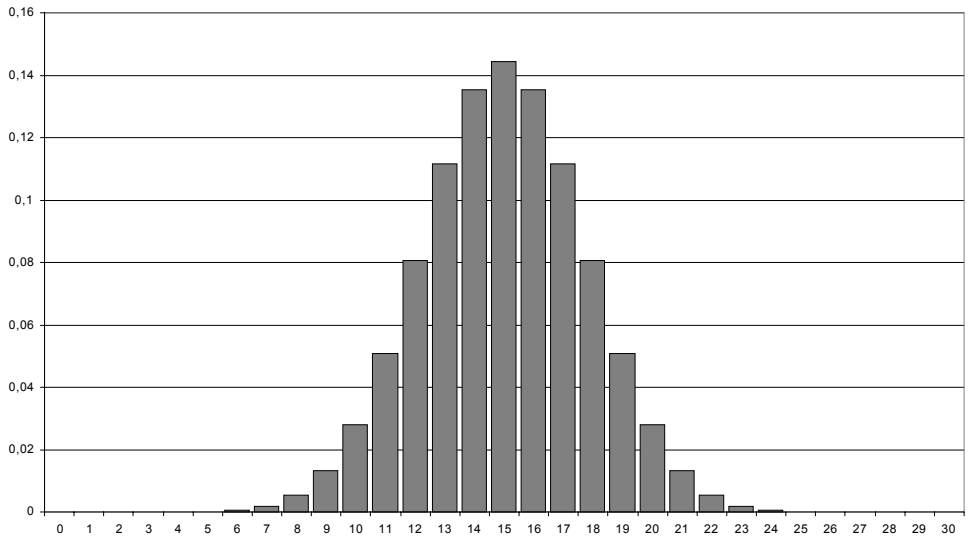
C			f		prop.	cumulate	
0	4T	tttt	1	1/16	.0625	.0625	1
1	3T, 1C	tttc, ttct, tctt, cttt	4	4/16	.2500	.3125	.9875
2	2T, 2C	ttcc, tcct, cctt, ctct, tctc, cttc	6	6/16	.3750	.6875	.6875
3	1T, 3C	tccc, ctcc, cctc, ccct	4	4/16	.2500	.9875	.3125
4	4C	cccc	1	1/16	.0625	1	.0625
			16		1.00		

Tabella 1.2

Come si può vedere, la somma di tutte le singole probabilità è pari a 1. E la sua rappresentazione grafica è:



Notiamo come la distribuzione delle probabilità in questo grafico mostri chiaramente la sua natura simmetrica. Se immaginiamo di lanciare una moneta per 100 volte o per 1000 volte, la rappresentazione grafica diventerebbe sempre più fitta avvicinandosi sempre più ad un'ipotetica curva simmetrica. Facciamo un esempio di rappresentazione usando 30 lanci della solita moneta:



Poiché ogni stanghetta del grafico, rappresenta la probabilità di un evento, ma è anche un'area del grafico, la somma di tutte le aree (e quindi delle probabilità) dev'essere 1.

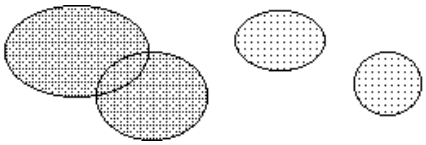
Diremo allora che tutte le distribuzioni che sommano a uno sono distribuzioni di probabilità e che tutte le curve la cui area è uno, sono curve di probabilità.

1.3 Regole per il calcolo con le probabilità

Una volta trovata la probabilità di un evento (classica, frequentista o soggettiva) le regole che si applicano sono praticamente le stesse e sono facilmente comprensibili se facciamo riferimento all'insiemistica.

1.3.1 Regola addizionale (OR, o)

La probabilità che due eventi accadano in alternativa l'uno all'altro, cioè che accada un evento A oppure un evento B corrisponde alla somma di 2 insiemi ed è quindi pari alla somma delle singole probabilità tolta l'eventuale intersezione, perché le due situazioni possibili sono:



e quindi:

$P(A) + P(B) - P(A \cap B)$

Se i due eventi non possono accadere in contemporanea, se cioé

$P(A \cap B) = 0$

allora è solo la somma delle singole probabilità:

$P(A) + P(B)$

Esempio 1

La probabilità di ottenere un numero pari con un lancio di un dado è data dalla probabilità di ottenere un due più quella di ottenere un quattro più quella di ottenere un sei e non vi è nulla da togliere perché i tre eventi sono fra loro incompatibili (ovvero disgiunti). Quindi:

$P(pari) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 3/6 = 1/3$

Allo stesso risultato arriveremmo considerando che 3 sono gli eventi che ci sono favorevoli (2, 4 e 6) su un totale di 6 e quindi 3/6 = 1/3.

Esempio 2

Qual è la probabilità di ottenere almeno 3 teste lanciando 4 monete (oppure lanciando 4 volte una moneta)?

Possiamo elencare tutti gli eventi che ci sono favorevoli (TTTC, TTCT, TCTT, CTTT, TTTT) e calcolare la probabilità con la regola generale, 5/36. Oppure possiamo considerare che vi sono due categorie che ci sono utili, l'apparire di 3 teste (in ordine qualsiasi) e l'apparire di 4 teste. Abbiamo visto in precedenza (Tabella 1.2) che vi sono 4 possibili eventi che cadono nella categoria 3T e 1 nella categoria 4T, che la probabilità totale di 3T è di .25 e quella di 4T è .0625, allora:

$P(3 \text{ o } 4\ T) = P(3t) + P(4t) = .25 + .0625 = .3125$

Esempio 3

Per calcolare la probabilità di ottenere almeno un 5 o un 6 lanciando 2 dadi, dobbiamo considerare che per ciascuno dei due dadi vale la situazione per ui la probabilità del 5 o del 6 è pari alla somma delle singole probabilità e quindi è:

$P(5 \text{ o } 6) = 1/6 + 1/6 = 2/6$

Quindi la probabilità generale dovrebbe essere pari alla somma di questa probabilità per i due dati, ovvero

$P(5{,}6 \text{ dado } 1) + P(5{,}6 \text{ dado } 2) = 2/6 + 2/6 = 4/6.$

Ma siccome i due eventi possono accadere contemporaneamente, allora dobbiamo sottrarre la probabilità dell'intersezione, cioè 4/6. Quindi ricapitolando tutto:

$P(A=5{,}6 \text{ dado } 1) = 2/6$

$P(B=5{,}6 \text{ dado } 2) = 2/6$

$P(A \cap B=\text{entrambi}) = 4/36$

$P(A) + P(B) - P(A \cap B) = 2/6 + 2/6 - 4/36 = (12+12-4)/36 = 20/36$

Il perché dell'intersezione diventa lampante se si considera la Tabella 1.3, dove sono elencate tutte le 36 possibilità. I numeri 5 e 6 possono comparire sul primo dado (ultime due colonne) assieme a uno qualunque dei valori del secondo dado e ci sono 12 eventi possibili a noi favorevoli oppure sul secondo dado (ultime 2 righe) assieme ad un qualunque valore del primo dado e ci sono altre 12 possibili eventi favorevoli, ma in questo modo, poiché gli eventi possono comparire assieme, gli eventi 55, 65, 56 e 66 vengono contati due volte (sia nella probabilità del primo dado sia in quella del secondo) e quindi dobbiamo operare un aggiustamento, sottraendoli una volta. E questi quattro valori sono esattamente quelli corrispondenti all'intersezione fra i due eventi.

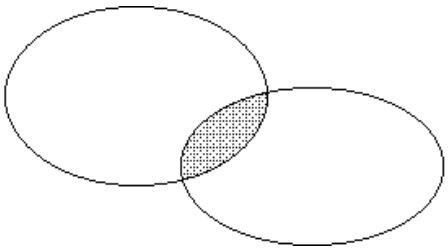
11	*21*	*31*	*41*	*51*	*61*
12	*22*	*32*	*42*	*52*	*62*
13	*32*	*33*	*43*	*53*	*63*
14	*24*	*34*	*44*	*54*	*64*
15	*25*	*35*	*45*	*55*	*65*
16	*26*	*36*	*46*	*56*	*66*

Tabella 1.3

Se usassimo il sistema classico, dovremmo contare tutti gli elementi a noi favorevoli, cioè 20, in rapporto al totale degli eventi, cioè 36, e quindi 20/36.

1.3.2 Regola moltiplicativa (AND, e)

La probabilità che accadano contemporaneamente 2 eventi corrisponde all'intersezione non nulla di due insiemi ed è pari al prodotto delle singole probabilità.



Consideriamo il precedente esempio del lancio di 2 dadi. Se su entrambi i dadi devono comparire un 5 o un 6, allora l'evento che ci interessa corrisponde all'intersezione fra i due eventi. Possiamo usare la regola generale per cui 4 sono gli eventi possibili in cui i numeri 5 e 6 possono comparire su entrambi i dadi su un totale di 36 eventi (cfr. Tabella 1.3) e quindi la probabilità di ottenere 5 o 6 contemporaneamente su due dadi è di 4/36. Ma consideriamo che possiamo anche giungere allo stesso risultato moltiplicando fra loro le singole probabilità:

$$P(A \text{ e } B) = \frac{2}{6} \cdot \frac{2}{6} = \frac{4}{36}$$

Questo è vero, però, se e soltanto se i due eventi sono fra di loro indipendenti.

Due eventi sono fra loro indipendenti quando l'accadere del primo evento non ha alcun influsso sull'accadere dell'altro. Nel caso di 2 dadi, il risultato del primo dado non ha alcun influsso sul secondo, mentre dopo aver estratto un numero della tombola, vengono modificate tutte le probabilità successive (tanto per cominciare perché la numerosità diminuisce a 89).

A questo punto possiamo rivedere l'evento P(4t). I quattro lanci sono fra loro indipendenti, quindi la probabilità di 4 teste è il prodotto delle singole probabilità. Siccome ciascuna è pari a 1/2,

$P(4t) = 1/2 * 1/2 * 1/2 * 1/2 = .5 * .5 * .5 * .5 = .5^4 = .0625$

Quale sarebbe allora la probabilità di rispondere correttamente ad un test con domande del tipo vero/falso se rispondo in modo assolutamente casuale?

$P(30 \text{ giusti}) = (1/2)^{30}$

1.3.3 Applicare entrambe le regole

Ovviamente possiamo applicare contemporaneamente entrambe le regole. Se voglio sapere la probabilità di ottenere 3 teste e 1 coda lanciando quattro volte una moneta, posso considerare che i quattro eventi sono fra di loro indipendenti e quindi un singolo evento contenente 3 teste e 1 croce si basa sulla probabilità dei singoli eventi. Poiché:

P(t) = P(c) = .5

e io ho 4 eventi simultanei, allora la loro probabilità sarà pari a

$.5^4 = .0625$

Ma io posso ottenere lo stesso evento in 4 possibili modi (tttc, ttct, tctt, cttt) e uno qualunque dei quattro mi va bene, quindi devo sommare le loro singole probabilità, ovvero:

.0625 + .0625 + .0625 + .0625 = .0625 * 4 = .25

1.3.4 Probabilità condizionale

Si chiama probabilità condizionale o condizionata la probabilità di un certo evento quando il suo apparire è condizionato da un altro evento. Facciamo un esempio.

In un sacchetto ci sono 5 palline bianche e 3 palline nere. Se estraggo due palline una immediatamente dopo l'altra, qual è la probabilità che siano entrambe bianche?

La probabilità che la prima pallina sia bianca è data dal numero di palline bianche sul totale delle palline, quindi 5/8. Se la prima pallina è bianca, la probabilità che anche la seconda è bianca è data dal numero di palline bianche (cioè 4) sul totale delle palline rimaste, cioè 7 e quindi 4/7. Se moltiplichiamo le due probabilità, perché i due eventi devono accadere in contemporanea, trovo che la probabilità sarà pari a:

$$\frac{5}{8} \cdot \frac{4}{7} = \frac{20}{56} = \frac{5}{14}$$

1.4 Calcolo combinatorio

Quando parleremo di distribuzione campionaria, useremo espressioni del tipo "tutti i possibili campioni di ampiezza n" e quando abbiamo affrontato la probabilità, abbiamo avuto necessità di calcolare "tutti i possibili eventi" che ci erano favorevoli rispetto agli eventi totali.

In entrambi i casi, si tratta di trovare il modo di enumerare determinate cose nei diversi possibili modi. Il calcolo combinatorio si occupa proprio di questo: di formalizzare alcuni tipi di enumerazione e di trovare delle formule che facilitino i calcoli. In alcuni casi, si tratta solo di identificare il tipo di enumerazione e di applicare la formula corretta.

Per enumerazione intendiamo fare riferimento ad un elenco di "eventi" generici come "a, b, c, d" dove *a*, *b*, *c* e *d* costituiscono degli elementi qualsiasi.

Per capirci, facciamo alcuni esempi:

- in quanti possibili modi 3 amici possono sedersi su un divano?
- quante possibili combinazioni posso ottenere lanciando due dadi?
- quanti numeri di 3 cifre posso fare con i simboli da 0 a 9?

Definiamo "disposizioni" le enumerazioni in cui l'ordine di presentazione/comparsa ha importanza e perciò *abcd* è diverso da *badc* e da *dabc* e chiamiamo "combinazioni" gli elenchi in cui invece l'ordine non ha importanza (e quindi *abc*, *acb*, *bca*… sono tutti analoghi e validi). Indicheremo le disposizioni con la lettera D e le combinazioni con la lettera C. Siccome dobbiamo anche indicare il numero totale di elementi e il numero di quelli da disporre, aggiungeremo due indicazioni: D(n,k) o C(n,k).

Inoltre distinguiamo fra enumerazioni con la possibilità che gli elementi si ripetano (aab, abb…) da quelle in cui ciò non è possibile. Se si tratta di estrazioni casuali, diremo "con reimmissione" se vi è la possibilità che un elemento si ripeta e "senza reimmissione" il caso contrario (ad es. l'estrazione dei numeri del lotto). Per distinguere queste due situazioni, useremo una *c* o una *s* (rispettivamente "con" e "senza" ripetizione); per cui le possibilità diventano:

D(n,k,c) = disposizione di *n* elementi presi *k* a *k con* ripetizione
D(n,k,s) = disposizione *senza* ripetizione
C(n,k,c) = combinazione *con* ripetizione
C(n,k,s) = combinazione *senza* ripetizione

1.4.1 Fattoriale

Per poter continuare ci serve di definire il concetto di "numero fattoriale". In matematica, la notazione 3! o 5! o più genericamente *n*!, indica un *numero fattoriale* ed è la rappresentazione sintetica di un'operazione più complessa:

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

In pratica, la moltiplicazione di tutti i numeri interi positivi decrescenti a partire da un qualunque *n*, è detta *fattoriale di n*.

Esempio:

$2! = 2 \cdot 1 = 2$

$3! = 3 \cdot 2 \cdot 1 = 6$

$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$

Per poter rispettare alcune proprietà matematiche, all'interno dei fattoriali vengono definiti due numeri particolari:

$0! = 1$

$1! = 1$

Tabella 1.4 – Tabella dei numeri primi

0!	*1*
1!	*1*
2!	*2*
3!	*6*
4!	*24*
5!	*120*
6!	*720*
7!	*5040*
8!	*40320*
9!	*362880*
10!	*3628800*
11!	*39916800*
12!	*479001600*
13!	*6227020800*
14!	*87178291200*
15!	*1307674368000*

1.4.2 Coefficiente binomiale

Sempre per poter usare sinteticamente il calcolo combinatorio, ci serve un'altra definizione, quella di *coefficiente binomiale*

$\binom{n}{k}$ che si legge “n su k” e che corrisponde a

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

si noti che fra *n* e *k* non esiste una barra divisoria, perché il coefficiente fattoriale non equivale ad una frazione (cioè *non significa* "n diviso k").

Esempi:

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4\cdot3\cdot2\cdot1}{2\cdot1\cdot2\cdot1} = \frac{4\cdot3}{2\cdot1} = \frac{12}{2} = 6$$

$$\binom{10}{4} = \frac{10!}{4!(10-4)!} = \frac{10\cdot9\cdot8\cdot7\cdot6!}{4\cdot3\cdot2\cdot1\cdot6!} = \frac{10\cdot9\cdot8\cdot7}{4\cdot3\cdot2\cdot1} = \frac{5\cdot3\cdot2\cdot7}{1\cdot1\cdot1\cdot1} = 210$$

Nel secondo esempio, ho evidenziato il fatto che si può eliminare il 6! al numeratore con quello al denominatore e che successivamente si possono ridurre il 10 con il 2, il 9 con il 3 e l’8 con il 4.

Anche per il coefficiente binomiale, definiamo:

$$\binom{n}{0} = 1 \text{ e } \binom{0}{0} = 1$$

1.4.3 Permutazione

In quanti possibili modi 3 amici possono sedersi su un divano?

Si tratta di una disposizione senza ripetizione perché la stessa persona non può sedere contemporaneamente in due posti diversi e perché l’ordine con cui si siedono è importante.

Possiamo rispondere a questa domanda, tramite uno schema che elenchi tutti i modi possibili. Cominciamo ad indicare i 3 amici con le iniziali del loro nome: Andrea, Bruno, Carlo.

Tabella 1.5

Modo	*Posizione*		
	1	*2*	*3*
1	*A*	*B*	*C*
2	*A*	*C*	*B*
3	*B*	*A*	*C*
4	*B*	*C*	*A*
5	*C*	*A*	*B*
6	*C*	*B*	*A*

Potete notare come, una volta scelto chi deve sedere al primo posto, gli altri due possono sedersi in due modi diversi, scambiandosi i posti (3 x 2). Oppure possiamo ragionare in un altro modo: ciascuno dei tre amici può essere selezionato per sedere al primo posto, ma al secondo posto può sedere uno solo dei due restanti e al terzo posto, deve forzatamente sedere la persona rimasta. Quindi:

Tabella 1.6

Posto	*1*	*2*	*3*
Possibilità	*3 x*	*2 x*	*1*

che in totale corrisponde a 6 (notate anche che è un 3!) e che effettivamente sono 6 le possibilità che avevamo elencato alla Tabella 1.5. Questa è una forma particolare di disposizione, in cui n=k e si chiama *Permutazione*.

$P_n=D(_{n,k=n,s})=n!$

1.4.4 Disposizioni senza ripetizione

Cinque amici, devono disporsi attorno ad un tavolo che può contenere solo 3 persone per volta: avremo la situazione D(5,3,s). Se sviluppiamo manualmente l'esempio, vediamo che:

Tabella 1.7

Posto	*1*	*2*	*3*
Possibilità	*5* *n*	*4* *n-1*	*3* *n-2*

al primo posto può stare uno qualunque dei 5 amici, al secondo uno qualunque dei 4 restanti e al terzo uno degli altri tre. Quindi 5·4·3=60 sono i diversi modi in cui questi 5 amici possono sedersi, 3 alla volta, attorno ad un tavolo.

Se sviluppiamo l'esempio in modo pratico, otteniamo di poter utilizzare le 3 lettere che rappresentano i 3 amici in 10 modi diversi, e al loro interno, ogni volta, ciascuna terna di lettere può generare 6 possibili modi che differiscono solo dall'ordine con cui si dispongono (è una permutazione di 3 elementi):

Tabella 1.8

			B		*C*	*D*
	C	*D*	*E*	*D*	*E*	*E*
A	*ABC*	*ABD*	*ABE*	*ACD*	*ACE*	*ADE*
	ACB	*ADB*	*AEB*	*ADC*	*AEC*	*AED*
	BAC	*BAD*	*BAE*	*CAD*	*CAE*	*DAE*
	BCA	*BDA*	*BEA*	*CDA*	*CEA*	*DEA*
	CAB	*DAB*	*EAB*	*DAC*	*EAC*	*EAD*
	CBA	*DBA*	*EBA*	*DCA*	*ECA*	*EDA*
B				*BCD*	*BCE*	*BDE*
				BDC	*BEC*	*BED*
				CBD	*CBE*	*DBE*
				CDB	*CEB*	*DEB*
				DBC	*EBC*	*EBD*
				DCB	*ECB*	*EDB*
C						*CDE*
						CED
						DCE
						DEC
						ECD
						EDC

In termini linguistici possiamo dire che è un "pezzo" di *n*! sviluppato solo per *k* volte. Secondo il calcolo combinatorio è più semplice pensare in questo modo:

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

e il perché si vede facilmente con l'esempio numerico:

$$\frac{5!}{(5-3)!}=\frac{5\cdot 4\cdot 3\cdot 2!}{2!}=5\cdot 4\cdot 3$$

perché il 2! si può elidere.

1.4.5 Combinazioni senza ripetizione

Gli stessi cinque amici di prima, devono disporsi attorno al solito tavolo a cui possono sedere solo tre per volta, ma questa volta decidono che l'ordine non è importante, per cui le disposizioni *abc*, *acb*, *bca*… sono tutte analoghe fra loro. In questo caso, quante possibilità hanno? Se riprendiamo la Tabella 1.8 ed eliminiamo, all'interno di ogni cella, tutte le non-differenze, otteniamo le sole 10 celle di partenza.

Tabella 1.9

	B			C		D
	C	*D*	*E*	*D*	*E*	*E*
A	*ABC*	*ABD*	*ABE*	*ACD*	*ACE*	*ADE*
B				*BCD*	*BCE*	*BDE*
C						*CDE*

In pratica, dalle 60 disposizioni possibili, abbiamo eliminato le disposizioni che differiscono solo per l'ordine con cui gli elementi sono disposti, ovvero, le permutazioni di 3 elementi. Quindi è come se usassimo:

$$\frac{D(n,k,s)}{P_k}$$

e poiché la formula, una volta sciolta diventerebbe $\frac{n!}{(n-k)!}\cdot\frac{1}{k!}$, la formula proposta dal calcolo combinatorio è:

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

che quindi, se la applichiamo al nostro esempio:

$$C(5,3,s)=\binom{5}{3}=\frac{5!}{3!(5-3)!}=\frac{5\cdot 4\cdot 3!}{3!\cdot 2!}=\frac{5\cdot 4}{2}=10$$

ci fornisce esattamente lo stesso valore che abbiamo trovato con tutte le possibili enumerazioni.

1.4.6 Disposizioni con ripetizione

In quanti possibili modi possono cadere 3 dadi (uno rosso, uno verde e uno blu)? Si tratta di disposizioni perché è importante l'ordine con cui otteniamo i risultati (l'1 sul rosso è diverso dall'1 sul verde…) ed è anche possibile che compaia lo stesso numero su due o su tutti e tre i dadi. Quindi lo indichiamo con D(6,3,c).

Immaginiamo che sul primo dado (quello rosso) sia comparso un 1. Vediamo cosa potrebbe succedere sugli altri due.

Tabella 1.10

		Dado verde					
		1	*2*	*3*	*4*	*5*	*6*
Dado blu	*1*	*11*	*21*	*31*	*41*	*51*	*61*
	2	*12*	*22*	*32*	*42*	*52*	*62*
	3	*13*	*23*	*33*	*43*	*53*	*63*
	4	*14*	*24*	*34*	*44*	*54*	*64*
	5	*15*	*25*	*35*	*45*	*55*	*65*
	6	*16*	*26*	*36*	*46*	*56*	*66*

Ci sono 36 possibili disposizioni fra il dado verde e quello blu; se le combiniamo con il dado rosso diventano 36 · 6 = 216.

In formula scriviamo:

$D(n,k,c) = n^k$

ovvero, per il nostro esempio

$D(6,3,c) = 6^3 = 216$

1.4.7 Combinazioni con ripetizione

Se i tre dadi fossero tutti bianchi e non mi importasse l'ordine, le possibilità sarebbero minori, perché dovremmo eliminare tutte le combinazioni che utilizzano gli stessi numeri. Sempre immaginando cosa può succedere sul secondo e terzo se il primo è uno, otteniamo:

Tabella 1.11

			Terzo dado						Totale
			1	*2*	*3*	*4*	*5*	*6*	
Primo dado	1	Secondo dado *1*	*111*	*112*	*113*	*114*	*115*	*116*	*6*
		2		*122*	*123*	*124*	*125*	*126*	*5*
		3			*133*	*134*	*135*	*136*	*4*
		4				*144*	*145*	*146*	*3*
		5					*155*	*156*	*2*
		6						*166*	*1*

Se il primo dado è 1, abbiamo 21 modi diversi. Se il secondo dado fosse 2, dalle prime 21 combinazioni dovremmo eliminare tutte quelle che presentano il valore 1 sul secondo e sul terzo dado , perché 122 equivale a 212 e a 221 e ci rimangono 15 (21-6) combinazioni. Se il primo dado è 3, dobbiamo eliminare tutte le combinazioni che contengono 1 o 2 sugli altri dadi e quindi 21-6-5=10. E così via per tutti gli altri numeri, per cui alla fine, abbiamo:

Tabella 1.12

Se il primo dado è	le combinazioni valide sono	
1	*21*	*21*
2	*21-6*	*15*
3	*21-6-5*	*10*
4	*21-6-5-4*	*6*
5	*21-6-5-4-3*	*3*
6	*21-6-5-4-3-2*	*1*

Totale	*56*

La formula è:

$$C(n,k,c)=\binom{n+k-1}{k}=\frac{(n+k-1)!}{k![(n+k-1)-k]!}=\frac{(n+k-1)!}{k!(n-1)!}$$

e quindi nel nostro esempio:

$$C(6,3,c)=\binom{6+3-1}{3}=\frac{(6+3-1)!}{3!(6-1)!}=\frac{8\cdot 7\cdot 6\cdot 5!}{3\cdot 2\cdot 1\cdot 5!}=8\cdot 7=56$$

1.4.8 Permutazioni con ripetizione

Un caso particolare di permutazione è dato dal seguente problema: in una collana di arte tribale vi sono 2 pietre bianche e 3 pietre azzurre. Quanti diversi tipi di collana si possono ottenere usando tutte le cinque pietre?

Ragioniamo in questo modo:

1) se consideriamo tutte le cinque pietre come diverse, avremo 5! possibili disposizioni (usiamo B1, B2, A1, A2 e A3);
2) ma a tutte queste possibilità, dobbiamo togliere le combinazioni delle pietre bianche, cioè 2!
3) e dobbiamo togliere anche quelle delle tre pietre azzurre, 3!

Ne consegue che il calcolo da farsi sarebbe:

$$\mathrm{P}_{\mathrm{b.a}}=\frac{n!}{b!a!}=\frac{5!}{2!\cdot 3!}=\frac{5\cdot 4\cdot 3!}{2\cdot 1\cdot 3!}=10$$

Se facciamo la prova concreta, otterremo:

1	*BBAAA*
2	*BABAA*
3	*BAABA*
4	*BAAAB*
5	*ABBAA*
6	*ABABA*
7	*ABAAB*
8	*AABBA*
9	*AABAB*
10	*AAABB*

Così se le pietre fossero 10, 5 rosse, 2 bianche e 3 azzurre, il calcolo sarebbe:

$$\mathrm{P}_{\mathrm{a.b.c}}=\frac{n!}{a!b!c!}=\frac{10!}{5!\cdot 2!\cdot 3!}=\frac{10\cdot 9\cdot 8\cdot 7\cdot 6\cdot 5!}{5!\cdot 2\cdot 1\cdot 3\cdot 2\cdot 1}=\frac{5\cdot 3\cdot 4\cdot 7\cdot 6}{1\cdot 1\cdot 1\cdot 1\cdot 1}=2520$$

1.4.9 Riepilogo

	Con ripetizione	*Senza ripetizione*
Disposizioni	$D(n,k,c)=n^k$	$D(n,k,s)=\dfrac{n!}{(n-k)!}$
Permutazioni	$P_{\text{b.a}}=\dfrac{n!}{b!a!}$	$P_{\text{n}}=D(n,k=n,s)=n!$
Combinazioni	$C(n,k,c)=\dfrac{(n+k-1)!}{k!(n-1)!}$	$C(n,k,s)=\dbinom{n}{k}=\dfrac{n!}{k!(n-k)!}$

1.5 Le principali distribuzioni di probabilità

Esistono delle distribuzioni di probabilità le cui caratteristiche sono conosciute, in quando conosciamo la formula matematica con cui possono essere rappresentate graficamente e quindi la loro forma e le principali informazioni statistiche. Queste distribuzioni di probabilità sono:

- la distribuzione binomiale
- la distribuzione ipergeometrica
- la distribuzione normale
- la distribuzione normale standardizzata
- la distribuzione di chi-quadro (χ^2)
- la distribuzione F di Snedecor
- la distribuzione t di Student
- la distribuzione campionaria

Essendo distribuzioni di probabilità note, per ciascuna, noi conosciamo non solo la media e la deviazione standard, ma anche la porzione di area corrispondente ad ogni valore che compone la distribuzione. E, viceversa, conoscendo una porzione di area possiamo risalire al valore che vi è associato. Vedremo in concreto questi passaggi al capitolo sulla distribuzione normale.

Se una statistica, si distribuisce secondo una di queste curve di probabilità, grazie a delle particolari tabelle, noi possiamo associare al valore della statistica un corrispondente valore di probabilità.

2 Cenni di teoria del campionamento [0.1]

La statistica descrittiva ci ha fornito le basi su cui fondare il resto dei ragionamenti statistici. Infatti, con queste prime conoscenze di statistica, noi possiamo solo descrivere la realtà che vogliamo studiare e, nel momento in cui l'abbiamo completamente descritta, se questa non cambia, il nostro compito potrebbe considerarsi terminato. Però, noi non potremo mai descrivere completamente, esaurientemente e definitivamente tutta la realtà. Siamo quindi costretti a isolare e considerare parti di realtà di volta in volta diversi e, anche in questo caso, non sempre ci è possibile studiare interamente quella parte di realtà. Ad es. se volessimo studiare la reazione degli italiani ad un certo evento, dovremmo studiare più di 55 milioni di individui, un lavoro enorme! Per questo motivo, si studia un piccolo insieme che si considera come un campione di quella realtà che vorremmo studiare.

Riprendiamo alcune definizioni già presentate nel capitolo sulla *Statistica descrittiva* e che sono importanti per questo argomento.

Definiamo come **popolazione** tutti i *casi statistici possibili* rispetto ad una particolare "variabile" che si vorrebbe misurare (ad es. per il peso umano, tutti gli individui; per l'atteggiamento nei confronti degli extra-comunitari da parte degli italiani, gli italiani), cioè, in termini insiemistici, l'universo (U) e chiamiamo **campione** quella parte della popolazione (di *numerosità inferiore*) su cui, effettivamente, andremo a realizzare la misurazione. Ciascuno degli elementi che costituisce il campione e quindi anche la popolazione, lo chiamiamo genericamente *unità statistica* o *caso statistico*.

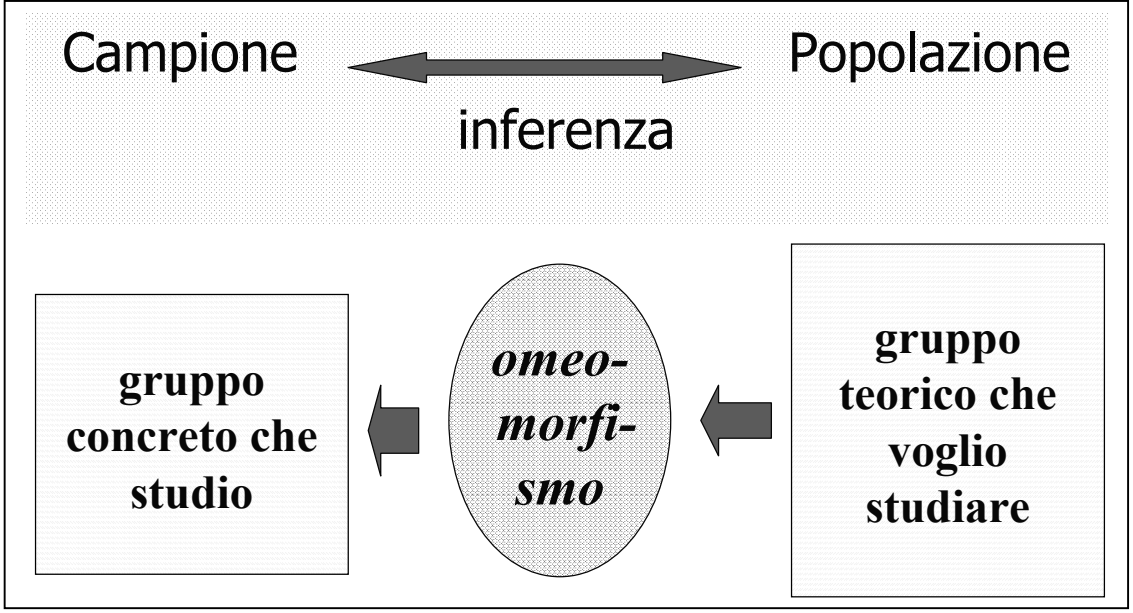
Il significato di *campione* è lo stesso che noi associamo ai *campioni omaggio* di shampoo o di profumo che troviamo nelle riviste o nei negozi. Se quell'esempio dimostrativo del prodotto, non costituisse nella nostra opinione un "esempio di come si comporta il prodotto nella sua generalità", non potremmo mai fidarci ad acquistarlo…; ogni acquisto potrebbe fornirci caratteristiche sempre diverse e imprevedibili di quel prodotto. Invece, noi usiamo il contenuto del campione omaggio ed estendiamo le caratteristiche che possiede all'intero prodotto, convinti e sicuri che le caratteristiche del prodotto sono rappresentate dal campione che noi abbiamo sperimentato.

Anche in senso statistico il campione deve essere *rappresentativo* della popolazione da cui lo estraiamo altrimenti non avremmo nessuna sicurezza che i risultati possano essere estesi a tutta la popolazione. E' quindi molto importante il modo in cui il campione viene estratto.

2.1 Rappresentatività

Il campione selezionato "dovrebbe" rappresentare "in piccolo" la popolazione che si vuol studiare… quindi il campione dev'essere **rappresentativo**, ovvero deve avere le stesse caratteristiche della popolazione (e nella stessa proporzione).

Sulla base del campione rappresentativo, noi estenderemo i dati ottenuti all'intera popolazione, tramite un processo di **inferenza statistica** (Fig.). In un certo senso, il campione, dovendo possedere le stesse caratteristiche della popolazione, dev'essere omeomorfo ad essa, ovvero deve avere la stessa forma, la stessa struttura.



Una volta selezionate le variabili che ci interessa di studiare (che saranno chiamate *variabili dipendenti*), possiamo individuare anche delle variabili che riteniamo importanti o che possono essere/produrre influenza sulla variabile che vogliamo studiare (queste variabili verranno chiamate *variabili indipendenti*). Il campione deve distribuirsi (in queste variabili) proporzionalmente alla popolazione, perché deve avere la stessa struttura.

2.2 Numerosità del campione

Siccome il campione è generalmente costituito da un gruppo di casi statistici estratti dalla popolazione ma di numerosità inferiore, il numero di unità che selezioniamo non è priva di importanza. Ovviamente dipende dal tipo di variabile che si vuol misurare, dall'ampiezza della popolazione di riferimento e da eventuali limiti materiali. Ad es. una ricerca d'opinione sugli italiani che utilizzi un campione di 500 individui, è meno rappresentativa di una che ne utilizzi 5000, mentre una ricerca sugli abitanti del Veneto potrebbe utilizzare anche un campione di soli 500 individui.

Un primo criterio minimale è quello di utilizzare una certa percentuale della popolazione di riferimento, ad es. il 10%, percentuale che può essere diminuita al 5% o all'1% se la popolazione è molto ampia. Un secondo criterio, è dato dai limiti materiali della popolazione. Ad es. una ricerca sulla soddisfazione del servizio all'interno di una casa di riposo per anziani, utilizzerà preferibilmente l'intera popolazione, mentre la stessa ricerca, riferita a tutte le case di riposo per anziani di una certa Regione, ne utilizzerà una percentuale.

Per quanto riguarda l'ampiezza minima, diremo che, per ogni possibile sottocampione, se la variabile da studiare è misurabile a livello di scala a intervallo, il limite minimo è 30 e i motivi per tale numero diverranno chiari nel capitolo sull'inferenza.

2.3 Modalità di estrazione

Il processo di estrazione del campione deve seguire delle regole che ci assicurino la sua rappresentatività. Nel processo di estrazione, un'unità statistica può essere riimmessa nella popolazione dopo essere stata estratta (e allora diremo che il campione è *con reimmissione* oppure *non esaustivo*) oppure può non essere ri-inserita (e allora diremo che il campione è *senza reinserimento* o *esaustivo*).

Possiamo suddividere i modi di estrazioni in:

A) **estrazione completamente casuale**: i casi statistici vengono estratti dalla popolazione in modo completamente casuale. Possiamo pensare alla popolazione come ad un gigantesco sacchetto contenente delle palline numerate, ciascuna corrisponde ad un caso statistico. Ogni pallina (salvo per il numero) è assolutamente uguale alle altre ed ha la stessa identica probabilità di essere estratta. In questo modo, il campione estratto dovrebbe essere uno tra i tanti possibili campioni estraibili e le probabilità che esso sia un

campione anomalo sono molto basse e diminuiscono all'aumentare dell'ampiezza del campione. Nella realtà, si possono usare 2 diversi metodi:

1) **estrazione casuale semplice**: numerare o elencare tutti i casi statitici possibili, quindi usare una tavola dei numeri casuali e selezionare le unità corrispondenti;
2) **estrazione casuale sistematica**: elencare i casi statistici, quindi selezionare i k-esimi elementi, dove k corrisponde alla numerosità della popolazione diviso l'ampiezza del campione da estrarre. Ad es. se la popolazione è di 1000 unità e voglio un campione di ampiezza 100, selezionerò il 10°, il 20°, il 30° elemento.

B) **estrazione casuale stratificata**: si suddivide la popolazione in gruppi che siano omogenei rispetto ad una variabile (indipendente) che è oggetto di studio. Quindi all'interno di questi strati si utilizza un'estrazione casuale (semplice o sistematica). Il campione può essere multi-stratificato e l'ampiezza dei singoli sottocampioni dovrà rispettare le proporzioni della popolazione.

3 Il test di chi-quadro [0.2]

3.1 Introduzione

Il test di chi-quadro (χ^2) è una *tecnica* di inferenza statistica che si basa sulla *statistica* di chi-quadro[1]. Si usa con variabili a livello di scala Nominale e/o Ordinale.

Lo scopo principale di questa tecnica è:

a) di verificare la casualità della distribuzione di una variabile categoriale;
b) l'indipendenza di due variabili nominali;
c) le differenze con un modello teorico.

Per lo scopo del punto b), la statistica di chi-quadro può essere considerata una statistica di associazione.

Per ora, ci limiteremo a considerare il primo aspetto (ovvero il punto *a*).

Da diversi anni, gli psichiatri hanno avanzato l'ipotesi che i pazienti affetti da schizofrenia nascano prevalentemente nei periodi invernali (Bradbury & Miller, 1985). Vogliamo vedere se anche i dati in nostro possesso ci portano a conclusioni analoghe. A questo scopo, usando le cartelle cliniche, raccogliamo le informazioni relative alla data di nascita di un certo numero di pazienti schizofrenici e li suddividiamo in categorie corrispondenti alle 4 stagioni dell'anno:

Patologia	*Primavera*	*Estate*	*Autunno*	*Inverno*	*Totale*
Soggetti schizofrenici	*125*	*130*	*153*	*228*	*636*

Tabella 3.1 - Dati fittizi

Se formulassimo un'ipotesi di assoluta uniformità, dovremmo aspettarci che ogni cella contenga più o meno la stessa percentuale di soggetti (poiché ci sono 4 celle, il 25% circa, cioè 159). E' ben difficile però ottenere esattamente questa cifra e si otterranno invece valori diversi che oscillano attorno a quello considerato uniforme. Valori molto vicini a quelli teorici avranno buone probabilità di essere delle "variazioni casuali", mentre valori molto diversi e lontani da quelli teorici avranno poche probabilità di essere considerati "fluttuazioni casuali". Serve quindi un criterio per decidere fino a che punto dobbiamo accettare come casuali le varie oscillazioni.

Il procedimento di calcolo che adotteremo è abbastanza simile a quello che abbiamo usato per la varianza e può essere riassunto concretamente così:

1. Calcoliamo il valore medio teorico (t) che dovremmo aspettarci all'interno di ogni cella se i 4 eventi fossero equiprobabili…	636 / 4 = 159
2. Quindi lo scarto della frequenza osservata (f) di ogni cella rispetto a quella teorica (t)	125 - 159 = -34
3. Questa differenza viene elevata a quadrato per diminuire le differenze piccole e aumentare quelle grandi…	$-34^2 = 1156$
4. Dividiamo infine per la frequenza teorica, in modo da standardizzare le distanze…	1156 / 159 = 7.27
5. Ripetiamo il procedimento per tutte le celle…	130 -> 5.29 153 -> 0.23

[1] Il nome chi-quadro si usa per indicare la distribuzione di probabilità, una particolare tecnica di inferenza statistica e un indice statistico. Alcuni autori usano il simbolo χ^2 (chi minuscola) per indicare la distribuzione di probabilità e X^2 (chi maiuscola) per indicare la statistica.

	228 -> 29.94
6. Sommiamo i vari risultati parziali…	42.73

Ci sono in questo procedimento due passaggi (terzo e quarto) che potrebbero essere complessi da capire: il quadrato della differenza rispetto al valore teorico e la sua divisione per il valore teorico. Elevare a quadrato una differenza (tecnica che abbiamo già applicato per il calcolo della varianza), ci permette di ottenere due effetti:

a) eliminare il segno negativo;
b) amplificare le differenze proporzionalmente alla loro grandezza (il quadrato di 2 è 4, il quadrato di 5 è 25 e quello di 10 è 100).

In questo modo, il numero che otteniamo è tanto più grande quanto maggiore è la differenza di partenza. Dividendo poi questo valore per la frequenza teorica, otteniamo una misurazione che, più o meno, equivale a dire: "quante frequenze teoriche stanno in questo scarto quadratico". Si utilizza quindi ciascuna frequenza teorica per esprimere lo scarto.

In pratica abbiamo costruito un numero che rappresenta *la somma ponderata degli scarti delle frequenze di ciascuna cella rispetto alla sua attesa teorica*.

E' semplice allora capire come, maggiore è il valore trovato e maggiore è lo scostamento delle frequenze osservate rispetto a quelle teoriche che ci dovremmo aspettare e maggiore sarà quindi la probabilità che la distribuzione non sia casuale ma in qualche modo influenzata da una delle variabili categoriali prese in considerazione.

E' altrettanto facile capire come il valore trovato dipenda (per la sua grandezza) anche dal numero di celle e dal numero di frequenze totali: quante più celle possiede la tabella, tanto maggiore sarà la probabilità che una di esse si comporti in modo anomalo; quanto più alto il totale, quanto più è probabile trovare valori elevati della statistica di chi-quadro.

Un lavoro analogo possiamo farlo su tabelle di contingenza (ossia tabelle a due entrate) che incrociano le frequenze con cui accadono assieme determinate categorie di due variabili. Ad es. la tabella di contingenza che incrocia i valori delle ipotetiche variabili A e B potrebbe essere:

	A_1	A_2	A_3
B_1	60	53	12
B_2	53	23	16
B_3	55	48	20

Tabella 3.2

3.2 Terminologia

Prima di proseguire, poniamo alcune basi terminologiche.

Abbiamo già visto che è possibile indicare i valori reali di una distribuzione, usando una lettera di variabile (generalmente la x) e una lettera indice (generalmente la i). Usando questa notazione, possiamo riscrivere la Tabella 3.2 in modo generico, in questo modo:

	A_1	A_2	A_3
B_1	n_{11}	n_{12}	n_{13}
B_2	n_{21}	n_{22}	n_{23}
B_3	n_{31}	n_{32}	n_{33}

Tabella 3.3

In questa notazione, n_{11} indica il contenuto della cella all'incrocio fra la riga 1 e la colonna 1, n_{32} la cella all'incrocio fra la riga 3 e la colonna 2; quindi n_{11}=60, n_{12}=53, n_{31}=55.

Più genericamente possiamo scrivere la tabella in quest'altro modo:

	A_1	A_2	A_3	tot.
B_1	n_{ij}	…	…	$n_{i.}$
B_2	…	…	…	…
B_3	…	…	…	…
tot.	$n_{.j}$	…	…	$n_{..}$

Tabella 3.4

dove n_{ij} indica le singole celle (al variare degli indici *i* e *j*), mentre $n_{i.}$, $n_{.j}$ e $n_{..}$ sono rispettivamente i totali di riga, i totali di colonna e il totale generale.

Alcune ovvie relazioni possono essere espresse in termini matematici all'interno di questa tabella, usando le stesse notazioni:

a) il totale della i-esima riga è pari alla somma delle celle di quella riga (con *c* che indica il numero delle colonne):

$$n_{i.} = \sum_{j=1}^{c} n_{ij}$$

b) analogamente per i totali della colonna j-esima (con *r* che indica il numero delle righe):

$$n_{.i} = \sum_{i=1}^{r} n_{ij}$$

c) il totale generale è pari alla somma di tutti i totali di riga, alla somma di tutti i totali di colonna e alla somma di tutte le celle:

$$N = n_{..} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} n_{i.} = \sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij}$$

3.3 La formula di chi-quadro

Se trasformiamo il procedimento di calcolo usato al par. 3.1 in formula, poiché abbiamo solo una riga, possiamo scrivere:

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_i - t_i)^2}{t_i}$$

dove indichiamo con *c* il numero di colonne, *i* è l'indice che assume tutti i valori fra 1 e *c*, n_i è il valore della cella (frequenza ottenuta) e t_i è il corrispondente valore teorico.

Se avessimo utilizzato una tabella a doppia entrata (come la Tabella 3.2), la formula sarebbe invece:

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r} \frac{(n_{ij} - t_{ij})^2}{t_{ji}}$$

in cui *i* e *j* indicano rispettivamente le righe e le colonne. In genere, però, la formula di chi quadro la si trova scritta in modi più generici, usando notazioni di derivazione anglosassone, ad es.:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

In questo contesto f_o significa frequenza osservata (*observed*) e f_e frequenza attesa (*expected*).

3.4 I valori teorici

Al paragrafo 3.1 abbiamo calcolato i valori teorici dividendo N per il numero di celle ($t_i=n_i/c$). Questo perché avevamo 1 sola variabile. Con 2 variabili, le cose cambiano un poco. Vediamo come.

Ipotizziamo di voler verificare se, in un campione di 42 soggetti di entrambi i sessi, la distribuzione dei livelli di educazione sia (H_1) o meno (H_0) dipendente dal sesso:

		Livello Educativo		
Sesso		Basso	Alto	Totale
	Maschi	*13*	*9*	*22*
	Femmine	*13*	*7*	*20*
Totale		*26*	*16*	*42*

Tabella 3.5 - Dati fittizi

Se precedentemente abbiamo utilizzato il totale generale e lo abbiamo diviso per il numero delle celle (cioè, in questo caso, 42 / 4 = 10.5), adesso dobbiamo tener conto che vi sono dei vincoli. Abbiamo solo 20 soggetti di sesso *femminile* e non potremo mai aspettarci, neppure teoricamente, che siano 21 (cioè 10.5 + 10.5); analogamente abbiamo solo 16 soggetti di livello educativo *Alto* (e non 21). I nostri valori teorici devono quindi tener conto di questi totali che "costringono" i risultati in certa direzione; per calcolare i valori attesi (per ogni cella) si utilizza una formula abbastanza semplice: si moltiplicano fra loro i totali di riga e di colonna di quella cella e si divide il risultato per il totale generale.

		Freq.	***Freq. teorica***
Maschi	Basso	*13*	*22 x 26 / 42 = 13.62*
	Alto	*9*	*22 x 16 / 42 = 8.38*
Femmine	Basso	*13*	*20 x 26 / 42 = 12.38*
	Alto	*7*	*20 x 16 / 42 = 7.62*

in formula:

1) $$t_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

La scelta di questo metodo di calcolo non è casuale. Infatti questa semplice formula di calcolo corrisponde alla probabilità teorica che si verifichino contemporaneamente 2 eventi fra di loro indipendenti ovvero la cui probabilità di comparsa di uno dei due non incide sulla probbilità del secondo e viceversa. Dallo studio della probabilità, sappiamo che tale valore è dato dal prodotto delle probabilità dei singoli eventi. Nel nostro caso, per la prima cella della tabella, si tratta di incrociare la probabilità di essere maschio [P(M)] con la probabilità di avere un basso livello educativo [P(B)]:

P(MB) = P(M) P(B)

La teoria della probabilità frequentista ci dice che la probabilità di un evento è data dalla frequenza con cui compare quell'evento, divisa per il totale degli eventi, quindi:

$$P(M) = \frac{22}{42} = 0.5238$$

$$P(B) = \frac{26}{42} = 0.6190$$

$$P(MB) = 0.5238 \cdot 0.6190 = 0.3243$$

Poiché abbiamo 42 soggetti, dobbiamo moltiplicare N per la probabilità della prima cella, al fine di stimare quanti soggetti dovrebbero stare in quella cella: $42 \cdot 0.3243 = 13.6188$

Se scriviamo tutto il procedimento in un colpo solo, vediamo facilmente che la formuletta 1 è ricavata da tutto questo ragionamento:

$$N \cdot P(MB) = N \cdot P(M) \cdot P(B) = 42 \cdot \frac{22}{42} \cdot \frac{26}{42} = \frac{22 \cdot 26}{42}$$

3.5 Un esempio

Partendo dai dati della Tabella 3.2, proviamo a calcolare un chi-quadro completo. Iniziamo calcolando i totali di riga e di colonna.

f_o

	A_1	A_2	A_3	tot.
B_1	60	53	12	125
B_2	53	23	16	92
B_3	55	48	20	123
tot.	168	124	48	34

Tabella 3.6

I valori teorici vengono calcolati come:

f_t

125 · 168 / 340	125 · 124 / 340	125 · 48 / 340
92 · 168 / 340	92 · 124 / 340	92 · 48 / 340
123 · 168 / 340	123 · 124 / 340	123 · 48 / 340

Tabella 3.7

Notate come, per ogni colonna, vi sia una parte della formula che non cambia (analogamente se consideriamo le formule per riga). Facendo i conti a mano, possiamo semplificarli così:

168 / 340 = 0.49	124 / 340 = 0.36	48 / 340 = 0.14
125 · 0.49	125 · 0.36	125 · 0.14
92 · 0.49	92 · 0.36	92 · 0.14
123 · 0.49	123 · 0.36	123 · 0.14

Tabella 3.8

[*Nel trascrivere i dati, ho arrotondato a 2 cifre decimali mentre sarebbe opportuno utilizzare tutti i decimali possibili*]

Otteniamo così le seguenti frequenze teoriche:

f_t

61.25	*45.00*	*17.50*
45.08	*33.12*	*12.88*
60.27	*44.28*	*17.22*

Tabella 3.9

Possiamo adesso applicare la formula per il calcolo del chi-quadro, dapprima calcolando, per ogni cella, il valore della formula e, successivamente sommando il tutto (volendo essere un esempio dettagliato, farò tutti i passaggi e userò tutti i decimali del visore di una normale calcolatrice):

$(60-61.25)^2/61.25$	$(53-45)^2/45$	$(12-17.5)^2/17.5$
$(53-45.08)^2/45.08$	$(23-33.12)^2/33.12$	$(16-12.88)^2/12.88$
$(55-60.27)^2/60.27$	$(48-44.28)^2/44.28$	$(20-17.22)^2/17.22$

$(-1.25)^2/61.25$	$(8)^2/45$	$(-5.5)^2/17.5$
$(7.92)^2/45.08$	$(-10.12)^2/33.12$	$(3.12)^2/12.88$
$(-5.27)^2/60.27$	$(3.72)^2/44.28$	$(2.78)^2/17.22$

(1.5625)/61.25	*(64)/45*	*(30.25)/17.5*
(62.7264)/45.08	*(52.6064)/33.12*	*(9.7344)/12.88*
(27.7729)/60.27	*(13.69)/44.28*	*(7.7284)/17.22*

0.0255102	*1.4222222*	*1.7285714*
1.3914463	*1.5883575*	*0.7557764*
0.460808	*0.3091689*	*0.4488037*

Sommando il contenuto di tutte le celle e arrotondando a due decimali, otteniamo un χ^2 di 8.13

E ora che abbiamo calcolato la statistica di chi-quadro, cosa ce ne facciamo?

Al paragrafo 3.1 avevamo scritto che la statistica di chi-quadro serviva per stabilire fino a punto potevamo accettare le frequenze ottenute come analoghe, simili, vicine a quelle teoriche e che più alto era il valore trovato, tanto più era improbabile che tale lontananza fosse casuale.

Dobbiamo a questo punto fare un procedimento di inferenza statistica.

3.6 La distribuzione chi-quadro

3.7 I gradi di libertà

Riprendiamo in considerazione la Tabella 3.2, con i suoi totali.

Il concetto di gradi di libertà nasce dal fatto che avendo 168 eventi nella categoria A1, dovendoli suddividere nelle 3 celle corrispondenti alle categorie di B, noi abbiamo libertà di mettere quanti eventi vogliamo in 2 sole celle… la terza è "costretta" a contenere gli eventi restanti. Lo stesso ragionamento viene fatto per A2, A3 e per ciascuno dei valori di B.

Nella tabella quindi vi sono delle celle (per convenzione le ultime) che non possono contenere "qualsiasi numero" ma solo quanto resta per poter sommare al totale degli eventi di quella categorie. Nella tabella che segue queste celle sono indicate con uno sfondo grigio.

	A_1	A_2	A_3	tot.
B_1	60	53	12	125
B_2	53	23	16	92
B_3	55	48	20	123
tot.	168	124	48	34

Tabella 3.10

Il numero delle celle "libere", corrisponde ai gradi di libertà (in inglese, *degree of freedom*, abbreviato in *df*). La formula generale, facilmente comprensibile dall'esempio precedente, è:

$$gl = (r-1)(c-1)$$

ossia numero di righe per numero di colonne, a ciascuno dei quali viene precedentemente sottratto uno.

3.8 L'inferenza

Una volta calcolato il valore finale di un chi-quadro si applica il solito meccanismo del livello di significatività, facendo riferimento alla *distribuzione di chi-quadro* e ai gradi di libertà implicati. Il valore di significatività trovato indica il rischio che noi corriamo, la probabilità che un determinato valore di chi-quadro sia casuale.

Ritornando all'esempio di Tabella 3.1, per sapere se il valore di chi-quadro da noi trovato (42.73) è significativo, consultiamo le tavole relative della distribuzione di chi-quadro; cerchiamo la riga corrispondente a 3 gradi di libertà (cioè 4-1) e quindi avanziamo alla ricerca di un valore che sia superiore a quello da noi trovato. Nessuno dei valori segnati sulla riga supera il valore di 42.73, quindi la probabilità a esso connessa è così piccola da essere inferiore allo .001.

Tabella 3.11- Valori critici di chi-quadro (estratto)

	.05	.01	.001
gl=1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.341	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515

La maggior parte dei programmi per computer, oltre a fornire il valore di chi-quadro e i gradi di libertà, fornisce anche un valore di probabilità o di significatività che permette di interpretare immediatamente il valore statistico calcolato senza utilizzare le tabelle. Se utilizzassimo un programma statistico per rifare lo stesso chi-quadro, otterremmo questi risultati:

	Patologia
Chi-square	42.73
Df	3
Sig.	.000

Tabella 3.12

Con 3 gradi di libertà (*df = degree of freedom*), un chi-quadro pari a 42.73 è da considerarsi molto significativo; in effetti la significatività è pari a .000 (che significa che vi è almeno una cifra diversa da zero a partire dal quarto decimale e che tale cifra non viene visualizzata per motivi di arrotondamento) ovvero vi è meno di 1 probabilità su 10.000 che i nostri dati siano così diversi tra loro per puro caso. Nel caso fittizio da noi considerato dovremmo quindi concludere che effettivamente nasce un numero di soggetti schizofrenici diverso rispetto alle stagioni di nascita e in particolare in *inverno*.

Senza bisogno di applicare le formule, si possono consultare delle apposite tavole, che forniscono la probabilità associata ad un certo valore di χ^2, per un dato grado di libertà.

Usando le tavole, dobbiamo seguire il seguente procedimento:

1- fissiamo un livello α (di solito $\alpha = .05$)
2- calcoliamo i gradi di libertà
3- troviamo sulle tavole la riga corrispondente ai gradi di libertà e la scorriamo fino alla colonna corrispondente al livello α
4- all'incrocio fra riga e colonna, troviamo il valore critico di χ^2 (χ^2_c)
5- se il nostro chi-quadro è inferiore al valore critico, accettiamo l'ipotesi H_0
6- se è superiore, accettiamo l'ipotesi alternativa

Tabella 3.13

$\chi^2 < \chi^2_c$	*accetto H_0*
$\chi^2 \geq \chi^2_c$	*accetto H_1*

Esempi:

$\chi^2 = 10.63$; gl=4; $\alpha = .05$

Consultando le tavole, trovo un χ^2 critico di 9.48

Poiché 10.63 è maggiore del valore critico, rifiuto H_0 e accetto H_1

3.9 Correzione di Yates

Riprendendo, invece, i dati di Tabella 3.3, già confrontando ad occhio i valori teorici e i valori ottenuti, possiamo aspettarci che il chi-quadro non sia significativo, poiché i due valori sono molto vicini fra loro. In effetti, se calcoliamo la statistica con un programma per computer, otteniamo:

Pearson Chi-square	*.155*
Continuity Correction	*.006*
Df	*1*
Sig.	*.694*

Tabella 3.14

Il risultato del chi-quadro (*Pearson Chi-Square*) è pari a 0.155 che, con 1 grado di libertà non risulta significativo: la probabilità esatta calcolata (*Sig.*) è pari infatti a 0.694, cioè: se decidessimo di accettare l'ipotesi H_1 che il sesso influisce sul livello economico di un individuo, correremmo un rischio di sbagliare del 69%; rischio che è considerato eccessivo e che ci induce ad accettare l'ipotesi opposta.

Poiché la tabella è composta da 4 celle in forma 2x2, viene calcolato un altro indice di chi-quadro (*Continuity correction*), conosciuto anche come "correzione di Yates", che permette di adeguare maggiormente la distribuzione del chi-quadro di una tabella 2x2 alla curva di chi-quadro.

Ora che abbiamo spiegato anche a livello intuitivo la statistica di chi-quadro, affrontiamo brevemente i criteri da considerare nella sua applicazione.

- Può essere usata con una o due variabili categoriali o ordinali;
- L'attribuzione di un caso (soggetto) ad una categoria/cella dev'essere univoca, ovvero un soggetto classificato in una cella non deve comparire contemporaneamente in un'altra: questo si chiama *indipendenza dei casi*;
- Non si può applicare (è rischioso) se più del 20% delle celle ha una "frequenza attesa" inferiore a 5 (solo nel caso di una tabella 2x2, si può utilizzare una formula alternativa chiamata "correzione di Yates"); oppure se una cella ha frequenza attesa inferiore a 1. In questo caso, se vi sono 3 o più categorie e se il "significato" logico di tali categorie lo permette, è possibile accorparne alcune in modo da ampliare la numerosità di quella particolare riga/colonna. Ad es. è possibile far confluire la categoria "convivente" con quella di "sposato" e la categoria "vedovo" con "divorziato" se ciò che importa nell'analisi è l'ampiezza del nucleo familiare;
- Quando il numero di celle è piccolo e il numero di casi è grande conviene "verificare" la validità del chi-quadro tramite l'uso del coefficiente C di contingenza.

Quando si usa il chi-quadro su tabelle di contingenza con più di 2 righe o colonne, e si trova un valore significativo di χ^2, si vorrebbe anche sapere quale cella o quali celle sono responsabili della significatività. Questa conoscenza aiuta molto nell'interpretare i risultati del test. Esistono delle tecniche abbastanza complesse, chiamate "tecniche di partizione" che permettono di andare a vedere come si comportano le celle o alcuni gruppi di celle rispetto a tutte le altre. Tralasciando queste tecniche di partizione, suggeriamo l'uso dei residui standardizzati, proposti da Haberman (1973):

$$R = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

Quando il residuo standardizzato di una cella supera il valore di 2, secondo Haberman, la cella si discosta dal suo valore teorico a sufficienza per essere considerata come una cella anomala, che ha contribuito a rendere significativo il test di chi-quadro.

Fino ad ora abbiamo utilizzato la statistica di chi-quadro per verificare se una determinata distribuzione era (oppure no) uniformemente distribuita. Per questo motivo, abbiamo calcolato i valori delle frequenze teoriche come rapporti ponderati delle righe e delle colonne e, per accettare l'ipotesi H_1, ci aspettavamo di trovare valori di chi-quadro molto elevati e statisticamente associati ad un basso valore di α.

Ma poiché la tecnica del chi-quadro confronta una distribuzione realmente ottenuta con una teorica, noi possiamo utilizzare questo test anche per verificare un nostro particolare modello di dati. In questo caso, però, un valore elevato di chi-quadro (quindi significativo), vorrebbe dire che la distribuzione reale dei nostri dati si discosta dalla distribuzione teorica che ci aspettavamo mentre un valore bassissimo o nullo, significherebbe che la nostra teoria spiega bene i dati da noi trovati.

Come esempio usiamo quello iniziale. Leggiamo un articolo in cui si afferma che nel periodo invernale (rispetto alle altre stagioni) nascono più soggetti che poi riveleranno disturbi di tipo schizofrenico. L'autore dell'articolo precisa anche che in genere, durante l'inverno, nella sua popolazione di riferimento, sono nati circa il 55% di tutti i soggetti con tali disturbi. Noi allora prendiamo i dati in nostro possesso e calcoliamo un normale test di chi-quadro, che ci risulta significativo. A questo punto ci chiediamo se le caratteristiche del nostro campione sono simili a quelle del campione dell'autore dell'articolo. Ricalcoliamo il chi-quadro, usando questa volta come frequenze teoriche i valori che ricaveremo dai dati dell'articolo, ad es. pari al 55% per l'inverno e al 20% per la primavera, al 10% per l'estate e al 15% per l'autunno. Se il chi-quadro così calcolato è *non significativo*, allora il nostro campione è simile a quello utilizzato nell'articolo di riferimento, se è significativo, allora non vi è somiglianza.

Se dovete calcolare un chi-quadro su dati già in forma tabellare, anziché usare un complesso programma statistico, è più semplice usare un programma apposito. Nel mondo di internet ve ne sono due facilmente utilizzabili: il primo è un programma in italiano per il sistema operativo Dos[2], mentre il secondo programma è in inglese, più completo, ed è disponibile in qualunque *mirror* di SimtelNet nella directory di statistica[3].

[2] http://psico.univr.it/germano/chiquadro.asp
[3] ad es. in http://sunsite.cnlab-switch.ch/ftp/mirror/simtelnet/msdos/statstcs/chi1_0.zip

4 La correlazione [0.4]

4.1 Cos'è la correlazione

La correlazione è un indice che misura l'associazione fra due variabili, più in particolare, misura il grado in cui due variabili si "muovono assieme". Esistono diversi indici di correlazione, applicabili a tipi diversi di variabili e a diversi livelli di misura. Prenderemo in considerazione la correlazione lineare prodotto-momento di Pearson, per capire il concetto di correlazione e vedremo quindi altri indici di correlazione.

Il concetto di correlazione è relativamente semplice, ma, da un punto di vista formale (ovvero matematico) ha molte relazioni con altre tecniche (come ad esempio la regressione lineare, i punti standard…). Il percorso che seguirò per spiegare questa tecnica statistica, è solo uno dei possibili, spero, il più semplice.

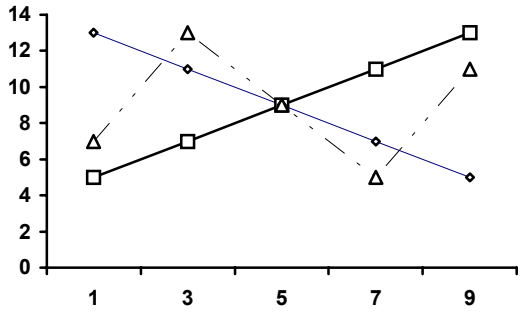
4.2 Correlazione lineare di Pearson

Immaginiamo di avere una serie di quattro variabili (del tutto fittizie, con valori scelti appositamente per evidenziare determinate relazioni), che chiameremo con le lettere finali dell'alfabetico, e i dati di alcuni soggetti (che chiameremo con le prime cinque lettere dell'alfabeto).

Tabella 4.1

	X	Y	Z	W
A	1	5	13	7
B	3	7	11	13
C	5	9	9	9
D	7	11	7	5
E	9	13	5	11
$\overline{X}$	5	9	9	9
s	2.828	2.828	2.828	2.828

Queste variabili sono state scelte in modo da avere uguale deviazione standard e media (per lo meno le variabili Y, Z e W). Come possiamo vedere, a piccoli valori di X, corrispondono piccoli valori di Y e grandi valori di Z, a valori grandi di X, corrispondono grandi valori di Y e piccoli valori di Z; non sembra esistere una vera relazione fra X e W. Possiamo rappresentare graficamente queste relazioni, in un grafico.



Il grafico evidenzia bene come:

- la relazione fra X e Y è una relazione lineare crescente;
- la relazione fra X e Z è lineare decrescente;
- la relazione fra X e W non è riconducibile ad una regola.

Se riscriviamo la Tabella 4.1 in modo da usare gli scarti dalla media (ovvero $x - \overline{X}$), possiamo notare qualcosa di ancora più significativo:

Tabella 4.2

	X	*Y*	*Z*	*W*
A	-4	-4	4	-2
B	-2	-2	2	4
C	0	0	0	0
D	2	2	-2	-4
E	4	4	-4	2

Quello che possiamo notare è che la relazione lineare crescente fra X e Y è caratterizzata dal fatto che tutti gli scarti dalla media hanno lo stesso segno, la relazione inversa fra X e Z corrisponde a scarti che hanno segno opposto, mentre la relazione non definita fra X e W ha scarti i cui segni si associano "casualmente".

Con questi dati possiamo tentare di costruire una statistica, che chiameremo indice di correlazione lineare. In teoria, questo indice, dovrebbe avere un valore positivo per indicare relazioni lineari positive (come quella fra X e Y), un valore negativo per relazioni lineari negative o inverse (X e Z) e un valore nullo per relazioni inesistenti o nulle (X e W). Inoltre dovremmo cercare di standardizzare l'indice affinché oscilli sempre fra valori predefiniti, qualunque siano i numeri che costituiscono le variabili. Una possibilità è quella che oscilli fra –1 e +1.

Un primo passo potrebbe essere quello di moltiplicare i valori delle variabili che vogliamo mettere in relazione e poi di sommare questi valori:

Tabella 4.3

XY	*XZ*	*XW*
16	-16	8
4	-4	-8
0	0	0
4	-4	-8
16	-16	8
40	-40	0

Se, a questo punto, dividiamo i totali per la numerosità, otteniamo qualcosa che assomiglia alla formula della varianza e che chiameremo covarianza:

$$\text{cov} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N}$$

E se dividiamo la covarianza per il prodotto delle deviazioni standard, otteniamo un valore standardizzato, che oscilla fra –1 e +1. Questa è una delle formule che esprime la *correlazione di Pearson*.

$$r = \frac{\text{cov}}{s_x s_y}$$

Tabella 4.4

	XY	*XZ*	*XW*
Cov	40	-40	0
Cov / n	8	8	0
$s_x s_y$	8	8	8

r	*1*	*-1*	*0*

4.3 Formule alternative

Una formula alternativa per la correlazione di Pearson è facilmente derivabile dalla precedente, se consideriamo che nella formula della covarianza abbiamo le somme degli scarti dalla media e che queste vengono poi divise per le deviazioni standard. La formula (la più facile da ricordare) esprime la correlazione come media dei prodotti dei punti z (cfr. la dimostrazione 6.1.1 in Appendice):

$$r = \frac{\sum z_x z_y}{N}$$

Una seconda formula alternativa, è:

$$r = \frac{\frac{\sum xy}{N} - \overline{XY}}{s_x s_y}$$

Una terza formula alternativa (generalmente usata per i calcoli, anche se è più complessa da ricordare), utilizza solo i dati grezzi (cfr. la dimostrazione 6.1.2 in Appendice) e può esprimersi in due modi leggermente diversi:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

E il calcolo finale della correlazione fra *x* e *y*, secondo le due formule, risulta:

Tabella 4.5

X	*Y*	*X2*	*Y2*	*XY*
1	*5*	*1*	*25*	*5*
3	*7*	*9*	*49*	*21*
5	*9*	*25*	*81*	*45*
7	*11*	*49*	*121*	*77*
9	*13*	*81*	*169*	*117*
25	*45*	*165*	*445*	*265*

$$r = \frac{265 - \frac{25 \cdot 45}{5}}{\sqrt{(165 - \frac{(25)^2}{5})(445 - \frac{(45)^2}{5})}} = \frac{265 - \frac{1125}{5}}{\sqrt{(165 - \frac{625}{5})(445 - \frac{2025}{5})}}$$

$$= \frac{265 - 225}{\sqrt{(165 - 125)(445 - 405)}} = \frac{40}{\sqrt{40 \cdot 40}} = \frac{40}{40} = 1$$

$$r = \frac{5 \cdot 265 - 25 \cdot 45}{\sqrt{[5 \cdot 165 - (25)^2][5 \cdot 445 - (45)^2]}} = \frac{1235 - 1125}{\sqrt{(825 - 625)(2225 - 2025)}}$$

$$= \frac{200}{\sqrt{200 \cdot 200}} = \frac{200}{200} = 1$$

4.4 Interpretazione

Non vi è un criterio matematico o statistico per interpretare la forza della relazione fra le due variabili. La prassi ha stabilito una serie di convenzioni:

Tabella 4.6

Valore di r	*Correlazione*	*Relazione*
0.00-0.20	*Piccola*	*Molto poco intensa, quasi inesistente*
0.20-0.40	*Bassa*	*Piccola, appena appena apprezzabile*
0.40-0.60	*Regolare*	*Considerevole*
0.60-0.80	*Alta*	*Intensa*
0.80-1.00	*Molto alta*	*Molto intensa*

Una particolare attenzione va posta nell'interpretare il significato stesso di correlazione.

Innanzitutto è necessario ricordare che la formula, generalmente utilizzata (quella di Pearson), è relativa ad una relazione lineare e che quindi tutte le forme diverse di relazione, possono produrre risultati anomali. Consideriamo i due grafici della figura seguente. Entrambi rappresentano dati che si distribuiscono in modo curvilineo; tuttavia, nel primo esempio, vi sono dati sufficienti per portare ad una correlazione lineare di .54 (una correlazione considerata considerevole).

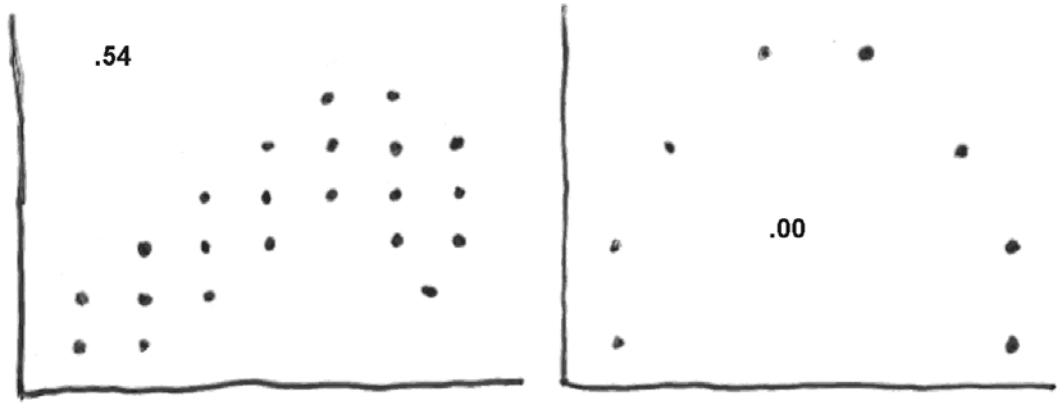


Figura 4.1 – Esempi di relazioni curvilinee e corrispondenti valori di r

Un secondo problema riguarda il rapporto di causalità dell'indice di correlazione, ci dice solo e soltanto che le due variabili hanno un andamento comune. Non ci dice mai che una variabile è la causa e che l'altra è l'effetto.

Un terzo problema è legato al fatto che, qualsiasi siano le variabili utilizzate, sempre si ottiene un qualunque valore di *r*. Tuttavia, non sempre questa correlazione avrà un significato logico. Ad esempio, io posso correlare il mio consumo giornaliero di acqua con il numero di barattoli di yogurt venduti giornalmente da un certo negozio; troverò certamente un valore di *r* che potrebbe anche essere diverso da zero. Sarà comunque una correlazione priva di significato logico. Questo tipo di correlazione è chiamata "casuale" o "spuria".

4.5 Mi posso fidare?

Un quarto problema dipende appunto dal fatto che ottengo sempre e comunque un valore di correlazione. Consideriamo che, quando calcoliamo la correlazione fra due va-

riabili, stiamo lavorando su un campione che è stato estratto casualmente da una popolazione. Potrei imbattermi in due situazioni opposte (illustrate in Figura 4.2):

1. il campione su cui calcolo la correlazione presenta, casualmente, una relazione lineare (i puntini cerchiati), mentre la popolazione da cui l'ho estratto non lo è;
2. il campione su cui calcolo la correlazione non ha una struttura lineare, mentre la popolazione sì.

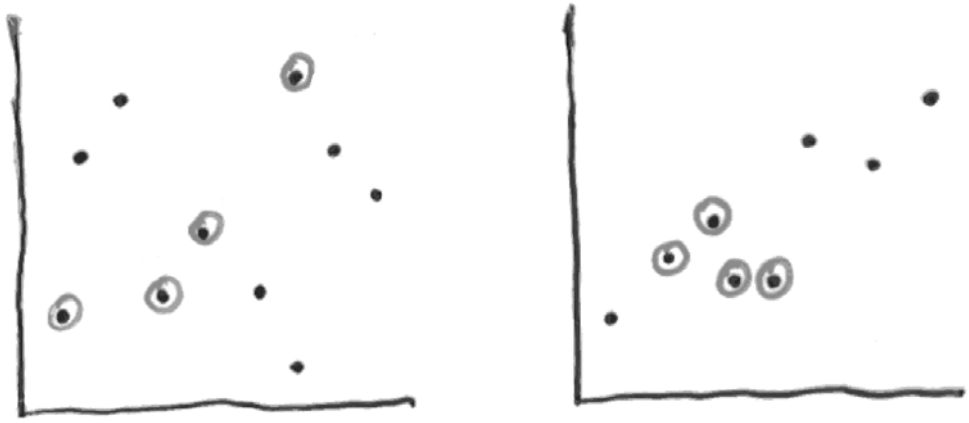


Figura 4.2

Si tratta a questo punto di attivare un processo di inferenza per sapere quanto possiamo fidarci della correlazione calcolata. Il ragionamento che sottostà all'inferenza è il seguente ed è comune a quello utilizzato con altre statistiche.

Ipotizziamo che la correlazione sia tratta da una popolazione che abbia correlazione nulla:

H_0: $\rho = 0$

Ovviamente formuliamo anche l'ipotesi alternativa:

H_1: $\rho \neq 0$

Come il solito, ragioniamo sulla base dell'ipotesi nulla.

Ipotizziamo una popolazione in cui la correlazione fra X e Y sia conosciuta e sia pari a 0 (o meglio $\rho = 0$). Estraggo un campione di ampiezza n e calcolo la correlazione fra le due variabili x e y. Estraggo un altro campione e, di nuovo, calcolo la correlazione. Ripeto il procedimento per un numero infinito di volte. Tutti i possibili valori di r calcolati su tutti i possibili campioni estratti da questa popolazione vanno a formare una distribuzione campionaria delle correlazioni, che avrà una sua propria media e una sua propria deviazione standard che tenderà ad approssimarsi alla distribuzione normale all'aumentare del valore di unità statistiche che uso per il calcolo (ovvero n):

Figura 4.3

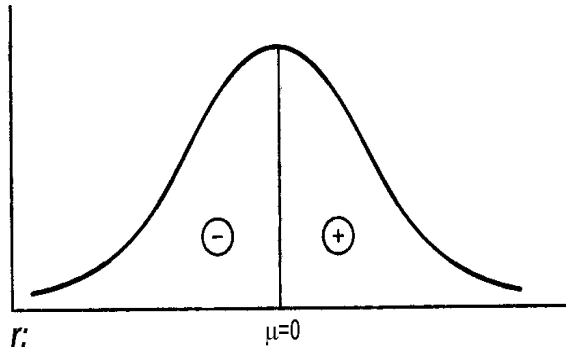


Tabella 4.7

$$\mu_r = 0 \qquad\qquad s_r = \sqrt{\frac{1-r^2}{n-2}}$$

E' ovvio aspettarsi che, sebbene il valore zero (corrispondente alla media) sia il più frequente, sia possibile ottenere anche correlazioni diverse da zero, sia negative sia positive. Tanto più le correlazioni saranno lontane da zero (cioè grandi, in valore assoluto) tanto meno saranno frequenti e probabili (sempre nell'ipotesi che siano calcolate su campioni provenienti da una popolazione con correlazione nulla).

Per sapere se la probabilità della nostra correlazione *r*, calcolata su un singolo campione, è sufficiente calcolare il punto z corrispondente a *r*:

$$r_t = \frac{r-\rho}{s_r}$$

r_t si distribuisce secondo la *distribuzione t di Student* con *n-2* gradi di libertà.

Senza bisogno di applicare le formule, si possono consultare delle apposite tavole, che forniscono la probabilità associata ad un certo valore di *r*, per un dato grado di libertà.

Tabella 4.8- Valori critici di correlazione di Pearson (estratto)

Mono-dir.	*.05*	*.025*	*.010*	*.005*
Bi-dir.	*.10*	*.05*	*.025*	*.01*
d=1	*.988*	*.997*	*.9995*	*.999*
2	*.900*	*.950*	*.980*	*.990*
3	*.805*	*.878*	*.934*	*.959*
4	*.729*	*.811*	*.882*	*.917*
5	*.669*	*.754*	*.833*	*.874*

Usando le tavole, dobbiamo seguire il seguente procedimento:

1- fissiamo un livello α (di solito α = .05)
2- stabiliamo l'ipotesi alternativa come monodirezionale o bidirezionale (quest'ultima è la scelta più comunemente usata)
3- calcoliamo i gradi di libertà (d=n-2)
4- troviamo sulle tavole la riga corrispondente ai gradi di libertà e la scorriamo fino alla colonna corrispondente al livello α

5- all'incrocio fra riga e colonna, troviamo il valore critico di r (r_c)
6- se la nostra correlazione è inferiore al valore critico, accettiamo l'ipotesi H_0
7- se è superiore, accettiamo l'ipotesi alternativa

Tabella 4.9

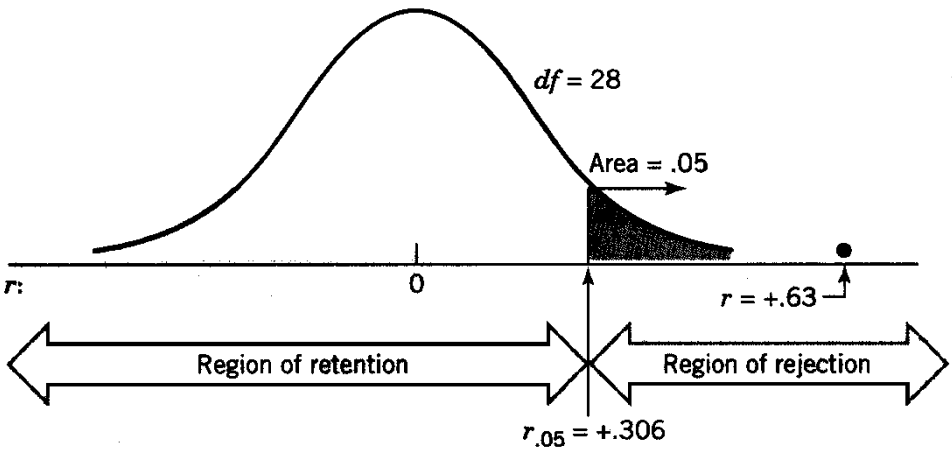
$r < r_c$	accetto H_0
$r \geq r_c$	accetto H_1

Accettare l'ipotesi nulla significa che, qualunque sia il valore di *r* che abbiamo trovato nel campione, esso è comunque pari a 0, poiché viene casualmente da una popolazione che ha correlazione zero. Non dobbiamo cercare di interpretare questa correlazione, perché è un errore di estrazione casuale del campione.

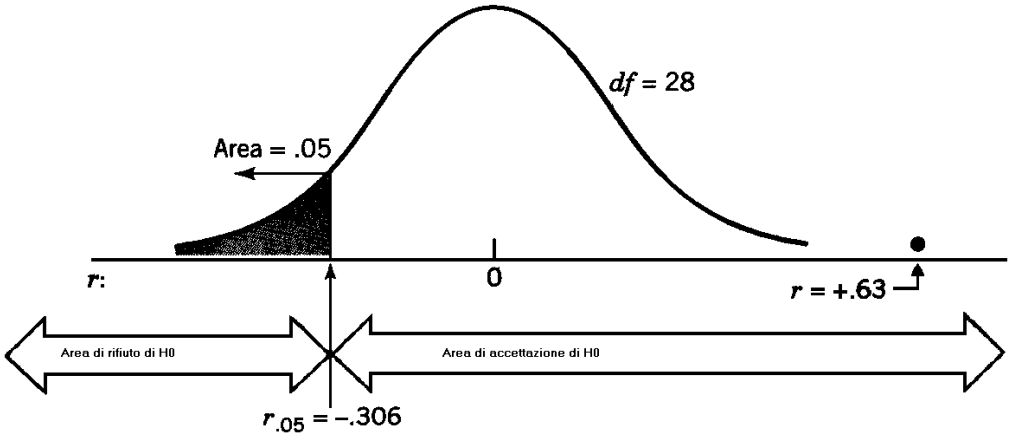
Accettare l'ipotesi alternativa, significa che il valore calcolato è effettivo poiché viene da un campione casualmente estratto da una popolazione che ha correlazione diversa da zero. A questo punto possiamo cercare di interpretare la correlazione trovata.

Esempi:
r=.63; N=30; gl=28; α = .05; ipotesi monodirezionale
Consultando le tavole, trovo un r critico di .306
Poiché .63 è maggiore del valore critico, rifiuto H_0 e accetto H_1

Fare un'ipotesi monodirezionale equivale a dire che noi ci aspettiamo che la correlazione sia non solo diversa da zero, ma anche positiva o negativa. In tal caso, l'area di probabilità corrispondente ad alfa, che ci indicherà l'area di accettazione o di rifiuto dell'ipotesi nulla, sarà posta su un solo lato della curva (cfr.).



Nell'ipotesi bidirezionale, ipotizziamo soltanto che la correlzaione sia diversa da zero, senza fare alcuna ipotesi sul suo andamento. In tal caso l'area di rifiuto dell'ipotesi nulla dovrà essere suddivisa fra i valori positivi e quelli negativi (cfr.).



Con i dati della Tabella 4.10, calcoliamo la correlazione lineare di Pearson e la sua significatività; per semplicità di calcolo, usiamo le formule alternative.

Le variabili x e y rappresentano rispettivamente l'"Abilità di comprensione di un testo scritto" e il "Quoziente di intelligenza".

Tabella 4.10 – Esempio di correlazione lineare

x	y	x^2	y^2	xy
46	126	2116	15876	5796
49	110	2401	12100	5390
48	103	2304	10609	4944
42	128	1764	16384	5376
46	111	2116	12321	5106
49	128	2401	16384	6272
43	104	1849	10816	4472
45	101	2025	10201	4545
49	111	2401	12321	5439
42	125	1764	15625	5250
40	113	1600	12769	4520
45	115	2025	13225	5175
48	100	2304	10000	4800
41	124	1681	15376	5084
43	101	1849	10201	4343
40	102	1600	10404	4080
47	129	2209	16641	6063
48	112	2304	12544	5376
48	128	2304	16384	6144
46	123	2116	15129	5658
905	2294	41133	265310	103833

Applicando la seconda formula alternativa, risulterà:

$$\frac{20 \cdot 103833 - 905 \cdot 2294}{\sqrt{\left(20 \cdot 41133 - 905^2\right)\left(20 \cdot 265310 - 2294^2\right)}} =$$

$$\frac{2076660 - 2076070}{\sqrt{(822660 - 819025)(5306200 - 5262436)}} =$$

$$\frac{590}{\sqrt{3635 \cdot 43764}} = \frac{590}{\sqrt{159082140}} = \frac{590}{12612.77685524} = 0.0467$$

che è un valore troppo basso per essere significativo. Quindi possiamo tranquillamente considerarlo come proveniente da una popolazione che ha correlazione nulla.

Se però volessimo fare una verifica, dovremmo cercare sulle tavole il valore critico di r per 20-2=18 gradi di libertà. Il valore dipenderà anche dal livello α che decidiamo di adottare e dal tipo di ipotesi alternativa: bidirezionale o mono-direzionale. Non avendo pre-conoscenze che ci portino ad esprimere un'ipotesi mono-direzionale, utilizziamo una un'ipotesi alternativa del tipo:

H_1: $\rho \neq 0$

Per 18 gradi di libertà, i valori critici di r sono:

α	.05	.01
r_c	.444	.561

Tutti i valori critici sono decisamente superiori al valore di r da noi calcolato che quindi non è significativo.

Attenzione: la correlazione si calcola sempre sui dati grezzi e mai usando la distribuzione di frequenza.

In effetti, se usassimo la tabella delle frequenze, potremmo imbatterci in alcuni problemi:

+ se le due variabili non hanno la stesso numero di categorie, come si può procedere al calcolo? Ovviamente è impossibile!

+ dal momento che le categorie in una tabella delle frequenze, sono state poste in ordine (crescente o decrescente), la correlazione sarebbe necessariamente sempre positiva.

4.6 La correlazione di Spearman

Vi sarete accorti che, negli esempi, abbiamo usato variabili che sono misurate a livello di scale intervallo o a rapporto. In effetti, la correlazione lineare di Pearson si applica su variabili quantitative. La stessa considerazione poteva scaturire considerando che, in una delle formule, abbiamo usato i punti z e la deviazione standard, statistiche che hanno significato se calcolate a livello di scale intervallo o a rapporto.

Esistono molte altre formule che stimano l'associazione di due variabili su scale diverse da quelle quantitative. Ad es. la correlazione a ranghi di Spearman (chiamata anche rho e indicata con r_s) utilizza dati misurati a livello di scala ordinale.

Poiché la scala ordinale non permette calcoli, dobbiamo prima di tutto trasformare la variabile in qualcosa di metrico e di lineare. La trasformazione implicata dalla rho di Spearman è l'ordinamento a ranghi (in inglese, *ranking*). I valori della variabile vengono dapprima ordinati fra di loro in modo crescente, quindi si procede ad associare a ciascun valore il rango che gli compete. Al valore più basso il rango 1, a quello immediatamente successivo il rango 2 e così via. Per mantenere la stessa struttura, a valori uguali dovremo associare ranghi uguali e quindi utilizzeremo la media dei ranghi. Facciamo un esempio: consideriamo le seguenti due variabili (x e y), di tipo ordinale.

X	Y
A	3
B	3
A	1

D	2
C	3
B	2

Per ciascuna, dobbiamo, per prima cosa, riordinare i valori:

A	A	B	B	C	D
1	2	2	3	3	3

e quindi assegnare i ranghi:

	A	*A*	*B*	*B*	*C*	*D*
Pos.	*1*	*2*	*3*	*4*	*5*	*6*
	$\frac{1+2}{2}$	$\frac{1+2}{2}$	$\frac{3+4}{2}$	$\frac{3+4}{2}$	*5*	*6*
rango	*1,5*	*1,5*	*3,5*	*3,5*	*5*	*6*
	1	*2*	*2*	*3*	*3*	*3*
Pos.	*1*	*2*	*3*	*4*	*5*	*6*
		$\frac{2+3}{2}$	$\frac{2+3}{2}$	$\frac{4+5+6}{3}$	$\frac{4+5+6}{3}$	$\frac{4+5+6}{3}$
rango	*1*	*2,5*	*2,5*	*5*	*5*	*5*

Infine usare i singoli ranghi al posto dei valori.

La trasformazione in ranghi ha prodotto una nuova variabile che "quantifica" la posizione dei singoli valori. A questo punto, usando i ranghi possiamo applicare la formula di Spearman per il calcolo dell'associazione, che fa uso delle differenze fra i ranghi:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

dove *d* è la differenza fra i ranghi e *n* è la numerosità.

X	*X rango*	*Y*	*Y rango*	*d*	d^2
A	*1.5*	*3*	*5*	*-3,50*	*12,25*
B	*3.5*	*3*	*5*	*-1,50*	*2,25*
A	*1.5*	*1*	*1*	*0,50*	*0,25*
D	*5*	*2*	*2.5*	*2,50*	*6,25*
C	*6*	*3*	*5*	*1,00*	*1,00*
B	*3.5*	*2*	*2.5*	*1,00*	*1,00*

$$r_s = 1 - \frac{6 \cdot 23}{6(6^2-1)} = 1 - \frac{138}{6 \cdot 35} = 1 - \frac{138}{210} = 1 - 0{,}657143 = 0{,}342857$$

Anche la correlazione di Spearman produce valori che oscillano fra –1 e +1 e anche per questa correlazione esistono dei test statistici per verificare se la correlazione calcolata è stata casualmente estratta da una popolazione con correlazione nulla. Anche le ipotesi relative all'inferenza sono analoghe:

H_0: $\rho = 0$

H_1: $\rho \neq 0$

I test statistici differiscono secondo la numerosità. Per n<30, è possibile calcolare direttamente la probabilità associati ai valori r_s, per valori di n>30, la distribuzione di r_s viene trasformata tramite la formula

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

che si approssima alla distribuzione t di Student con n-2 gradi di libertà.

La correlazione di Spearman, può essere usata con variabili a intervallo (previa trasformazione in ranghi) quando la numerosità del campione è inferiore a 30.

4.6.1 Tavola dei valori critici di rho

rho= coefficiente di correlazione a ranghi di Spearman

g.l = n-2

Mono-	.05	.025	.01	.005
Bi-	.10	.05	.02	.01
4	1.000			
5	.900	1.000	1.000	
6	.829	.886	.943	1.000
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.683	.783	.833
10	.564	.648	.746	.794
12	.506	.591	.712	.777
14	.456	.544	.645	.715
16	.425	.506	.601	.665
18	.399	.475	.564	.625
20	.377	.450	.534	.591
22	.359	.428	.508	.562
24	.343	.409	.485	.537
26	.329	.392	.465	.515
28	.317	.377	.448	.496
30	.306	.364	.432	.478

4.7 Altri tipi di correlazione

Solo a livello informativo, diremo che esistono svariate altre formule di calcolo per stimare il grado di associazione fra due variabili. Alcune sono di tipo lineare, mentre altre presuppongono relazioni di tipo non lineare.

la maggior parte di queste formule sono state costruite in modo da oscillare fra –1 e +1, per poter rendere comprensibili e confrontabili i risultati ottenuti.

5 Inferenza sulla media [0.1]

5.1 Distribuzione campionaria delle medie

Consideriamo una qualunque variabile *x* presente in una popolazione di 25 casi statistici. Assegniamo a questa variabile *x* dei valori qualsiasi:

39	98	50	75	86
12	44	73	41	13
70	67	33	27	66
98	11	41	78	99
29	41	69	30	83

La media di questi 25 valori è 54,92. Estraiamo casualmente 1 campione di 10 valori e calcoliamo la sua media, estraiamo un altro campione di 10 valori e calcoliamo la media anche di questo. Ripetiamo questo procedimento per tutti i possibili campioni estraibili. Quante medie avremmo?

Si tratta di calcolare la combinazione di 25 elementi presi 10 a 10 senza ripetizione e quindi:

$$C(25,10,s) = \frac{25!}{10!\cdot(25-10)!} = \frac{25!}{10!\cdot 15!} = 3268760$$

Succede quindi che ci ritroveremmo con più di 3 milioni di medie. Possiamo considerare queste medie come una distribuzione di dati, in cui i valori sono rappresentati da ciascuna delle medie. Questa distribuzione è chiamata *distribuzione campionaria delle medie* (perché è formata da medie calcolate sui campioni estratti casualmente) e ha alcune caratteristiche particolari:

- tende a distribuirsi secondo la curva della normale;
- la sua media tende ad essere uguale alla media della popolazione ($\mu_{\bar{x}} = \mu$);
- la sua deviazione standard tende a diminuire all'aumentare della numerosità dei singoli campioni ($\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$).

Se la rappresentiamo graficamente otteniamo qualcosa di analogo a:

[grafico della distribuzione campionaria]

Se si distribuisce normalmente, media, mediana e moda coincideranno.

5.2 Questo gruppo appartiene a questa popolazione?

5.3 Distribuzione campionaria delle differenze delle medie

5.4 Questi due gruppi appartengono alla stessa popolazione?

6 Appendici [0.3]

6.1 Correlazione

6.1.1 Dimostrazione 1

$$r = \frac{\sum (x-\overline{X})(y-\overline{Y})}{Ns_x s_y} = \frac{\sum z_x z_y}{N}$$

Partiamo dalla formula di arrivo e ricordiamo che $z = \frac{x-\overline{X}}{s}$, possiamo quindi riscrivere la formula, come:

$$\frac{\sum z_x z_y}{N} = \frac{\sum \left(\frac{x-\overline{X}}{s_x}\right)\left(\frac{y-\overline{Y}}{s_y}\right)}{N} = \frac{1}{N}\sum\left(\frac{x-\overline{X}}{s_x}\frac{y-\overline{Y}}{s_y}\right)$$

e poiché le due deviazioni standard sono delle costanti, per una delle proprietà della sommatoria possiamo riscrivere il tutto come:

$$\frac{1}{Ns_x s_y}\sum (x-\overline{X})(y-\overline{Y}) = \frac{\sum (x-\overline{X})(y-\overline{Y})}{Ns_x s_y}$$

6.1.2 Dimostrazione 2

Partiamo sempre da questa formula

$$r = \frac{\sum (x-\overline{X})(y-\overline{Y})}{Ns_x s_y} = \frac{\sum z_x z_y}{N}$$

e riscriviamo le deviazioni standard usando la formula alternativa $s = \sqrt{\frac{\sum x^2}{N} - \overline{X}^2}$:

$$r = \frac{\sum (x-\overline{X})(y-\overline{Y})}{Ns_x s_y} = \frac{\sum (x-\overline{X})(y-\overline{Y})}{N\sqrt{\frac{\sum x^2}{N} - \overline{X}^2}\sqrt{\frac{\sum y^2}{N} - \overline{Y}^2}}$$

Osserviamo poi che il denominatore può essere sviluppato come:

$$\sum\left[(x-\overline{X})(y-\overline{Y})\right] = \sum\left(xy - \overline{X}y - x\overline{Y} + \overline{X}\overline{Y}\right)$$

e per una delle proprietà della sommatoria, diventare:

$$\sum xy - \overline{X}\sum y - \overline{Y}\sum x + N\overline{X}\overline{Y}$$

se esprimiamo anche le medie tramite i dati grezzi, otteniamo

$$\sum xy - \frac{\sum x}{N}\sum y - \frac{\sum y}{N}\sum x + N\frac{\sum x}{N}\frac{\sum y}{N}$$

Nell'ultimo membro, la N al numeratore si annulla con una delle N al denominatore e diventa esattamente la versione positiva del secondo e terzo membro, elidendosi ancora. Risulta perciò:

$$\sum\left[(x-\overline{X})(y-\overline{Y})\right] = \sum xy - \frac{\sum x \sum y}{N} = N\sum xy - \sum x \sum y$$

Torniamo al denominatore della formula principale ed esprimiamo anche qui la media in termini di dati grezzi

$$N\sqrt{\frac{\sum x^2}{N}-\overline{X}^2}\sqrt{\frac{\sum y^2}{N}-\overline{Y}^2}=N\sqrt{\frac{\sum x^2}{N}-\left(\frac{\sum x}{N}\right)^2}\sqrt{\frac{\sum y^2}{N}-\left(\frac{\sum y}{N}\right)^2}$$

all'interno di ogni radice, possiamo esprimere la media anche come

$$\left(\frac{\sum x}{N}\right)^2=\frac{\left(\sum x\right)^2}{N^2}$$

e se raccogliamo $\frac{1}{N}$ all'interno di ogni radice, otteniamo:

$$N\sqrt{\frac{1}{N}\left(\sum x^2-\frac{(\sum x)^2}{N}\right)}\sqrt{\frac{1}{N}\left(\sum y^2-\frac{(\sum y)^2}{N}\right)}$$

che moltiplicati fra loro e portati fuori dalla radice si elidono con N

$$N\sqrt{\frac{1}{N^2}\left(\sum x^2-\frac{(\sum x)^2}{N}\right)\left(\sum y^2-\frac{(\sum y)^2}{N}\right)}=\sqrt{\left(\sum x^2-\frac{(\sum x)^2}{N}\right)\left(\sum y^2-\frac{(\sum y)^2}{N}\right)}$$

Se ora uniamo le due parti, otteniamo:

$$\frac{\sum xy-\frac{\sum x\sum y}{N}}{\sqrt{\left(\sum x^2-\frac{(\sum x)^2}{N}\right)\left(\sum y^2-\frac{(\sum y)^2}{N}\right)}}$$

che è esattamente una delle due alternative.

7 Bibliografia [0.4]

Areni, A., Ercolani, A.P. & Scalisi, T.G. (1994*). Introduzione all'uso della statistica in psicologia*. Milano: LED.

Areni, A. & Scalisi, T.G. (1985). *Esercizi di statistica per la ricerca psicologica : problemi ed esercizi svolti e commentati*. Milano: Masson.

Arslan, C. (1993). *Statistica psicologica*. Milano: Guerini.

Bonanno, I. (1990). *Manuale di statistica psicometrica. Breve esposizione teorica ed esemplificazioni*. Roma: Edizioni Kappa.

Botella, J., Léon, O. G. & San Martín, R. (1993). *Análisis de datos en psicología I*. Madrid: Pirámide.

Burigana, L. & Lucca, A. (1975). *Fondamenti della misurazione*. Padova: Cleup.

Cohen, L., & Holliday, M. (1996). *Practical statistics for students. An introductory text*. London: Paul Chapman.

Cottini, L. (1999). *La statistica nella ricerca psicologica ed educativa. Esercitazioni di base*. Firenze: Giunti.

Cristante, F., Lis, A., & Sambin, M. (1977). *Probabilità e inferenza nelle scienze psicologiche*. Padova: Cleup.

Cristante, F., Lis, A., & Sambin, M. (1982). *Statistica per psicologi*. Firenze: Giunti.

Cristante, F., Lis, A., & Sambin, M. (1992). *Fondamenti teorici dei metodi statistici in psicologia*. Padova: Upsel-Domeneghini.

Cristante, F., Lucca, A., & Sambin, M. (1974). *Complementi e problemi di statistica psicometrica*. Padova: Cleup.

Dazzi, C., & Pedrabissi, L. (1999). *Fondamenti ed esercitazioni di statistica applicata ai test*. Bologna, Pàtron.

Enzensberger, H.M. (1997). *Der Zahlenteufel*. Wien: C. Hanser. Trad. it. *Il mago dei numeri*. Milano: Mondadori, 2000

Ercolani, A.P., Areni, A. & Cinanni, V. (1999). *Problemi risolti di statistica applicata alla psicologia*. Milano: LED.

Flores D'Arcais, G.B. (s.d.). *Metodi statistici per la ricerca psicologica*. Firenze: Giunti-Barbèra.

Freedman, D., Pisani, R. & Purves, R. (1998). *Statistics*. Norton, 3rd ed. Trad. it. *Statistica*. Milano: McGraw-Hill.

Greene, J. & D'Oliveira, M. (1999). *Learning to use statistical tests in psychology*. ???: ???. Trad. It. *Statistica per psicologi : come scegliere il test adeguato*. Milano: McGraw-Hill Libri Italia.

Lombardo, E. (1993). *I dati statistici in pedagogia : esplorazione e analisi*. Firenze: La Nuova Italia.

Luccio, R. (1996). *Tecniche di ricerca e analisi dei dati in psicologia*. Bologna: Il Mulino.

McQueen, R.A. & Knussen, C. (1999). *Research methods in psychology : a pratical introduction*. London: Prentice Hall Europe.

Minium, E. W., Clarke, R. C., & Coladarci, T. (1998). *Elements of statistical reasoning*. New York: Wiley & Sons.

Minium, E. W., King, B. M., & Bear, G. (1993). *Statistical reasoning in psychology and education* (3. ed.). New York: Wiley & Sons.

Odifreddi, P. (1999). *Il Vangelo secondo la Scienza. Le religioni alla prova del nove*. Torino: Einaudi.

Runyon, R.P. & Haber, A. (1976). *Fundamentals of behavioral statistics*. Reading: Addison-Wesley, 3rd ed. Trad. it. *Fondamenti di statistica*. Amsterdam: Inter European Editions.

Siegel, S., & Castellan, N. J. Jr. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill. Trad. it. *Statistica non parametrica* (2.ed.). Milano: McGraw-Hill Italia.

Tomat, L., Nicotra, E., & Pedon, A. (1996). *Complementi ed esercizi di statistica per psicologi*. Saonara, PD: Logos Edizioni.

Vidotto, G., Xausa, E., & Pedon, A. (1996). *Statistica per psicologi*. Bologna: Il Mulino.

Visauta Vinacua, B. & Batallé Descals, P. (1991). *Métodos estadísticos aplicados. Tomo I: Estadística descriptiva*. Barcelona: PPU.

Vogt, W. P. (1993). *Dictionary of statistics and methodology*. Newbury Park, CA: Sage. Trad. it. *Dizionario di tecniche e metodologia per la ricerca psicologica* (ed. italiana ampliata a cura di Sandro Nicole). Roma: Edizioni Kappa.

Zuliani, A. (1976). *Statistica per la ricerca educativa*. Torino: SEI.