Christina Bachmann
Riccardo Luccio
Emilia Salvadori

La verifica della significatività dell'ipotesi nulla in Psicologia

Firenze University Press

Monografie
Biomedica
2

Monografie
Biomedica

1. *Il pupazzo di garza*, a cura di Massimo Papini e Debora Tringali, 2004

Christina Bachmann
Riccardo Luccio
Emilia Salvadori

La verifica della significatività dell'ipotesi nulla in Psicologia

Firenze University Press
2005

La verifica della significatività dell'ipotesi nulla in psicologia / Christina Bachmann, Riccardo Luccio, Emilia Salvadori. – Firenze : Firenze university press, 2005.
(Monografie. Biomedica, 2)
http://digital.casalini.it/8884532264
Stampa a richiesta disponibile su http://epress.unifi.it

ISBN 88-8453-226-4 (online)
ISBN 88-8453-227-2 (print)
150.724 (ed. 20)
Psicologia-Ricerche-Metodo

Avvertenza

Un libro, anche scritto a più mani, è un'opera unitaria, in cui gli autori discutono tra di loro quanto stanno singolarmente scrivendo, trovano delle sintesi quando le loro opinioni divergono, anche solo per stile, e si donano amorevolmente, al termine degli scambi di idee, balsami e bende. Detto questo, ognuno poi ha le proprie idiosincrasie, le proprie più specifiche competenze, le proprie *petites madeleines* – avete notato quanto diversamente profuma il mio passato dal tuo, dal suo, dal loro? E il lavoro si suddivide, anche solo per motivi pratici, in tre. Così di Christina Bachmann sono il secondo e il quarto capitolo, di Emilia Salvadori il terzo e il quinto capitolo, di Riccardo Luccio il primo e il sesto capitolo, e le brevi conclusioni.

Ch. B., R. L. e E. S.

In copertina:
Fortunato Depero, *La casa del mago* (1920), Torino, Collezione privata.

Editing di Baldo Conti e Leonardo Raveggi
Impaginazione di Fulvio Guatelli

Università degli Studi di Firenze
Firenze University Press
Borgo Albizi, 28, 50122 Firenze, Italy
http://epress.unifi.it/

Printed in Italy

Indice

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*,
ISBN 88-8453-226-4,

Capitolo 1
La verifica della significatività dell'ipotesi nulla

1.1. Un ibrido non felice

Nell'analisi dei dati delle ricerche psicologiche il paradigma prevalente è la cosiddetta Verifica della Significatività dell'Ipotesi Nulla (VeSN — dall'inglese *NHST, Null Hypothesis Significance Testing*). La VeSN è il prodotto della fusione tra due approcci: uno di Sir R.A. Fisher (1925), il *p-value approach* (PVA); e uno di J. Neyman e E.S. Pearson (1933), the *fixed alpha approach* (FAA).

Di fatto, i due approcci hanno dei tratti in comune, al di là di differenze terminologiche e di quelle che sono, come vedremo, delle differenze sostanziali. Così entrambi condividono l'utilizzo dell'ipotesi nulla H_0 (peraltro non chiamata così da Fisher). Ciò che inoltre li accomuna è l'utilizzo di un valore critico di probabilità p (*p-value*) e di un livello critico α per determinare la probabilità del verificarsi di eventi dovuti al caso o ad errori di campionamento.

Di contro, gli stessi autori sopracitati non amavano particolarmente né l'altrui approccio, né l'ibrido risultante. Così, secondo Fisher (1955), il FAA è un approccio "russo" (inteso come "sovietico"), interessato soltanto all'efficienza, come nei piani economici quinquennali; e ognuno sa a quali risultati brillanti questi piani poi di fatto portavano. Ma secondo Jerzy Neyman (citato da Stegmüller, 1973), il PVA era "peggio che inutile".

È importante sottolineare che non sono molti i domini scientifici in cui viene applicata la VeSN, questo curioso prodotto dell'ingegnosità dell'uomo. È il caso ovviamente della psicologia, nonché della medicina (clinica e epidemiologica), della sociologia, dell'agricoltura, ma la maggior parte delle scienze (fisica, biologia, chimica, astronomia, e così via), peraltro, analizza i propri dati in modo alquanto differente. Esse preferiscono verificare dei *modelli* (vedi Cap. 6).

Nella maggior parte dei manuali di statistica applicata alla psicologia, la VeSN è usualmente presentata come *il* modo di analizzare i dati sperimentali. Si afferma che i test di significatività mirano ad ottenere informazioni riguardanti una certa caratteristica della popolazione che non è direttamente osservabile. Ciò che viene osservato sono le caratteristiche del campione, le quali permettono di fare inferenze sulla popolazione. Raramente viene detto esplicitamente che la verifica delle ipotesi è solo uno dei metodi che il ricercatore ha a disposizione per fare inferenze.

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*, ISBN 88-8453-226-4,

Secondo l'approccio della VeSN (per una recente trattazione italiana del problema, vedi Pisati, 2002), il ricercatore formula due ipotesi mutuamente escludentisi: l'ipotesi nulla H_0 implica che il trattamento non abbia avuto effetto, mentre con l'ipotesi alternativa, detta abitualmente *sostantiva*, H_1, si ipotizza che ci sia stato un effetto. L'ipotesi nulla è chiamata anche ipotesi della non-relazione e non-differenza (Bakan, 1966; Cohen, 1988; Hinkle, Wiersma e Jurs, 1994); l'ipotesi sostantiva, detta anche *sperimentale*, indica che c'è una differenza tra le medie delle popolazioni dalle quali i campioni sono stati estratti, senza specificarne il verso (ipotesi bidirezionale), o specificandone la direzione (ipotesi monodirezionale). (Ciò vale, evidentemente, nel caso di disegni di ricerca che prevedano il confronto tra medie. Nel caso si abbiano disegni di tipo diverso — ad esempio correlazionali — l'ipotesi sarà differente, ma il senso del discorso non muta).

La logica che sottende la decisione statistica è di tipo falsificazionista: si parte dal presupposto che l'ipotesi nulla sia vera e si cerca di falsificarla. Solo se la probabilità *p* associata ai dati, ammesso che H_0 sia vera, è troppo bassa, si rifiuta H_0 in favore di H_1. Il limite entro il quale decidere se accettare o rifiutare H_0 viene fissato arbitrariamente, e si tratta del livello di probabilità α, detto anche livello di significatività critico, che rappresenta la probabilità teorica di rifiutare H_0 quando è vera. Se α viene fissato a 0,05, ciò significa che si ha la probabilità di rifiutare l'ipotesi nulla quando è vera in 5 casi su 100.

Calcolata la probabilità *p* associata ai dati osservati, ammesso che H_0 sia vera, con la statistica più appropriata, questa si confronta con α: se *p* è minore o uguale a α si rifiuta H_0, se *p* è maggiore di α si accetta H_0. Si può incorrere in due tipi di errore. Se si rifiuta H_0 quando è vera si commette un errore di I tipo (*falso positivo, falso allarme*), cioè si afferma erroneamente che il trattamento ha avuto effetto quando invece non lo ha avuto. Se invece si accetta H_0 quando è falsa si commette un errore di II tipo *(falso negativo,* detto anche *omissione*, con probabilità β), cioè si afferma che non c'è stato un effetto dovuto al trattamento mentre in realtà c'è stato.

Più è basso il livello di α e più è basso il rischio di commettere l'errore di I tipo, ma si badi che il livello di α non è indipendente da β, e cioè dalla probabilità di commettere un errore di II tipo; il rischio quindi è di sacrificare quella che è definita la *potenza* del test (e cioè, $1 - \beta$; vedi il Cap. 5). Ma, come vedremo meglio, ben pochi ricercatori si preoccupano di commettere questo tipo di errore, e di conseguenza nel tempo si è quasi smesso di prestare attenzione alla potenza statistica.

A chiarimento di quanto sopra, nei corsi di statistica per psicologi gli studenti vengono di solito posti di fronte, con piccole varianti, a questo schema (la cui efficacia didattica è peraltro indiscutibile).

Lo schema permette di visualizzare il processo decisionale. Si afferma che solo Dio sa se in natura sia vera H_0 o H_1. (Nell'Est europeo, se del caso, le eredità del *Diamat*, il materialismo dialettico, impongono di sostituire a Dio la Natura; cfr. Milosevic, 1995). Lo statistico allora gioca contro Dio (la Natura), e cerca di indovinare, sulla base della probabilità che assegna ai dati osservati, posto che H_0 sia vera, cosa c'è nella mente di Dio.

		Dio: H_0	Dio: H_1
Statistico	H_0	Rifiuto corretto	Falso negativo Errore di II tipo β
	H_1	Falso positivo Errore di I tipo α	Hit Potenza $(1 - \beta)$

Dato lo schema, allo studente viene allora detto che un processo inferenziale "genuinamente fisheriano" richiede i seguenti passaggi:

1. assegnare un valore di probabilità ad α (usualmente 0,05);
2. assumere che H_0 sia vera;
3. determinare la probabilità associata ai dati osservati ammesso che H_0 sia vera [cioè, p (dati | H_0)];
4. se $p > \alpha$, accettare H_0 (e il lavoro finirà con tutta probabilità nel cestino della carta straccia);
5. altrimenti, rifiutare H_0 e accettare H_1.

Lo studente non viene incoraggiato ad approfondire i problemi relativi ad errori di II tipo, potenza del test, e così via.

Ora, se il procedimento indicato può essere considerato effettivamente fisheriano, tutto l'apparato ha ben pochi rapporti con Fisher, ed è viceversa largamente dovuto a Neyman e Pearson (1933), i quali peraltro suggerivano una strategia decisionale sostanzialmente differente. È allora opportuno cercare di fare chiarezza, partendo dalla storia di questi straordinari personaggi che abbiamo evocato.

1.2. L'approccio fisheriano

1.2.1. Chi era Fisher

Partiamo proprio da Ronald Aylmer Fisher (1890-1962), non a torto definito frequentemente il "principe degli statistici" (un'ottima sua biografia è stata scritta dalla figlia Joan Box Fisher, 1978).

Fisher era nato da una famiglia benestante presso Londra, e fin da ragazzo dimostrò una straordinaria attitudine per la matematica. Peraltro, la morte della madre e la rovina economica del padre resero molto difficili gli anni della sua adolescenza. Egli studiò a Cambridge, al Gonville and Caius College, soprattutto matematica ed astronomia, ma si interessò anche alla biologia, e si appassionò (purtroppo) ai problemi dell'eugenica, coltivando delle detestabili idee politiche, ben entro i confini del razzismo.

Dopo la laurea, conseguita nel 1912, Fisher svolse diversi lavori da statistico e da insegnante di matematica nelle scuole secondarie, e cominciò, sia pure tra diverse incomprensioni, a farsi conoscere nel campo della statistica accademica, sinché, nel 1919, iniziò la sua carriera universitaria accettando la cattedra di statistica alla Rothamsted Agricultural Experimental Station, un importante istituto sperimentale di agraria, rifiutando nel contempo l'offerta di Karl Pearson, allora "Galton professor" all'University College di Londra (e incontrastato dominatore della statistica inglese, e non solo), di occupare l'importantissima posizione di statistico capo dei Laboratori Galton. (Per un vivace ritratto di Karl Pearson, vedi Williams, Zumbo, Ross e Zimmerman, 2003).

Si trattava chiaramente di una mano tesa da parte di Pearson, che se da un lato indicava una grande stima nei confronti del collega di tanto più giovane, dall'altro voleva essere un superamento delle polemiche che avevano diviso i due negli anni precedenti. Ma Fisher, dimostrando che l'asprezza del carattere sarebbe sempre stata un suo tratto distintivo, rifiutò l'offerta, portando a livello di totale rottura i rapporti con il collega più anziano.

La controversia tra i due, che ci interessa direttamente, dato che l'inimicizia di Karl Pearson per Fisher verrà ereditata dal figlio Egon, autore con Neyman dell'approccio alternativo al problema che qui trattiamo, e non sarà estranea allo sviluppo di questi due modi diversi di concepire l'analisi dei dati, nasce infatti nel 1917, quando Pearson criticò in un suo lavoro il concetto di "verosimiglianza" (*likelihood*), che era stato elaborato da Fisher per la prima volta nel 1915, e che dal 1921 in avanti sarebbe stato da quest'ultimo considerato uno dei suoi più importanti contributi alla statistica. Fisher respinse aspramente le critiche di Pearson, e la sua irritazione si accrebbe quando seppe che nel 1918 Pearson era stato uno dei *referees* che avevano fatto rifiutare un suo lavoro inviato alla Royal Society (anche se di fatto Pearson, trattandosi di problemi di statistica applicata alla genetica, aveva dato parere favorevole per ciò che riguardava gli aspetti statistici, mentre aveva chiesto un giudizio da parte di persona più esperta per quel che riguardava gli aspetti biologici).

Ma il rifiuto del 1919 creò una situazione di guerra aperta. Mentre Fisher, alla stazione Rothamsted metteva a punto l'analisi della varianza (ANOVA — il suo contributo più noto, se non il più importante, alla statistica del XX secolo), e pubblicava altri studi di fondamentale importanza, tra cui merita di essere segnalato quello sul concetto di informazione, del 1925, Pearson iniziava la sua offensiva aperta contro di lui. In un articolo del 1922 lo accusò di avere usato il χ^2 (una statistica da Pearson stesso creata) in modo erroneo, e approfittò della sua posizione di presidente della Royal Statistical Society e di *editor* di *Biometrika*, la più importante rivista di statistica di allora, per bloccare la pubblicazione di molti articoli di Fisher, che nel 1925 si dimise per protesta dalla Società. Ma Fisher, evidentemente, non rimase con la "penna in

mano", e la polemica tra i due crebbe sino a livelli insostenibili, ed ebbe termine solo con la morte di Karl Pearson, nel 1936.

C'è un'ironia della sorte nel fatto che quando Pearson, nel 1933, andò in pensione, la sua cattedra venisse assegnata proprio a Fisher. Ma poiché il destino deve sempre creare situazioni complesse, e possibilmente sgradevoli, la cattedra fu divisa a metà: la parte di Eugenica fu per Fisher, ma quella di statistica andò al figlio di Karl, Egon Sharpe Pearson, che, come vedremo meglio nel prossimo paragrafo, si era già segnalato per la sua ostilità, largamente ricambiata, nei confronti di Fisher. Di più, anche il Dipartimento venne diviso, e Pearson fu fatto Direttore del nuovo dipartimento di Statistica Applicata.

Gli anni '30 furono quindi anni di conflitto, esacerbato anche dalle inaccettabili concezioni politiche generali che Fisher man mano sviluppava. In particolare, suscitò violente polemiche il suo darwinismo sociale, in base a cui sosteneva ad esempio l'opportunità di sterilizzare i deboli mentali, e la necessità che la società rendesse difficile l'esistenza, con un sistema premiante in termini di reddito e benefici, agli individui sterili e meno sani. Queste idee non erano peraltro nuove per Fisher. Già nel 1924, in uno dei suoi meno degni lavori, aveva sostenuto la necessità di sterilizzare i deboli mentali, tentando di dimostrare la falsità del cosiddetto "principio di Handy-Weinberg" (dal nome dei due studiosi che lo avevano scoperto indipendentemente nel 1908), secondo cui è impossibile con la sterilizzazione eliminare da una popolazione i geni recessivi. È peraltro corretto dire che si trattava di idee che se oggi ci appaiono inaccettabili, all'epoca erano largamente diffuse, almeno tra i genetisti. Peraltro Fisher non ebbe mai paura di esprimere idee politicamente indecenti, fino al suo tentativo di screditare le basi statistiche delle ricerche che cominciavano a dimostrare la relazione causale tra fumo e cancro ai polmoni (1958).

Nel 1943, Fisher decise infine di lasciare Londra per Cambridge, dove fu chiamato come "Balfour Professor" di Genetica, e dove rimase fino al 1959, due anni dopo la pensione, seguitando ad insegnare in attesa che venisse nominato un suo successore. Fu tra l'altro qui nominato Presidente del Gonville and Caius College, dove aveva studiato. Si trasferì poi all'Università di Adelaide, in Australia, dove morì nel 1962.

Tra le numerosissime opere di Fisher, quelle che forse hanno avuto il maggior impatto nel mondo della ricerca sono state *Statistical Methods for Research Workers* (1925), che ebbe un numero infinito di riedizioni, e che ha radicalmente indirizzato in modo nuovo, sino a tutt'oggi, l'analisi dei dati non solo in biologia e in agricoltura, ma anche in psicologia e in sociologia; e *The Design of Experiments* (1935), che è stato la guida della maggior parte degli statistici nell'applicazione dell'ANOVA.

1.2.2. Il p-value approach

L'enunciazione più chiara del *p-value approach* di Fisher si trova proprio nel più volte citato *Statistical Methods for Research Workers*. Qui, nel primo capitolo introduttivo (p. 9), Fisher afferma che un problema della statistica, "su cui fino ad ora sono stati compiuti pochi progressi, è alla base dei test di significatività, con i quali noi possiamo

esaminare se i dati sono o meno in armonia con le ipotesi che vengono avanzate". Se noi traiamo un campione da un universo, le caratteristiche del campione possono farci prevedere quali sono le caratteristiche dell'universo; se quindi un secondo campione viola le nostre aspettative, "possiamo inferire che è stato tratto da una diversa popolazione […] Test critici di questo tipo possono essere chiamati test di significatività, e se tali test sono disponibili possiamo scoprire se il secondo campione è o meno significativamente diverso dal primo" [p. 44].

Poco oltre, compare questo limite di probabilità di 0,05, che trionferà poi nei decenni successivi come signore incontrastato dell'inferenza statistica. Dalle tabelle della curva soggiacente alla distribuzione normale, Fisher mostra che a un valore di unità di deviazione standard di 1,96 corrisponde a 2 code una probabilità di 0,05 (o di uno su venti). "È conveniente assumere questo punto come limite per giudicare se una deviazione va considerata o meno significativa. Le deviazioni che eccedono il doppio della deviazione standard sono così considerate formalmente significative" [p. 48]. Si osservi il contrasto tra il "conveniente" e il "formale".

Nel Cap. 4, dedicato al χ^2, al § 21 viene presentato il problema della verifica dell'indipendenza tra variabili nelle tabelle di contingenza, in termini in cui la sola differenza che possiamo trovare rispetto ai trattamenti di oggi è probabilmente data dalla straordinaria chiarezza e sequenzialità delle argomentazioni di Fisher (che per una curiosa leggenda metropolitana seguita a essere definito "di difficile lettura", se non "oscuro").

Nel precedente § 20, Fisher ha mostrato l'uso del χ^2 per verificare la bontà dell'adattamento di dati osservativi a una funzione. In questo paragrafo, affronta "una classe particolare e importante di casi in cui l'accordo tra aspettativa ed attesa che può essere sottoposta a verifica comprende dei test di *indipendenza*. Se lo stesso gruppo di individui viene classificato in due (o più) modi diversi, per esempio persone inoculate e non inoculate, nonché attaccate e non attaccate da una malattia, può essere necessario sapere se le due classificazioni sono indipendenti" (p. 81, corsivo di Fisher). Fisher presenta poi degli esempi numerici (che oggi eviteremmo con cura: l'N complessivo del suo primo esempio è di 18.483 individui, e ben sappiamo quanto il χ^2 sia sensibile alla grandezza del campione!), in cui mostra come calcolare le frequenze attese e come confrontarle con quelle osservate, ai fini del calcolo della statistica, nonché le formule abbreviate. Fisher giunge così a un esempio in cui esamina una tabella 2×2, e afferma: "I valori attesi sono calcolati dal totale osservato, e le quattro classi [le quattro frequenze di cella; NdR] devono avere una somma concordante, e se i valori di tre classi sono date in modo arbitrario il quarto valore è perciò determinato, di qui n=3, χ^2=10,87, e la probabilità di eccedere tale valore è compresa tra 0,01 e 0,02*; se assumiamo* p = *0,05 come limite della deviazione significativa, diremo che in questo caso le deviazioni dall'aspettativa sono chiaramente significative*." [pp. 82-83, corsivo nostro].

Nell'accennare in altre parti dei *Methods* al problema della significatività, Fisher non si sposta da questa posizione. La probabilità va posta a 0,05 perché è conveniente farlo, ed è un livello sufficiente per prendere una decisione ragionevole.

Si osservi, peraltro, che nel corso della sua attività scientifica Fisher ebbe modo di cambiare opinione almeno su un punto non secondario del suo approccio. Infatti, se nel 1925 appare chiaro che la probabilità va posta *a priori* a 0,05, negli anni '50

(Fisher, 1955, 1956) egli afferma esplicitamente che l'esatto livello di significatività deve essere calcolato *dopo* che si è effettuato il test. Il livello di significatività è quindi una proprietà dei dati. Anche in questo senso il *p-value approach* si differenzia dall'approccio di Neyman e Pearson, che vedremo ora.

1.3. Il *fixed-alpha approach*

1.3.1. Egon Pearson e la sua controversia con Fisher

Come abbiamo avuto modo di accennare, la controversia tra Egon Sharpe Pearson e Ronald Aylmer Fisher esplose ben prima che i due entrassero in concorrenza diretta per la successione alla cattedra di Karl Pearson. Vi fu da un lato il desiderio del primo di difendere il padre dai furibondi attacchi che Fisher seguitava a rivolgergli, dall'altro l'insofferenza del secondo per qualsivoglia critica gli venisse rivolta. Lo scontro iniziò negli anni '20, con due recensioni che Pearson (1927, 1929) scrisse alle prime due edizioni del volume di Fisher (1925) sugli *Statistical Methods for Research Workers.*

L'immagine che ci viene solitamente tramandata di Egon Pearson è quella di un personaggio abbastanza grigio, schiacciato da un lato affettivamente da un padre straripante, che lo iperprotesse e gli prefigurò una carriera straordinaria, con pochi meriti da parte del giovane Egon; e dall'altro schiacciato intellettualmente da quell'altra grande figura che fu Jerzy Neyman, certamente uno dei più grandi matematici applicati del '900, a cui si devono in larghissima misura le basi teoriche dell'approccio che va sotto i loro due nomi. In realtà, Egon Pearson fu una personalità abbastanza forte, certo non prepotente quanto il padre, ma perfettamente in grado di difendersi ed attaccare ove necessario. E certamente i suoi contributi originali alla statistica, seppure non geniali quanto quelli di Neyman, furono tutt'altro che disprezzabili. Ma di sicuro Pearson fu un grande organizzatore del lavoro scientifico, e forse ancor più di Fisher contribuì a collocare la scuola statistica britannica in quella posizione di eccellenza che ancor oggi occupa.

È comunque indubbio che il padre lo agevolò non poco. Iscrittosi all'Università, al Trinity College di Cambridge, nel 1914, evitò il servizio militare per un soffio al cuore (era appena scoppiata la I Guerra Mondiale), ma volle rendersi utile al Paese servendo come civile all'Ammiragliato, per cui interruppe gli studi. Riuscì però a conseguire la laurea nel 1919 in una speciale sessione per militari, a cui non avrebbe avuto molto diritto di partecipare, ed entrò come ricercatore in astrofisica a Cambridge, cominciando ad interessarsi alla teoria degli errori.

Tanto bastò perché nel 1921 Karl Pearson riuscisse a fare entrare il figlio come *lecturer* nel suo Dipartimento di Statistica Applicata all'University College di Londra. In realtà, la storia dice che di suo in quel periodo di lezioni non ne fece, facendo invece le lezioni al posto del padre nel corso di questi. E così il padre lo fece diventare nel 1924 vice direttore di *Biometrika*, la rivista più importante di statistica dell'epoca, che dirigeva, e che era nelle mani di Karl Pearson uno straordinario strumento di potere.

La svolta ci fu nel 1925, quando Jerzy Neyman giunse all'University College grazie a una borsa Rockfeller biennale. Di questo arrivo di Neyman parleremo più distesamente in seguito, ma è certo che tra i due nacque una profonda amicizia, cementata da una grande affinità intellettuale. Soprattutto, però, il lavoro con Neyman diede una nuova sicurezza al giovane Pearson, anche se di certo si guardò bene dal tagliare il cordone ombelicale con il padre.

Nel 1926 Pearson pensò bene di entrare direttamente nell'agone contro Fisher, e lo fece con una perfida, brevissima recensione al saggio sugli *Statistical Methods* del 1925 di questi, che appena pubblicato ebbe da subito una risonanza enorme. La recensione di Pearson è apparentemente rispettosa. Il compito che Fisher si è assunto è "molto difficile", la sua stesura del testo è "diligente", ma è "un po' dubbio" che il lettore possa poi applicare i suoi risultati a situazioni "un po' diverse" dagli esempi che porta. Soprattutto, poi, è difficile "seguire le sue dimostrazioni basate sul concetto di gradi di libertà […] che appaiono fondarsi su analogie". Di più, un metodo "consolidato da tanto tempo" come quello del rapporto di correlazione, è incomprensibilmente "accennato di sfuggita". Ma "se i vecchi metodi sono trattati solo sommariamente", esempi e nuove tabelle sono di "considerevole interesse".

Ce n'era a sufficienza per fare infuriare Fisher, specie se si tiene conto del fatto che la sua polemica sempre più aspra con Pearson padre aveva avuto al suo centro proprio il problema dei gradi di libertà, specie nel χ^2, e della correlazione. La risposta (Fisher, 1927) non si fece attendere. Manca qui qualsiasi cortesia formale. Fisher non si aspetta che Pearson "sia d'accordo", ma fa presente che già nel 1922 aveva dimostrato tutti i limiti del coefficiente di correlazione, e le tre pagine dedicategli nel libro (errore semmai di "commissione e non di omissione") servivano a evitare agli studiosi perdite di tempo e a cadere nelle incongruenze di cui erano rimasti vittime tanti "illustri biometristi"; e l'allusione a Karl Pearson non poteva essere più trasparente.

Pearson attese l'uscita della seconda edizione (1928) degli *Statistical Methods* per tornare all'attacco, e questa volta lo fece dalle colonne di *Nature*, ben più influenti di quelle di *Science Progress*, su cui era comparsa la prima polemica (Pearson, 1929a). Qui la recensione non era firmata, ma l'autore era riconoscibilissimo. Il problema al centro dell'interesse di Pearson era quello dell'applicabilità di tanti test statistici concepiti per variabili distribuite normalmente quando venivano usati su campioni di cui non era nota l'appartenenza a distribuzioni normali, o che comunque si dipartivano dalla normalità. È questo il problema della robustezza delle statistiche, a cui Pearson si stava comunque appassionando, e sui cui stava stimolando l'interesse dell'amico Jerzy Neyman, rientrato in Polonia, un problema tutt'oggi di grande attualità, ad esempio nell'analisi delle strutture di covarianza.

Questa volta Fisher divenne letteralmente furioso. Invano tentò di far da paciere William Sealy Gossett, questa curiosa figura di "birraio" (diresse per alcuni decenni la famosa Birreria Guinness), chimico con la passione per la statistica, famoso ancora adesso con lo pseudonimo di "Student" — il test t di Student è forse, con il χ^2 la statistica più usata nella ricerca inferenziale. Student aveva buoni rapporti con entrambe le parti in causa (Pearson, 1990, gli avrebbe dedicato una bellissima biografia); vi fu, tra lui e Fisher, un frenetico scambio di corrispondenza, e infine Gossett pensò di

chiudere l'incidente con una lettera a *Nature* (Gossett, 1929), in cui, in modo un po' contorto, sosteneva che (i) il recensore (Pearson) aveva sollevato il fatto che un lettore disattento avrebbe potuto credere che Fisher sostenesse che il suo lavoro si poteva applicare tranquillamente anche a distribuzioni non normali; (ii) che il recensore, usando il verbo "ammettere" (*admit*), anziché ad esempio "sottolineare" (*stress*), fa credere che Fisher sia caduto nell'equivoco; (iii) che comunque egli riteneva che ad esempio la distribuzione di Student potesse essere usata ragionevolmente bene anche per dati non distribuiti normalmente; e infine, (iv) che era comunque opportuno che Fisher stesso potesse indicare, dandone le basi teoriche, quali modifiche dovessero essere fatte alle sue tabelle nel caso di scostamenti dalla normalità.

Ma Fisher non si quietò affatto, e in una sua lettera a *Nature* (Fisher, 1929) se la prese equamente non solo con Pearson, ma anche con Student. In che senso avrebbe mai dovuto dire come correggere le sue statistiche e le sue tabelle per condizioni di non normalità? Non ci si rendeva conto che non ci sarebbe mai stato un limite alle correzioni da apportare? Non si trattava tanto di "stoltificazione" di qualsiasi metodo statistico, ma di "abbandono della teoria degli errori". E, in ogni caso, nella pratica della ricerca erano problemi che non si ponevano.

La palla tornava allora a Pearson, che (1929b) ribadiva, con un esempio tratto da un articolo su *Biometrika* sulla lunghezza di uova di alcune specie di uccelli, che il caso in cui si avevano scostamenti anche sensibili dalla media anche in biologia era tutt'altro che eccezionale, e i metodi di Fisher non potevano darne conto.

Si osservi che in tutto questo l'atteggiamento di Egon Pearson era ben diverso da quello del padre Karl. Come nota Reid (1997), temperamento timido e introverso, sempre tormentato da un acuto sentimento di inferiorità, specie nei confronti del padre, nel corso della controversia, come negli anni a venire, Egon soffriva del tormento di scoprire che il padre poteva avere torto; ma date anche le sue modeste basi matematiche, non tutto capiva di quello che Fisher scriveva, e comunque lo odiava per i continui attacchi al padre, che non cessarono neppure dopo la morte di questi. Eppure sentiva che Fisher spesso aveva ragione.

Si può facilmente immaginare quanto poco sereno dovesse essere il rapporto tra Egon Pearson e Fisher quando nel 1933, come si è avuto modo di accennare, al ritiro di Karl Pearson la sua cattedra venne divisa tra i due. Peraltro, grazie soprattutto alla collaborazione con Neyman, Egon Pearson non era più il figlio di papà che aveva fatto carriera solo grazie agli appoggi paterni, e poteva vantare una serie di contributi scientifici di tutto valore.

Gli anni che vanno dal 1926 al 1933 sono i più produttivi della sua attività scientifica. Sono gli anni della collaborazione con "Student", che tentò invano di rendere più sereni i rapporti con Fisher, ma che certamente insegnò a Egon infinitamente di più di quanto era riuscito a insegnargli il padre. Ma soprattutto sono gli anni dell'intensa corrispondenza con Jerzy Neyman, che si coronerà nel 1933 con il famoso lavoro sul "lemma" (vedi oltre). Neyman tornò in Inghilterra nel 1934, per poi lasciarla definitivamente per Berkeley nel 1938. Curiosamente, al suo ritorno la collaborazione tra i due si affievolì. Tra l'altro, alla sua morte Neyman (1981) ha fatto presente che negli anni della maggiore collaborazione l'iniziativa era quasi sempre stata di Pearson.

Questi si interessò ora soprattutto ad aggiornare le opere del padre, specie le famose *Tables for Statisticians and Biometricians.* Uno sguardo alle sue opere più significative, raccolte nel 1966, ci mostra che, anche se non mancarono dei contributi pregevoli, il suo periodo di grazia si era comunque concluso. Ne rimase l'immagine di un gentiluomo mite e affabile con tutti e da tutti apprezzato (salvo, naturalmente, che dall'infame Fisher), grande organizzatore del lavoro scientifico, propalatore instancabile dei metodi statistici applicati alla ricerca empirica.

1.3.2. Jerzy Neyman

L'altro grande personaggio all'interno della controversia, e a giudizio di chi scrive forse il più grande e sotto ogni aspetto il più rispettabile dei tre, era un matematico moldavo, Jerzy Splawa Neyman (1894-1981; il primo cognome, Splawa, di origine nobiliare, fu da lui abbandonato negli anni '20).

Neyman nacque a Bendary, allora appartenente alla Russia, da una famiglia cattolica polacca, e studiò a Kharkov, dove la madre in difficoltà economiche si era trasferita nel 1906 dopo la morte del marito. All'Università studiò dal 1912 al 1917 fisica e matematica, appassionandosi soprattutto alla teoria della misura di Lebesgue, su cui, ancora studente, scrisse i primi saggi scientifici. Ma la lettura che dovette segnare la sua vita fu *Grammar of Science* di Karl Pearson, segnalatagli da A. Bernstein, suo professore di Probabilità, un'opera che per il suo antidogmatismo lo doveva prima turbare profondamente, poi affascinare. Tra l'altro, questo libro, di immensa diffusione, caratterizzato anche da un'impostazione materialistica, ricevette dei giudizi estremamente lusinghieri da parte di Lenin.

Dopo la laurea rimase all'Università di Kharkov come assistente, e iniziò ad appassionarsi di statistica; ma lo scoppio della guerra e la Rivoluzione russa resero le condizioni di vita particolarmente dure. Lo scoppio della guerra russo-polacca fece quindi precipitare la situazione. Considerato polacco dai russi, fu prima arrestato e quindi preferì trasferirsi con la moglie, sposata due anni prima, nel 1921 in Polonia. Qui con l'aiuto di Sierpinski, dopo qualche lavoro saltuario come statistico, nel 1923 riuscì a entrare all'Università di Varsavia come assistente. Nel 1925, poi, come abbiamo già avuto modo di dire, grazie a una borsa biennale Rockfeller potè raggiungere Londra, e soprattutto poté andare a lavorare con quel Karl Pearson che tanto lo aveva intellettualmente affascinato.

L'impatto, peraltro, non fu dei più felici. Da molti anni Karl Pearson aveva smesso di aggiornarsi scientificamente, troppo preso dalla gestione del suo immenso potere, e Neyman rimase sconvolto dall'ignoranza dei più recenti risultati in campo matematico che regnava all'University College. L'aspetto positivo fu però dato dall'amicizia che poté stabilire con Egon Pearson, premessa di una straordinaria collaborazione tra i due.

In ogni caso, Neyman ritenne più opportuno utilizzare la sua borsa per andare a Parigi, e lì poté incontrare, a un livello di gratificazione molto più elevato, un altro idolo dei suoi anni universitari, Lebesgue, oltre ad altri grandi matematici dell'epoca,

come Borel e Hadamard. Lì fu raggiunto da una lettera di Egon Pearson, che lo invitava a collaborare con lui su problemi di statistica. Pearson lo raggiunse anche a Parigi nel 1927 per un breve periodo, e ancora in Polonia, dove Neyman tornò nel 1928, per brevi scambi diretti di idee, oltre all'intensa corrispondenza. La collaborazione si rivelò straordinariamente proficua. Neyman poteva offrire a Pearson quelle basi matematiche in cui il giovane inglese si mostrava carente, e questi a sua volta viveva in un osservatorio privilegiato, avendo quotidianamente sotto gli occhi quelli che erano i problemi più vivi del dibattito statistico del tempo. In particolare, non era tanto l'influenza del padre (le cui carenze Egon cominciava a scoprire), quanto quella di "Student", e soprattutto quella del nemico Fisher, che Pearson studiava accanitamente tra infinite difficoltà matematiche, che stimolavano i due a quel loro straordinario impegno.

Nel 1933 i due lavori sul "lemma" erano una realtà. Nello stesso anno, diventato direttore del Dipartimento di Statistica Applicata dell'University College, Pearson poteva invitare Neyman a Londra, dove giunse nel 1934, e che abbandonò nel 1938 per l'Università di California a Berkeley. Negli anni di Londra Neyman fornì dei contributi fondamentali sul campionamento e sugli intervalli di fiducia. E soprattutto un lavoro di straordinaria importanza, nel 1937, sulla stima statistica.

Neyman non lasciò più Berkeley, e vi fondò nel 1955 un Dipartimento di statistica. Peraltro, lasciata Londra, i contributi più importanti di Neyman furono destinati alle applicazioni della statistica, dalle elezioni alla medicina alla meteorologia.

1.3.3. Il lemma di Neyman e Pearson

Di fatto, lo schema presentato sopra non è fisheriano, ma deriva dalla FAA di Neyman-Pearson. I concetti di ipotesi alternativa, errore di I e di II tipo, e di potenza statistica non sono frutto della teorizzazione fisheriana, ma sono stati formulati da Neyman & Pearson (1933). Al contrario di Fisher, questi autori vedevano i test di significatività come un metodo per selezionare una ipotesi tra due ipotesi possibili e non come procedura di *testing* di una sola ipotesi. Si osservi che, come già rilevato, nell'ottica fisheriana l'ottenere un p associato ai dati, ammesso che H_0 sia vera, e cioè $p(\text{dati} \mid H_0)$, inferiore al livello di α prefissato, porta a rifiutare l'ipotesi nulla, poiché un valore di probabilità così basso rende implausibile la condizione della sua verità. Ma p non fornisce nessuna indicazione sulla verità di H_1.

Nella FAA, di contro, in primo luogo è fissato *a priori* il valore di α (di qui il nome dell'approccio). Infatti Neyman e Pearson ritengono che il primo problema è quello di evitare gli errori di I tipo, ma prestano altrettanta attenzione agli errori di II tipo. Il loro procedimento prevede che la decisione che viene presa, che non è del tipo "tutto o nulla", sia quella che preserva la massima potenza possibile del test, $1 - \beta$, una volta fissato α. Va peraltro osservato che questa diversa impostazione non riguarda solo aspetti di natura statistico-matematica, ma trova le sue radici in un modo profondamente diverso di concepire il ruolo che questi studiosi attribuivano alla loro disciplina, alla ricerca scientifica, e in generale alla visione del mondo. E in parte certe differenze derivavano anche da problemi più personali.

Vediamo allora di presentare l'approccio di Neyman e Pearson in quella che è la sua forma più nota, e cioè il loro famoso "lemma". Di questo lemma daremo una presentazione per quanto possibile poco tecnica, anche se cercheremo di mantenerci a livello di rigore.

1.3.3.1. La ripartizione dello spazio dei parametri

Il punto di partenza è quello della definizione di una popolazione data da una variabile casuale continua ξ, che si distribuisce sulla base di una funzione di densità di probabilità $z(\xi; \theta)$. I valori che questa variabile assume li indicheremo con x. Si osservi che θ è il parametro, o i parametri, della funzione z e appartiene allo spazio dei parametri Ω, e cioè $\theta \in \Omega$.

Vediamo questo cosa significa, nel caso di due tra le funzioni di densità di probabilità di riscontro più frequente in psicologia, e cioè la funzione normale e la funzione di Poisson. Nel primo caso, vi sono due parametri, che sono rispettivamente la media μ e la deviazione standard σ, e quindi abbiamo:

$$z(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}, \tag{1.1}$$

e quindi la funzione ha due parametri, media μ e deviazione standard σ.

Nel caso della distribuzione di Poisson, abbiamo invece un solo parametro, λ, essendo la funzione

$$z(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}. \tag{1.2}$$

(Con l'aumentare di λ la distribuzione diventa sempre più simmetrica).

Il passo successivo consiste nella possibilità di operare una partizione nello spazio dei parametri Ω. Immaginiamo che tale partizione ci consenta di dividere questo spazio in un sottospazio C e nel sottospazio complementare $A = \Omega - C$. A può essere a sua volta ripartito in altri sottospazi, e chiameremo uno qualsiasi di questi sottospazi C'. Chiamiamo i due sottospazi C e A, perché il primo per motivi che saranno più chiari in seguito rappresenta la regione critica, il secondo la regione di accettazione dell'ipotesi nulla.

Ora, in cosa consiste la decisione statistica? Noi, è noto, non abbiamo a disposizione la popolazione, ma solo dei campioni. Facciamo allora il caso più semplice: abbiamo estratto un campione, che rappresenteremo come un vettore $\mathbf{x} = [X_1, X_2, \ldots, X_n]$, per cui dovrà valere la nostra funzione di densità di probabilità $z(\mathbf{x}; \theta)$. Nota bene: noi ci poniamo il problema se questo campione appartiene a questa popolazione; per esempio, se questo campione, che ha assunto un sonnifero, dorma in modo uguale

o diverso dalla quota della popolazione che il sonnifero non lo ha assunto. Per fare questo, partiamo dall'assunto che la legge *z* valga allo stesso modo e per il campione e per la popolazione, altrimenti il confronto sarebbe insensato. Ma se la legge è la stessa, quello che cambia è il parametro (i parametri). Nell'esempio fatto, il problema che ci poniamo è se chi assume il sonnifero dorme di più di chi non lo assume. Posto che popolazione e campione si distribuiscano entrambi normalmente, andremo allora a vedere il parametro media μ nella popolazione e nel campione.

In altri termini, il valore del parametro sarà lo stesso in tutti i sottospazi di Ω in assenza di trattamento, o nel caso che per ipotesi il trattamento sia inefficace. Quest'ultima ipotesi è quella che abbiamo chiamato ipotesi nulla H_0 nel primo paragrafo del capitolo. Possiamo formalizzare quanto sopra in questi termini: viene stabilita una funzione di ripartizione di Ω $t(\mathbf{x})$ tale per cui, posto che H_0 sia vero, la probabilità condizionale che θ appartenga a *C* è uguale alla probabilità condizionale che appartenga a qualsivoglia altro sottospazio *C'* di Ω, e questa probabilità è appunto l'α prefissato; e cioè:

$$p\left[t(\mathbf{x}) \in C \middle| \mathrm{H}_0\right] = p\left[t(\mathbf{x}) \in C' \middle| \mathrm{H}_0\right] = \alpha. \tag{1.3}$$

Ciò posto, nel caso che sia invece vera l'ipotesi sostantiva H_1, il nostro parametro verrà ad avere un valore diverso nel sottospazio *C* rispetto a quello che ha negli altri sottospazi, e l'applicazione della funzione di ripartizione deve portare a un valore di probabilità condizionale più alto quando siamo in *C* rispetto a quando siamo in un qualsiasi altro sottospazio *C'*:

$$p\left[t(\mathbf{x}) \in C \middle| \mathrm{H}_1\right] > p\left[t(\mathbf{x}) \in C' \middle| \mathrm{H}_1\right]. \tag{1.4}$$

1.3.3.2. La stima dei parametri

Prima di passare al lemma propriamente detto, il nostro problema è però quello di introdurre rapidamente i concetti di stima dei parametri e di massima verosimiglianza. Evidentemente non potremo trattare che molto superficialmente un argomento così complesso, ma quel che ci interessa è chiarire qualche concetto, tra cui quello di massima verosimiglianza, sviluppato da Fisher (1921a), ed estremamente importante per il lemma di Neyman-Pearson.

La *funzione di verosimiglianza* può essere definita come la funzione di densità di probabilità congiunta di *n* variabili estratte casualmente (e quindi indipendenti) dall'universo (e quindi identicamente distribuite), definita nello spazio Ω. Se la funzione di densità di probabilità è $z(\mathbf{x}; \theta)$, la funzione di verosimiglianza, funzione di densità di probabilità congiunta, sarà quindi, date le *n* variabili estratte $(x_1, x_2, \ldots, x_n)$,

$$L(x;\theta) = z(x_1;\theta)z(x_2;\theta)...z(x_n;\theta) = \prod_{i=1}^{n} z(x_i;\theta). \tag{1.5}$$

Si osservi che di solito si utilizza, per motivi soprattutto di semplicità di calcolo, non la funzione L di verosimiglianza, ma la sua trasformazione logaritmica G, e cioè:

$$G(x;\theta) = \ln L(x;\theta) = \sum_{i=1}^{n} \ln z(x_i;\theta). \tag{1.6}$$

Tutto ciò cosa significa? Il nostro compito è quello di trovare la migliore stima per il parametro che ci interessa. In questo caso, si badi, θ è la variabile, mentre i diversi valori di x estratti possono essere considerati costanti. La migliore stima del parametro è allora quella per cui è massima la probabilità che i valori che costituiscono il vettore della variabile casuale campionaria **x** assumano i valori particolari (x_1, x_2, …, x_n). Si tratta allora di trovare il valore di θ tale per cui G(**x**; θ) è massimo. Tale valore, indicato come $\hat{\theta}$, è detto "stima di massima verosimiglianza", ed è diverso da qualsiasi altro possibile valore del parametro $\breve{\theta}$. In altri termini,

$$G(\mathbf{x};\hat{\theta}) \geq G(\mathbf{x};\breve{\theta}), \tag{1.7}$$

per qualsivoglia $\breve{\theta}$ diverso da $\hat{\theta}$.

Per trovare allora $\hat{\theta}$ dobbiamo ricorrere all'analisi. Noi sappiamo che il punto di massimo di una funzione è quello in cui si annulla la sua derivata prima, e in cui la sua derivata seconda è negativa. In altri termini, ci basterà trovare il valore di θ tale per cui

$$G'(\mathbf{x};\theta) = 0, \tag{1.8}$$

$$G''(\mathbf{x};\theta) < 0. \tag{1.9}$$

Possiamo così ora affrontare il lemma di Neyman-Pearson.

1.3.3.3. Il lemma

Veniamo quindi al lemma di Neyman-Pearson propriamente detto. Estraiamo, come si è detto, dalla popolazione ξ, che ha funzione di densità di probabilità z(ξ; θ), un campione **x** = [X_1, X_2, …, X_n], per cui varrà comunque la stessa funzione di densità di probabilità, z(**x**; θ). Formuliamo quindi le nostre ipotesi, nulla e sostantiva, in forma un po' diversa da quella a cui siamo abituati, e cioè

$$\mathrm{H}_0 : \theta = \theta_0 \tag{1.10}$$

$$\mathrm{H}_1 : \theta \neq \theta_1 \tag{1.11}$$

In realtà, ogni impressione di stranezza scompare, solo che poniamo mente al fatto che θ potrebbe essere μ. In altri termini, l'ipotesi nulla afferma che la migliore stima del parametro corrisponde al valore del parametro dell'universo, mentre l'ipotesi sostantiva afferma che ne differisce.

Come stabilirlo? Detta *G* (o *L*, il che ai nostri fini è lo stesso) la funzione di verosimiglianza, assumiamo di poter stabilire un valore costante k_α, tale per cui la probabilità di errore di primo tipo sia uguale ad α. In altri termini, definite le funzioni di verosimiglianza per l'ipotesi nulla e l'ipotesi sostantiva, rispettivamente $G(\mathbf{x}; \theta_0)$ e $G(\mathbf{x}; \theta_1)$, α deve essere uguale alla probabilità condizionale che il loro rapporto sia minore di k_α, posto che H_0 sia vero:

$$p\left[\frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)} < k_\alpha \,|\, \mathrm{H}_0\right] = \alpha. \tag{1.12}$$

Allora la regione critica *C* dello spazio dei parametri potrà essere definita come:

$$C = \left\{\mathbf{x} : \frac{L(\mathbf{x};\theta_1)}{L(\mathbf{x};\theta_0)} \geq k_\alpha\right\}. \tag{1.13}$$

Analogamente, la regione di accettazione può essere definita come

$$A = \left\{\mathbf{x} : \frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)} > k_\alpha\right\}. \tag{1.14}$$

La dimostrazione del lemma, peraltro non difficile, va oltre i nostri scopi, pertanto rinviamo il lettore interessato a un testo di statistica matematica (p.e. Orsi, 1985, p. 410 sgg;), oltre che, ovviamente, all'articolo originale di Neyman e Pearson (1933).

Esempio 1.1

Chiariamo quanto detto con un esempio. Supponiamo che la funzione di densità di probabilità in gioco sia la funzione di Poisson (vedi sopra, la 1.2). Supponiamo anche che precedenti ricerche abbiano dimostrato che gli infortuni sul lavoro in una industria si distribuiscono secondo una poissoniana con λ = 0,12. Ora, secondo i sindacati nell'ultimo anno l'azienda ha risparmiato sulla sicurezza, portando quindi il valore di λ a 0,18. Dobbiamo allora verificare le seguenti ipotesi statistiche:

$$\mathrm{H}_0 : \lambda = \lambda_0 = 0,12 \tag{1.15}$$

$$\mathrm{H}_1 : \lambda = \lambda_1 = 0,18 \tag{1.16}$$

Immaginiamo che vengano rilevati gli infortuni avvenuti nel corso di questo ultimo anno. Tenuto conto della (1.2) e della (1.5), la nostra funzione di verosimiglianza sarà uguale a

$$L(\mathbf{x};\lambda) = \frac{\lambda^{\sum x_i}}{\prod x_i !} e^{(-n\lambda)}. \tag{1.17}$$

Vediamo allora di calcolare il rapporto di verosimiglianza:

$$\frac{L(\mathbf{x};\lambda_1)}{L(\mathbf{x};\lambda_0)} = \frac{\lambda_1^{\sum x_i}}{\prod x_i !} \frac{\prod x_i !}{\lambda_0^{\sum x_i}} \frac{e^{(-n\lambda_1)}}{e^{(-n\boldsymbol{\lambda}_0)}}. \tag{1.18}$$

Noi però conosciamo i valori di λ_0 e λ_1, rispettivamente 0,12 e 0,18, e possiamo sostituirli nella 1.18. Sulla base della 1.13, possiamo allora impostare la disuguaglianza tale per cui il valore del rapporto sia maggiore o uguale a una costante *c*:

$$\left(\frac{0,16}{0,18}\right)^{\sum x_i} e^{(-0,8n)} \geq c. \tag{1.19}$$

Passando al logaritmo, otteniamo:

$$\sum x_i \log 2 - 0,8n = \sum x_i\, 0,6931 - 0,8n \geq c. \tag{1.20}$$

Di qui:

$$\sum x_i \geq \frac{\log c + 0,8n}{0,6931} = k_\alpha. \tag{1.21}$$

Si osservi che, come detto sopra, il problema della verosimiglianza è quello di determinare la probabilità che i valori che costituiscono il vettore della variabile casuale campionaria **x** assumano i valori particolari (x_1, x_2, …, x_n), per i quali vale il parametro θ. Possiamo allora sostituire a $\sum x_i$ la $\sum X_i$ e determinare

$$p\left(\sum x_i \geq \frac{\log c + 0,8n}{0,6931} = k_\alpha\right). \tag{1.22}$$

Ora noi sappiamo che questa probabilità deve essere uguale ad α, che abbiamo già fissato a 0,05. Pertanto, una volta stabilita la grandezza del campione (n), risolvendo la (1.21) saremo in grado di conoscere il valore di c. Supponiamo così di avere $n = 20$: c dovrà allora essere uguale a 0,465.

1.4. La verifica della significatività dell'ipotesi nulla (VeSN)

Il lettore si starà chiedendo cosa c'entra tutto questo con la VeSN, per come è stata da noi presentata all'inizio del nostro capitolo. Riprendiamo allora lo schema del processo decisionale, quello in base a cui schema si afferma che il processo inferenziale, in genere detto "fisheriano" richiede i seguenti passaggi: (i) assegnare un valore di probabilità ad α (usualmente 0.05); (ii) assumere che H_0 sia vera; (iii) determinare la probabilità associata ai dati osservati ammesso che H_0 sia vera [cioè, p (dati | H_0)]; (iv) se $p > \alpha$, accettare H_0; (v) altrimenti, rifiutare H_0 e accettare H_1.

Cosa c'è di genuinamente fisheriano in questo schema? Certo non il concetto di due ipotesi contrapposte, nulla e sostantiva, anche se in un certo senso si possono ritenere implicite nel discorso di Fisher. Fisheriano è invece il fatto di affidarsi esclusivamente al valore di probabilità associato ai dati, posto che l'ipotesi sulla non deviazione della forma della loro distribuzione sia corretta, per accettare questa ipotesi o rifiutarla. In realtà, Fisher non dice che dobbiamo cercare di accettare l'ipotesi nulla (se vogliamo usare questa terminologia), ma ci invita direttamente a *verificare* l'ipotesi sostantiva, sulla base della probabilità dell'area di rifiuto.

In questo senso è vero che lo schema decisionale è fisheriano. Il fatto è che la tabella della decisione statistica con questo schema non c'entra nulla, mentre è chiaramente ispirato direttamente da Neyman e Pearson. Qui infatti il compito è diverso: individuare l'area critica per cui si abbia la massima potenza associata all'accettazione dell'ipotesi sostantiva. E come è facile capire, si tratta di due approcci sostanzialmente diversi.

Può essere utile cercare di fornire uno schema che metta in evidenza le principali differenze tra i due approcci. Seguendo in parte Huberty (1993), il lettore può utilmente servirsi della Tabella 1.1.

Nell'approccio AFA, non si ha quindi un giudizio di tutto o di nulla: si ha una considerazione pragmatica della situazione nella sua complessità, che può portare a tre principali esiti: (i) l'accettazione dell'ipotesi sostantiva, (ii) la verifica della necessità di aumentare l'area critica, ad esempio aumentando la grandezza del campione, o (iii) l'accettazione dell'ipotesi nulla (cfr. Neyman, 1950). Tutto ciò dà al ricercatore non solo una maggiore libertà intellettuale di affrontare i dati, ma gli consente anche di padroneggiare situazioni spesso assai complesse nella loro interezza, senza essere imbrigliato da rigidi schematismi, o, per usare l'espressione di Gigerenzer (1998), "rituali di calcolo".

Fisher (PVA)	**Neyman-Pearson (FAA)**
Approccio a probabilità variabile	Approccio ad α prefissato
Test di significatività	Test d'ipotesi
Porre H_0	Porre H_0 e H_1
Specificare la statistica da utilizzare (T) e la sua distribuzione	Specificare la statistica da utilizzare (T) e la sua distribuzione
Raccogliere i dati e calcolare il valore di T	Specificare l'errore di primo tipo α e determinare la regione critica C (regione estrema della distribuzione)
Determinare il valore di p (probabilità condizionale associata ai dati osservati, posto che H_0 sia vera)	Raccogliere i dati e calcolare il valore di T
Rifiutare H_0 se p è piccolo, altrimenti tollerare H_0	Rifiutare H_0 a favore di H_1 se T si trova nella regione d'esclusione della distribuzione, altrimenti tollerare H_0
Concludere l'esperimento	Ripetere se possibile su nuovi campioni

Tabella 1.1 – *Differenze tra l'approccio fisheriano e quello di Neyman-Pearson*

Si osservi che questo ha portato a un notevole utilizzo dell'AFA in applicazioni della statistica anche lontane dalla ricerca scientifica, e in particolare nelle applicazioni industriali, particolarmente nel controllo di qualità.

Un esempio classico è quello della cosiddetta "sberlatura" nell'industria farmaceutica. Il prodotto, ad esempio compresse, viene trasportato su un nastro davanti all'operaio che effettua il controllo, e che lo vede illuminato a luce radente contro uno sfondo adatto. In questo caso, la produzione si sviluppa in continuità, e il controllo deve servire a eliminare le compresse difettose. È evidente che l'interesse dell'industria (e dei consumatori) non è tanto quello di eliminare compresse prive di difetti (falsi allarmi, errori di I tipo), quanto quello di evitare che entrino in commercio com-

presse difettose (omissioni, errori di II tipo). Periodicamente vengono effettuati dei campionamenti, e se il livello α può essere relativamente elevato (anche 0,1), qui è il livello β che va tenuto molto basso, 0,01, o anche e di molto inferiore, se il difetto può portare a danni alla salute del consumatore. Un risultato significativo in uno dei campionamenti porta di conseguenza al blocco della produzione, e all'individuazione della causa che ha portato alla comparsa del difetto.

Come si vede, i due approcci sono sostanzialmente diversi. Come tutto ciò sia stato fuso nell'immaginario collettivo in un unico approccio, è quanto vedremo nel prossimo capitolo.

Capitolo 2
Il dibattito attuale

2.1. La messa in questione della VeSN

2.1.1. La nascita del paradigma

Nel capitolo precedente abbiamo visto come si sono generati i due approcci fondamentali alla verifica delle ipotesi statistiche, quello fisheriano e quello derivato da Neyman e Pearson. Ricordiamo molto rapidamente che secondo Fisher non si faceva una distinzione tra ipotesi nulla e ipotesi sostantiva. Schematizzando, lo sperimentatore estraeva un campione da un universo di cui ipotizzava nota la funzione di distribuzione e, quindi, calcolava la probabilità che il campione estratto appartenesse effettivamente a tale distribuzione. Se questa probabilità era troppo bassa, scartava l'ipotesi fatta dell'appartenenza del campione alla distribuzione. Secondo Neyman e Pearson, invece, lo sperimentatore assegnava un valore α prefissato (usualmente 0,05) alla probabilità dell'ipotesi nulla H_0. Stabiliva poi le funzioni di verosimiglianza per H_0 e per l'ipotesi sostantiva H_1, e calcolava il valore del rapporto tra le due funzioni. Se la probabilità che questo fosse inferiore a un certo valore critico *k* era inferiore a α, scartava H_0 a favore di H_1. Come si vede, la procedura in entrambi i casi era abbastanza diversa da quella oggi codificata nei manuali di analisi dei dati.

La VeSN è una teoria che contiene concetti di entrambi gli approcci. Per esempio, abbiamo visto come la verifica dell'ipotesi nulla H_0 sia ripresa da Fisher, mentre gli errori di I e II tipo siano ripresi da Neyman e Pearson. Generalmente ci si riferisce alla VeSN come a "la" teoria, senza menzionare chi fossero gli autori dei rispettivi concetti (Gigerenzer & Murray, 1987).

Si osservi che, come nota Sterling (1959), se si esaminano le quattro principali riviste internazionali di psicologia sperimentale, tra il 1930 e il 1940 solo in quattro articoli si trovano procedure statistiche che comportano la verifica di ipotesi e che possono essere ricondotte più o meno direttamente alla VeSN; tra il 1940 e il 1955 si afferma la VeSN come vero e proprio paradigma, nel senso di Kuhn (1962); in questi quindici anni, l'80% degli articoli di ricerca pubblicati su queste quattro riviste analizza i dati utilizzando la VeSN.

Come nasce allora questo paradigma e come mai nasce non scegliendo l'un approccio o l'altro, ma fondendoli nell'ibrido che abbiamo descritto nel primo capitolo?

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*, ISBN 88-8453-226-4,

Al suo affermarsi contribuì di certo in misura notevole un altro famoso manuale di Fisher, oltre al già ampiamente citato *Statistical Methods* del 1925, e cioè *The Design of Experiments*, del 1935. Come abbiamo già avuto modo di osservare, a dispetto della sua fama di oscurità Fisher, specie come quando in questi manuali si rivolgeva al pubblico dei ricercatori e non a quello più specialistico dei matematici e degli statistici, era di una chiarezza cristallina e di una straordinaria persuasività, al contrario di Neyman e Pearson, la cui fama era grande tra gli statistici, ma la cui notorietà tra i ricercatori era più fondata sul sentito dire che sulla conoscenza diretta.

2.1.2. La controversia attuale

Il problema della VeSN ha generato nel tempo molti dibattiti critici, da un lato sull'uso di questi test e sull'interpretazione dei risultati da parte dei ricercatori, dall'altro sulla logica stessa che ne è alla base, e ciò a partire ancora dagli anni '30. Nel precedente paragrafo, abbiamo visto la clamorosa affermazione di questo paradigma nel mondo della psicologia sperimentale, ma anche qui non sono mancati contrasti. Da un lato, molti psicologi matematici, con Duncan Luce in testa, non hanno accettato questa impostazione, considerata un vero ostacolo al progresso scientifico, e hanno fondato una loro rivista, il *Journal of Mathematical Psychology*, per sfuggire alle pesanti politiche editoriali dei prevalenti giornali psicologici che privilegiavano massicciamente l'uso della VeSN. Ma anche nel mondo proprio della psicologia sperimentale non mancarono le voci di dissenso. Si osservi che, nel quindicennio di cui parliamo, la psicologia sperimentale era dominata imperialisticamente dal comportamentismo, ma proprio in questo ambito si levarono le voci più critiche. Burrhus Frederic Skinner, la voce più radicale del comportamentismo americano, fondò una sua rivista, il *Journal of Experimental Analysis of Behavior*, proprio per sfuggire alle forche caudine della VeSN. (È un fatto che forse andrebbe spiegato agli epigoni italiani del comportamentismo, che della VeSN sono anche le vestali più fedeli).

Nonostante le critiche teoriche reiterate negli anni successivi, sino praticamente ad oggi, nella pratica delle applicazioni della statistica i cambiamenti sono stati pochi (Cohen, 1994). Le rinnovate critiche negli anni '90 da parte di alcuni psicologi (Cohen, 1994; Schmidt, 1996) hanno portato alla pubblicazione di un numero speciale del *Journal of Experimental Education* (Vol. 61 (4)) e di *Psychological Science* (Vol. 8 (1)) e di un volume dedicato alla prospettive future dei test di significatività in psicologia, con contributi sia a favore sia contrari (Harlow, Mulaik & Steiger, 1997; per una rassegna delle critiche alla VeSN, vedi Nickerson, 2000; Sullivan, 2000).

Non deve stupire che questo tipo di approccio sia divenuto tanto popolare da essere considerato la base dell'inferenza statistica nelle scienze sociali, poiché offrire una scelta forzata del tipo accetto/rifiuto era molto attraente per chi cercava uno schema deterministico, meccanico e obiettivo. Come notava Yates (1951, p. 33), "i ricercatori hanno spesso considerato l'esecuzione di un test di significatività come l'obiettivo ultimo di un esperimento. I risultati sono significativi o non lo sono, e questo è tutto".

I problemi legati alla VeSN hanno dunque generato una gigantesca quantità di critiche, e molti autori (Carver, Cohen, Gigerenzer, Hunter, Meehl, Schmidt, Schulman, Tryon, e così via; prenderemo più avanti in esame dettagliatamente le loro critiche) hanno proposto di bandire questo tipo di analisi. Tra i critici più virulenti, va segnalato Carver (1978, 1993), che definisce la VeSN una "forma corrotta del metodo scientifico" (1993, p. 288).

La controversia sulla VeSN è nata per il cattivo uso che è stato fatto dei test d'ipotesi, sia da parte dei ricercatori sia da parte degli editori delle riviste scientifiche sulle quali i ricercatori intendevano pubblicare i loro lavori. La polemica investe tutti i settori disciplinari in cui la VeSN fa la parte del leone nell'analisi dei dati. Si pensi che dal 1988 al 1991 l'*American Journal of Epidemiology* ha rifiutato tutti i lavori in cui la VeSN era utilizzata come unico tipo di analisi dei dati. Oggi questi lavori sono di nuovo accettati, ma soltanto se le analisi sono supportate da dati supplementari sulla grandezza dell'effetto, l'analisi di potenza, e così via. E la polemica non accenna a placarsi: si veda la discussione tutt'ora aspra che si è accesa in econometria (cfr. Ziliak & McCloskey, 2004).

Abelson (1995), un autore che ha sempre mantenuto una posizione molto equilibrata sulla controversia, ha tuttavia dovuto ammettere che l'utilità della VeSN è solo parziale. Secondo Abelson, sono cinque i criteri per giudicare la qualità delle affermazioni che possono farsi sui risultati di una ricerca: la grandezza (*magnitude*) dell'effetto, la sua articolazione (*articulation*), la sua generalità (*generality*), quanto è degno di interesse (*interestingness*) e la sua credibilità (*credibility*). È tutto questo che rende una ricerca acronicamente MAGIC. Ora, i valori di *p* hanno interesse solo per quel che riguarda la credibilità, in quanto legati all'accettazione o rifiuto dell'ipotesi nulla. Anche questo è comunque solo un aspetto marginale della credibilità, che si fonda soprattutto sulla coerenza teorica e sulla plausibilità dei processi psicologici postulati.

2.2. A difesa della VeSN

Non vorremmo peraltro dare l'impressione che la VeSN sia considerata, almeno a livello degli statistici, un approccio ormai del tutto obsoleto. La VeSN ha tutt'ora dei difensori, tra i quali merita di essere segnalato soprattutto Chow (1996, 1998, 1999), secondo cui le critiche sono una ripetitiva "litania" che ha riproposto per un quarantennio argomenti sempre uguali (e, a giudizio di Chow, sostanzialmente inconcludenti). Come osserva Abelson (1997), a detta di tanti critici sembra che la VeSN sia una specie di stupido ronzino che si può fare andare avanti solo frustandolo a morte. In realtà, il problema probabilmente è più legato alla necessità di farne uso con una certa accortezza, avendo presente i limiti e soprattutto i rischi di un suo utilizzo acritico.

Sullivan (2000) ha recentemente passato in rassegna i principali argomenti portati a difesa della VeSN, che possono essere raggruppati nelle seguenti categorie: (i) decisioni a livello nominale o ordinale; (ii) mancanza di analisi alternative dei dati;

(iii) responsabilità dei ricercatori per errori e fraintendimenti; (iv) mancanza di forme alternative di analisi in varie situazioni. Vediamo con ordine questi punti.

(1) L'utilità della VeSN quando si voglia procedere a livello di categorie non ordinate (scale nominali) e relazioni asimmetriche (scale ordinali): in questi casi, palesemente, quel che conta non è una non evidenziabile grandezza dell'effetto, ma semplicemente l'affermazione o meno della direzionalità dell'effetto. È evidente che allora la falsificazione dell'ipotesi nulla è sufficiente.

(2) L'insoddisfazione dei ricercatori per le analisi alternative proposte: non sempre è possibile ottenere una potenza soddisfacente, dati i vincoli che la ricerca pone (ad esempio, in termini di numero di soggetti) e spesso il gioco non sembra valere la candela. Del resto, il grosso delle ricerche psicologiche banalmente non si presta allo sviluppo di modelli (alla predicazione di funzioni, ad esempio in relazione al tempo), come avviene in altre scienze, prima tra tutte la fisica. Anche lo studio, ad esempio, di complesse strutture di covarianza, come nei modelli di equazioni strutturali, appare estraneo al grosso della ricerca, almeno in psicologia sperimentale.

(3) La considerazione per cui la VeSN, secondo il modo in cui venne originariamente concepita, è un metodo logico e ben fondato di analisi statistica, ed è colpa dei ricercatori se poi è stato male utilizzato ed ha dato luogo a tanti errori e fraintendimenti; peraltro, questa controargomentazione ci sembra molto debole. Di fatto, la ricerca psicologica è condotta da questi ricercatori, ed evidentemente, a partire dalla loro formazione, la VeSN sembra particolarmente studiata per far commettere a *loro* certi errori.

(4) Il fatto che si tratta di un metodo debole non deve in ogni caso fare dimenticare che in molte situazioni è l'unica analisi realmente disponibile.

Come afferma Huberty (1987, p. 7), "non c'è niente di sbagliato nei test sulla significatività in sé. Se usati come guide ed indicatori, invece che come modi per arrivare a risposte definitive, essi vanno bene".

McLean (2001), ad esempio, è a favore dell'uso della VeSN, purché intesa come una forma di selezione tra modelli. Secondo questo autore la conoscenza del mondo è in termini di modelli, che possono essere sia individuali sia condivisi da molti. Anche le analisi statistiche sono modelli, in particolare modelli di probabilità. La ricerca scientifica consiste essenzialmente nel formulare un modello su alcuni aspetti del mondo reale, nel fare previsioni sulla base di tale modello, raccogliere dati per poi accettare, modificare e rifiutare il modello. Attraverso la VeSN facciamo una selezione tra modelli, in quanto operiamo una scelta tra un modello con i parametri specificati dall'ipotesi nulla e un altro (ipotesi alternativa) con parametri specificati dai dati campionari.

2.3. UN CATALOGO DI INCONVENIENTI

Proponiamo allora un breve catalogo di errori, equivoci, fraintendimenti conseguenti ad un'assunzione acritica della VeSN. In altri termini, questa serie di errori non segue

necessariamente dalla VeSN, ma assumere questa come unico criterio nell'analisi dei dati di ricerca conduce facilmente a cadervi.

2.3.1. La potenza del test

Il primo dei problemi, e per qualche autore il più importante (Cohen, 1977, 1988) è quello di trascurare la potenza del test utilizzato, che dovrebbe essere determinata prima della raccolta dei dati, se non altro per fissare la grandezza ottimale dei campioni utilizzati, ma che non viene praticamente mai fissata dal ricercatore — su questo tema torneremo più diffusamente nel Cap. 5. È particolarmente importante rendersi conto che il fissare α implica fissare nello stesso tempo β — le due probabilità sono legate tra loro. Si badi infatti che H_0 e H_1 corrispondono a eventi a "discriminabilità imperfetta" (se la discriminabilità fosse perfetta, non ci sarebbe allora bisogno di una decisione statistica), ma ciò significa che le corrispondenti distribuzioni sono in parte sovrapposte.

Ora, fissando α noi determiniamo un'area di rifiuto nella distribuzione di H_0 (in una ipotesi a una coda, monodirezionale, si tratta dell'area a destra di un asse di decisione — vedi Fig. 2.1). Ma l'asse di decisione interseca anche la distribuzione corrispondente a H_1, e l'area alla sua sinistra corrisponde a β.

Si osservi che i parametri delle distribuzioni H_0 e H_1 e la distanza tra di esse (equivalente alla misura della grandezza dell'effetto, come *d* di Cohen; su questo indice, e in genere sulla grandezza dell'effetto, vedi il Cap. 3) sono dati. Così, il solo modo di trovare un β ottimale è agire su *n*, la grandezza del campione.

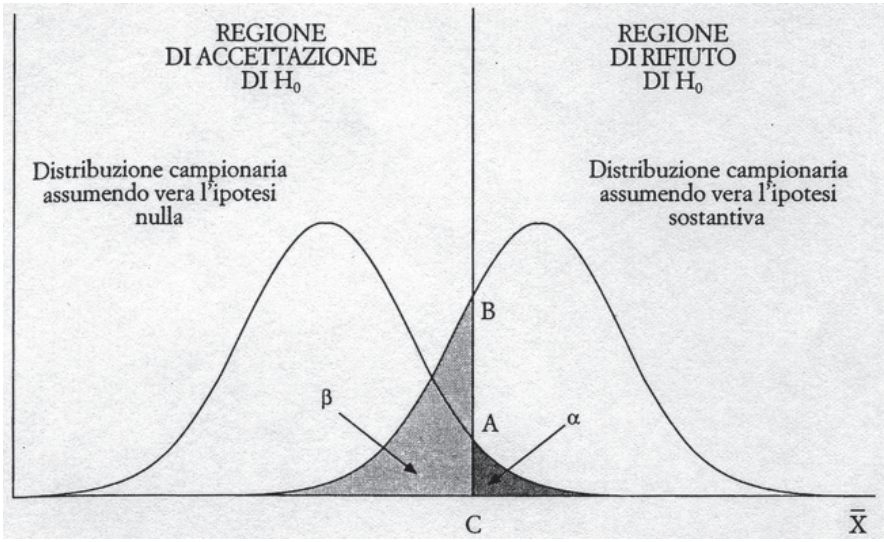

Figura 2.1 – *Le due curve corrispondono alle distribuzioni relative all'ipotesi nulla, H_0, e all'ipotesi sostantiva, H_1*

Facciamo un esempio. Supponiamo di conoscere già μ_0 e μ_1, e cioè le medie rispettivamente degli universi corrispondenti a H_0 e H_1, nonché le loro deviazioni standard σ_0 e σ_1. È allora banale il fatto che le aree di rifiuto relative all'ipotesi nulla e all'ipotesi sostantiva saranno quelle corrispondenti ai punti *z* (monodirezionali) corrispondenti. Se quindi fisseremo α a 0,05 e β a 0,2, i punti *z* saranno rispettivamente 1,645 e – 0,845. Si osservi, infatti, che l'area di rifiuto è quella di sinistra per H_0 (*z* positivo) e quella di destra per H_1 (*z* negativo).

Potremo allora impostare il seguente sistema di equazioni:

$$\begin{cases} \dfrac{M_0 - \mu_0}{\dfrac{\sigma_0}{\sqrt{N}}} = 1,645 \\ \dfrac{M_1 - \mu_1}{\dfrac{\sigma_1}{\sqrt{N}}} = -0,845 \end{cases} . \qquad (2.1)$$

Di fatto, è questo un modo simmetrico a quello di Neyman-Pearson, che abbiamo visto nel Cap. 1, di impostare il problema della potenza. Di qui, il calcolo della grandezza ottimale di un campione per avere una potenza adeguata è immediato, rimanendo come unica incognita la radice della grandezza del campione.

Evidentemente, questo calcolo a priori può essere fatto solo nel caso eccezionale in cui si conoscano già le caratteristiche delle distribuzioni H_0 e H_1. Nella pratica, sarà invece importante determinare la potenza a posteriori, e questo sulla base della determinazione della grandezza dell'effetto. Per vedere come si procede in proposito, rimandiamo ai Cap. 3 e 5.

Ma questo è soltanto *uno* dei problemi.

2.3.2. Rigidità della decisione binaria

La convenzione di fissare una soglia critica che determini la scelta tra due alternative opposte, quali il rifiuto o meno dell'ipotesi nulla, appare in tutta la sua fallacia se riflettiamo brevemente su talune situazioni che spesso i ricercatori si trovano ad affrontare di fronte agli esiti delle proprie analisi dei dati. Molti di noi si sono trovati, nella pratica della ricerca, a dover accettare l'ipotesi nulla poiché avevano ottenuto valori di α di 0,052. È impossibile, in tali circostanze ignorare il dubbio che in realtà l'effetto cercato sia talmente prossimo alla significatività da poter essere realmente influente. Altresì l'impossibilità di sostenere questa ipotesi ci fa comprendere chiaramente quali siano i limiti imposti da una decisione che, nella pratica, si risolve in un tutto o nulla. È piuttosto sconfortante pensare che numerosi lavori che avrebbero potuto fornire

un valido ed importante contributo all'avanzamento delle conoscenze della comunità scientifica (e non solo) non siano mai stati pubblicati o resi noti perché invalidati dal quel 0,002 di scarto rispetto al valore critico di rifiuto.

Peraltro, come nota Abelson (1997), in psicologia è comunque a volte necessario dover prendere questo tipo di decisioni categoriche, pur tenendo comunque conto del fatto che può trattarsi di decisioni incerte e provvisorie. Sarebbe peraltro insensato inferire una differenza tra due gruppi sulla base del fatto che relativamente all'intervento di una certa variabile i risultati in un gruppo sono significativi, e non nell'altro. Per esempio, noi possiamo rilevare che in un gruppo di donne meridionali gli effetti di una dieta sono statisticamente significativi, misurando il calo di peso prima e dopo la dieta, per esempio con un t per misure ripetute, e ottenendo un valore di t corrispondente a una probabilità $p = 0{,}045$; la stessa dieta, in un gruppo di donne settentrionali, porta a un t con valore associato $p = 0{,}055$. Il primo valore, significativo, ci induce a decidere per l'efficacia della dieta, il secondo ci induce a decidere per la sua inefficacia. Ma sarebbe assolutamente sciocco dedurre da questo una differenza tra donne meridionali e settentrionali, che non sono state per nulla confrontate in questo disegno.

Un argomento sovente portato per sottolineare il rischio insito nel trarre inferenze categoriche si riferisce all'utilizzo sempre più diffuso delle meta-analisi. È evidente che le meta-analisi hanno senso solo se si ritiene che i risultati delle ricerche che vengono per così dire sommarizzate con questo metodo si distribuiscono lungo un continuum. Al di là di altri inconvenienti che la meta-analisi presenta (per esempio, l'abuso che ne viene spesso fatto in campo clinico; cfr. Luccio e Salvadori, 2002), è evidente che con la meta-analisi si perdono proprio quelle informazioni rilevanti peculiari di ogni singola ricerca, che fanno sì che le diverse ricerche meta-analizzate non sono mere ripetizioni le une delle altre, e che nella maggior parte dei casi richiedono proprio un giudizio categorico: l'effetto, in questo caso, c'è o non c'è.

2.3.3. Arbitrarietà di α

In psicologia, come del resto nella maggior parte delle discipline sia mediche che sociologiche, il valore di α (livello critico di significatività) viene convenzionalmente fissato a 0,05 (5%), e a volte a 0,01 (1%). Questo valore, poi confrontato con il livello di probabilità osservato (p) per decidere se rifiutare o meno l'ipotesi nulla, è quindi stabilito a priori e fissato in modo del tutto arbitrario dal ricercatore. Non vi sono motivazioni teoriche che testimonino la preferibilità di questa scelta rispetto ad alternative altrettanto plausibili: cosa ci impedisce di fissare α a 0,01, piuttosto che 0,03 o a 0,07? Come dicono Rosnow e Rosenthal (1989, p. 1277), "Dio ama lo 0,06 esattamente quanto lo 0,05". Di fatto, spesso si afferma che è più opportuno tenere più basso il livello in ricerche che abbiano potenzialmente serie conseguenze sul piano della salute o economico, mentre nelle normali ricerche di laboratorio di psicologia il problema non è così rilevante.

Questo è ovviamente insensato: maggiore è la rilevanza della ricerca, maggiore dovrebbe essere l'attenzione per la potenza del test, e quindi per β, e per la replicabilità dei risultati! Ma questi sono problemi indipendenti dal livello di α, che è solo uno dei fattori in gioco.

2.3.4. Sopravvalutazione di p

Il valore di *p* dovrebbe avere come unico significato quello di essere confrontato con il valore di α per giungere alla decisione statistica di accettare o rifiutare l'ipotesi nulla, eppure i ricercatori spesso commettono un ulteriore *bias* nell'interpretazione dei loro risultati. È comune ad una logica erronea di interpretazione ritenere che più il valore di *p* risulta basso, più alta è l'inaccettabilità dell'ipotesi nulla. Questo modo di interpretare i risultati conferisce a *p* il ruolo di misura del grado di inaccettabilità dell'ipotesi nulla, considerandolo come fosse un indice di grandezza dell'effetto, il che è errato. Innanzitutto *p* viene calcolato rispetto all'assunto che l'ipotesi nulla sia vera. Inoltre, tutto quello che possiamo dedurre dal suo valore è che sia maggiore o minore di α, mentre la grandezza di questa differenza non ci fornisce nessuna informazione aggiuntiva rispetto all'ipotesi nulla. Il valore di *p* rappresenta la misura del livello di dubbio che dovremmo avere nello scegliere di accettare l'ipotesi nulla; la differenza in termini di decimali non è una distanza matematica.

Alle radici di questa cattiva interpretazione del significato di *p* vi è probabilmente un altro vizio di fondo messo in luce chiaramente da Carver (1993) e Daniel (1998); e cioè l'idea che *p* misuri il grado in cui il risultato è stato ottenuto per caso. In altri termini, il ricercatore è portato a ritenere che avere un *p*, poniamo, di 0,03, significhi che esiste solo il 3% di probabilità che i risultati siano stati ottenuti per caso, mentre vi è il 97% di probabilità che si tratti di un effetto "reale". Ora, questo assunto è assolutamente arbitrario. Ripetiamo che la probabilità è calcolata in base all'ipotesi dell'inesistenza dell'effetto, e cioè assumendo la verità dell'ipotesi nulla: in altri termini, assumendo che l'ipotesi nulla abbia probabilità 1 di essere vera. E *p* ha l'unico scopo di portarci ad accettare o rifiutare il fatto che l'ipotesi nulla ha probabilità 1 di essere vera. È quindi evidente che *p* non può in alcun modo dirci qualcosa sulle quote di casualità che possono esserci nell'effetto rilevato.

2.3.5. Significatività statistica contro significatività sostanziale

Un ulteriore significato erroneo che si tende ad attribuire a *p* consiste nell'interpretare questo valore di significatività osservato come indicatore dell'entità dell'effetto indagato. Spesso cioè il valore di *p* viene considerato come una misura del valore della ricerca, dove un valore di *p* inferiore a 0,01 è "più significativo" di un *p* inferiore a 0,05. Ognuno di noi ha letto articoli in cui incauti autori utilizzavano espressioni infelici come "risultati molto significativi", riportando, a supporto di questa interpretazione, valori di *p* molto inferiori a 0,05. La significatività statistica non corrisponde affatto alla signifi-

catività sostanziale, ovvero alla grandezza dell'effetto studiato. Un esito statisticamente significativo ci lascia presumere che esista un effetto, ma non ci dice niente in merito alle caratteristiche dello stesso. Infatti, il valore di *p* non risente soltanto dell'effetto considerato (θ) ma anche della grandezza del campione; inoltre, la relazione stessa che esiste tra θ e *p* non è lineare, ad un lieve incremento dell'effetto θ può corrispondere un decremento di *p* proporzionalmente molto maggiore o viceversa. Presumere che vi sia una relazione tra i valori di θ e *p* senza conoscerne adeguatamente l'andamento e senza prendere in considerazione l'influenza dell'ampiezza del campione costituisce un'inferenza arbitraria ed erronea, ma piuttosto comune nella pratica di ricerca odierna.

È probabile che l'infelice termine "significatività" abbia contribuito a generare equivoci, fraintendimenti, e in ultima analisi malusi. Esso tende ad essere letto come "importanza" in assoluto, il che è arbitrario (cfr. Schafer, 1993). I più critici verso questo termine, come ad esempio Meehl (1997), lo definiscono senza mezze misure "canceroso" e "fuorviante". È stato così da molte parti suggerito, e particolarmente da parte dei membri più autorevoli dell'AERA, *American Educational Research Association* (come Thompson, 1996, 1997), di indicare sempre il termine "significatività" accompagnandolo con l'aggettivo "statistica", in modo da rendere chiaro al lettore (ma, diremmo, anche all'autore della ricerca) che l'effetto alone che accompagna il termine *significativo* va eliminato. Esso non significa qui *rilevante*, ma si riferisce esclusivamente alla caratteristica statistica del risultato.

È stato obiettato (Robinson e Levin, 1997; Levin e Robinson, 1999) che le riviste scientifiche non si rivolgono a un lettore comune, ma a un pubblico accademico, e pertanto questa raccomandazione non sembrerebbe giustificata, salvo che per riviste più divulgative (*magazines*, e non *journals*, come *Psychology Today* o, in Italia, *Psicologia contemporanea*). Secondo Levin e Robinson, più sensato, per i *journals*, sarebbe far cadere il "significativo", piuttosto che lo "statistico" e si potrebbe dire senza rischio di fraintendimenti o di sovrainterpretazioni che, ad esempio, ci sono delle "differenze statistiche" e non delle "differenze significative".

Il procedimento che essi propongono è a "due gradini": "gli autori dovrebbero *prima* indicare se l'effetto osservato è statisticamente improbabile (per esempio, la differenza è maggiore di quella che ci si sarebbe dovuti attendere in base al caso?), e *solo in questo caso* dovrebbero indicare di quanto è *grande* o *importante* (è una differenza che *fa* la differenza?)" (Robinson e Levin, 1997, p. 22).

Peraltro, questo comporta con chiarezza, come nota Cahan (2000), un'assunzione tutt'altro che dimostrata, e cioè una forte relazione negativa tra significatività statistica ed entità dell'errore casuale contenuto negli effetti osservati. Secondo questo ragionamento, infatti, dire che ha senso parlare della grandezza dell'effetto solo se si ha una significatività statistica equivale a dire che solo in questo caso esiste un effetto "vero" (chiamiamolo τ) alla base dell'effetto osservato *S*, mentre l'effetto τ_0, quello ipotizzato con H_0, se è vero come sostengono Robinson e Levin che non vale neppure la pena di misurarne l'entità, implica che *S* è dovuto pressoché integralmente all'errore casuale. Levin lo aveva del resto già osservato nel 1993 (p. 379): "La significatività statistica dovrebbe precedere qualsiasi discussione sulla grandezza dell'effetto. Parlarne in presenza di risultati non significativi statisticamente non ha senso. Se non è reale, chiamiamola zero".

Come peraltro osserva Cahan, la significatività statistica è funzione della differenza tra effetto osservato ed effetto ipotizzato con H_0, e cioè $S - \tau_0$, e, nel caso che τ_0 sia posto uguale a 0, è funzione di *S*. L'errore casuale è invece funzione di $S - \tau$, e quindi, a meno che τ e τ_0 non coincidano, significatività statistica e errore casuale risultano funzione di variabili differenti.

Il fatto è che Robinson e Levin non si rendono conto che non necessariamente gli effetti non significativi sono dovuti interamente o soprattutto al caso, anche se è vero che effetti dovuti interamente al caso non sono significativi. E di contro, nota Cahan, nulla esclude che negli effetti significativi vi sia una quota di errore casuale. Insomma, il processo a due gradini assegna alla significatività statistica una sorta di ruolo da "certificato *kosher*" (in yiddish, la purezza secondo i precetti del culto ebraico) della grandezza dell'effetto, che risulta accettabile purché i risultati siano significativi. Ma è un assunto non garantito.

In una gustosa replica in stile Sholem Aleichem[1], Levin e Robinson (2000) raccontano la storia dell'Onnipotente, sia Egli benedetto, e di Rabbi Sholem dello *shtetl* (villaggio) di Anatevka. L'Onnipotente promette un premio a Rabbi Sholem se egli dimostrerà che la maggioranza delle case del suo villaggio è *kosher* — e il Rabbi fa un campionamento da cui estrae dieci case, e dimostra che 6 di queste lo sono. Quindi, conclude, il 60% delle case del villaggio sono *kosher* e io ho diritto al premio. No, ribatte l'Onnipotente, applicando il test della binomiale ne risulta una probabilità di 0,38, e quindi non puoi attribuire nessun significato alla grandezza dell'effetto che hai trovato.

La morale è chiara: "Pur se si può sostenere che un test inferenziale statistico e una stima della grandezza dell'effetto sono operazioni tecnicamente prive di relazioni — come fa il professor Cahan — noi sosteniamo che in una indagine primaria di ricerca le due operazioni sono concettualmente e funzionalmente collegate" (p. 35). Confessiamo di non essere convinti da questa argomentazione: non riusciamo a capire come un potenziale conflitto tra i risultati che si possono ottenere "tecnicamente" (e cioè, applicando operazioni matematicamente significanti ai dati che si ottengono), e la coerenza teorica (che non altrimenti interpreteremmo il "concettualmente e funzionalmente") possa essere risolto a favore di quest'ultima. Ora, Cahan dimostra convincentemente, e Levin e Robinson non lo smentiscono, che questa mancanza di relazione "tecnica" può condurre al conflitto. Per questi ultimi, quindi, se i dati entrano in conflitto con la teoria, abbandona i dati e tieni la teoria. Curioso, no?

2.3.6. Sovrainterpretazione dei risultati

Nel test d'ipotesi, l'ipotesi nulla H_0 contiene la premessa che gli effetti in essa contenuti siano nulli. L'inghippo sta nel pensare che dire ipotesi nulla significhi che il para-

[1] Pseudonimo (letteralmente, "La pace sia con voi") dello scrittore ebreo di origine ukraina Sholem Rabinovitch (1859-1916), forse il più grande autore in lingua yiddish, che ha descritto nei suoi racconti la vita delle comunità ebraiche degli *shtetele*, i villaggi rurali dell'Europa orientale.

metro in essa specificato sia zero. Questo non è corretto in tutti i casi. Secondo Fisher (1935, 1955) l'ipotesi nulla è quella che ci aspettiamo di annullare o di falsificare. Si potrebbe pensare che sia conservativa, poiché presuppone che non ci sia effetto, alcuni casi dimostrano però il contrario. Uno di questi è il caso del confronto tra varianze, dove l'ipotesi nulla assume che le varianze siano uguali.

Nel caso in cui l'ipotesi nulla venga accettata, si assume che la stima campionaria di θ equivalga esattamente al parametro "vero" θ*; al contrario, se l'ipotesi nulla viene rifiutata si tende ad assumere la totale mancanza di effetto. Abbiamo discusso già al punto precedente la non ragionevolezza di questo modo di pensare. Difficilmente un effetto, per quanto blando, potrà essere considerato inesistente, ed altrettanto arbitrario è comunque ritenere che la corrispondenza tra θ stimato e reale possa essere perfetta.

2.3.7. L'accettazione dell'ipotesi nulla: una decisione imbarazzante

Dicevamo all'inizio di questo lavoro che, nel caso in cui il ricercatore dovesse rilevare che la sua $p(\text{dati} \mid H_0)$ è maggiore del valore prefissato di α, e cioè nella maggior parte dei casi del fatidico 0,05, il suo lavoro finirà probabilmente nel cestino della carta straccia. Non è una battuta di spirito. Tutto lascia ritenere che i ricercatori si autocensurino in proposito, e non sottopongano neppure lavori con questi esiti, che vengono normalmente detti "negativi" (Hubbard e Armstrong, 1997), per la pubblicazione.

I dati in proposito sono impressionanti, e noti già da tempo. Così, Wilson, Smoke & Martin (1973), esaminando gli articoli pubblicati da *American Journal of Sociology*, *American Sociological Review* e *Social Forces* nel 1969-1970 hanno rilevato che l'ipotesi nulla veniva respinta nell'80,3% dei casi in cui si usava la VeSN. Greenwald (1975), esaminando l'annata 1972 del *Journal of Personality and Social Psychology*, ha trovato che qui lo stesso avveniva per l'87,9% degli articoli. Ma dati analoghi si trovano negli articoli di marketing. Per Hubbard e Armstrong (1992, 1994) qui la percentuale sale al 92,2%, per Lindsay (1994) è dell'84,2%. In generale, Sterling, Rosenbaum e Weinkam (1995) hanno dimostrato che nel 95,6% dei lavori psicologici pubblicati in cui si utilizza la VeSN l'ipotesi nulla viene respinta, mentre nei giornali medici la percentuale scende all'85,4%.

Vi è un'ovvia motivazione alla resistenza che dimostrano i ricercatori a proporre per la pubblicazione (e le riviste ad accogliere) accettazioni dell'ipotesi nulla. Una ricerca viene abitualmente compiuta per dimostrare l'efficacia di un trattamento, e l'accettazione dell'ipotesi nulla dimostra di frequente che il presupposto era errato. Ma se questa può essere la motivazione di base, ciò non significa che non sia errata. Di fatto, la sola domanda corretta che occorrerebbe porsi è: questo risultato costituisce un progresso in termini di accrescimento delle conoscenze, oppure no?

È evidente che un risultato anche "negativo", se sensato, costituisce un progresso. Come minimo, indica agli altri ricercatori che questa strada non va ulteriormente battuta, evitando così la perdita di tempo e denaro. Ma attenzione: come ancora nel 1940 osservava Lundquist, di fatto gli errori di I tipo possono essere per la ricerca più dannosi di quelli di II tipo, proprio perché hanno come conseguenza il blocco di una

certa linea di ricerca. E allora, perché ritenere che il rischio di commetterli sia minore, rispetto agli errori di II tipo? Molto spesso, però, l'accettazione dell'ipotesi nulla dice una serie di cose interessanti. Dimostrare la non validità di un'alternativa può aiutare a chiarire i presupposti teorici di un problema, quanto indicare una via percorribile.

Si afferma che l'accettazione dell'ipotesi nulla è una scelta più "debole" del suo rifiuto, perché si limita ad affermare non la verità dell'ipotesi nulla, ma l'impossibilità di respingerla. In effetti, questo ragionamento, che pure si trova in una versione o nell'altra in tanti testi di metodologia, dimostra una volta di più l'incapacità di vedere le cose in termini complessivi, e di rendersi conto della specularità del rapporto tra H_0 e H_1. Se la potenza del test è sufficientemente elevata, all'accettazione dell'ipotesi nulla deve corrispondere una probabilità molto bassa di commettere l'errore di II tipo, e quindi di rifiutare erroneamente H_1.

Così, ancora una volta è la potenza del test statistico utilizzato che può dirci se questa accettazione è sensata o no. Hubbard e Armstrong (1992) hanno allora esaminato la potenza nelle (poche) ricerche in cui si accettava l'ipotesi nulla, e hanno verificato che le probabilità medie di rilevare effetti rispettivamente piccoli, medi e grandi, secondo la classificazione di Cohen (1977; vedi Cap. 5) erano rispettivamente di .35, .89, and .99. In altri termini, si poteva escludere con ragionevole certezza che il trattamento potesse avere un grande effetto (p = 0,01) o anche solo un effetto di grandezza media (p = 0,11), mentre dei dubbi rimangono relativamente a un effetto piccolo. (Sul concetto di grandezza dell'effetto, e sul modo di misurarlo, vedi il Cap. 3; sui suoi rapporti con la potenza, vedi il Cap. 5). Si osservi che di solito la potenza media nelle ricerche pubblicate in cui si rifiuta l'ipotesi nulla è molto più bassa.

Come osservava McNemar già nel 1960, il fatto che la pubblicazione di ricerche che portano all'accettazione dell'ipotesi nulla sia scoraggiata porta a conseguenze di notevole spiacevolezza. I ricercatori sono portati ad escludere i risultati non significativi, selezionando solo quelli significativi, e a scartare i dati che non confermano le teorie. Un caso ben noto in una "disciplina" vicina alla psicologia (non diciamo scienza), e cioè in parapsicologia, si verificò con le stupefacenti dimostrazioni di telepatia e chiaroveggenza ottenute da Rhine e dai suoi collaboratori alla Duke University negli anni '30, e dovute semplicemente alla selezione delle serie sperimentali in cui i risultati ottenuti erano significativi (su tutta la squallida vicenda, cfr. Luccio, 1982). È straordinario il fatto che questa serie di clamorosi artefatti venga tutt'ora presentata dai sostenitori della parapsicologia come una dimostrazione inoppugnabile dell'esistenza del paranormale!

Stando così le cose, apparirebbe opportuno più coraggio da parte dei ricercatori nel cercare di pubblicare un maggiore numero di ricerche in cui si accetta l'ipotesi nulla. Ovviamente, la potenza del test andrebbe pubblicata in parallelo con i risultati. Tra l'altro, come osservano Lane e Quinones (1997), l'esplosione della pubblicazione elettronica dei risultati delle ricerche dovrebbe consentire facilmente questo maggiore ardimento da parte di autori ed editori, dato l'abbattimento dei costi che la pubblicazione in rete comporta.

Si osservi, tra l'altro, a proposito della politica di pubblicazione dei giornali (e quindi, indirettamente, degli autori) che si ha addirittura una riluttanza a pubblicare

lavori in cui il livello di *p* sia *solo* di 0,05, privilegiando il livello di 0,01 o meno, come candidamente ammise Melton (1962), nel congedarsi dopo 12 anni di direzione del *Journal of Experimental Psychology*.

2.3.8. Eccezionalità dell'ipotesi nulla in natura

Nella pratica universalmente riconosciuta del test d'ipotesi, l'ipotesi nulla (H_0) contiene la premessa che gli effetti in essa contenuti siano nulli. Questo significa che si presuppone che la differenza delle medie della popolazione sia pari a zero, che la correlazione sia zero, che la proporzione di maschi e di femmine in una qualsiasi popolazione sia del 50%. Basta fermarsi un attimo a riflettere su queste formulazioni per comprendere quanto siano insensate. Non vi è motivo alcuno per presupporre che maschi e femmine debbano essere equamente distribuiti o per pensare che prendendo a caso due campioni le medie delle misurazioni effettuate debbano essere assolutamente identiche (Bakan, 1966). Sarà, altresì, altamente improbabile che nel mondo reale tutto questo possa avvenire. Tutto questo comporta una riflessione alquanto inquietante, dato che l'ipotesi nulla è, per sua natura, praticamente sempre falsa, a cosa ci serve rifiutarla? Fino a che punto può essere informativo rifiutare una informazione che sappiamo, dotati di buonsenso, essere falsa? Come dice argutamente Loftus (1996), rifiutare l'ipotesi nulla è come rifiutare l'ipotesi che la luna sia fatta di formaggio.

A dimostrazione della pratica impossibilità di trovare un'ipotesi nulla reale in natura, in una ricerca (di fatto, una provocazione) molto citata Bakan (1966) estrasse un campione di popolazione di 60.000 persone, che poteva essere classificato in vari modi diversi, lo sottopose a vari test, e riuscì a dimostrare che comunque si riscontravano differenze tra le classi in cui la popolazione era suddivisa, indipendentemente dal criterio di classificazione: abitanti del Maine contro abitanti del resto del paese, abitanti a Est contro abitanti a Ovest del Mississipi, abitanti del Sud contro abitanti del Nord, e così via. Le differenze potevano essere molto piccole, ma comunque sussistevano, e con un campione così ampio non c'è da meravigliarsi se il valore di *p* era sempre significativo.

Ma il problema è realmente così importante? C'è chi, come Levin & Robinson (1999) ne dubita. In effetti, come osserva Tukey (1991, p. 100), nessun ricercatore assume che la differenza tra due medie o la correlazione tra due variabili sia esattamente 0: "È folle chiedersi, 'Gli effetti di A e B sono differenti?' Essi sono sempre differenti — a livello di qualche decimale". Il ricercatore tende ad escludere che la differenza (o la correlazione) sia molto piccola, tanto da non potersi ragionevolmente escludere che sia dovuta al caso. Di fatto, come osservano Mulaik, Raju e Harshman (1997), il problema è quello delle deviazioni dall'ipotesi nulla per la grandezza del campione che stiamo esaminando. Quel che noi valutiamo è se per campioni di tale grandezza le differenze sono o meno dovute al caso. E come nota Rindskopf (1997, p. 321), "a meno che il campione non sia grande […] i test per l'ipotesi nulla verificano approssimativamente quel che è giusto in molte situazioni comuni". Di più, secondo Chow (1996, 1998) questo problema non ha comunque nulla a che vedere con la verifica delle ipotesi in senso stretto.

2.3.9. Il problema dell'ipotesi alternativa

Come già spiegato nel Cap. 1, nell'approccio fisheriano l'ipotesi alternativa è implicita, non viene espressamente specificata. Fisher (1935, 1955) parla solo di ipotesi nulla, che può essere vera oppure no. Risulta quindi evidente che se l'ipotesi nulla è falsa, ne deve esistere almeno una alternativa.

Sono Neyman e Pearson ad aver concettualizzato l'ipotesi alternativa, sostenendo che l'ipotesi nulla e l'ipotesi alternativa si escludono a vicenda e che una delle due deve necessariamente essere vera. Ricordiamo che siamo in un'ottica probabilistica tale per cui i termini "vero" e "falso" stanno per "probabilmente vero" e "probabilmente falso", e sono strettamente connessi con i concetti di errore di I tipo e errore di II tipo.

Generalmente nell'ipotesi nulla è specificato un parametro sotto forma di valore atteso, mentre nell'ipotesi alternativa il valore solitamente non è specificato. In realtà non è così. Al momento in cui formuliamo la nostra ipotesi alternativa, il parametro sconosciuto verrà stimato in base al campione. Sarebbe quindi più corretto dire che il test permette di selezionare tra due ipotesi: una (l'ipotesi nulla) suggerita dalla teoria, l'altra (l'ipotesi alternativa) derivata dai dati campionari (McLean, 2001).

2.3.10. Influenza della grandezza del campione

È ormai universalmente noto e riconosciuto che la grandezza del campione utilizzato influenza la significatività dei risultati ottenuti. Già Neyman & Pearson nel 1933 mettevano in guardia da questo errore affermando chiaramente che la probabilità di rifiutare l'ipotesi nulla è funzione di diversi fattori, tra cui il numero di osservazioni effettuate. Gli altri fattori erano l'essere il test mono- o bi-direzionale, il livello di significatività prefissato, la deviazione standard e l'entità della deviazione dall'ipotesi nulla. Nel 1938 Berkson affermava che "quando i numeri dei dati sono decisamente grandi, i p tendono ad essere piccoli" (p. 526). Analogamente, Nunnally (1960, p. 643) osservava che, se l'ipotesi nulla non viene rifiutata, probabilmente è perché viene utilizzato un campione troppo piccolo, laddove l'utilizzo di un campione sufficientemente ampio costituisce la garanzia di rifiutare la stessa.

Negli anni si sono susseguite molte ed autorevoli voci a mettere in guardia i ricercatori da questo importante rischio. Così, Kerlinger (1979, p. 318) notava che "la significatività statistica dice poco o nulla sulla grandezza di una differenza o di una relazione. Con un ampio numero di soggetti […] i test di significatività mostrano una significatività statistica anche quando la differenza tra le medie è minima o banale, e lo stesso dicasi del coefficiente di correlazione". Come mostra Daniel (1998), un r di 0,6 (quindi decisamente alto) non sarà significativo con un campione di 10 soggetti, mentre con 500 soggetti basterà avere un r di 0,088 (che corrisponde allo 0,077 % della varianza spiegata) per ottenere una significatività con $p < 0{,}05$. E con 10.000 soggetti sarebbe sufficiente un r addirittura di 0,0196: una varianza spiegata dello 0,039 ‰!

Hays (1981, p. 293) scrive che "virtualmente qualunque ricerca può giungere a risultati significativi se si usano abbastanza soggetti". Anche Cohen (1994) affronta l'argomento, affermando che se un campione è sufficientemente ampio (e quindi ha sufficiente potenza) qualunque effetto sarà significativo. Rifiutare quindi H_0 significherà solo avere un campione abbastanza ampio da permetterci di farlo.

Un altro chiaro esempio ci è dato dall'ANOVA. Come ogni studente sa bene, il valore di F varia al variare di due gradi di libertà, quelli relativi al trattamento (*between*) e quelli relativi all'errore (*within*). Ora, limitandoci per semplicità a un disegno a una via per campioni indipendenti, il valore di F è dato dal rapporto tra MS_b e MS_w, che a loro volta sono dati dai rapporti tra le somme dei quadrati degli scarti dalle relative medie (SS_b e SS_w) e i rispettivi gradi di libertà. Se noi abbiamo un trattamento a k livelli, i gradi di libertà (che nel calcolo dell'F sono comunque al denominatore) saranno sempre $k-1$, indipendentemente dalla grandezza dei campioni. I gradi di libertà within sono invece dati da $k(n-1)$, dove n è la grandezza di ogni singolo campione, e questo valore è invece al numeratore. È allora chiaro che aumentando la grandezza dei campioni si avrà un aumento proporzionalmente minore di MS_w rispetto a MS_b, e l'F si "gonfierà" per il puro effetto dell'aumento di n.

Ben noto è poi a ogni ricercatore il problema legato al χ^2. Ora, noi sappiamo che il χ^2 cresce in misura considerevole con l'aumentare della grandezza del campione (cfr. Long, 1983, p. 75). Di fatto, avendo due campioni di grandezza rispettivamente N_1 e N_2, il rapporto tra i due χ^2 è, a parità di condizioni,

$$\chi_2^2 = \chi_1^2 \frac{N_2 - 1}{N_1 - 1}. \tag{2.2}$$

Così, per esempio, se N_1 è uguale a 100 e N_2 è uguale a 1000, a parità di condizioni un χ^2 del primo campione di 1,64, a cui corrisponderebbe con un grado di libertà un p di 0,20, diventerebbe nel secondo campione 16,55, significativo con $p < 0{,}001$! Ora, questo probabilmente non costituisce un problema per tanti ricercatori che trattano con campioni di ampiezza ragionevole, o che di contro sono alla disperata ricerca della significatività. E non escludiamo che sia uno dei motivi alla base della straordinaria popolarità di questa statistica, semplice da calcolarsi anche a mano, e prodiga di, diciamo così, soddisfazioni. Lo è, però, e notevole, se si deve lavorare con grandi campioni, e lo scopo è quello non della falsificazione, ma della verifica dei modelli. Se passiamo, cioè, dalla logica della VeSN alla logica della bontà dell'adattamento.

È questo ad esempio il caso dei modelli di equazioni strutturali, di cui prototipico è il LISREL (cfr. Primi, 2002). È anche il caso, seppure in minor misura, dei modelli log-lineari. In tutti questi casi, l'uso del χ^2 rende molto spesso impossibile accettare il modello, proprio per la grandezza dei campioni usati, frequentemente di diverse centinaia di soggetti. E della stessa difficoltà soffrono molti altri indici di bontà dell'adattamento, a cominciare da quello forse più noto e di più largo impiego nell'analisi delle strutture di covarianza, e cioè il GFI.

Non riporteremo tutti i contributi di questi decenni sulla grandezza del campione, sia per necessità di brevità, sia poiché vorremmo poter dare per assodata e chiara a tutti una evidenza che riteniamo sia ormai da considerarsi patrimonio noto della comunità scientifica.

2.3.11. Il problema di Bonferroni

Fissare il livello α a 0,05 equivale a dire che si ha una probabilità di 0,95 di giungere a una conclusione "non significativa". Supponiamo però ora di verificare due ipotesi nulle indipendenti. In questo caso, se il livello α è per entrambe lo stesso, la probabilità che nessuna delle due sia significativa è $0,95 \times 0,95 = 0,90$. E se noi sottoponiamo a verifica 15 ipotesi nulle indipendenti, la probabilità che non ne sia significativa nessuna è $0,95^{15} = 0,46$. Ma ciò significa che vi è una probabilità di 0,54, e cioè superiore al caso, di dover respingere come significativa almeno un'ipotesi nulla. E se si trattasse di 20 ipotesi nulle, la probabilità sarebbe addirittura 1. È questo il famoso problema di Bonferroni (1936; per un recente trattamento, con una chiara indicazione degli esiti catastrofici a cui conduce non tenerne conto in ambito medico, cfr. Bland, 2000).

La correzione proposta allora classicamente da Bonferroni consiste semplicemente nel fissare un α corretto, $\alpha_c = \alpha / k$, dove k è il numero delle ipotesi nulle sottoposte a verifica. Così, nell'esempio posto prima delle 15 ipotesi nulle, si avrebbe $\alpha_c = 0,05 / 15 = 0,003$.

La correzione di Bonferroni ha peraltro suscitato anche delle critiche, perché è apparsa a diversi autori, specie in ambito biometrico, eccessivamente conservatrice. In realtà, le critiche si capiscono meglio alla luce proprio di quanto scrive uno dei più tenaci avversari della correzione, Perneger (1998), che osserva che essa è corretta nel "quadro originale della teoria dei test statistici proposta da Neyman e Pearson (1928)", teoria che si proponeva di "aiutare le decisioni in situazioni ripetitive". La correzione di Bonferroni "segue la logica originale dei test statistici a sostegno di decisioni ripetute, ma è di scarso aiuto nel determinare quel che i dati dicono in uno studio particolare". Appunto, QED.

Di fatto, rileviamo che, senza alterare la logica del ragionamento di Bonferroni, esistono comunque forme di correzione meno conservatrici, come quella proposta da Hochberg (1998).

Per mostrare il ridicolo in cui si cade quando non si vogliano correggere i dati sulla base della correzione di Bonferroni, citeremo un caso famoso, quello della ricerca di Michel Gauquelin (1955; ma vedi anche Gauquelin, Gauquelin e Eysenk, 1979) sugli influssi astrologici sulla personalità, che avrebbero portato a dimostrare il ben noto (ad astrologi ed associati) "effetto Marte", e cioè una correlazione positiva tra la presenza della casa astrologica di Marte nelle carte astrologiche di un individuo e la predisposizione allo sport. Ci scusiamo per l'eventuale imprecisione della terminologia astrologica che impieghiamo, ma è questa una disciplina di cui poco sappiamo e ancora meno vorremmo apprendere. Ciò che qui preme rilevare è che questo effetto sarebbe stato riscontrato sulla base della bellezza di 60 confronti, con 12 campioni di professionisti e

5 "case". È quindi chiaro che sarebbe stato assolutamente improbabile che Gauquelin, non applicando la correzione di Bonferroni (ma anche correzioni meno conservative), non avesse trovato alcune correlazioni significative. Con la correzione di Bonferroni si dovrebbe avere un α_c di 0,0008. E si badi che qui non può valere l'obiezione di Perneger: siamo proprio di fronte, infatti, a "decisioni ripetute".

2.3.12. La replicabilità dei risultati

Il fatto che l'ipotesi nulla sia falsificata viene, in molte ricerche, interpretato come una riprova della replicabilità dei risultati ottenuti. È questa quella che Carver (1978) ha chiamato la "fantasia della replicabilità". Non capiamo in base a quale logica affermare che H_0 è falsa possa dire qualcosa sul fatto che i risultati sono replicabili. A meno che non si ritenga che tale falsificazione dimostri senza ombra di dubbio che i due campioni (per esempio) che si stanno studiando appartengono per il solo fatto di questa falsificazione a due diversi universi. Ma per poterlo affermare, occorrerebbe che la potenza del test statistico fosse soddisfacente (almeno di .80, cfr. Cohen 1988). Solo in questo caso potremmo sentirci autorizzati a pensare che, nelle stesse condizioni, una replica attingerebbe sempre a due universi distinti, e agli stessi universi dei due campioni studiati.

In realtà, come nota Thompson (1993), sono numerose le tecniche che consentono di valutare con una buona attendibilità se la ricerca ha fornito dei risultati replicabili, ma nessuno passa attraverso la verifica della significatività dell'ipotesi nulla. I test d'ipotesi infatti non valutano la probabilità che i risultati del campione descrivano la popolazione, ma partono già dall'assunto che H_0 sia vera, e cioè che H_0 descriva esattamente la popolazione, e solo successivamente esaminano la probabilità del campione rispetto alla popolazione (Thompson, 1996). Scrive Cohen (1994, p. 997) sui test di significatività statistica: "non ci dicono quello che vogliamo sapere e noi tanto lo vogliamo sapere che, per disperazione, crediamo nonostante tutto che ce lo dicano!".

Tra le tecniche più potenti e più usate oggi a disposizione, che assumono solo l'indipendenza dei dati nell'estrazione del campione, ma non richiedono assunzioni sulla distribuzione, vi sono quelle di ricampionamento (*resampling*), come il *bootstrap* e il *jackknife*, con le quali da un solo campione si formano nuovi campioni con la esclusione di un valore alla volta, o con un campionamento con rimpiazzo (cfr. Efron e Gong, 1983). Di questo parleremo più distesamente nel Cap. 6.

2.3.13. Intensità e direzione dell'effetto

Nella VeSN, la significatività non può affermare altro che l'esistenza e la direzione di una differenza. Se la media del campione soggetto al trattamento A è maggiore di quella del campione soggetto al trattamento B e se l'ipotesi nulla è falsificata, evidentemente possiamo affermare che A produce un aumento dell'effetto rispetto a B.

Ma un aumento di quanto? Questo non possiamo dirlo. La falsificazione dell'ipotesi nulla ci permette così infatti di affermare esclusivamente una relazione ordinale. La valutazione dell'intensità dell'effetto (della grandezza della differenza) è peraltro nella maggior parte dei casi un elemento di importanza cruciale per dare senso alla ricerca. Come possa essere misurato lo vedremo nel prossimo capitolo, ma è paradossale che l'attenzione dei ricercatori, focalizzata sul valore di probabilità, li induca a trascurare questa misura, per cui hanno già a disposizione tutti i dati.

2.3.14. Significatività statistica e significatività pratica

Come afferma McLean (2001), se uno degli argomenti più dibattuti è il ruolo dei test d'ipotesi in statistica e il loro significato, al dibattito corrisponde una notevole confusione anche nelle menti degli studenti di discipline statistiche. Didatticamente, poi, la cosa acquista degli aspetti drammatici. Anche gli studenti che riescono a risolvere quesiti riguardanti test d'ipotesi durante un compito in classe, spesso hanno notevoli difficoltà nello spiegare concetti quali il livello di significatività, lasciando il docente nel dubbio se alla base vi sia di una difficoltà di espressione o di comprensione (si veda a proposito Shaughnessy, 1983; Falk, 1986; Pollard & Richardson, 1987).

Ammesso che una ricerca fornisca dei risultati statisticamente significativi, ciò non significa che quello stesso risultato abbia senso nel mondo reale o sia davvero "importante" o "informativo" anche soltanto rispetto alle premesse dello studio stesso. In altre parole, non è detto che un risultato statisticamente significativo sia anche un risultato che abbia una significatività pratica (Sullivan, 2000). Non si tratta di un problema nuovo. Già nel 1931 Tyler metteva in guardia dal confondere tra significatività statistica e significatività "sociale".

Thompson (2003) distingue tra tre tipi di significatività: statistica, pratica e clinica. La significatività statistica non dice se i risultati sono importanti. Infatti ci sono eventi rari e insoliti (con p molto basso) che non hanno nessuna rilevanza e altri molto frequenti e probabili che sono importantissimi. È compito del ricercatore tenere sempre ben presente cosa è importante o rilevante ai fini della ricerca e della divulgazione, senza affidarsi alla mera significatività statistica di un risultato, perdendo la capacità di ragionare sul senso delle proprie premesse e conclusioni.

2.3.15. La logica del test di significatività e la logica formale

Questo problema è affrontato con chiarezza proprio da Cohen (1994) e, per quanto sia stato ampiamente descritto da vari autori in tempi sia precedenti che successivi a questo articolo, faremo riferimento agli esempi discussi da questo autore per chiarire meglio questo aspetto.

Precisiamo, comunque, che curiosamente i sillogismi sono presentati da Cohen (ma anche da altri autori) con la premessa maggiore sotto forma di implicazione logica

(se, allora), il che a rigore non è corretto, e può portare a ridicoli paradossi (solo che si tenga conto che una implicazione è sempre vera, purché la premessa sia falsa).

La logica che sottende il rifiuto dell'ipotesi nulla e, di conseguenza, l'accettazione dell'ipotesi alternativa vorrebbe richiamarsi alle regole del ragionamento sillogistico deduttivo. Tuttavia questo modo di considerare la logica che sottende il rifiuto dell'ipotesi nulla è errato, non vi sono, infatti, i presupposti teorici per potersi richiamare al sillogismo ed alle regole della logica formale.

Cerchiamo di chiarire meglio questa affermazione. Nella sua forma classica il sillogismo ha una struttura del tipo:

Tutti gli A sono B
S non è B

∴ Allora, S non è A

laddove, secondo le ben note regole del modus tollens, la negazione del conseguente comporta la negazione dell'antecedente.

Apparentemente il ragionamento del test d'ipotesi potrebbe essere sintetizzato da un sillogismo del tipo:

Per ogni ipotesi nulla vera, il fenomeno non può avere luogo.
Il fenomeno ha avuto luogo.

∴ Allora, l'ipotesi nulla è falsa.

Se così fosse, avremmo un ragionamento corretto dal punto di vista della logica formale. Tuttavia questo modo di procedere non corrisponde esattamente al ragionamento del test d'ipotesi, poiché esso è, per sua natura, un tipo di ragionamento probabilistico. Dovremmo, infatti, riformulare il sillogismo precedente alla luce di questa caratteristica imprescindibile del test d'ipotesi e ne verrebbe fuori quanto segue:

Per ogni ipotesi nulla vera, il fenomeno è altamente improbabile.
Il fenomeno ha avuto luogo.

∴ Allora, l'ipotesi nulla è altamente improbabile.

Appare chiaro, a questo punto, che dando al sillogismo la forma probabilistica esso non può più essere considerato valido, vengono violate le regole della logica formale. Se nella premessa viene introdotta la probabilità, e quindi questo termine perde in assolutezza, la negazione dell'antecedente non conterrà più la certezza della negazione del conseguente. La caratteristica di probabilità associata inizialmente al conseguente dovrebbe poi essere attribuita, per negazione dello stesso, al conseguente, laddove tutto questo non era certamente contenuto nelle premesse. La cosa è facilmente comprensibile se le due premesse sono:

Per ogni cittadino italiano, essere Presidente della Repubblica è molto improbabile.
Ciampi è Presidente della Repubblica.

La conclusione insensata sarebbe:

∴ Allora, è molto improbabile che Ciampi sia cittadino italiano.

2.3.16. Il problema della linearità dei modelli

Il classico modello lineare nella statistica fisheriana (ad esempio, nell'ANOVA) consiste nell'assumere che il dato che viene osservato sia una composizione lineare di almeno tre effetti: l'appartenenza del dato a un determinato universo (rappresentata dalla media μ: su questa assunzione, e sulla metafisica soggiacente, che si deve, con scarse variazioni successive, a Quételet, 1835, cfr. Hawking, 1990), l'influenza del *j*-esimo trattamento (α_j) e l'errore proprio dell'*i*-esimo soggetto (e_i):

$$y_{i,j} = \mu + \alpha_j + e_i .$$

Ora, questo modello (con la conseguente partizione della somma dei quadrati) regge proprio in quanto lineare. Ma nulla garantisce tale linearità. Non solo, in molti casi possiamo essere assolutamente certi che i rapporti (per esempio, nell'interazione tra trattamenti) non sono affatto lineari.

Perché il modella regga, peraltro, sono necessarie almeno altre tre assunzioni supplementari. La prima dice che gli errori si distribuiscono normalmente. La seconda afferma che gli errori non sono correlati tra loro. La terza, infine, che la varianza delle distribuzioni d'errore sono uguali per ogni condizione sperimentale.

Anche qui, nulla garantisce che le cose vadano in questo modo. È per esempio ben noto che nelle ricerche panel, con rilevazioni prese a distanza di tempo, gli errori presentano una correlazione tra di loro più o meno elevata. Micceri (1989) ha studiato 440 grandi campioni su cui erano state condotte ricerche utilizzando i più classici test parametrici, dall'ANOVA al *t* di Student, e ha rilevato che nessuno di questi si distribuiva normalmente.

2.3.17. La formazione post-hoc *dei campioni*

Abbiamo esitato a inserire questo paragrafo, tanto macroscospica è l'insensatezza di chi cade in questo errore. Peraltro, dovendo seguitare a riscontrare questo tipo di errore, chiediamo l'indulgenza del lettore se abbiamo ritenuto di doverlo segnalare. Supponiamo che si voglia condurre una ricerca in cui si vuole verificare se il quoziente intellettivo influenza significativamente la memoria a breve termine. Abbiamo allora un gruppo di adolescenti a cui sottoponiamo, per esempio, la scala

Wechsler, e ne rileviamo tutti i valori di QI. Dividiamo quindi il campione in due sottocampioni, composti da ragazzi rispettivamente con QI elevato e con QI basso, sulla base della mediana dei punteggi rilevati. A questo punto, però, presi da uno scrupolo (idiota), decidiamo di verificare con un test di significatività (per esempio, il solito *t* di Student) se le differenze tra i QI medi dei due sottogruppi sono effettivamente significative. Caspita, lo sono per davvero! E con un livello di probabilità più che soddisfacente. Tutti contenti, procediamo allora nella nostra ricerca. Ora, è evidente che, a meno di una ridicola esiguità dei campioni, un confronto tra le medie di due gruppi, con tutti i valori di un gruppo superiori a tutti i valori del secondo gruppo, non può non risultare significativo, ma il dato è anche privo di qualsiasi valore informativo.

L'esempio è lampante. Ma l'errore logico alla base di questo esempio è lo stesso in tutti i casi di selezione *post-hoc* dei campioni. Ogni test di significatività su campioni formati *post-hoc* va pertanto escluso.

2.4. Le raccomandazioni dell'American Psychological Association

Sulla scia del dibattito in merito alla legittimità dell'applicazione dei test di significatività, e soprattutto grazie all'ampio dibattito suscitato dalla pubblicazione del celebre articolo di Cohen "*The earth is round* ($p < 0.05$)" nel 1994, il Consiglio Scientifico dell'*American Psychological Association* (APA) ritenne di dover intervenire per dire una parola chiarificatrice in proposito. Tra l'altro, nel 1996, alla convenzione dell'APA si scatenò un dibattito talmente acceso su questo problema, da dare l'impressione, come disse argutamente Abelson (1997), che "un gruppo di attivisti radicali" avesse preso in ostaggio "10 statistici e 6 direttori di riviste, al canto di 'Sostieni il bando totale dei test [di significatività]!' e 'Annulla la nulla!'". Venne nominata una commissione di esperti, denominata *Task Force on Statistical Inference* (TFSI), a cui l'APA affidò il mandato di studiare approfonditamente la questione e di fornire consigli ed alternative possibili. La Task Force era così composta da esperti in varie discipline, in modo da garantire il confronto tra un'ampia gamma di conoscenze e punti di vista. Ne facevano parte statistici, insegnanti di statistica, direttori di riviste, autori di libri di statistica, esperti di informatica e vari altri specialisti in materia.

Nei due anni seguenti la Task Force si è incontrata due volte. La prima riunione, avvenuta il 14 e 15 dicembre del 1996 all'aeroporto di Newark, dette origine ad un report introduttivo in cui vennero esplicitate le linee guida che la commissione avrebbe utilizzato nella propria indagine.

Innanzitutto la Task Force individuò due oggetti di studio interconnessi ma distinti:

1. Il ruolo dei test di significatività dell'ipotesi nulla nella ricerca psicologica.
2. Le modifiche occorse nel tempo in merito al trattamento dei dati in psicologia.

All'interno di questa seconda questione la commissione individuò, inoltre, quattro punti focali che sarebbero stati studiati in modo approfondito:

- l'utilizzo di approcci che incrementino la qualità nell'utilizzo dei dati e proteggano dalle potenziali interpretazioni errate di risultati quantitativi;
- la necessità, talvolta eccessiva e fuorviante, di studi che siano in grado di produrre teorie;
- l'utilizzo di disegni di ricerca e strategie di analisi che siano parsimoniosi e sufficienti;
- l'attenzione di fronte alle analisi dei dati computerizzate, laddove programmi molto avanzati e sofisticati aiutano i ricercatori nel loro lavoro ma vengono talvolta santificati, col rischio che coloro che li utilizzano non abbiano il controllo sui processi matematici sottesi o perdano di vista la ragionevolezza dell'esito rispetto al punto di partenza.

In seguito al secondo incontro, avvenuto come detto due anni dopo, la Task Force auspicò una serie di azioni da seguire per migliorare lo stato delle cose, suggerendo in primis una revisione della sezione relativa alla statistica dell'*American Psychological Association Publication Manual* (APA, 1994). Il Consiglio Scientifico dell'APA deliberò affinché, prima che la Task Force si occupasse della revisione del manuale, le considerazioni che la stessa aveva steso fino a quel momento fossero pubblicate quanto prima sulla rivista *American Psychologist* (Wilkinson e TFSI, 1999). Gli aspetti che più ci interessano di questo *report* sono relativi al metodo e alla presentazione dei risultati.

Per ciò che attiene al metodo, la TFSI faceva le seguenti raccomandazioni:

1. Disegno della ricerca: deve essere reso il più chiaro e comprensibile possibile e, laddove esistano più obiettivi, devono essere rese evidenti le priorità tra essi esistenti.
2. Popolazione: deve essere definita in ogni suo aspetto e caratteristica. La popolazione di riferimento può essere costituita non solo dai partecipanti ma da un insieme di stimoli o di ricerche. Inoltre se vi sono gruppi di controllo e di confronto anch'essi devono essere definiti con chiarezza.
3. Campione: descrivere le procedure di campionamento e chiarire bene criteri di inclusione o esclusione.
4. Assegnamento: può essere casuale o meno. Il primo garantisce la possibilità di compiere inferenze causali forti, senza l'interferenza di variabili intervenienti o confondenti. Nel secondo caso dovrà essere posta particolare attenzione al controllo di questi fattori incontrollati, descrivendo accuratamente anche i metodi utilizzati per attenuare queste fonti di errore.
5. Misurazione: vanno definite accuratamente le *variabili*, le loro relazioni con gli obiettivi dello studio ed i metodi con cui vengono misurate. Se viene utilizzato un *questionario* ne devono essere riportate le proprietà psicometriche. Deve essere adeguatamente descritta anche la *procedura* di misurazione e le metodiche prescelte per ovviare a problemi di *noncompliance*, *dropout*, decesso dei soggetti o altri fattori. Inoltre deve essere adeguatamente esplicitata la *grandezza del campione* ed il processo che ha portato a stabilire questa grandezza.

Passiamo alla descrizione delle questioni affrontate nella sezione relativa ai risultati:

1. Complicazioni: devono essere accuratamente riportate tutte le complicazioni che hanno avuto luogo durante la raccolta dei dati, cercando di fornire anche

suggerimenti utili ad evitare questi problemi in futuro. Deve inoltre essere verificata, statisticamente e graficamente, l'influenza che questo tipo di anomalie (dati mancanti, *outliers*) potrebbero avere sull'andamento dei propri dati.

2. Analisi dei dati: la prima e più importante raccomandazione è quella di scegliere *analisi semplici*, *parsimoniose* e *sufficienti*, senza andare necessariamente a cercare metodi sofisticati che probabilmente poco aggiungerebbero alla nostra conoscenza del fenomeno. È inoltre importante la *scelta dei programmi di analisi dei dati*. Anche in questo caso si raccomanda di utilizzare programmi di cui si conoscano le procedure di analisi e di cui si sia in grado di verificare e comprendere a pieno i risultati. Un altro punto è la verifica degli *assunti*: è necessario assicurarsi che gli assunti richiesti per uno specifico tipo di analisi siano soddisfatti dai propri dati (i residui, ad esempio, dovranno essere esaminati con attenzione). Per quanto riguarda la *verifica di ipotesi*, è auspicabile rinunciare alla scelta dicotomica tra accettare o rifiutare l'ipotesi nulla e riportare semplicemente il valore di *p* o, ancor meglio, gli *intervalli di fiducia*. Il consiglio è anche quello di riportare indici che forniscano una stima della *grandezza dell'effetto*. Torneremo tra breve su questi ultimi concetti per descriverli in modo più approfondito. Ulteriori consigli in merito alla analisi dei dati riguardano l'accuratezza nella gestione della *molteplicità* delle variabili considerate e dei legami tra esse esistenti e l'attenzione che va posta nell'affermare relazioni di *causalità* tra le proprie variabili, tanto più quando si lavora con disegni di ricerca non randomizzati. Infine si raccomanda l'utilizzo delle *figure* nell'esposizione dei propri risultati poiché maggiormente comprensibili e piacevoli all'occhio del lettore. Figure possibilmente semplici, eventualmente accompagnate da tabelle, preferibilmente corredate dalla rappresentazione grafica degli intervalli di fiducia, laddove possibile.

Per esigenze di brevità, non riporteremo analiticamente quanto esposto dalla Task Force in merito alla discussione dei risultati. I consigli fondamentali riguardano l'accuratezza nella valutazione della credibilità, generalizzabilità e cogenza delle interpretazioni che si possono trarre dai propri risultati e la parsimonia con cui debbano essere elargite speculazioni che spesso sembrano illudersi di spiegare più di quanto un unico studio possa avere la pretesa di fare.

All'interno di questo panorama, comunque, quelle che a noi interessa approfondire sono le tematiche più vicine alla VeSN.

Particolarmente interessante, a questo proposito, ci è sembrato l'invito ad evitare decisioni dicotomiche di accettazione o rifiuto dell'ipotesi nulla e l'esortazione esplicita ai ricercatori a non utilizzare espressioni infelici quali "*accettare l'ipotesi nulla*". Piuttosto si suggerisce di riportare non tanto, o meglio non solo, il valore di *p*, ma anche gli intervalli di fiducia. Questi costituiscono, secondo la Task Force, degli ottimi indicatori della stabilità dei risultati attraverso studi diversi, proprio perché possono essere confrontati direttamente. Non solo, la Task Force raccomanda fortemente anche l'utilizzo di indici che forniscano una stima della grandezza dell'effetto. Sull'importanza di questo punto la Task Force è molto chiara: riportare la grandezza dell'effetto è essenziale per valutare la bontà di una ricerca, poiché garantisce la possibilità di valutare la stabilità dei risultati rispetto ai campioni, ai disegni della ricerca ed alle analisi condotte.

Capitolo 3
La grandezza dell'effetto

3.1. Il problema della grandezza dell'effetto

Come alternativa all'utilizzo delle procedure di VeSN, sono stati studiati diversi tipi di indici che vengono genericamente definiti come "misure della grandezza dell'effetto" (*effect magnitude measures,* Maxwell e Delaney, 1990) e forniscono il grado con cui la variabile dipendente è controllata, predetta o spiegata dalla variabile indipendente. Il vantaggio di questi indici, rispetto a quelli forniti dalle analisi di verifica della significatività, consiste nella loro minore dipendenza dalla grandezza del campione. In realtà questi indici possono essere più propriamente suddivisi in due categorie (Nix & Barnette, 1998) in base al fatto che misurino: (i) la forza della relazione tra le variabili dipendenti e indipendenti (misure di associazione); (ii) la grandezza dell'effetto in senso stretto.

Vediamo adesso nel dettaglio come vengono calcolati alcuni indici e le loro caratteristiche più importanti.

3.2. Misure di associazione

Le misure di associazione vengono solitamente utilizzate per esaminare proporzioni di varianza (Maxwell e Delaney, 1990) o per controllare quanto della variabilità della variabile dipendente è associato alla variazione della variabile indipendente.

3.2.1. L'associazione nell'ANOVA

Le misure della grandezza dell'effetto che si utilizzano per l'analisi della varianza stimano il grado di associazione tra l'effetto (sia esso un effetto principale oppure una interazione) e la variabile dipendente. Possiamo pensare ad esse come a delle misure di correlazione tra l'effetto e la variabile dipendente. Se il valore dell'indice è elevato al quadrato possiamo, inoltre, interpretarlo come la proporzione di varianza della variabile dipendente che viene spiegata dall'effetto stesso.

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia,*
ISBN 88-8453-226-4,

Nell'ANOVA vengono solitamente utilizzati quattro tipi di misure della grandezza dell'effetto:

1. Eta quadro (η^2)
2. Eta quadro parziale (η_p^2)
3. Omega quadro (ω^2)
4. Correlazione intraclasse (ρ_I)

Gli indici ω^2 e ρ_I sono stime del grado di associazione calcolate nella popolazione, mentre η^2 e η_p^2 sono stime calcolate nel campione.

È importante sapere che l'indice η_p^2 viene fornito in output da alcune versioni del programma SPSS (versione 9.0 e 8.0), mentre le ultime versioni (dal 10.0 in poi) forniscono l'indice η^2. In entrambi i casi per ottenere questi indici nell'output è necessario, durante la specificazione delle opzioni per il modello di ANOVA, eseguire la procedura: *Options* → *Estimates of effect size*.

L'**eta quadro** (η^2), detto *correlation ratio*, rappresenta la proporzione di varianza totale che è da attribuire all'effetto ed è dato dal rapporto tra la varianza dovuta all'effetto ($SS_{effetto}$) e la varianza totale (SS_{totale}). La formula per il calcolo dell'η^2 è la seguente:

$$\eta^2 = \frac{SS_{effetto}}{SS_{totale}} \tag{3.1}$$

ESEMPIO 3.1

Poniamo che venga svolta una ricerca sulla popolazione giovanile e che si voglia verificare se età e classe sociale sono due fattori che influiscono sul rendimento scolastico. Nella tabella 3.1 riportiamo i risultati dell'ANOVA.

Fonte di variazione	**Somma dei quadrati**	***gl***	**Media dei quadrati**	***F***	***p***
Modello corretto	280,000	5	56,000	3,055	0,036
Intercetta	2400,000	1	2400,000	130,909	0,000
Età	24,000	1	24,000	1,309	0,268
Classe Sociale	112,000	2	56,000	3,055	0,072
Età × Classe Sociale	144,000	2	72,000	3,927	0,038
Errore	330,000	18	18,333		
Totale	3010,000	24			
Totale Corretto	610,000	23			

Tabella 3.1 – *Risultati ANOVA*

Sostituendo nella formula 3.1 i valori della tabella 3.1, possiamo calcolare l'indice η^2 per ciascuna delle variabili indipendenti e per le interazioni tra di esse.

È importante sapere che per il calcolo di questo indice non si utilizza come SS_{totale} il valore della somma dei quadrati corrispondente a *Totale* nella tabella di *output* dell'ANOVA, bensì si utilizza il valore corrispondente a *Totale corretto*. Vediamo come procedere:

$$\eta^2_{età} = \frac{24}{610} = 0,039$$

$$\eta^2_{classe} = \frac{112}{610} = 0,184$$

$$\eta^2_{età \times classe} = \frac{144}{610} = 0,236$$

Come si può vedere, il procedimento risulta estremamente semplice ed immediato. Nella tabella 3.2 riassumiamo i valori di η^2 calcolati.

Effetto	$SS_{effetto}$	SS_{totale} (Totale Corretto)	η^2
Età	24	610	0,039
Classe Sociale	112	610	0,184
Età × Classe Sociale	144	610	0,236

Tabella 3.2 – *Calcolo di* η^2 *utilizzando le misure fornite dai risultati dell'ANOVA*

Come abbiamo visto nella tabella 3.1, l'interazione 'età × classe sociale' era significativa; utilizzando l'indice η^2 possiamo dire che questa interazione spiega circa il 24% della variabilità totale del punteggio di rendimento scolastico. Per rendere la cosa più chiara possiamo rappresentare graficamente le proporzioni di varianza totale attribuibili a ciascun effetto. Allo scopo ci serviremo di un grafico a torta (fig. 3.1). Il cerchio rappresenta la varianza totale, gli spicchi la percentuale di varianza attribuita a ciascun effetto o all'errore; questa percentuale è, a sua volta, data dal valore di η^2. In questo esempio la varianza dovuta all'errore spiega più della metà della varianza totale, cioè della variabilità riscontrata nei punteggi della variabile dipendente, mentre l'unico effetto significativo (l'interazione 'età × classe sociale') spiega il 24% della stessa.

Uno dei problemi legati all'utilizzo di η^2 è dato dal fatto che il valore di questo indice rispetto ad uno degli effetti dipende dal numero e dalla grandezza degli altri effetti che vengono indagati dall'ANOVA. Se, ad esempio, avessimo incluso nel dise-

gno della nostra ricerca una terza variabile indipendente, probabilmente la grandezza dell'effetto attribuita all'interazione 'età × classe sociale' sarebbe stata inferiore, mentre la varianza attribuita alla stessa interazione sarebbe rimasta inalterata.

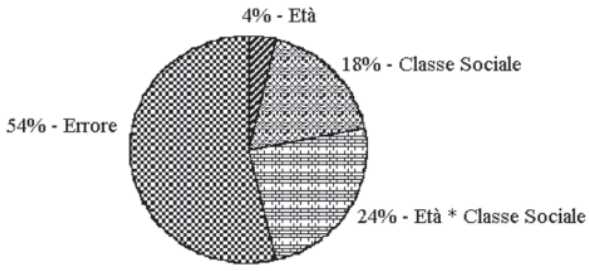

Figura 3.1 – *Percentuali di varianza spiegata dai singoli fattori, dall'interazione e dall'errore, calcolate sulla base dell'indice* η^2.

L'**eta quadro parziale** (η_p^2) differisce dall'η^2 poiché al denominatore la formula non utilizza la varianza totale (SS_{totale}), bensì la somma tra la varianza dovuta all'effetto ($SS_{effetto}$) e la varianza dovuta all'errore (SS_{errore}). In formula:

$$\eta_p^2 = \frac{SS_{effetto}}{\left(SS_{effetto} + SS_{errore}\right)} \tag{3.2}$$

ESEMPIO 3.2

Procediamo al calcolo dell'indice η_p^2 utilizzando i dati dell'esempio precedente. Partendo dai dati della tabella 3.1 e sostituendo i valori che ci interessano nella formula 3.2, otterremo un η_p^2 per ciascuna delle variabili indipendenti e per l'interazione:

$$\eta_{p,età}^2 = \frac{24}{(24+330)} = \frac{24}{354} = 0,068$$

$$\eta_{p,classe}^2 = \frac{112}{(112+330)} = \frac{24}{442} = 0,253$$

$$\eta_{p,età\times classe}^2 = \frac{112}{(144+330)} = \frac{24}{474} = 0,304.$$

Anche in questo caso il procedimento è molto semplice e rapido. Riassumiamo i risultati nella tabella 3.3.

Effetto	$SS_{effetto}$	SS_{errore}	$SS_{effetto} + SS_{errore}$	η_p^2
Età	24	330	354	0,068
Classe Sociale	112	330	442	0,253
Età × Classe Sociale	144	330	474	0,304

Tabella 3.3 – *Calcolo di* η_p^2 *utilizzando le misure fornite dai risultati dell'ANOVA*

In questo caso, la rappresentazione grafica delle percentuali di varianza prevede che venga costruito un grafico a torta per ciascuno degli effetti indagati (fig. 3.2). In questo troviamo anche un limite dell'indice η_p^2, cioè il fatto che i valori che assume nei singoli effetti non possono essere sommati. È infatti possibile che la loro somma sia anche superiore ad 1. Questo perché i valori assunti da questo indice non rappresentano la quantità di varianza della variabile dipendente spiegata dalle variabili indipendenti considerate complessivamente.

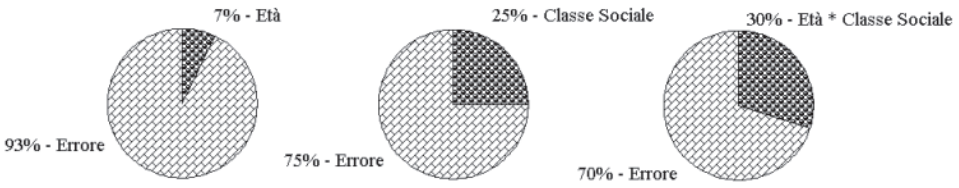

Figura 3.2 – *Percentuali di varianza spiegata dai singoli fattori, dall'interazione e dall'errore, calcolate sulla base dell'indice* η_p^2.

L'indice **omega quadro** (ω^2) fornisce una stima della quantità di varianza spiegata dalla variabile indipendente nella popolazione. La formula che fa riferimento ai parametri della popolazione è la seguente:

$$\omega^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_{S/A}^2} \tag{3.3}$$

dove σ_A^2 è la varianza di trattamento e $\sigma_{S/A}^2$ è la varianza d'errore.

Come si può comprendere osservando questa formula, se la varianza di trattamento è uguale a 0 allora ω^2 è, a sua volta, uguale a 0, il che significa, giustamente, assenza di effetto. D'altro canto se la varianza di trattamento è superiore a 0 allora ω^2 varia tra 0 ed 1. Infine se la varianza d'errore è uguale a 0 allora ω^2 è uguale ad 1, ovvero il trattamento è l'unico fattore a determinare il cambiamento ed ha un effetto, per così dire, "totale".

La formula esposta in precedenza è basata, come abbiamo detto, sui parametri della popolazione che però solitamente sono ignoti e dovranno quindi essere stimati a partire dai dati campionari. Per fare questo, cioè per calcolare una stima dell'indice ω^2, indicata dal simbolo $\hat{\omega}^2$, ci possiamo servire di una delle due seguenti formule:

$$\hat{\omega}^2 = \frac{\left[SS_{effetto} - \left(df_{effetto}\right)\left(MS_{errore}\right)\right]}{MS_{errore} + SS_{totale}} \tag{3.4}$$

$$\hat{\omega}^2 = \frac{(k-1)(F-1)}{(k-1)(F-1)+kn} \tag{3.5}$$

Nella seconda formula *k* corrisponde al numero dei gruppi, *F* al valore del test ed *n* alla numerosità dei gruppi. Ricordiamo che queste formule non possono essere utilizzate per i disegni a misure ripetute.

ESEMPIO 3.3

Basandoci sui dati dell'esempio seguito fino ad ora, vediamo come calcolare questo indice. Utilizzando i dati della tabella 3.1 e sostituendoli in modo adeguato nella formula 3.4, possiamo facilmente calcolare i valori dell'indice $\hat{\omega}^2$.

Come nel caso dell'indice η^2, anche qui il valore corrispondente ad SS_{totale} non è il valore della somma dei quadrati *Totale* nella tabella di *output* dell'ANOVA, bensì si utilizza il valore *Totale corretto*. Vediamo come procedere:

$$\hat{\omega}^2_{età} = \frac{\left[24 - (1)(18,333)\right]}{18,333 + 610} = \frac{5,667}{628,333} = 0,009$$

$$\hat{\omega}^2_{classe} = \frac{\left[112 - (2)(18,333)\right]}{18,333 + 610} = \frac{112 - 36,666}{628,333} = \frac{75,334}{628,333} = 0,119$$

$$\hat{\omega}^2_{età \times classe} = \frac{\left[144 - (2)(18,333)\right]}{18,333 + 610} = \frac{144 - 36,666}{628,333} = \frac{107,334}{628,333} = 0,171$$

Riassumiamo i valori ottenuti nella tabella 3.4.

La rappresentazione grafica di questo indice deve essere compiuta costruendo tanti grafici a torta quanti sono gli effetti indagati, in modo analogo a come abbiamo proceduto nel caso dell'indice η^2_p .

Effetto	$SS_{effetto}$	$gl_{effetto}$	MS_{errore}	$SS_{totale\ (Corretto)}$	$\hat{\omega}^2$
Età	24	1	18,333	610	0,009
Classe Sociale	112	2	18,333	610	0,119
Età × Classe Sociale	144	2	18,333	610	0,171

Tabella 3.4 – *Calcolo di* $\hat{\omega}^2$ *utilizzando le misure fornite dai risultati dell'ANOVA*

Dato che, come abbiamo detto in precedenza, entrambi gli indici eta quadri sono calcolati sui dati campionari, mentre l'indice ω^2 è calcolato sulle stime dei dati della popolazione, il suo valore sarà sempre più piccolo degli eta quadri. Vediamo, infatti, nella tabella 3.5 che questo andamento emerge con chiarezza dal confronto diretto tra gli indici da noi calcolati.

Effetto	η^2	η_p^2	$\hat{\omega}^2$
Età	0,039	0,068	0,009
Classe Sociale	0,184	0,253	0,119
Età × Classe Sociale	0,236	0,304	0,171

Tabella 3.5 – *Confronto tra i valori degli indici* η^2, η_p^2 *e* $\hat{\omega}^2$ *ottenuti partendo dagli stessi dati*

Il vantaggio legato a questo indice sta nel fatto che esso non dipende né dalla dimensione di *F*, né dalla probabilità di *F*, supera perciò i principali limiti del test *F* di Fisher.

L'indice di **correlazione intraclasse** (ρ_I) fornisce una stima del grado di associazione tra la variabile indipendente e la variabile dipendente nella popolazione per un modello *random effects*. In formula:

$$\rho_I = \frac{\left(MS_{effetto} - MS_{errore}\right)}{\left[MS_{effetto} + \left(df_{effetto}\right)\left(MS_{errore}\right)\right]} \tag{3.6}$$

Il quadrato di questo indice può essere utilizzato come stima della quantità di varianza della variabile dipendente spiegata dalla variabile indipendente.

Esempio 3.4

Il modello utilizzato negli esempi precedenti è *fixed effects*, quindi non è possibile calcolare l'indice ρ_I partendo dai dati presi in esame. Poniamo di svolgere una ricerca per verificare

l'efficacia di un trattamento farmacologico nei disturbi da attacchi di panico. Suddividiamo il campione di pazienti in tre gruppi: pazienti a cui viene somministrato il farmaco, pazienti a cui viene somministrato il placebo e pazienti senza trattamento. Durante il periodo di trattamento viene misurato, per ciascun paziente, il numero di attacchi di panico che hanno avuto luogo. Riassumiamo nella tabella 3.6 i risultati dell'ANOVA univariata a tre gruppi.

	Somma dei quadrati	***gl***	**Media dei quadrati**	***F***	*p*
Tra i gruppi	2401,111	2	1200,556	162,726	0,000
Entro i gruppi	309,867	42	7,378		
Totale	2710,978	44			

Tabella 3.6 – *Risultati ANOVA*

Avendo a disposizione tutti i dati necessari e sostituendoli nella formula 3.6, possiamo facilmente ricavare il valore dell'indice ρ_I.

$$\rho_I = \frac{(1200,556 - 7,378)}{[1200,556 + (2)(7,378)]} = \frac{1193,178}{1200,556 + 14,756} = \frac{1193,178}{1215,312} = 0,98$$

Da questo è possibile effettuare una ulteriore trasformazione:

$$\rho_I^2 = 0,98^2 = 0,96.$$

Questo risultato significa che il 96% della varianza della variabile dipendente, quindi il numero di attacchi di panico, è spiegata dall'associazione con la variabile indipendente, cioè dal fatto di appartenere ad un gruppo di trattamento e non ad un altro.

3.2.2. Associazione e correlazione

I coefficienti di correlazione possono essere considerati, essi stessi, delle misure di grandezza dell'effetto. Come vedremo meglio nel paragrafo relativo all'interpretazione della grandezza dell'effetto, il coefficiente di correlazione può essere utilizzato per stimare la quota di miglioramento che un trattamento ha comportato (Rosenthal e Rubin, 1982), oppure r^2 può essere utilizzato per stimare la percentuale di varianza che viene spiegata dal fatto di appartenere al gruppo sperimentale.

Consideriamo adesso un caso particolare di coefficiente di correlazione, il coefficiente punto-biseriale. È necessario però avere ben chiaro fin dall'inizio che quanto

verrà detto a proposito di questo coefficiente vale anche nel caso di tutti gli altri tipi di coefficienti di correlazione che si possono calcolare su vari tipi di dati.

Nel caso in cui si abbia una variabile indipendente dicotomica ed una variabile dipendente su scala metrica si ricorre al coefficiente di correlazione punto-biseriale (r_{pb}). In formula:

$$r_{pb} = \frac{(M_1 - M_2)}{s}\sqrt{\frac{n_1}{n}\frac{n_2}{n}} \tag{3.7}$$

Il coefficiente r_{pb} può essere calcolato anche a partire dal valore del χ^2 con $gl = 1$. In formula:

$$r_{pb} = \sqrt{\frac{\chi^2_{(1)}}{N}} \tag{3.8}$$

Il coefficiente r_{pb} si può calcolare a partire dal valore del test t di Student:

$$r_{pb} = \sqrt{\frac{t^2}{t^2 + gl}} \tag{3.9}$$

Infine, questo indice può essere calcolato utilizzando il valore del test F proveniente da una ANOVA univariata con due gruppi.

$$r_{pb} = \sqrt{\frac{F_{(1,?)}}{F_{(1,?)} + gl_{errore}}} \tag{3.10}$$

Il coefficiente di correlazione r di Pearson può essere, a sua volta, utilizzato come misura della grandezza dell'effetto e calcolato a partire dai valori dei test t di Student, F e χ^2 utilizzando le stesse formule presentate per il coefficiente di correlazione r_{pb}.

Infine, è possibile ricavare i coefficienti r e r_{pb} anche da due misure di grandezza dell'effetto in senso stretto che verranno presentate nel dettaglio nel paragrafo seguente. Per completezza dell'esposizione riportiamo comunque in questo paragrafo le formule che permettono di ricavare r e r_{pb} rispettivamente dall'indice d di Cohen e dall'indice g di Hedges. Come in precedenza, le formule sono identiche per entrambi i coefficienti di correlazione.

Per convertire d in r si utilizza la seguente formula:

$$r = \frac{d}{\sqrt{d^2 + 4}} \tag{3.11}$$

Quando il campione è poco numeroso oppure le numerosità dei due campioni sono molto diverse, è preferibile utilizzare una formula leggermente più complicata ma più precisa che è stata messa a punto da Aaron, Kromrey e Ferron (1998):

$$r = \frac{d}{d^2 + \sqrt{\frac{N^2 - 2N}{n_1 n_2}}} \tag{3.12}$$

Infine, per convertire l'indice *g* di Hedges in *r* si ha la seguente formula:

$$r = \frac{g}{\sqrt{g^2 + 4\left(\frac{gl_w}{N}\right)}} \tag{3.13}$$

ESEMPIO 3.5

Poniamo di svolgere una ricerca per vedere se il livello di ansia nelle situazioni sociali si diversifica in base al genere. Abbiamo un campione di 12 soggetti (6 maschi e 6 femmine) e somministriamo a ciascun soggetto un questionario sull'ansia avente punteggio da 0 a 30; laddove a punteggio alto corrisponde un alto livello di ansia e viceversa.

Riassumiamo i dati raccolti nella tabella 3.7.

Punteggio Ansia	**Genere** Maschio = 1 Femmina = 2
13	1
9	1
29	2
11	1
18	2
14	2
16	2
7	1
6	1
17	2
12	2
1	1

Tabella 3.7 – *Punteggi ottenuti al questionario sull'ansia*

Naturalmente 12 soggetti sarebbero un campione troppo piccolo se volessimo realmente condurre una ricerca ed effettuare una qualsiasi delle analisi che stiamo per svolgere; si consideri, quindi, il seguente modo di procedere, scorretto dal punto di vista puramente metodologico, funzionale semplicemente a rendere l'esempio breve e comprensibile.

Il gruppo dei maschi ha $M = 9{,}67$, mentre il gruppo delle femmine ha $M = 17{,}67$. La deviazione standard dell'intero campione è $s = 6{,}1$. A questo punto abbiamo a disposizione tutte le informazioni che permettono di calcolare l'indice di correlazione r_{pb} utilizzando la formula 3.7. Vediamo come procedere:

$$r_{pb} = \frac{(17{,}67 - 9{,}67)}{6{,}1} \times \sqrt{\frac{6}{12} \times \frac{6}{12}} = \frac{8}{6{,}1} \times \sqrt{0{,}5^2} = 1{,}311 \times 0{,}5 = 0{,}66.$$

Supponiamo adesso di voler analizzare i dati della ricerca utilizzando il test χ^2. Per farlo dovremo, prima di tutto, rendere dicotomica la variabile dipendente. È possibile suddividere i punteggi ottenuti al questionario in due categorie (alti e bassi) utilizzando come soglia il valore della mediana calcolata sull'intero campione. La mediana ha valore 12,5, quindi suddividiamo i punteggi in due gruppi:

Punteggio Alto: $x > 12{,}5$

Punteggio Basso: $x \leq 12{,}5$

A questo punto riassumiamo i dati in una tabella 2×2 (vedi tabella 3.8).

		Genere	
		Maschi	Femmine
Punteggio	Alto	1	5
	Basso	5	1

Tabella 3.8 – *Frequenze osservate dei punteggi "Alto" e "Basso", gruppi "Maschi" e "Femmine"*

Anche se formalmente è scorretto, poiché le frequenze di cella sono troppo basse per poter effettuare questo tipo di analisi, poniamo che si utilizzi il test χ^2 per analizzare questa tabella. Il valore che otteniamo è $\chi_{(1)}^2 = 5{,}333$ con $p = 0{,}021$. Proviamo a ricavare il coefficiente di correlazione dal valore del test χ^2 utilizzando la formula 3.8:

$$r_{pb} = \sqrt{\frac{5{,}333}{12}} = \sqrt{0{,}4444} = 0{,}66.$$

Vediamo adesso cosa accade se decidiamo di utilizzare il test *t* di Student per analizzare i dati e ricavarne il coefficiente di correlazione. Effettuando l'analisi *t* di Student, otteniamo un valore $t = 2{,}977$ con $p = 0{,}014$ e $gl = 10$. Sostituendo questi valori nella formula 3.9 calcoliamo il valore corrispondente di r_{pb}.

$$r_{pb} = \sqrt{\frac{2{,}977^2}{2{,}977^2 + 10}} = \sqrt{\frac{8{,}86}{8{,}86 + 10}} = \sqrt{\frac{8{,}86}{18{,}86}} = \sqrt{0{,}469} = 0{,}68.$$

Infine, è possibile analizzare i dati anche mediante una ANOVA univariata a due gruppi e ricavare dal valore di F il coefficiente di correlazione r_{pb}. Seguendo questo procedimento, otteniamo un valore di $F_{(1,10)} = 8{,}86$ con $p = 0{,}014$ e $gl_{errore} = 10$. Sostituendo opportunamente questi valori nella formula 3.10, ricavo il valore di r_{pb}.

$$r_{pb} = \sqrt{\frac{8,86}{8,86 + 10}} = \sqrt{\frac{8,86}{18,86}} = \sqrt{0,469} = 0,68.$$

Da notare la sostanziale identità dei valori sostituiti nelle formule per il test t di Student e del test F. Ciò risulta chiaro se si considera che, come è noto, il rapporto tra t di Student ed F è dato da

$$F = t^2.$$

Come era prevedibile, sulla base degli esempi svolti, risulta chiaro che sia partendo dai dati grezzi, sia partendo da altri test statistici si giunge ad un valore di r sostanzialmente identico. Piccole variazioni ottenute, in questo caso, nel valore di r_{pb} sono da attribuirsi alla ristrettezza del campione ed alla sua sostanziale inadeguatezza a questo tipo di analisi.

Non ci dilungheremo adesso sul significato da attribuirsi ai valori dell'indice r ottenuti, piuttosto rimandiamo il lettore al paragrafo relativo all'interpretazione della grandezza dell'effetto, dove potrà trovare esaurienti spiegazioni in merito.

Infine, pur avendo fornito le formule che permettono di ricavare il coefficiente di correlazione da due misure di grandezza dell'effetto in senso stretto, d e g, non presenteremo in questa sezione un esempio numerico, poiché è preferibile che prima il lettore abbia acquisito maggior informazioni in merito a questi indici. Per gli esempi numerici rimandiamo, quindi, il lettore al paragrafo relativo agli indici di grandezza dell'effetto in senso stretto.

3.2.3. Associazione e tabelle di contingenza

La misura di associazione più largamente utilizzata nelle tabelle di contingenza è l'indice C, il coefficiente di contingenza di Pearson. In formula:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{3.14}$$

Il valore minimo che questo indice può assumere è 0 ed indica assenza di effetto; il valore massimo di C tende ad 1.

Come abbiamo già visto in precedenza, anche il coefficiente di correlazione può essere utilizzato per determinare la grandezza dell'effetto a partire dal χ^2 delle tabelle di contingenza 2×2 mediante la formula:

$$r_{pb} = \sqrt{\frac{\chi^2_{(1)}}{N}} \tag{3.8}$$

Anche questa formula vale per tutti i coefficienti di correlazione calcolabili tra i vari tipi di variabili: r, r_{pb} e ϕ. Vi è però il vincolo che il suo utilizzo è limitato alle tabelle di contingenza 2×2, cioè ad un χ^2 avente $gl = 1$.

Infine, un ulteriore indice che viene utilizzato per il χ^2 delle tabelle di contingenza è il ϕ' di Cramer, in formula:

$$\phi' = \sqrt{\frac{\chi^2}{N(gl_{smaller})}}, \tag{3.15}$$

Se k è il numero più piccolo, tra colonne o di righe, allora:

$$gl_{smaller} = k - 1. \tag{3.16}$$

Questo indice inoltre può essere utilizzato con tabelle di ogni dimensione.

Esempio 3.6

Poniamo di svolgere una ricerca in cui siamo interessati a comprendere se vi sia relazione tra pena di morte ed etnia dell'imputato. Riassumiamo nella tabella 3.9 i dati.

		Etnia imputato	
		Bianco	Nero
Pena di morte	Sì	20	80
	No	60	15

Tabella 3.9 – *Frequenze osservate della variabile "Pena di morte", livelli "Sì" e "No" gruppi "Imputato Bianco" e "Imputato Nero"*

Utilizzando il test χ^2 per tabelle di contingenza otteniamo il risultato $\chi_{(1)}^2 = 62{,}171$ con $p < 0{,}001$. Sapendo che $N = 175$, possiamo procedere al calcolo dell'indice C sostituendo i valori noti nella formula 3.14.

$$C = \sqrt{\frac{62{,}171}{62{,}171 + 175}} = \sqrt{\frac{62{,}171}{237{,}171}} = \sqrt{0{,}262} = 0{,}512.$$

Procedendo in modo sostanzialmente analogo, possiamo servirci della formula 3.8 per ricavare il valore del coefficiente di correlazione:

$$r = \sqrt{\frac{62,171}{175}} = \sqrt{0,355} = 0,596.$$

Se calcolassimo l'indice ϕ' di Cramer dai dati presentati, la formula di questo indice si ridurrebbe a quella utilizzata per il coefficiente di correlazione, poiché in questo caso la componente gl_{minore}, che al numeratore moltiplica per N, ha valore 1 (essendo $k = 2$).

Poniamo, quindi, di effettuare una nuova ricerca che sia volta ad indagare se vi sia una relazione tra etnia e tipo di reati per i quali si è processati. Riassumiamo i dati raccolti nella tabella 3.10.

		Etnia imputato		
		Bianco	Nero	Cinese
Tipo di reato	Omicidio	10	40	65
	Furto	50	45	20
	Stupro	20	10	5

Tabella 3.10 – *Frequenze osservate della variabile "Tipo di reato" su tre livelli "Omicidio", "Furto" e "Strupro" nei tre gruppi "Imputato Bianco", "Imputato Nero" e "Imputato Cinese"*

Utilizzando il test χ^2 per tabelle di contingenza otteniamo il risultato $\chi_{(4)}^2 = 64,168$ con $p < 0,001$. Sapendo che $N = 265$ e $gl_{minore} = 2$ (essendo $k = 3$, vedi formula 3.16) possiamo procedere al calcolo dell'indice ϕ' di Cramer sostituendo i valori noti nella formula 3.15.

$$\phi' = \sqrt{\frac{64,168}{265 \times 2}} = \sqrt{\frac{64,168}{530}} = \sqrt{0,121} = 0,348.$$

Nell'esempio presentato la tabella era di tipo 3×3, quindi il numero di righe era identico al numero delle colonne. Vogliamo comunque ricordare che, se avessimo utilizzato una tabella con un diverso numero di righe e colonne, ad esempio una tabella 5×6, per il calcolo della componente gl_{minore} avremmo utilizzato come k il numero più piccolo tra colonne e righe, quindi 5, e gl_{minore} avrebbe assunto valore 4 (cioè, $k - 1$).

Per l'interpretazione degli indici calcolati, rimandiamo il lettore al paragrafo relativo all'interpretazione della grandezza dell'effetto.

3.2.4. Associazione e regressione

Una misura di associazione ampiamente utilizzata nell'analisi di regressione è il coefficiente di determinazione R^2. Questo indice fornisce una quantificazione della pro-

porzione della variazione totale della variabile dipendente che è determinata dalla sua relazione con la variabile indipendente.

In formula:

$$R^2 = \frac{SS_{effetto}}{SS_{totale}} \tag{3.17}$$

Il coefficiente R^2 può assumere valori che variano tra 0 (assenza di effetto, poiché tutta la variazione spiegata è dovuta all'errore) ed 1 (effetto 'totale' o predizione perfetta, tutta la variazione spiegata è dovuta alla variabile indipendente).

Questo indice è particolarmente agevole poiché viene fornito direttamente nell'output dai programmi che eseguono analisi di regressione.

Esempio 3.7

Poniamo di svolgere una ricerca per verificare se il livello di scolarità (misurato in anni di frequenza scolastica senza ripetenze) influenza l'intolleranza razziale. Poniamo di misurare l'intolleranza mediante un questionario avente punteggio da 0 a 40, laddove a punteggio alto corrisponde una forte intolleranza.

Visualizziamo nella tabella 3.11 una parte dei risultati dell'analisi di regressione effettuata sui dati raccolti.

Modello	**Somma dei quadrati**	***gl***	**Media dei quadrati**	***F***	*p*
Regressione	940,9	1	940,9	48,074	0,000
Residuo	254,433	13	19,572		
Totale	1195,333	14			

Tabella 3.11 – *Risultati Analisi di Regressione*

Utilizzando la formula 3.17 e sostituendo al suo interno i valori ottenuti, possiamo procedere, molto facilmente, al calcolo dell'R^2. (Ricordiamo nuovamente che, avvalendosi dei più comuni programmi di analisi statistica, il coefficiente di determinazione viene fornito automaticamente nell'*output* e non è necessario calcolarlo manualmente).

Otteniamo:

$$R^2 = \frac{940,9}{1195,333} = 0,787.$$

Questo significa che il 78,7% della variazione della variabile intolleranza razziale è spiegata (in senso statistico) dal livello di scolarità delle persone.

Esiste, inoltre, il cosiddetto coefficiente di non determinazione, corrispondente a $1 - R^2$, che rappresenta la proporzione della varianza della variabile dipendente che non può essere attribuita alla sua relazione con la variabile indipendente. Nel nostro caso il coefficiente di non determinazione assume valore 1 – 0,787 = 0,213; ciò significa che il 21,3% della varianza della variabile intolleranza non è spiegata dal livello di scolarità delle persone.

3.3. MISURE DI GRANDEZZA DELL'EFFETTO

Le misure della grandezza dell'effetto in senso stretto *(measures of effect size*) prevedono una analisi delle differenze tra le medie. Ogni indice che sia una differenza, grezza o standardizzata, tra medie costituisce una misura della grandezza dell'effetto.

Cohen (1988) ha ampiamente studiato questo tipo di misurazioni ed ha più volte messo in guardia la comunità scientifica riguardo all'esigenza che venissero adottate nella pratica della ricerca. Cohen fornisce una importante definizione della grandezza dell'effetto come del "*grado in cui, in assenza di implicazioni di causalità, il fenomeno studiato è presente nella popolazione, ossia il grado in cui l'ipotesi nulla è falsa*". Lo stesso autore ci suggerisce anche che "*l'ipotesi nulla prevede sempre che la grandezza dell'effetto sia zero*" (Cohen, 1988).

Vediamo adesso quali indici di grandezza dell'effetto in senso stretto sono disponibili, suddividendoli in base al disegno della ricerca ed al tipo di analisi condotte sui dati raccolti.

3.3.1. Due campioni indipendenti

Prendiamo in considerazione le misure di grandezza dell'effetto che si utilizzano con due gruppi indipendenti: sono tutti indici basati su differenze standardizzate tra le medie dei due campioni. Gli indici maggiormente utilizzati sono:

- *d* di Cohen
- *g* di Hedges
- Δ di Glass

Cohen (1988) ha definito l'**indice *d*** come il rapporto tra la differenza tra le medie ($M_1 - M_2$) e la deviazione standard (σ) di uno dei due gruppi. In formula:

$$d = \frac{M_1 - M_2}{\sigma} \tag{3.18}$$

Cohen specifica che la deviazione standard di uno dei due gruppi, e quindi questa versione della formula, può essere utilizzata solo quando le varianze degli stessi sono

omogenee. Per convenzione la sottrazione tra le medie viene compiuta in modo che la differenza sia positiva quando indica un miglioramento o va nella direzione prevista, e che sia negativa quando indica un peggioramento o va in direzione opposta al previsto. L'indice *d* è una misura descrittiva.

Nella pratica (Rosnow e Rosenthal, 1996), dato che l'omogeneità delle varianze non è un prerequisito sempre garantito, si utilizza al posto di σ la deviazione standard *pooled* (σ_{pooled}). In formula:

$$d = \frac{M_1 - M_2}{\sigma_{pooled}} \tag{3.19}$$

La deviazione standard raggruppata (*pooled*) è data dalla radice quadrata della media dei quadrati delle singole deviazioni standard (Cohen, 1988). In formula:

$$\sigma_{pooled} = \sqrt{\frac{\sigma_1^{\,2} + \sigma_2^{\,2}}{2}} \tag{3.20}$$

Come è chiaro, il calcolo dell'indice *d* si basa sui dati campionari, che, come sempre accade, possono fornire soltanto una stima dei parametri della popolazione e sono soggetti ad errori di campionamento. Per correggere questo tipo di *bias*, soprattutto quando ci troviamo in presenza di campioni di numerosità ridotta, Hedges e Olkin (1985) consigliano l'utilizzo di una formula corretta:

$$d_c = d\left(1 - \frac{3}{\left[4\left(N_1 + N_2\right) - 9\right]}\right) \tag{3.21}$$

L'indice *d* può essere calcolato anche partendo dal valore del test *t* di Student per due campioni indipendenti (Rosenthal & Rosnow, 1991). In formula:

$$d = \frac{t\left(n_1 + n_2\right)}{\sqrt{gl}\sqrt{\left(n_1 n_2\right)}} \tag{3.22}$$

La formula contiene i gradi di libertà (*gl*) del test *t*. Questa prima formula viene utilizzata quando la numerosità dei due campioni è diversa, e gli *n* di entrambi vengono quindi inseriti nel calcolo. Altresì, nel caso in cui i due campioni siano costituiti da un numero identico di partecipanti è possibile utilizzare la formula semplificata:

$$d = \frac{2t}{\sqrt{gl}} \tag{3.23}$$

L'indice *d* può essere calcolato basandosi sul valore del test *F* per l'analisi della varianza univariata con due gruppi, mediante la seguente formula:

$$d = \frac{2\sqrt{F_{(1,?)}}}{\sqrt{gl_{errore}}} \tag{3.24}$$

L'indice di Cohen può essere calcolato, con molta semplicità, anche a partire dal coefficiente di correlazione *r* (Friedman, 1968). In formula:

$$d = \frac{2r}{\sqrt{1 - r^2}} \tag{3.25}$$

Come abbiamo visto in precedenza, le formule che includono indici di correlazione sono da considerarsi utilizzabili con tutti gli indici di correlazione esistenti. In questo senso la formula sovraesposta per ricavare *d* a partire da *r* è valida anche nel caso che abbia a disposizione un indice di correlazione r_{pb} oppure un indice di correlazione r_{phi} (ϕ) per variabili dicotomiche.

Infine, è possibile ricavare l'indice *d* anche partendo da un altro indice che abbiamo già menzionato e vedremo nello specifico tra breve: l'indice *g* di Hedges. In formula:

$$d = g\sqrt{\frac{N}{gl_w}} \tag{3.26}$$

ESEMPIO 3.8

Poniamo di svolgere una ricerca in cui si sia interessati a verificare se vi è una differenza di genere nella ricerca di sensazioni forti. La variabile dipendente verrà misurata mediante una scala avente un punteggio che varia da 0 a 40, laddove a punteggio alto corrisponde una forte ricerca di sensazioni. Riassumiamo nella tabella 3.12 i dati raccolti.

Genere	**Ricerca di Sensazioni**	**Genere**	**Ricerca di Sensazioni**
2	15	2	14
2	10	2	8
1	32	1	26
2	18	1	24
1	39	1	19
1	28	2	9
1	27	1	21
2	20	2	22
2	13	2	15
1	38	1	35

Tabella 3.12 – *Punteggi ottenuti al questionario sulla ricerca di sensazioni (Genere: Maschio = 1; Femmina =2)*

Il campione dei maschi ha un punteggio medio pari a 28,9±6,9, il campione delle femmine ha un punteggio medio pari a 14,4±4,65.

Dobbiamo, prima di tutto, ricavare la deviazione standard *pooled*, utilizzando la formula 3.20 e sostituendo i valori di deviazione standard noti.

$$\sigma_{pooled} = \sqrt{\frac{6,9^2 + 4,65^2}{2}} = \sqrt{\frac{47,61 + 21,62}{2}} = \sqrt{\frac{69,23}{2}} = \sqrt{34,615} = 5,88.$$

A questo punto possiamo procedere, molto semplicemente, al calcolo dell'indice *d* utilizzando la formula 3.19 come segue:

$$d = \frac{28,9 - 14,4}{5,88} = \frac{14,5}{5,88} = 2,47.$$

Vediamo adesso come calcolare l'indice *d* partendo dai valori di alcuni test statistici. In tutti gli esempi che seguono continueremo a fare riferimento ai dati presentati in precedenza.

Procedendo al calcolo del *t* di Student sui dati, otteniamo un valore $t = 5,51$ con $gl = 18$ e $p < 0,001$. Sostituendo i valori ottenuti nella formula 3.23, è possibile calcolare il valore dell'indice *d*.

$$d = \frac{2 \times 5,51}{\sqrt{18}} = \frac{11,02}{4,24} = 2,59.$$

Effettuando una ANOVA univariata a due gruppi sui dati, ottengo un valore $F_{(1.18)} = 30,359$ con $p < 0,001$. Sostituendo i valori ottenuti nella formula 3.24, possiamo calcolare il valore dell'indice *d*.

$$d = \frac{2\sqrt{30,359}}{\sqrt{18}} = \frac{2 \times 5,51}{4,24} = \frac{11,02}{4,24} = 2,59.$$

Infine, il coefficiente di correlazione punto-biseriale (r_{pb}) calcolato sui dati ha valore 0,77. Utilizzando la formula 3.25 è possibile ricavare *d* dal coefficiente di correlazione.

$$d = \frac{2 \times 0,77}{\sqrt{1 - 0,77^2}} = \frac{1,54}{\sqrt{1 - 0,59}} = \frac{1,54}{\sqrt{0,41}} = \frac{1,54}{0,64} = 2,41.$$

L'indice ***g* di Hedges** (1981) è una misura inferenziale, e deve il suo nome a Gene V. Glass, uno dei pionieri della meta-analisi. In formula:

$$g = \frac{M_1 - M_2}{S_{pooled}} \tag{3.27}$$

dove:

$$S_{pooled} = \sqrt{MS_{within}} \tag{3.28}$$

Come si vede, la formula dell'indice *g* prevede l'utilizzo della media dei quadrati entro i gruppi (MS_{within}) derivata dal procedimento di analisi della varianza per due gruppi.

L'indice *g* può essere anch'esso calcolato a partire dal valore del test *t* di Student per due campioni indipendenti (Rosenthal & Rosnow, 1991). Come avevamo visto anche per l'indice *d*, esistono due formule per derivare l'indice *g* dal *t* di Student, rispettivamente: la prima formula si utilizza quando le numerosità dei due campioni sono diverse, la seconda formula si utilizza altresì quando gli *n* dei due campioni sono uguali.

$$g = \frac{t\sqrt{n_1 + n_2}}{\sqrt{n_1 n_2}} \tag{3.29}$$

$$g = \frac{2t}{\sqrt{N}} \tag{3.30}$$

L'indice *g* può essere ricavato dall'indice di correlazione *r*, mediante la formula:

$$g = \frac{2r\sqrt{\frac{gl_w}{N}}}{\sqrt{1 - r^2}}. \tag{3.31}$$

Infine, come abbiamo visto che dall'indice *g* è possibile ricavare l'indice *d* di Cohen, è chiaramente vero anche il contrario cioè che da *d* si può ricavare *g*. In formula:

$$g = d\sqrt{\frac{N}{gl}}\ . \tag{3.32}$$

ESEMPIO 3.9

Facciamo nuovamente riferimento ai dati della ricerca presentata nell'esempio 3.8 per il calcolo dell'indice *d*.

Per calcolare l'indice *g* utilizzando la formula 3.27 dovremo, prima di tutto, ricavare S_{pooled} seguendo la formula 3.28 che richiede l'utilizzo della media dei quadrati entro i gruppi (MS_{within}) derivata dall'ANOVA per due gruppi. Riassumiamo nella tabella 3.13 i risultati dell'ANOVA.

	Somma dei quadrati	*gl*	Media dei quadrati	*F*	*p*
Tra i gruppi	1051,25	1	1051,25	30,359	0,000
Entro i gruppi	623,3	18	34,628		
Totale	1674,55	19			

Tabella 3.13 – *Risultati ANOVA*

Sostituendo i valori appropriati nella formula 3.28 ricaviamo il valore di S_{pooled}:

$$S_{pooled} = \sqrt{34,628} = 5,88 .$$

A questo punto possiamo calcolare *g* mediante la formula 3.27.

$$g = \frac{28,9 - 14,4}{5,88} = \frac{14,5}{5,88} = 2,47.$$

Da notare l'assoluta identità dei valori utilizzati per il calcolo di *g* con i valori utilizzati per il calcolo dell'indice *d* (vedi Esempio 3.8). Non soltanto il risultato, cioè il valore dell'indice, è lo stesso, ma anche i valori sostituiti all'interno delle due formule sono identici.

Procediamo adesso al calcolo dell'indice *g* partendo dai valori di altri test statistici. Continuiamo a fare riferimento ai dati del medesimo esempio.

Sappiamo già che il test *t* di Student ha valore $t = 5,51$ con $gl = 18$ e $p < 0,001$, quindi mediante la formula 3.30 ricaviamo *g*.

$$g = \frac{2 \times 5,51}{\sqrt{20}} = \frac{11,02}{4,47} = 2,46.$$

Analogamente sappiamo già il valore del coefficiente $r_{pb} = 0.77$, con $gl = 18$. Utilizzando la formula 3.31 possiamo facilmente ricavare *g*.

$$g = \frac{2 \times 0,77\sqrt{\frac{18}{20}}}{\sqrt{1 - 0,77^2}} = \frac{1.54\sqrt{0,9}}{\sqrt{1 - 0,5929}} = \frac{1,54 \times 0,95}{\sqrt{0,4071}} = \frac{1,46}{0,64} = 2,29.$$

Infine, partendo dal valore noto dell'indice *d* vediamo come procedere per calcolare l'indice *g* mediante la formula 3.32.

$$g = 2,47\sqrt{\frac{20}{18}} = 2,47\sqrt{0,9} = 2,47 \times 0,95 = 2,35.$$

Vediamo, infine, l'indice **Δ di Glass** (1976) definito dal rapporto tra la differenza tra le medie del gruppo sperimentale e del gruppo di controllo ($M_1 - M_2$) e la deviazione standard del gruppo di controllo. In formula:

$$\Delta = \frac{M_1 - M_2}{\sigma_{control}} \tag{3.33}$$

ESEMPIO 3.10

Poniamo di svolgere una ricerca per verificare l'efficacia di un farmaco antidepressivo di nuova generazione nel trattamento dei pazienti affetti da depressione maggiore. Il campione dei pazienti viene casualmente ed equamente suddiviso in due gruppi: al primo gruppo viene somministrato il farmaco (gruppo sperimentale), al secondo gruppo viene somministrato un placebo (gruppo di controllo). Dopo un periodo di trattamento, ad entrambi i gruppi viene somministrato un questionario che indaga la depressione; il questionario ha un punteggio che varia da 0 a 10, laddove a punteggio alto corrisponde una maggiore tendenza depressiva.

I punteggi medi dei due gruppi sono: per il gruppo sperimentale M_S = 5,4±1,5, per il gruppo di controllo M_C = 8,1±1,8. Sostituendo i risultati ottenuti nella formula 3.33 possiamo calcolare il valore dell'indice Δ con estrema semplicità**,**

$$\Delta = \frac{8,1 - 5,4}{1,8} = \frac{2,7}{1,8} = 1,5.$$

3.3.2. Due campioni dipendenti

In letteratura troviamo molte indicazioni contrastanti su come debba essere calcolata la grandezza dell'effetto nel caso di due gruppi dipendenti. Prendiamo in considerazione un esempio tipico di disegno della ricerca per misure ripetute.

$$\begin{array}{ccc} O_{c1} & & O_{c2} \\ O_{e1} & x & O_{e2} \end{array}$$

I partecipanti vengono assegnati casualmente ad una delle due condizioni: sperimentale (*e*) o di controllo (*c*). Al tempo 1 ($O_{.1}$) a tutti i partecipanti viene somministrato un pre-test. Al gruppo sperimentale viene somministrato un trattamento (*x*). Viene quindi effettuata una seconda misurazione su entrambi i gruppi al tempo 2 ($O_{.2}$) ed il periodo di tempo intercorso tra le due rilevazioni è uguale per entrambi i gruppi.

Il calcolo della grandezza dell'effetto è, per sua natura, basato sul confronto tra la media del gruppo sperimentale e la media del gruppo di controllo. Nel disegno a misure ripetute preso in considerazione, la media del gruppo sperimentale sarà data dai punteggi della seconda misurazione sul gruppo sperimentale (O_{e2}), il problema

sta nella scelta di quale punteggio utilizzare come controllo, dato che potrebbe essere utilizzata ciascuna delle tre restanti misurazioni. Potremmo, infatti, calcolare la grandezza dell'effetto confrontando O_{e2} con il punteggio prima del trattamento del gruppo sperimentale stesso (O_{e1}), con il punteggio al tempo 1 del gruppo di controllo (O_{c1}), oppure con il punteggio al tempo 2 del gruppo di controllo (O_{c2}). Da notare che di queste tre possibilità di confronto, due costituiscono in realtà un confronto tra due campioni indipendenti ($O_{e2} - O_{c1}$ e $O_{e2} - O_{c2}$) mentre solo l'eventuale confronto tra O_{e2} e O_{e1} è da ritenersi un vero e proprio confronto tra gruppi dipendenti.

Wilson, Becker e Tinker (1995) calcolano la grandezza dell'effetto mediante l'indice *d* di Cohen, confrontando il punteggio post-trattamento del gruppo sperimentale (O_{e2}) con il punteggio al tempo 2 del gruppo di controllo (O_{c2}). Questo procedimento si riduce, come abbiamo detto, al confronto tra due gruppi indipendenti e l'utilizzo dell'indice *d* risulta quindi adeguato.

Supponiamo, invece, di voler calcolare l'indice *d* di Cohen confrontando il punteggio post-trattamento del gruppo sperimentale (O_{e2}) con il punteggio dello stesso gruppo prima del trattamento (O_{e1}), come posso calcolare la deviazione standard *pooled*? Ci sono due alternative:

1. Utilizzare il valore del test *t* per misure ripetute per ricavare l'indice *d* di Cohen;
2. Utilizzare le deviazioni standard delle due medie originali.

In realtà entrambi questi approcci presentano degli svantaggi.

Dato che il test *t* per misure ripetute prende in considerazione la correlazione tra i due punteggi, il valore di questo test sarà sempre maggiore del valore del *t* di Student per gruppi indipendenti. Di conseguenza, il valore dell'indice *d* partendo dal test *t* per misure ripetute sarà sempre maggiore dello stesso valore ottenuto a partire dal *t* per campioni indipendenti, ed anche rispetto allo stesso valore ottenuto, invece, utilizzando nella formula le deviazioni standard originali dei due gruppi.

Nonostante queste osservazioni, Rosenthal (1991) suggerisce di utilizzare il valore del test *t* per misure ripetute per ricavare l'indice di grandezza dell'effetto. Al contrario, Dunlap, Cortina, Vaslow e Burke (1996) suggeriscono in modo piuttosto convincente di utilizzare le deviazioni standard dei due gruppi per ricavare l'indice di grandezza dell'effetto nei disegni a misure ripetute. Questi autori sostengono che se la deviazione standard *pooled* viene corretta sulla base della quota di correlazione tra le misurazioni, allora il valore che otteniamo per l'indice di grandezza dell'effetto sarà sovrastimato rispetto alla grandezza reale. L'entità della sovrastima dipende dalla grandezza della correlazione tra i punteggi. Dunlap e colleghi (1996) dimostrano, ad esempio, che se la correlazione arriva ad essere di 0,8, il valore dell'indice *d* calcolato mediante il test *t* per misure ripetute è più del doppio del valore che lo stesso indice assume se calcolato a partire dalle deviazioni standard originali.

Lo stesso problema si porrebbe se decidessimo di utilizzare, per stimare l'indice *d*, il test *F* a partire da una ANOVA per misure ripetute a due gruppi.

Sembrerebbe quindi preferibile utilizzare le deviazioni standard originali per calcolare l'indice di grandezza dell'effetto *d*, piuttosto che ricorrere ai valori dei test *t* o *F* per misure ripetute. In realtà anche questa modalità è discutibile poiché considerare le deviazioni standard di due gruppi dipendenti come se derivassero da due gruppi

indipendenti comporta un problema legato all'errore casuale. Mentre per due gruppi indipendenti l'errore tende ad annullarsi, nel caso di gruppi dipendenti esso tenderà a sommarsi, poiché trattandosi degli stessi soggetti se vi è stato un errore al tempo 1 esso rimarrà costante al tempo 2.

Fino ad ora abbiamo considerato la possibilità di calcolare la grandezza dell'effetto per campioni dipendenti mediante l'indice *d*, una alternativa interessante è data dall'utilizzo dell'indice *f* di Cohen (1988). In formula:

$$f = \frac{\mu_y}{\sigma_y} \tag{3.34}$$

dove μ_y è la media delle differenze tra le medie e σ_y è dato da:

$$\sigma_y = \sqrt{\sigma_A^2 + \sigma_B^2 - 2r\sigma_A^2\sigma_B^2} \tag{3.35}$$

ESEMPIO 3.11

Poniamo di svolgere una ricerca per verificare se un corso mirato a stimolare l'assertività ha come effetto un miglioramento nelle abilità comunicative di un gruppo di dirigenti d'azienda. Riassumiamo nella tabella 3.14 i dati raccolti.

Prima del corso	**Dopo il corso**	**Prima del corso**	**Dopo il corso**
15	18	19	22
14	25	14	24
20	23	12	25
16	21	11	26
18	19	20	22
19	20	18	20
13	20	17	19
13	22	19	21
12	24	16	20
17	25	13	23
11	25	13	23
15	18	15	25
16	23	18	21
18	20	15	21
18	22	15	23

Tabella 3.14 – *Punteggi ottenuti al questionario sull'assertività prima e dopo il corso*

Vediamo adesso come si diversifica la grandezza dell'effetto utilizzando le tre strategie presentate in precedenza:

1. Calcolare l'indice *d* partendo dal valore del test *t* per misure ripetute;
2. Calcolare l'indice *d* partendo dalle deviazioni standard dei due gruppi;
3. Calcolare l'indice *f*.

1. Effettuando un test *t* per misure ripetute sui dati raccolti abbiamo come risultato $t_{(29)}$ = 8,203 con $p < 0{,}001$, Utilizzando la formula 3.23, possiamo ricavare il valore dell'indice *d*.

$$d = \frac{2 \times 8,203}{\sqrt{29}} = \frac{16,406}{5,38} = 3,05.$$

2. Volendo calcolare l'indice *d* mediante la formula 3.19, dovremo prima di tutto procedere al calcolo della deviazione standard *pooled*. Dato che le deviazioni standard dei due campioni sono, rispettivamente, $\sigma_{tempo1} = 2.68$ e $\sigma_{tempo2} = 2.26$, utilizzando la formula 3.20, avremo:

$$\sigma_{pooled} = \sqrt{\frac{2,68^2 + 2,26^2}{2}} = \sqrt{\frac{7,18 + 5.11}{2}} =$$

$$= \sqrt{\frac{12,29}{2}} = \sqrt{6,145} = 2,48.$$

A questo punto, sapendo che le medie dei due campioni sono, rispettivamente, M_{tempo1} = 15,67 e M_{tempo2} = 22, mediante la formula 3.19, procedo al calcolo di *d*.

$$d = \frac{22 - 15,67}{2,48} = \frac{6,33}{2,48} = 2,55.$$

3. Infine, per procedere al calcolo dell'indice *f* di Cohen, dovrò prima di tutto calcolare σ_y. La correlazione tra i due campioni dipendenti è $r = -0,461$. A questo punto, mediante la formula 3.35, ricaviamo il valore di σ_y.

$$\sigma_y = \sqrt{2,68^2 + 2,26^2 - 2 \times (-0,461) \times 2,68^2 \times 2,26^2)} =$$

$$= \sqrt{7,18+5,11+5,58} = \sqrt{17,87} = 4,23.$$

Sapendo che $\mu_y = 6,33$, procediamo al calcolo di *f* con la (3.34),

$$f = \frac{6,33}{4,23} = 1,49.$$

3.3.3. Altri indici di grandezza dell'effetto

Cohen (1988) ha creato anche altri indici di grandezza dell'effetto.

3.3.3.1. Indice q

L'indice *q* di Cohen (1988) viene utilizzato per ottenere una grandezza dell'effetto delle differenze tra coefficienti di correlazione.

Poniamo di avere due popolazioni e di estrarre da ciascuna popolazione un campione. Dopo aver misurato la correlazione tra le variabili all'interno dei due campioni (r_1 e r_2), vediamo di capire se e quanto si differenzino gli *r* delle due popolazioni. In questo caso la semplice differenza tra gli indici non è una misura affidabile, dovremo quindi ricorrere alla trasformazione *z* di Fisher degli indici *r* (Cohen, 1988). Per questa trasformazione si utilizza la formula:

$$z = \frac{1}{2}\log_e \frac{1+r}{1-r} \tag{3.36}$$

In realtà esiste una tavola (Cohen, 1988) che permette di effettuare molto più semplicemente questa trasformazione, poiché fornisce già il valore *z* corrispondente a ciascun valore di *r*. Riproponiamo questa tavola nella tabella 3.15.

Effettuata la trasformazione, l'indice *q* può essere così calcolato:

$$q = z_1 - z_2 \tag{3.37}$$

ESEMPIO 3.12

Poniamo di svolgere una ricerca per verificare se tra coppie di gemelli troviamo una maggiore affinità riguardo ad alcune variabili psicologiche rispetto a coppie di fratelli non omozigoti. A tutti i partecipanti viene somministrato un questionario di personalità. Nel campione di gemelli (N = 60) la correlazione tra fratelli rispetto alle variabili indagate è di $r = 0{,}65$, nel campione di fratelli non omozigoti (N = 60) la correlazione è di $r = 0{,}42$.

Utilizzando la formula 3.36, è possibile trasformare gli indici *r* in *z*:

$$z_{gemelli} = \frac{1}{2}\log_e \frac{1+0,65}{1-0,65} = \frac{1}{2}\log_e \frac{1,65}{0,35} = \frac{1}{2}\log_e 4,71 = 0,5 \times 1,55 = 0,775;$$

$$z_{fratelli} = \frac{1}{2}\log_e \frac{1+0,42}{1-0,42} = \frac{1}{2}\log_e \frac{1,42}{0,58} = \frac{1}{2}\log_e 2,45 = 0,5 \times 0,896 = 0,448.$$

Andando a controllare i valori ottenuti nella tabella 3.15 possiamo facilmente constatare che il risultato è giusto. Ricordiamo che avremmo potuto utilizzare, molto più semplicemente, questa tabella per ricavare il valore di *z* partendo da *r*.

r	z	r	z	r	z	r	z
0,01	0,010	0,26	0,266	0,51	0,563	0,76	0,996
0,03	0,030	0,28	0,288	0,53	0,590	0,78	1,045
0,05	0,050	0,30	0,310	0,55	0,618	0,80	1,099
0,07	0,070	0,32	0,332	0,57	0,648	0,82	1,157
0,09	0,090	0,34	0,354	0,59	0,678	0,84	1,221
0,11	0,110	0,36	0,377	0,61	0,709	0,86	1,293
0,13	0,131	0,38	0,400	0,63	0,741	0,88	1,376
0,15	0,151	0,40	0,424	0,65	0,775	0,90	1,472
0,17	0,172	0,42	0,448	0,67	0,811	0,92	1,589
0,19	0,192	0,44	0,472	0,69	0,848	0,94	1,738
0,21	0,213	0,46	0,497	0,71	0,887	0,96	1,946
0,23	0,234	0,48	0,523	0,73	0,929	0,98	2,298

Tabella 3.15 – *Trasformazione del coefficiente di correlazione r in z*

Utilizzando la formula 3.37 ricaviamo il valore di *q* corrispondente.

$$q = 0,775 - 0,448 = 0,327.$$

Per valutare la grandezza dell'effetto dell'indice *q* rimandiamo il lettore al paragrafo relativo all'interpretazione della grandezza dell'effetto.

3.3.3.2. Indice h

L'indice *h* (Cohen, 1988) viene utilizzato per la grandezza dell'effetto nelle differenze tra proporzioni. Poniamo di prendere in considerazione due popolazioni indipendenti (ad esempio alta e bassa estrazione sociale), di estrarre da ciascuna popolazione un campione e di misurare su entrambi i campioni una qualsiasi variabile dipendente dicotomica (ad esempio successo *vs* insuccesso scolastico). Per ciascuno dei due campioni otterremo una proporzione di successo (P_1 e P_2) e volendo stabilire se vi è una relazione tra l'appartenenza ad una delle due popolazioni e la proporzione di successi scolastici risulta chiaro che, in un certo senso, torniamo a ragionare in termini correlazionali.

Come abbiamo visto per l'indice *q*, anche qui non possiamo accontentarci della differenza tra le proporzioni poiché non sarebbe un indice affidabile e, di nuovo, do-

vremo ricorrere ad una trasformazione non lineare di P (Cohen, 1988). Questa volta la formula di trasformazione si avvale dell'arcoseno:

$$\phi = 2 \arcsin \sqrt{P} \tag{3.38}$$

È allora possibile ricavare con semplicità l'indice di grandezza dell'effetto mediante la formula:

$$h = \phi_1 - \phi_2 \tag{3.39}$$

Per semplificare la trasformazione delle proporzioni P in ϕ, riportiamo nella tabella 3.16 i valori corrispondenti dei due indici come indicati da Cohen (1988).

P	ϕ	P	ϕ	P	ϕ	P	ϕ
0,00	0,000	0,25	1,047	0,50	1,571	0,75	2,094
0,02	0,284	0,27	1,093	0,52	1,611	0,77	2,141
0,04	0,403	0,29	1,137	0,54	1,651	0,79	2,190
0,06	0,495	0,31	1,181	0,56	1,691	0,81	2,240
0,08	0,574	0,33	1,224	0,58	1,731	0,83	2,292
0,10	0,644	0,35	1,266	0,60	1,772	0,85	2,346
0,12	0,707	0,37	1,308	0,62	1,813	0,87	2,404
0,14	0,767	0,39	1,349	0,64	1,855	0,89	2,465
0,16	0,823	0,41	1,390	0,66	1,897	0,91	2,532
0,18	0,876	0,43	1,430	0,68	1,939	0,93	2,606
0,20	0,927	0,45	1,471	0,70	1,982	0,95	2,691
0,22	0,976	0,47	1,511	0,72	2,026	0,97	2,793
0,24	1,024	0,49	1,551	0,74	2,071	0,99	2,941
						1,00	3,142

Tabella 3.16 – *Trasformazione delle proporzioni P in* ϕ

ESEMPIO 3.13

Torniamo all'esempio preso in considerazione all'inizio di questo paragrafo. Poniamo di voler verificare se esiste una relazione tra l'estrazione sociale ed il successo scolastico. Una volta estratti a caso due campioni di alta e bassa estrazione sociale, misuriamo la proporzione di successi o insuccessi scolastici per ciascuno dei due. Riassumiamo nella tabella 3.17 i dati raccolti.

		Estrazione sociale	
		Alta	Bassa
Esito scolastico	Successo	65	34
	Insuccesso	15	46
	Tot	80	80

Tabella 3.17 – *Frequenze osservate della variabile "Esito scolastico" nei due gruppi "Estrazione sociale alta" e "Estrazione sociale bassa"*

La proporzione di successo per il campione ad alta estrazione sociale è $P_A = 65/80 = 0{,}81$, mentre la proporzione di successo per il campione a bassa estrazione sociale è $P_B = 34/80 = 0{,}42$.

Utilizzando la formula 3.38 possiamo effettuare la trasformazione delle proporzioni di successo P in ϕ per ciascuno dei due campioni:

$$\phi_A = 2 \arcsin \sqrt{0,81} = 2 \arcsin 0,9 = 2 \times 1,12 = 2,24;$$

$$\phi_B = 2 \arcsin \sqrt{0,42} = 2 \arcsin 0,65 = 2 \times 0,705 = 1,41.$$

A questo punto, mediante la formula 3.39 è possibile ricavare il valore dell'indice h.

$$h = 2,24 - 1,41 = 0,83.$$

Anche in questo caso, per comprendere il significato del valore ottenuto dell'indice, rimandiamo il lettore al paragrafo relativo all'interpretazione della grandezza dell'effetto.

3.3.3.3. Indice w

L'indice w viene utilizzato per valutare la grandezza dell'effetto nel test χ^2, sia quando questo test è impiegato nella bontà dell'adattamento, sia quando deriva dall'analisi delle tabelle di contingenza.

Nel caso che si utilizzi il test χ^2 per verificare la bontà dell'adattamento abbiamo un unico insieme di categorie contenenti frequenze o proporzioni e vogliamo confrontarle con uno stesso gruppo di categorie, di cui però conosciamo l'andamento sulla base dell'ipotesi nulla. Ad esempio possiamo ipotizzare l'equidistribuzione delle frequenze nelle categorie, oppure una distribuzione normale delle frequenze nelle categorie; questo dipenderà dal tipo di rilevazioni e dall'obiettivo della ricerca.

Nel caso del test χ^2 per tabelle di contingenza, le frequenze vengono altresì calcolate simultaneamente su più livelli di più variabili. Qui l'ipotesi nulla prevede che vi sia indipendenza (non vi sia associazione) tra le variabili prese in considerazione.

In entrambi i casi lavoriamo sulle celle: le categorie nel caso della bontà dell'adattamento e le categorie congiunte nel caso delle tabelle di contingenza. Per ciascuna cella abbiamo due proporzioni: una prevista dall'ipotesi nulla ed una prevista dall'ipotesi alternativa. L'indice *w* misura la discrepanza tra queste due proporzioni nelle varie celle, in formula:

$$w = \sqrt{\sum_{i=1}^{m} \frac{\left(P_{1i} - P_{0i}\right)^2}{P_{0i}}} \tag{3.40}$$

dove P_{0i} è la proporzione nella cella *i*-esima posta dall'ipotesi nulla, P_{1i} è la proporzione nella cella *i*-esima posta dall'ipotesi alternativa e *m* è il numero delle celle.

ESEMPIO 3.14.1

Consideriamo il caso dell'indice *w* nella bontà dell'adattamento. Poniamo di voler verificare se gli appartenenti ad una popolazione sono equamente distribuiti nelle diverse classi sociali (definite, poniamo in base al reddito).

Riassumiamo i dati ottenuti nella tabella 3.18.

Classi sociali					
Reddito molto basso	Reddito basso	Reddito medio	Reddito alto	Reddito molto alto	**Totale**
15	35	60	30	10	150

Tabella 3.18 – *Frequenze osservate della variabile "Classe sociale" su cinque livelli*

Dato che l'ipotesi nulla prevede che i partecipanti siano equamente distribuiti nelle cinque categorie di reddito, dovremo confrontare i dati ottenuti con la distribuzione teorica presentata nella tabella 3.19.

Classi sociali					
Reddito molto basso	Reddito basso	Reddito medio	Reddito alto	Reddito molto alto	**Totale**
30	30	30	30	30	150

Tabella 3.19 – *Frequenze teoriche della variabile "Classe sociale" su cinque livelli*

Per poter effettuare il calcolo dell'indice *w* dobbiamo prima di tutto trasformare i dati, espressi in frequenze, in proporzioni. Riassumiamo nella tabella 3.20 le trasformazioni in proporzioni sia per i dati osservati che per quelli teorici.

Proporzioni	**Classi sociali**				
	Reddito molto basso	Reddito basso	Reddito medio	Reddito alto	Reddito molto alto
Osservate	$\frac{15}{150} = 0,1$	$\frac{35}{150} = 0,23$	$\frac{60}{150} = 0,4$	$\frac{30}{150} = 0,2$	$\frac{10}{150} = 0,07$
Teoriche	$\frac{30}{150} = 0,2$	$\frac{30}{150} = 0,2$	$\frac{30}{150} = 0,2$	$\frac{30}{150} = 0,2$	$\frac{30}{150} = 0,2$

Tabella 3.20 – *Proporzioni osservate e teoriche della variabile "Classe sociale"*

A questo punto possiamo procedere al calcolo di *w*, mediante la formula 3.40:

$$w = \sqrt{\frac{(0,1-0,2)^2}{0,2} + \frac{(0,23-0,2)^2}{0,2} + \frac{(0,4-0,2)^2}{0,2} + \frac{(0,2-0,2)^2}{0,2} + \frac{(0,07-0,2)^2}{0,2}} =$$

$$= \sqrt{\frac{(-0,1)^2}{0,2} + \frac{(0,3)^2}{0,2} + \frac{(0,2)^2}{0,2} + \frac{(0)^2}{0,2} + \frac{(-0,13)^2}{0,2}} =$$

$$= \sqrt{\frac{0,01}{0,2} + \frac{0,0009}{0,2} + \frac{0,04}{0,2} + 0 + \frac{0,0169}{0,2}} =$$

$$= \sqrt{0,05 + 0,0045 + 0,2 + 0 + 0,0845} = \sqrt{0,339} = 0,58.$$

Per valutare il significato di questo tipo di risultato rimandiamo il lettore al paragrafo 3.5 sull'interpretazione della grandezza dell'effetto.

ESEMPIO 3.14.2

Consideriamo il caso dell'indice *w* per le tabelle di contingenza. Poniamo di svolgere una ricerca per verificare l'associazione tra genere e capacità di risolvere problemi matematici nei bambini di quinta elementare. Riassumiamo i dati raccolti nella tabella 3.21.

		Maschi	**Femmine**
Problemi di matematica	Risolti correttamente	7	20
	Risolti non correttamente	23	10

Tabella 3.21 – *Frequenze osservate della variabile "Problemi di matematica" gruppi "Maschi" e "Femmine"*

Anche in questo caso dobbiamo trasformare le frequenze in proporzioni (vedi tabella 3.22).

		Maschi	**Femmine**	**Tot**
Problemi di matematica	Risolti correttamente	$\frac{7}{60} = 0,12$	$\frac{20}{60} = 0,33$	0,45
	Risolti non correttamente	$\frac{23}{60} = 0,38$	$\frac{10}{60} = 0,17$	0,55
	Tot	0,5	0,5	1

Tabella 3.22 – *Proporzioni osservate della variabile "Problemi di matematica" gruppi "Maschi" e "Femmine"*

Dato che l'ipotesi nulla prevede che non vi sia associazione tra le due variabili, dobbiamo costruire una tabella delle proporzioni attese sulla base di H_0. Riassumiamo questi valori nella tabella 3.23.

		Maschi	**Femmine**	**Tot**
Problemi di matematica	Risolti correttamente	$0,45 \times 0,5 = 0,225$	$0,45 \times 0,5 = 0,225$	0,45
	Risolti non correttamente	$0,55 \times 0,5 = 0,275$	$0,55 \times 0,5 = 0,275$	0,55
	Tot	0,5	0,5	1

Tabella 3.23 – *Proporzioni teoriche della variabile "Problemi di matematica" gruppi "Maschi" e "Femmine"*

A questo punto procediamo al calcolo di *w*, mediante la formula 3.40.

$$w = \sqrt{\frac{(0,12 - 0,225)^2}{0,225} + \frac{(0,33 - 0,225)^2}{0,225} + \frac{(0,38 - 0,275)^2}{0,275} + \frac{(0,17 - 0,275)^2}{0,275}} =$$

$$= \sqrt{\frac{(-0,105)^2}{0,225} + \frac{(0,105)^2}{0,225} + \frac{(0,105)^2}{0,275} + \frac{(-0,105)^2}{0,275}} =$$

$$= \sqrt{\frac{0,011}{0,225} + \frac{0,2011}{0,225} + \frac{0,011}{0,275} + \frac{0,2011}{0,275}} = \sqrt{0,05 + 0,05 + 0,04 + 0,04} = \sqrt{0,18} = 0,42.$$

Come in precedenza, per valutare il significato di questo tipo di risultato rimandiamo il lettore al paragrafo sull'interpretazione della grandezza dell'effetto.

Per quanto l'indice *w* sia una misura della grandezza dell'effetto per le tabelle di contingenza, abbiamo visto in precedenza che esistono altri indici utilizzabili a questo scopo, in particolare tra le misure di associazione. Vediamo, quindi, che legami esistono tra l'indice *w* e le misure di associazione per le tabelle di contingenza.

La tabella 3.24 mostra le formule che permettono di trasformare un indice nell'altro, mentre la tabella 3.25 mostra la corrispondenza tra i valori che i vari indici possono assumere.

Misura di associazione	**Formula**	**Trasformazione in *w***	**Trasformazione di *w* nella misura di associazione**
C	$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$	$w = \sqrt{\frac{C^2}{1 - C^2}}$	$C = \sqrt{\frac{w^2}{w^2 + 1}}$
ϕ	$\phi = \sqrt{\frac{\chi^2_{(1)}}{N}}$	$w = \phi$	$cw = \phi$
ϕ'	$\phi' = \sqrt{\frac{\chi^2}{N(df_{smaller})}}$	$w = \phi' \sqrt{df_{smaller}}$	$\phi' = \frac{w}{\sqrt{df_{smaller}}}$

Tabella 3.24 – *Formule per la trasformazione delle misure della grandezza dell'effetto per tabelle di contingenza: indice w e misure di associazione C, ϕ e ϕ'*

		ϕ'				
w	***C***	***k* = 2 ⇔ ϕ**	***k* = 3**	***k* = 4**	***k* = 5**	***k* = 6**
0,10	0,100	0,10	0,071	0,058	0,050	0,045
0,20	0,196	0,20	0,141	0,115	0,100	0,089
0,30	0,287	0,30	0,212	0,173	0,150	0,134
0,40	0,371	0,40	0,283	0,231	0,200	0,179
0,50	0,447	0,50	0,354	0,289	0,250	0,224
0,60	0,514	0,60	0,424	0,346	0,300	0,268
0,70	0,573	0,70	0,495	0,404	0,350	0,313
0,80	0,625	0,80	0,566	0,462	0,400	0,358
0,90	0,669	0,90	0,636	0,520	0,450	0,402

Tabella 3.25 – *Trasformazione delle misure della grandezza dell'effetto per tabelle di contingenza: indice w e misure di associazione* C, ϕ *e* ϕ'

ESEMPIO 3.15

Prendiamo di nuovo in considerazione i dati dell'esempio precedente. Trattandosi di una tabella 2×2 ci limiteremo a confrontare l'indice *w* con gli indici ϕ e *C* per tabelle 2×2. Lo stesso procedimento di trasformazione potrebbe essere eseguito per l'indice ϕ' qualora si utilizzassero tabelle con un maggior numero di righe o di colonne. Effettuando una analisi mediante il test χ^2 sui dati proposti, si ottiene un valore del test corrispondente a $\chi_{(1)}^2 = 11{,}38$ con $p = 0{,}001$.

Utilizzando le formule 3.14 e 3.8 è possibile calcolare l'indice *C* e l'indice ϕ.

$$C = \sqrt{\frac{11,38}{11,38 + 60}} = \sqrt{\frac{11,38}{71,38}} = \sqrt{0,16} = 0,4;$$

$$\phi = \sqrt{\frac{11,38}{60}} = \sqrt{0,19} = 0,44.$$

Utilizzando poi la formula per la trasformazione di *w* in *C*, contenuta nella tabella 3.24, possiamo ricavare di nuovo il valore di *w* partendo dal valore calcolato di *C*.

$$w = \sqrt{\frac{0,4^2}{1 - 0,4^2}} = \sqrt{\frac{0,16}{1 - 0,16}} = \sqrt{\frac{0,16}{0,84}} = \sqrt{0,19} = 0,44.$$

3.3.3.4. Indice f

L'indice *f* (Cohen, 1988) viene utilizzato come misura della grandezza dell'effetto nell'ANOVA e nell'ANCOVA. Esso può essere considerato come una estensione dell'utilizzo dell'indice *d* di Cohen ai casi in cui i gruppi sono più di due, essendo i livelli della variabile indipendente più di due ($k > 2$).

L'indice *f* può essere calcolato mediante la formula:

$$f = \frac{\sigma_m}{\sigma} \tag{3.41}$$

dove σ corrisponde alla deviazione standard di una qualsiasi delle popolazioni, partendo (come nel caso dell'indice *d*) dall'assunto che esse siano uguali tra loro, e σ_m, per gruppi di uguale numerosità, è dato dalla formula:

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}(m_i - m)^2}{k}} \tag{3.42}$$

dove gli m_i sono i singoli valori delle medie dei *k* gruppi estratti dalla popolazione ed *m* è la media della popolazione.

L'indice *f* può assumere valori prossimi allo 0 quando le medie dei gruppi sono tutte simili tra loro e quindi gli scarti saranno molto bassi, oppure può assumere un valore tanto più grande quanto più σ_m cresce rispetto a σ.

Esempio 3.16

In una ricerca sull'effetto di un nuovo trattamento dei disturbi d'ansia, il campione viene suddiviso in tre gruppi: farmaco, placebo e controllo, costituito da pazienti non trattati. Dopo una settimana dall'inizio del trattamento ai partecipanti viene somministrato un questionario per rilevarne il livello di ansia. I dati relativi all'ansia sono riportati nella tabella 3.26.

Il gruppo di pazienti ai quali è stato somministrato il farmaco ha una media pari a 6,6 con deviazione standard di 2,2, mentre il gruppo al quale è stato somministrato il placebo ha una media pari a 11,2 con deviazione standard di 3,1, ed infine il gruppo di controllo, in assenza di trattamento, ha una media di 16,5, con deviazione standard di 2,5. L'intero campione, non suddiviso in gruppi, ha una media di 11,4, con deviazione standard di 0,48.

Utilizzando la formula 3.42 possiamo ricavare il valore di σ_m. Vediamo come procedere:

$$\sigma_m = \sqrt{\frac{(6,6 - 11,4)^2 + (11,2 - 11,4)^2 + (16,5 - 11,4)^2}{3}} = \sqrt{\frac{(-4,8)^2 + (-0,2)^2 + (5,1)^2}{3}} =$$

$$= \sqrt{\frac{23,04 + 0,04 + 26,05}{3}} = \sqrt{\frac{49,13}{3}} = \sqrt{16,38} = 4,05.$$

A questo punto, mediante la formula 3.41 calcoliamo il valore dell'indice *f*.

$$f = \frac{4,05}{2,5} = 1,62.$$

Al denominatore è stato deciso di utilizzare la deviazione standard del gruppo in assenza di trattamento, questo per due motivi: sia perché si tratta di un vero e proprio gruppo di controllo, quindi ci possiamo aspettare che sia il valore più prossimo a quello della popolazione di origine, sia perché, semplicemente, è il valore centrale.

Prima di procedere alla decisione di quale tra le deviazioni standard utilizzare, è stata, comunque, effettuata una verifica dell'omogeneità delle varianze che, appunto, risultano non differenziarsi tra loro in modo statisticamente significativo. Consigliamo al lettore di eseguire sempre la verifica delle omogeneità delle varianze, prima di procedere ad utilizzare questa formula. Laddove le varianze risultassero diverse tra loro in modo statisticamente significativo, non sarebbe certamente corretto utilizzare questo tipo di formula.

Farmaco	**Placebo**	**Nessun trattamento**
5,00	12,00	17,00
4,00	14,00	18,00
9,00	8,00	15,00
7,00	7,00	16,00
8,00	16,00	18,00
5,00	14,00	14,00
5,00	10,00	19,00
11,00	11,00	20,00
9,00	9,00	15,00
6,00	6,00	15,00
3,00	8,00	14,00
8,00	12,00	11,00
8,00	13,00	18,00
6,00	13,00	20,00
5,00	15,00	17,00

Tabella 3.26 – *Punteggi ottenuti al questionario sull'ansia*

Come abbiamo detto in apertura di questo paragrafo, l'indice *f* è legato all'indice *d* non solo concettualmente, ma anche matematicamente. Cohen (1988) ha infatti dimostrato il legame tra *f* e *d* in una serie di possibili situazioni:

1. Per $k = 2$ la relazione tra *f* e *d* è:

$$d = 2f \tag{3.43}$$

2. Per $k > 2$, due medie (la maggiore e la minore) definiscono *d* e le restanti medie ($k - 2$) si trovano al centro della distribuzione (*variabilità minima*):

$$d = f\sqrt{2k} \tag{3.44}$$

3. Per $k > 2$, due medie (la maggiore e la minore) definiscono *d* e le restanti medie ($k - 2$) sono equamente spaziate all'interno della distribuzione (*variabilità intermedia*):

$$d = 2f\sqrt{\frac{3(k-1)}{k+1}} \tag{3.45}$$

4. Per $k > 2$, due medie (la maggiore e la minore) definiscono *d* e le restanti medie ($k - 2$) si trovano ai margini della distribuzione (*variabilità massima*):

Quando *k* è pari vale la (3.43), se invece è dispari,

$$d = f\frac{2k}{\sqrt{k^2 - 1}} \tag{3.46}$$

Infine, è possibile ricavare l'indice *f* direttamente dal valore dell'indice η^2, in formula:

$$f = \sqrt{\frac{\eta^2}{1 - \eta^2}}\ , \tag{3.47}$$

e viceversa:

$$\eta^2 = \sqrt{\frac{f^2}{1 + f^2}} \tag{3.48}$$

Concludendo, se prendiamo in considerazione la formula (3.48) di trasformazione di η^2 in *f*, e ricordando la formula di η^2:

$$\eta^2 = \frac{SS_{effetto}}{SS_{totale}} \tag{3.1}$$

mediante una serie di passaggi matematici di estrema semplicità possiamo giungere ad una formula per il calcolo dell'indice *f* che permette di utilizzare direttamente i dati contenuti nell'output dell'ANOVA (proprio come avviene per il calcolo di η^2):

$$f = \sqrt{\frac{SS_{effetto}}{SS_{errore}}} \tag{3.49}$$

Nella tabella 3.27 presentiamo alcuni valori di η^2 in funzione di f ed alcuni valori di f in funzione di η^2.

f	η^2	η^2	f
0,00	0,0000	0,00	0,000
0,05	0,0025	0,01	0,101
0,10	0,0099	0,02	0,143
0,15	0,0220	0,03	0,176
0,20	0,0385	0,04	0,204
0,25	0,0588	0,05	0,229
0,30	0,0826	0,06	0,253
0,35	0,1091	0,07	0,274
0,40	0,1379	0,08	0,295
0,45	0,1684	0,09	0,314
0,50	0,2000	0,10	0,333
0,55	0,2322	0,15	0,420
0,60	0,2647	0,20	0,500
0,65	0,2970	0,25	0,577
0,70	0,3289	0,30	0,655
0,75	0,3600	0,40	0,816
0,80	0,3902	0,50	1,000
0,85	0,4194	0,60	1,225
0,90	0,4475	0,70	1,528
0,95	0,4744	0,80	2,000
1,00	0,5000	0,90	3,000

Tabella 3.27 – *Valori di* η^2 *in funzione di* f *e valori di* f *in funzione di* η^2

ESEMPIO 3.17

Prendendo nuovamente in considerazione l'esempio presentato all'inizio di questo paragrafo, effettuiamo una ANOVA univariata sui dati raccolti. Riassumiamo nella tabella 3.28 i risultati dell'ANOVA.

Partendo dai valori contenuti nella tabella dei risultati dell'ANOVA e sostituendoli nella formula 3.49 possiamo ricavare, molto facilmente, il valore dell'indice f.

$$f = \sqrt{\frac{731,244}{287,733}} = \sqrt{2,54} = 1,6.$$

	Somma dei quadrati	*Gl*	Media dei quadrati	*F*	*p*
Tra i gruppi	731,244	2	365,622	53,369	0,000
Entro i gruppi	287,733	42	6,851		
Totale	1018,978	44			

Tabella 3.28 – *Risultati dell'ANOVA*

3.3.3.5. Indice f²

L'indice f^2 (Cohen, 1988) viene utilizzato per il calcolo della grandezza dell'effetto nella regressione multipla ed in altri metodi multivariati. Dato che questi metodi sono comunque basati sulla distribuzione *F*, in realtà quello che viene utilizzato è sempre l'indice *f*. In questo caso si lavora con l'indice f^2 poiché i metodi multivariati si basano su proporzioni di varianza spiegata, ma dovremo sempre tenere presente che questo indice è sostanzialmente identico al già discusso indice *f*. Per il calcolo dell'indice f^2 si utilizza la formula

$$f^2 = \frac{PV_{effetto}}{PV_{errore}} \tag{3.50}$$

dove $PV_{effetto}$ corrisponde alla proporzione di varianza della variabile dipendente spiegata dal fattore preso in considerazione, mentre PV_{errore} corrisponde alla proporzione di varianza spiegata dall'errore o varianza residua. Osservando questa formula risulta evidente la sostanziale identità degli indici *f* ed f^2.

Per comprendere a pieno l'utilizzo di questo indice, prendiamo in considerazione alcune situazioni tipiche dell'analisi di regressione:

• *Analisi di regressione semplice*: una variabile dipendente *Y* ed una variabile indipendente *B*.

$$f^2 = \frac{R^2_{Y.B}}{1 - R^2_{Y.B}} \tag{3.51}$$

dove $R^2_{Y.B}$ è la proporzione di varianza spiegata dal fattore *B* e $1 - R^2_{Y.B}$ è la varianza residua.

• *Analisi di regressione multipla* (caso 1): una variabile dipendente *Y* e due variabili indipendenti *A* e *B*; siamo interessati alla proporzione di varianza spiegata da *B* senza tenere in considerazione *A*.

$$f^2 = \frac{R^2_{Y.A,B} - R^2_{Y.A}}{1 - R^2_{Y.A,B}}. \tag{3.52}$$

• *Analisi di regressione multipla* (caso 2): una variabile dipendente Y e tre variabili indipendenti A, B e C; siamo interessati alla proporzione di varianza spiegata da B senza tenere in considerazione A, ma tenendo conto anche della variabile C.

$$f^2 = \frac{R^2_{Y.A,B} - R^2_{Y.A}}{1 - R^2_{Y.A,B,C}}. \tag{3.53}$$

ESEMPIO 3.18.1

Poniamo di svolgere una ricerca per verificare se l'età dei genitori è un fattore che influenza il livello di autostima dei bambini. Il livello di autostima viene misurato mediante una scala che ha punteggio da 0 a 40; laddove a punteggio alto corrisponde alta autostima e viceversa.

Effettuando una analisi di regressione semplice sui dati raccolti otteniamo un valore di R^2 = 0,67. La varianza residua o coefficiente di non determinazione, corrispondente a $1 - R^2$, è (1 − 0,67) = 0,33. Sostituendo i valori ottenuti nella formula 3.51 possiamo ricavare facilmente il valore di f^2.

$$f^2 = \frac{0{,}67}{0{,}33} = 2{,}03.$$

ESEMPIO 3.18.2

Poniamo adesso di introdurre una seconda variabile indipendente nel disegno di ricerca presentato: il livello di scolarità dei genitori (misurato in anni). In realtà, però, vogliamo verificare che proporzione di varianza viene spiegata dalla scolarità dei genitori escludendo l'effetto dovuto alla loro età.

Effettuando una analisi di regressione multipla sui dati si ottengono i risultati riassunti nella tabella 3.29.

	Coefficienti non standardizzati		**Coefficienti standardizzati**		
Modello	B	Errore standard	Beta	*t*	*p*
Costante	− 1,1	2,5		− 0,440	0,662
Età genitori	0,458	0,147	0,386	3,124	0,003
Scolarità genitori	0,981	0,226	0,538	4,349	0,000

Tabella 3.29 – *Risultati Analisi di Regressione Multipla*

Dovendo utilizzare la formula 3.52 per il calcolo dell'indice f^2, evidentemente dovremo ricavare gli elementi contenuti nella stessa, elementi che purtroppo non vengono forniti nell'output dell'analisi di regressione multipla. Per ricavare $R^2_{Y.A,B}$, cioè il coefficiente di determinazione per un modello con due variabili indipendenti, possiamo avvalerci della seguente formula:

$$R^2_{Y.A,B} = \beta^*_A \times r_{Y.A} + \beta^*_B \times r_{Y.B}, \tag{3.54}$$

dove β^* è il coefficiente di regressione standardizzato per ciascuna variabile ed r è il coefficiente di correlazione per ciascuna variabile indipendente con la variabile dipendente.

Procedendo al calcolo dei coefficienti di correlazione sui dati della nostra ricerca abbiamo che il coefficiente di correlazione tra età dei genitori e autostima è $r_{Y.A}$= 0,819 e il coefficiente di correlazione tra scolarità dei genitori ed autostima è $r_{Y.B}$ = 0,848.

Sostituendo i coefficienti di correlazione ed i coefficienti di regressione standardizzati (vedi tabella 3.29) nella formula 3.54, possiamo ricavare $R^2_{Y.B}$.

$$R^2_{Y.A,B} = 0,386 \times 0,819 + 0,538 \times 0,848 = 0,316 + 0,456 = 0,772.$$

e, di conseguenza, possiamo calcolare anche i restanti membri della formula.

$$R^2_{Y.A} = 0,386 \times 0,819 = 0,316,$$

$$1 - R^2_{Y.A,B} = 1 - 0,772 = 0,228.$$

Sostituendo i valori ottenuti nella formula 3.52, otterrò il valore dell'indice f^2.

$$f^2 = \frac{0,772 - 0,316}{0,228} = \frac{0,456}{0,228} = 2.$$

Esempio 3.18.3

Prendiamo, infine, in considerazione il caso più complesso: una analisi di regressione multipla con tre variabili indipendenti. Poniamo di introdurre nel disegno della ricerca sopra esposto una terza variabile indipendente: età del bambino. A questo punto, però, l'interesse è quello di verificare la proporzione di varianza spiegata dalla variabile scolarità dei genitori escludendo l'effetto dell'età degli stessi, ma tenendo conto anche della variabile età del bambino. Il procedimento sarà analogo a quello seguito per l'esempio 2 e ci avvarremo nuovamente della formula 3.54 per ricavare i dati utili al calcolo dell'indice f^2. Riassumiamo nella tabella 3.30 i risultati dell'analisi di regressione multipla.

Modello	Coefficienti non standardizzati		Coefficienti standardizzati		
	B	Errore standard	Beta	*t*	*p*
Costante	− 1,262	2,514		− 0,502	0,618
Età genitori	0,568	0,194	0,479	2,933	0,005
Scolarità genitori	1,062	0,244	0,582	4,346	0,000
Età bambino	− 0,504	0,577	− 0,146	− 0,874	0,387

Tabella 3.30 – *Risultati dell'analisi di regressione multipla*

Riassumiamo in tabella 3.31 i risultati delle correlazioni tra le variabili indipendenti e la variabile dipendente:

	Autostima
Età genitori	0,819
Scolarità genitori	0,848
Età bambino	0,749

Tabella 3.31 – *Coefficienti di correlazione r tra le variabili indipendenti e la variabile dipendente*

Sostituendo i valori ottenuti nella formula 3.53 possiamo procedere al calcolo dell'indice f^2.

$$f^2 = \frac{\left[\left(0,479 \times 0,819\right) + \left(0,582 \times 0,848\right)\right] - \left(0,479 \times 0,819\right)}{1 - \left[\left(0,479 \times 0,819\right) + \left(0,582 \times 0,848\right) + \left(-0,146 \times 0,749\right)\right]} =$$

$$= \frac{\left[0,392 + 0,493\right] - 0,392}{1 - \left[0,392 + 0,493 - 0,109\right]} = \frac{0,493}{1 - 0,776} = \frac{0,493}{0,224} = 2,2.$$

Per la valutazione dei valori dell'indice f^2 ottenuti nei tre esempi precedenti, rimandiamo il lettore al paragrafo relativo all'interpretazione della grandezza dell'effetto.

3.4. La BESD di Rosenthal e Rubin

Rosenthal e Rubin (1982) hanno proposto un metodo per visualizzare ed interpretare la grandezza dell'effetto in modo semplice e chiaro: tale metodo è denominato *Binomial Effect Size Display* (BESD). L'obiettivo dell'introduzione di questo metodo non è stato tanto quello di risolvere la controversia in merito a quali indici di grandezza dell'effetto siano più adatti, bensì quello di fornire uno strumento che fosse facilmente comprensibile ed utilizzabile sia dai ricercatori che dagli studenti, e che nel contempo si potesse facilmente applicare a numerosi contesti.

r^2	r	Incremento della quota di successo		Differenza nella quota di successo
		da	a	
0,01	0,10	0,45	0,55	0,10
0,04	0,20	0,40	0,60	0,20
0,09	0,30	0,35	0,65	0,30
0,16	0,40	0,30	0,70	0,40
0,25	0,50	0,25	0,75	0,50
0,36	0,60	0,20	0,80	0,60
0,49	0,70	0,15	0,85	0,70
0,64	0,80	0,10	0,90	0,80
0,81	0,90	0,05	0,95	0,90
1,00	1,00	0,00	1,00	1,00

Tabella 3.32 – *Quote di incremento dei successi a partire da alcuni valori di r ed r^2*

La BESD si concentra su un quesito fondamentale: quale effetto produce l'introduzione di un certo trattamento sulla quota di successo (laddove per successo intendiamo un miglioramento, una guarigione, la stima corretta della lunghezza di un segmento, ecc.)? Lo scopo di questo metodo è quello di rendere evidente il cambiamento nella quota di successo da attribuire all'introduzione di un certo trattamento.

Per effettuare una stima del miglioramento, Rosenthal e Rubin (1982) partono dai valori dei coefficienti r ed r^2 per calcolare i corrispondenti valori di accrescimento della quota di successi. Il calcolo delle quote di incremento dei successi a partire dal valore di r è allora molto semplice. La quota di successo di partenza, corrispondente quindi

alla quota di successi nel gruppo di controllo, prima o in assenza del trattamento, è data dalla formula:

$$0,50 - \frac{r}{2} \tag{3.55}$$

La quota di successi dopo il trattamento, quella del gruppo sperimentale, è data invece dalla formula:

$$0,50 + \frac{r}{2} \tag{3.56}$$

Nella tabella 3.32 riportiamo i valori indicati dagli autori di r e r^2.

Osservando l'ultima colonna della tabella è facile notare come la differenza nella quota di successo tra prima e dopo sia identica al valore del coefficiente di correlazione r.

ESEMPIO 3.19

Poniamo di svolgere una ricerca per verificare l'impatto di un intervento chirurgico sulla sopravvivenza di pazienti affetti da una grave cardiopatia. Il coefficiente di correlazione calcolato sui dati raccolti è r_{pb} = 0,33. Sostituendo questo valore rispettivamente nelle formule 3.55 e 3.56 otteniamo i seguenti valori:

$$0,50 - \frac{0,33}{2} = 0,50 - 0,165 = 0,335;$$

$$0,50 + \frac{0,33}{2} = 0,50 + 0,165 = 0,665.$$

Questo significa che la quota di successo, quindi la probabilità sopravvivenza dei pazienti, incrementa dal 33,5% al 66,5% per coloro che vengono sottoposti all'intervento.

L'utilizzo di questo metodo è, inoltre, favorito dall'esistenza di una serie di formule che consentono di trasformare i test statistici più utilizzati in coefficienti di correlazione. Abbiamo in precedenza già mostrato e discusso queste formule, ma per completezza di esposizione riproponiamo nella tabella 3.33 quelle di maggior interesse.

I due autori (Rosenthal e Rubin, 1982) suggeriscono che la procedura BESD è più appropriata quando le varianze dei due gruppi sono simili.

Potrebbe sembrare che il metodo BESD possa essere utilizzato solo quando la variabile dipendente è dicotomica (successo *vs* insuccesso).

In realtà il metodo BESD è spesso una rappresentazione realistica della grandezza dell'effetto del trattamento anche quando la variabile dipendente è su scala metrica a patto che la sua varianza sia approssimativamente la stessa in due gruppi di numerosità quasi uguale.

Test	Trasformazione in *r*	Note
t	$r = \sqrt{\frac{t^2}{t^2 + gl}}$	
F	$r = \sqrt{\frac{F_{(1,?)}}{F_{(1,?)} + gl_{errore}}}$	Da utilizzare solo con ANOVA univariata per due campioni ($gl = 1$).
χ^2	$r = \sqrt{\frac{\chi^2_{(1)}}{N}}$	Da utilizzare solo con tabelle di contingenza 2 × 2 ($gl = 1$).
d	$r = \frac{d}{\sqrt{d^2 + 4}}$	

Tabella 3.33 – *Calcolo del coefficiente r a partire da t, F, χ^2 e dall'indice d*

Supponiamo che Y sia una variabile dipendente avente uguale varianza in due gruppi sperimentali di numerosità uguale. Se Y fosse binomiale, con uguale varianza nei due gruppi sperimentali, allora in un gruppo la media sarebbe *p* mentre nell'altro gruppo la media sarebbe (1 – *p*); così come previsto dalla BESD. Poniamo che Y* sia un trasposizione della variabile Y in forma dicotomica, uguale a 1 quando Y è superiore alla mediana, e a –1 quando è inferiore.

Se calcoliamo la grandezza dell'effetto mediante la variabile Y* la BESD è il metodo di elezione, dato che si tratta di una variabile dicotomica avente le medie dei due gruppi sperimentali ugualmente maggiori e minori di 0,5. Il quesito è: quanto può essere diversa la correlazione ρ tra il trattamento e Y rispetto alla correlazione ϕ tra il trattamento e Y*? A tal fine dobbiamo ricavare la relazione esistente tra i due coefficienti e verificare quanto variano l'uno rispetto all'altro.

Poniamo $\phi = 1 - 2T$, con T funzione di ρ per due diverse approssimazioni:

1. Se Y è distribuita normalmente, T corrisponde al valore di *p* ad una coda associato a:

$$\frac{\rho}{\sqrt{1 - \rho^2}} \tag{3.57}$$

2. Se Y segue la distribuzione *t*, T corrisponde al valore di *p* ad una coda associato a:

$$\frac{\rho}{\sqrt{1-\rho^2}} \times \frac{\sqrt{gl-2}}{gl}. \tag{3.58}$$

Nella tabella 3.34 riportiamo i valori di corrispondenza tra ρ e ϕ calcolati seguendo il metodo appena esposto (Rosenthal e Rubin, 1982).

Variabile dipendente continua	**Variabile dipendente dicotomica**	
Indice ρ	**Indice** ϕ	
	Distribuzione normale	Distribuzione *t*
0,05	0,04	0,06
0,10	0,08	0,13
0,15	0,12	0,19
0,20	0,16	0,25
0,25	0,20	0,31
0,30	0,25	0,38
0,35	0,29	0,44
0,40	0,34	0,50
0,45	0,39	0,55
0,50	0,44	0,61
0,55	0,49	0,66
0,60	0,55	0,72
0,65	0,61	0,76
0,70	0,67	0,81
0,75	0,74	0,86
0,80	0,82	0,90
0,85	0,89	0,93
0,90	0,96	0,96
0,95	0,998	0,99

Tabella 3.34 – *Valori dei coefficienti di correlazione calcolati su variabili dicotomiche, variabili continue distribuite normalmente e continue aventi distribuzione t*

Esaminando la tabella si comprende con chiarezza che solitamente i valori di ρ e ϕ sono piuttosto simili. Ne consegue che avendo a disposizione un valore ρ esso può essere rappresentato mediante la BESD, senza che vi sia una differenza apprezzabile rispetto ad avere un valore ϕ calcolato sulla stessa variabile dicotomizzata e rappresentato con la BESD.

Un accorgimento potrebbe, altresì, essere quello di correggere il valore della correlazione ρ utilizzando al suo posto uno dei due valori di ϕ in base al tipo di distribuzione dei dati (normale *vs t*).

Esempio 3.20

Poniamo di svolgere una ricerca per verificare se l'introduzione di un nuovo metodo di insegnamento ha una influenza sulle abilità di lettura acquisite dai bambini al termine della prima classe elementare. Riassumiamo i dati nella tabella 3.35.

Punteggi abilità di lettura			
Nuovo metodo	Metodo tradizionale	Nuovo metodo	Metodo tradizionale
23	20	27	22
20	15	27	22
21	21	30	23
25	16	28	14
22	17	26	24
25	17	29	20
20	18	30	25
23	16	28	25
24	19	31	26
22	20	19	23
25	15	24	19
21	19	25	20
25	20	26	21
26	18	24	24
29	21	25	20

Tabella 3.35 – *Punteggi ottenuti al test di abilità di lettura in due gruppi (sperimentale e controllo)*

Il coefficiente di correlazione r_{pb} tra trattamento e variabile dipendente continua (punteggio di abilità verbale) è 0,625. Utilizzando il punteggio di mediana di tutto il campione, cioè 22,5, trasformiamo la variabile dipendente da continua a dicotomica: assegniamo tutti coloro che hanno un punteggio inferiore alla mediana alla categoria "punteggio basso", viceversa tutti coloro che hanno punteggio superiore alla mediana faranno parte della categoria "punteggio alto". Riassumiamo i dati trasformati nella tabella 3.36.

		Tipo di insegnamento	
		Tradizionale	Nuovo
Punteggio	Basso	23	7
	Alto	7	23

Tabella 3.36 – *Frequenze della variabile "Punteggio di abilità di lettura", livelli "Alto" e " Basso", gruppi "Insegnamento Tradizionale" e "Insegnamento Nuovo"*

Il coefficiente di correlazione ϕ tra trattamento e variabile dipendente è 0,533.

Vediamo adesso come avremmo potuto ricavare il coefficiente ϕ partendo dal coefficiente r_{pb}. Supponiamo ad avere una distribuzione normale e di poter quindi utilizzare la formula 3.57 per calcolare il valore di T. Vediamo come procedere:

$$\frac{0,625}{\sqrt{1-0,625^2}} = \frac{0,625}{\sqrt{1-0,39}} = \frac{0,625}{\sqrt{0,61}} = \frac{0,625}{0,78} = 0,8.$$

Andando a controllare il valore di p corrispondente al valore di $z = 0,8$ sulla tavola della distribuzione normale, possiamo ricavare molto semplicemente il valore di T = 0,2119 e, da questo, il valore del coefficiente $\phi = 1 - 2T = 1 - 0,4238 = 0,5762$.

La lieve differenza ottenuta tra i due valori del coefficiente ϕ è da attribuirsi alla composizione anomala dei dati presi in considerazione per formulare l'esempio.

3.5. L'INTERPRETAZIONE DELLA GRANDEZZA DELL'EFFETTO

Non esiste una regola universalmente accettata per interpretare la grandezza dell'effetto, ma esiste una tradizione di ricerca che può essere d'aiuto. A seconda degli interrogativi che la ricerca si pone, può capitare che un grosso effetto sia poco rilevante mentre uno piccolo sia molto importante (Rosenthal, 1993). Spesso la letteratura ci offre dei punteggi di riferimento, che verranno riportati qui di seguito, ma che devono essere utilizzati con cautela, in modo adeguato e solo in talune circostanze. Ritorneremo in seguito sul dibattito relativo all'opportunità di questi punteggi.

In questo paragrafo, forniremo inizialmente i valori di riferimento che la letteratura riporta in merito ad alcuni degli indici presentati in precedenza, dopodiché discuteremo l'utilità della relazione tra l'indice *d* ed il coefficiente di correlazione *r* ai fini dell'interpretazione dei risultati, ed infine esporremo un metodo messo a punto da Cohen e basato sulla percentuale di sovrapposizione delle distribuzioni dei due campioni.

3.5.1. Valori di riferimento per singoli indici

La tabella 3.37 riassume i valori di riferimento presenti in letteratura per la maggior parte degli indici di grandezza dell'effetto presi in considerazione nei paragrafi precedenti.

		Grandezza dell'effetto		
Tipo di analisi	**Test**	**Piccolo**	**Medio**	**Grande**
t test per campioni indipendenti	***d***	0,20	0,50	0,80
t test per campioni dipendenti	***f***	0,10	0,25	0,40
ANOVA	***f***	0,10	0,25	0,40
	w^2	0,01	0,06	0,15
ANCOVA	***f***	0,10	0,25	0,40
Correlazione	***r***	0,10	0,30	0,50
Regressione multipla e metodi multivariati	f^2	0,02	0,15	0,35
Chi quadrato	***w***	0,10	0,30	0,50
Differenze tra coefficienti di correlazione	***q***	0,10	0,30	0,50
Differenze tra proporzioni	***h***	0,20	0,50	0,80

Tabella 3.37 – *Entità della grandezza dell'effetto degli indici più diffusi*

Per quanto concerne, altresì, l'indice ϕ' di Cramer le valutazioni della grandezza dell'effetto dipendono dal numero di gradi di libertà del test χ^2, e sono riassunti nella tabella 3.38 (Cohen, 1988).

Vedremo in seguito che queste indicazioni, seppure riportate dalla maggior parte dei lavori sugli indici di grandezza dell'effetto, sono da considerarsi grossolane ed imprecise, e vi sono altri metodi da preferire, soprattutto nel caso si utilizzi proprio l'indice *d*.

	Grandezza dell'effetto		
gl	**Piccolo**	**Medio**	**Grande**
1	0,10	0,30	0,50
2	0,07	0,21	0,35
3	0,06	0,17	0,29

Tabella 3.38 – *Entità della grandezza dell'effetto dell'indice* ϕ' *di Cramer*

3.5.2. Relazione tra l'indice d *ed il coefficiente* r

Generalmente è consigliato utilizzare come riferimento i valori di *d* e *r* ritenuti convenzionali (Cohen, 1992a e 1992b). Riportiamo questi valori nella tabella 3.39. Inizialmente questi valori non erano stati motivati empiricamente, ma sembra che gli effetti medi trovino una buona corrispondenza con la media degli effetti riscontrati in vari ambiti della psicologia (Cooper e Findley, 1982; Haase, Waechter e Solomon, 1982; Sedlmeier e Gigerenzer, 1989).

		d	***r***
Grandezza dell'effetto	**Piccolo**	0,2	0,1
	Medio	0,5	0,3
	Grande	0,8	0,5

Tabella 3.39 – *Entità della grandezza dell'effetto utilizzando gli indici d ed r*

L'utilizzo di punteggi di riferimento non è l'unico modo in cui la relazione esistente tra *d* ed *r* può essere utilizzata per interpretare la grandezza dell'effetto. Infatti, come abbiamo visto nei paragrafi precedenti, l'indice *d* di Cohen può essere convertito nel coefficiente di correlazione *r* e viceversa. Ad esempio, avendo un valore *d* = 0,8 ed utilizzando la formula:

$$r = \frac{d}{\sqrt{d^2 + 4}} \tag{3.11}$$

possiamo facilmente ricavare il valore di *r* = 0,371.

Il quadrato del coefficiente di correlazione (r^2) corrisponde alla quota di varianza della variabile dipendente spiegata dall'appartenenza ad uno dei gruppi formati dalla variabile indipendente. Quindi, con un valore dell'indice d = 0,8 e un valore di $r^2 = (0{,}371)^2 = 0{,}137$, la quota di varianza della variabile dipendente spiegata dall'appartenenza al gruppo sperimentale o di controllo è del 13,7%.

Nella tabella 3.40 riportiamo nel dettaglio i valori di corrispondenza della relazione tra l'indice d ed i coefficienti r ed r^2.

Indicazioni di Cohen	***d***	***r***	r^2
Effetto grande	2,0	0,707	0,500
	1,8	0,669	0,448
	1,6	0,625	0,390
	1,4	0,573	0,329
	1,2	0,514	0,265
	1,0	0,447	0,200
	0,8	0,371	0,138
Effetto medio	0,7	0,330	0,109
	0,6	0,287	0,083
	0,5	0,243	0,059
Effetto piccolo	0,4	0,196	0,038
	0,3	0,148	0,022
	0,2	0,100	0,010
	0,1	0,050	0,002
	0,0	0,000	0,000

Tabella 3.40 – *Valori degli indici d, r ed r^2 per vari livelli di grandezza dell'effetto*

3.5.3. Sovrapposizione delle distribuzioni

Cohen (1988) ha sviluppato un metodo per l'interpretazione della grandezza dell'effetto che si basa sul livello di sovrapposizione delle distribuzioni dei punteggi dei due campioni (quello sperimentale e quello di controllo). Nella tabella 3.41 riportiamo i valori dell'indice d ed i corrispondenti valori di non sovrapposizione.

Indicazioni di Cohen	Grandezza dell'effetto	Percentile del gruppo di controllo	Percentuale di non sovrapposizione
Effetto grande	2,0	97,7	81,1%
	1,8	96,4	77,4%
	1,6	94,5	73,1%
	1,4	91,9	68,1%
	1,2	88	62,2%
	1,0	84	55,4%
	0,8	79	47,4%
Effetto medio	0,7	76	43,0%
	0,6	73	38,2%
	0,5	69	33,0%
Effetto picolo	0,4	66	27,4%
	0,3	62	21,3%
	0,2	58	14,7%
	0,1	54	7,7%
	0,0	50	0%

Tabella 3.41 – *Valori corrispondenti dell'indice d, del percentile del gruppo di controllo e della percentuale di non sovrapposizione per grandezza dell'effetto*

La trasformazione del valore *d* in percentile può essere interpretata utilizzando la tavola statistica della distribuzione normale standard. Il valore dell'indice *d* corrisponde al valore di *z* ed, una volta trovata sulla tavola la porzione di area corrispondente, ad essa viene sommato 0,5 (corrispondente alla porzione di area della metà curva di sinistra). Il risultato di questa somma ci fornisce la posizione percentile a cui si troverà un partecipante del gruppo sperimentale rispetto alla distribuzione percentile del gruppo di controllo.

Per comprendere meglio questa procedura facciamo degli esempi numerici. Per un effetto grande, $d = 1,8$, l'area sottostante la curva normale sarà data dalla somma di 0,4641 + 0,5 = 0,9641. Questo, in altri termini, significa che il trattamento fa sì che un qualsiasi partecipante salga dal 50° al 96° percentile del gruppo di controllo.

Vediamo ora cosa accade con un effetto piccolo, $d = 0,1$. Si osservi che l'area sottostante la curva normale sarà 0,0398 + 0,5 = 0,5398; in questo caso il trattamento muove un qualsiasi partecipante dal 50° al 54° percentile del gruppo di controllo.

Risulta, quindi, evidente il diverso impatto che la grandezza dell'effetto comporta in vista di un reale cambiamento nella prestazione dei partecipanti.

È inoltre possibile trasporre i valori dell'indice *d* sotto forma di percentuale di non sovrapposizione tra la distribuzione dei punteggi del gruppo sperimentale e la distribuzione del gruppo di controllo (Cohen, 1988). In questo caso, seguendo la tabella, possiamo vedere come ad un effetto grande, $d = 1{,}8$, corrisponde un valore percentuale di non sovrapposizione delle due distribuzioni del 77,4%.

Se dovessimo rappresentare graficamente questo dato, potremmo immaginare per le due distribuzioni una configurazione come quella nella figura 3.3, con una modesta sovrapposizione tra le due curve.

Al contrario, per un effetto piccolo, $d = 0{,}1$, abbiamo una percentuale di non sovrapposizione pari al 7,7%, questo significa che le due distribuzioni sono sostanzialmente sovrapposte, e potrebbero essere rappresentate graficamente come mostrato nella figura 3.4.

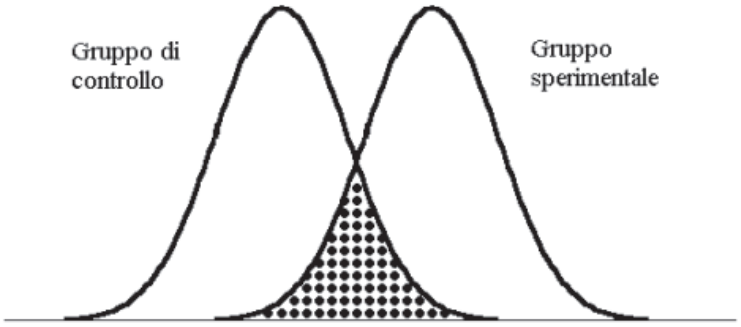

Figura 3.3 – *Distribuzioni dei punteggi del gruppo sperimentale e del gruppo di controllo aventi una percentuale di non sovrapposizione circa del 77,4%*

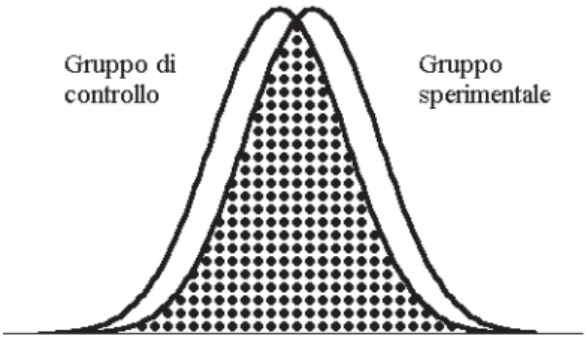

Figura 3.4 – *Distribuzioni dei punteggi del gruppo sperimentale e del gruppo di controllo aventi una percentuale di non sovrapposizione circa del 7,7%*

3.5.4. La strada più semplice non sempre è la migliore

Nonostante i numerosi tentativi di fornire dei punteggi di *cut off* o dei *range* che consentano di interpretare i valori ottenuti mediante gli indici di grandezza dell'effetto, molti ricercatori ammoniscono contro un utilizzo rigido di questi metodi e ricordano che la grandezza dell'effetto prima di tutto deve essere direttamente ed esplicitamente interpretata in relazione agli effetti messi in luce dalle ricerche precedenti (Wilkinson e APA Task Force on Statistical Inference, 1999).

Non dobbiamo poi sottovalutare che l'importanza da attribuire ad un valore della grandezza dell'effetto dipende fortemente dall'area di indagine. In questo senso Welkowitz, Ewen e Cohen (1982) dicono chiaramente che non dovrebbero essere utilizzati dei valori convenzionali per valutare la grandezza dell'effetto, qualora si disponga di norme specifiche per il proprio oggetto di indagine.

Thompson (2001) ha chiaramente dichiarato che, se i ricercatori interpretano la grandezza dell'effetto utilizzando dei *cut off* fissi, quindi con la stessa rigidità con cui è stato utilizzato il valore $\alpha = 0{,}05$ nella verifica delle ipotesi, commettono lo stesso stupido errore utilizzando semplicemente una scala diversa. Lo stesso autore (Thompson, 2002) ricorda inoltre che i benefici connessi all'utilizzo e l'interpretazione della grandezza dell'effetto non derivano tanto dall'impiego di punteggi come parametri di confronto per definire un effetto "grande", "medio" o "piccolo", bensì dalla possibilità di confrontare direttamente il valore ottenuto con il valore dello stesso indice ottenuto in ricerche precedenti.

Concludiamo prospettando una situazione paradigmatica della cautela che dobbiamo utilizzare nella ricerca e, soprattutto, nel commentare i risultati ottenuti. Poniamo che un esperimento abbia fornito un risultato statisticamente significativo, ma che l'effetto trovato risulti piccolo, tanto piccolo che sarebbe necessario interrogarsi in merito alla sua reale importanza nel determinare un certo esito. Certamente questo tipo di situazione andrebbe contestualizzata e valutata in merito all'area di indagine ed al significato teorico del risultato; qui ci limitiamo a suggerire un modo adeguato dal punto di vista metodologico di riportare il risultato ottenuto:

"*Nonostante che si sia rilevata con l'ANOVA una differenza statisticamente significativa tra le medie* ($F_{(...)} = \ldots, p = \ldots$)*, la grandezza dell'effetto risulta modesta. L'indice* η_p^2 *è 0,02, e questo significa che il fattore X spiega solo il 2% della varianza totale*".

Capitolo 4
Gli intervalli di fiducia

4.1. Gli intervalli di fiducia

Abbiamo più volte detto che uno dei mezzi consigliati per supplire all'insufficienza del puro valore assegnato alla probabilità associata alla statistica utilizzata nella VeSN consiste nel fornire contemporaneamente il valore degli intervalli di fiducia (o di confidenza, come spesso viene detto italianizzando l'inglese *confidence*). È allora opportuno in questo capitolo affrontare in modo più completo la problematica degli intervalli di fiducia.

Nella stima dei parametri si stabilisce, in termini probabilistici, il valore numerico di uno o più parametri incogniti della popolazione a partire dai dati campionari. Gli indici statistici utilizzati per stimare i parametri della popolazione da cui il campione è estratto sono detti stimatori. La stima dei parametri della popolazione può essere effettuata secondo due modalità:

- *stima puntuale* (*point estimation*): si risolve in un valore assunto a rappresentare un parametro della popolazione e può essere ottenuto servendosi del metodo dei minimi quadrati, del metodo dei momenti o del metodo della massima verosimiglianza;
- *stima intervallare* (*interval estimation*): si calcola la probabilità che tali parametri si collochino entro due valori limite.

Rispetto alla stima puntuale, che permette di assegnare un valore ad un parametro incognito, la stima intervallare fornisce informazioni sulla probabilità che il valore che viene assegnato al parametro $\hat{\theta}$ sia vicino al valore vero del parametro θ. Il primo autore a proporre questo metodo fu Neyman nel 1935.

La stima intervallare individua un intervallo di valori che con ragionevole fiducia contiene θ, fissando un limite inferiore e un limite superiore tali che la probabilità che il parametro sia compreso tra il limite inferiore e il limite superiore sia uguale a $1 - \alpha$. I due limiti sono detti *limiti di fiducia* o *limiti di confidenza*, mentre l'intervallo tra i due limiti è detto *intervallo di fiducia* o *intervallo di confidenza*.

L'ampiezza dell'intervallo viene determinata dalla probabilità che quell'intervallo di fiducia racchiuda un certo parametro; questa probabilità è detta *coefficiente di fiducia* o *confidenza* (detto anche *livello di confidenza*) *cf* e rappresenta il grado di fiducia che associamo alla stima. Viene indicato con $1 - \alpha$, dove α è un numero molto piccolo, generalmente 0,05 o 0,01.

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*,
ISBN 88-8453-226-4,

Gli stimatori intervallari permettono di specificare il margine di errore dovuto inevitabilmente al processo di campionamento, che dipende sia dal livello di fiducia desiderato sia dalla numerosità del campione. Sia:

$$es = \sqrt{\frac{s^2}{n}} \tag{4.1}$$

dove:
- es = errore standard
- s^2 = varianza campionaria
- n = ampiezza campionaria

Per le popolazioni finite si definisce un *fattore di correzione*

$$fc = \frac{N - n}{N - 1} \tag{4.2}$$

dove:
- N = ampiezza della popolazione
- n = ampiezza campionaria

e che trasforma la precedente formula (4.1) in

$$es = \sqrt{\frac{s^2}{n} \cdot \frac{(N - n)}{(N - 1)}} \tag{4.3}$$

Quando l'ampiezza della popolazione è molto maggiore dell'ampiezza campionaria il fattore di correzione si avvicina sempre più ad 1 fino a poter essere trascurato.

A parità di coefficiente di fiducia, più aumenta la dimensione del campione n, più si restringe il margine di errore, e di conseguenza più è ristretto l'intervallo di fiducia per la stima (aumentando la precisione della stima). A parità di n, più diminuisce il coefficiente di fiducia, più l'intervallo diventa ristretto. Aumentando il coefficiente di fiducia aumenta anche l'ampiezza dell'intervallo, rendendo però poco informativo l'intervallo stesso.

Se il coefficiente di fiducia è dato da $1 - \alpha$, ne consegue che α è dato da $1 - cf$. α perciò rappresenta il rischio che siamo disposti a correre, ossia la probabilità che l'intervallo di fiducia non contenga il valore del parametro sconosciuto.

Gli intervalli di fiducia possono essere applicati per ogni parametro da stimare, quindi non solo per la media (μ), ma anche per la varianza (σ^2), la proporzione (π) , la mediana, le frequenze, le percentuali, ecc.

4.2. Gli intervalli di fiducia della media

Quando dobbiamo stimare la media di una popolazione (parametro μ) conoscendo il valore della varianza (σ^2) della popolazione o della varianza (s^2) del campione, utilizziamo come stimatore l'indice statistico M del campione.

Sappiamo che il campione con media M è uno dei possibili campioni estratti dalla popolazione e che fa parte della distribuzione campionaria delle medie (DCM), ne consegue che assumiamo che la media della distribuzione campionaria delle medie sia uguale alla media della popolazione dalla quale il campione è estratto ($\mu_M = \mu$). Il parametro incognito diventa perciò μ_M.

La variabile casuale M però differisce da μ, solitamente si distribuisce intorno a μ assumendo valori maggiori o minori. È più corretto quindi stimare un intervallo di valori plausibili per μ, entro il quale riteniamo che θ sia contenuto:

$$M - \text{errore} < \mu < M + \text{errore}$$

Per procedere con le spiegazioni sul calcolo degli intervalli di fiducia della media riteniamo utile distinguere il caso dei campioni grandi dal caso dei campioni piccoli, con σ noto e σ non noto.

4.2.1. Intervallo di fiducia della media per campioni grandi

Anche quando la popolazione non è normale, per il teorema del limite centrale la DCM assume forma normale se il campione è numeroso ($n \geq 30$).

Sappiamo che l'intervallo tra la media ed una deviazione standard (σ) comprende il 34,13% dei dati, percentuale che corrisponde anche alla probabilità di estrarre un campione con media compresa tra μ_M e $\mu_M + \sigma$. Allo stesso modo la probabilità di estrarre un campione con media compresa tra μ_M e $\mu_M - 1$ sarà del 34,13%. Di conseguenza possiamo affermare che la probabilità di estrarre una media compresa nell'intervallo $\mu_M \pm 1\sigma_M$ è del 68,26% (0,3413 + 0,3413 = 0,6826, ossia 68.26%), ma anche che si ha una probabilità del 68,26% che, avendo estratto un campione con media M, l'intervallo $\mu_M \pm 1\sigma_M$ contenga la media μ_M (e quindi μ).

In altre parole possiamo affermare che si ha una fiducia al 68,26% che la media della popolazione μ sia compresa nell'intervallo.

In pratica, vediamo come si procede nel caso che σ sia noto. Supponiamo di avere una popolazione normale con deviazione standard σ nota e di voler stimare un intervallo di fiducia che contenga la media della popolazione μ. L'equazione sarà:

$$P(\lambda_1 < \theta < \lambda_2) = 1 - \alpha \qquad (4.4)$$

ossia la probabilità che il parametro ignoto sia compreso entro un limite inferiore e un limite superiore è uguale al coefficiente di confidenza.

Si fissa α, ossia il rischio che si accetta di correre nel dichiarare che il parametro cade nell'intervallo entro i limiti calcolati quando non è vero. Se fissiamo $\alpha = 0{,}05$, avremo una fiducia del 95% che l'intervallo contenga il parametro. In questo caso ogni coda comprenderà il 2,5% della distribuzione (vedi fig. 4.1).

Figura 4.1 – *Regione di rifiuto nella distribuzione normale standardizzata*

Si calcola di quanto si è sopra e sotto la media in unità di deviazione standard. Avendo campioni grandi ($n \geq 30$) si tratterà di una distribuzione normale standardizzata, potremo perciò utilizzare i punti z:

$$z = \frac{\overline{X} - \mu}{\sigma_M} \tag{4.5}$$

dove σ_M è l'errore standard della media:

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \tag{4.6}$$

Si trova uno z critico tale che:

$$p\left(-z_{crit} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +z_{crit} \right) = 1 - \alpha \tag{4.7}$$

Nella tabella 4.1 riportiamo i valori critici di z corrispondenti a $\alpha/2$ nel caso della distribuzione normale standardizzata:

$1 - \alpha$	z_{crit}
0,90	1,645
0,95	1,96
0,99	2,576

Tabella 4.1 – *Valori critici di z corrispondenti a* $\alpha/2$

L'unica incognita è la media della popolazione μ, perciò la disequazione tra parentesi può essere riscritta:

$$P\left(\overline{\mathrm{X}} - z_{crit}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{\mathrm{X}} + z_{crit}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.8}$$

Sostituiamo alla variabile $\overline{\mathrm{X}}$ la media del campione M e otteniamo l'intervallo di fiducia:

$$M - z_{crit}\frac{\sigma}{\sqrt{n}}; \quad M + z_{crit}\frac{\sigma}{\sqrt{n}} \tag{4.9}$$

Ne consegue che :

$$M - z_{crit}\frac{\sigma}{\sqrt{n}} \leq \mu \leq M + z_{crit}\frac{\sigma}{\sqrt{n}} \tag{4.10}$$

La parte sinistra ($M - z_{crit}\sigma_M$) rappresenta il limite inferiore, la parte destra ($M + z_{crit}\sigma_M$) il limite superiore. Solo i campioni con medie comprese nell'intervallo vengono considerati come appartenenti alla popolazione, al contrario dei campioni con medie maggiori o minori di tali limiti di fiducia.

Esempio 4.1

Vediamo un esempio con dei dati concreti. Supponiamo di avere una popolazione con $N = 64$ e $\sigma^2 = 100$. Trattandosi di una normale standardizzata con $\alpha = 0{,}05$ avremo uno $z_{crit} = \pm 1{,}96$, corrispondente a $p = 0{,}025$. Sostituiamo nella disequazione:

$$M - 1,96\sigma_M \leq \mu_M \leq M + 1,96\sigma_M. \tag{4.11}$$

Dato che

$$\frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{M - \mu}{1,25}; \tag{4.12}$$

allora

$$p(-1,96 \leq \frac{M - \mu}{1,25} \leq +1,96) = 0,95 \tag{4.13}$$

Risolvendo l'equazione avremo che

$$p(-1,96 \times 1,25 \leq M - \mu \leq +1,96 \times 1,25) = 0,95;$$

$$p(\mu - 1,96 \times 1,25 \leq M \leq \mu + 1,96 \times 1,25) = 0,95;$$

$$p(\mu - 2,45 \leq M \leq \mu + 2,45) = 0,95.$$

Questo significa che nel 95% dei casi l'intervallo con estremi $M - 2{,}45$ e $M + 2{,}45$ contiene il valore vero incognito di μ. Infatti, se invece di risolvere le disequazioni

$$-1,96 \leq \frac{M - \mu}{1,25} \quad \text{e} \quad \frac{M - \mu}{1,25} \leq +1,96$$

rispetto a M le risolviamo rispetto a μ, otteniamo

$$\mu < M + 1,96(1,25) \quad \text{e} \quad \mu < M - 1,96(1,25)$$

e quindi

$$p(-1,96 \leq \frac{M - \mu}{1,25} \leq +1,96) = 0,95;$$

$$p(M - 1,96(1,25) \leq \mu \leq M + 1,96(1,25)) = 0,95;$$

$$p(M - 2,45) \leq \mu \leq M + 2,45) = 0,95.$$

Quest'ultima formula nella pratica è più utile della precedente, perché generalmente l'incognita è μ e abbiamo a disposizione solo i dati osservati del campione.

Nel caso che σ non sia noto, se il campione è sufficientemente grande si può stimare a partire dalla deviazione standard (s) dei campioni. La distribuzione campionaria delle medie assume la forma t di Student con $n - 1$ gradi di libertà, che tende a somigliare alla normale all'aumentare della grandezza del campione.

$$t = \frac{M - \mu}{\frac{s}{\sqrt{n-1}}} \tag{4.14}$$

Ricordiamo che la distribuzione t di Student ha forma campanulare ed è simmetrica rispetto alla media, ma la sua dispersione è maggiore di quanto si riscontra nella normale. Rispetto alla normale, la probabilità è più addensata sulle code, e la probabilità di osservare valori in un intervallo centrato sulla media è minore.

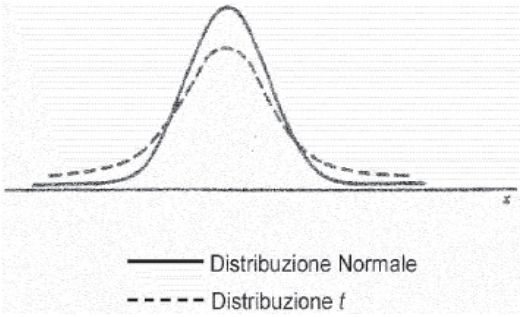

Figura 4.2 – *Confronto tra distribuzione normale e distribuzione t di Student*

All'aumentare di *n,* quindi dei gradi di libertà, la distribuzione diventa sempre più simile alla normale, e *s* una stima più precisa di σ (vedi fig. 4.3). La disequazione diventa:

$$M - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} \leq \mu_M \leq M + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}, \tag{4.15}$$

dove il valore critico di t, analogamente a quanto avviene per *z*, è uguale in valore assoluto a sinistra e a destra della distribuzione, ma differisce solo per segno.

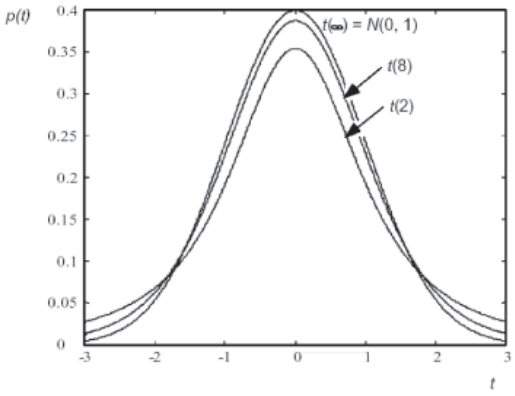

Figura 4.3 – *Distribuzione t di Student al crescere dei gradi di libertà*

Peraltro, se il campione è sufficientemente grande (e in genere si ritiene tale un campione con $n > 30$) si possono tranquillamente usare i valori relativi alla normale, e in luogo dei valori critici di *t* usare i punti *z* corrispondenti.

Esempio 4.2

Supponiamo di avere un campione di 50 soggetti (sufficiente per usare la distribuzione normale) con media 20 e deviazione standard 2,2, e di porre $\alpha = 0{,}05$.

Calcoliamo il limite inferiore:

$$20 - 1{,}96 \frac{2{,}2}{\sqrt{49}} = 20 - 1{,}96(0{,}31) = 19{,}39$$

Calcoliamo il limite superiore:

$$20 + 1{,}96 \frac{2{,}2}{\sqrt{49}} = 20 + 1{,}96(0{,}31) = 20{,}61$$

L'intervallo di fiducia sarà:

$$19{,}39 \leq \mu_M \leq 20{,}61$$

Avremo una probabilità del 95% che l'intervallo [19,39; 20,61] includa la media della popolazione da cui è estratto il campione.

È possibile determinare l'ampiezza del campione necessaria per stimare un certo parametro della popolazione utilizzando i punti z. Occorre trovare il valore z che corrisponde ad un intervallo di fiducia $[\mu - e; \mu + e]$, tenendo conto della formula dell'errore standard

$$e = \mu - M = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \ . \tag{4.16}$$

La grandezza campionaria n si trova partendo dalla seguente formula

$$M = \mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}; \tag{4.17}$$

di conseguenza

$$n = \frac{z^2_{\alpha/2} \sigma^2}{e^2} \tag{4.18}$$

ESEMPIO 4.3

Sapendo che lo scarto quadratico medio dello spessore di alcune lamine di acciaio è 0,766 mm, si determini l'ampiezza di un campione che, con fiducia 0,90, stimi la media dello spessore con errore non superiore a 0,15 mm.

I dati a nostra disposizione sono:

1. $(1 - \alpha) = 0{,}90$;
2. $\alpha = 0{,}10$;
3. $\alpha/2 = 0{,}05$;
4. $z_{\alpha/2} = 1{,}645$;
5. $\sigma = 0{,}766$;
6. $e = 0{,}15$.

Applichiamo la (4.18), sostituendo

$$n = \frac{z^2_{\alpha/2}\sigma^2}{e^2} = \frac{1{,}645^2 \times 0{,}766^2}{0{,}15^2} = \frac{2{,}706 \times 0{,}587}{0{,}0225} = 70{,}596.$$

Con un grado di fiducia del 90% possiamo affermare che un campione di $n \geq 71$ stima la media con errore $e \leq 0{,}15$.

4.2.2. Intervallo di fiducia della media per campioni piccoli

Se la dimensione del campione è inferiore a 30 ($n < 30$), la distribuzione campionaria della media (DCM) assume forma t di Student con $n - 1$ gradi di libertà. La distribuzione t di Student varia in funzione di α e dei *gl*. Al posto dello z critico cercheremo quindi il t critico sulla tabella 4.2 con $n - 1$ gradi di libertà.

Ecco come si procede in pratica. Se fissiamo $\alpha = 0{,}05$, ogni coda comprenderà il 2,5% della distribuzione. Si calcola di quante deviazioni standard si è sopra e sotto la media per avere il livello di fiducia stabilito.

Con campioni piccoli ($n < 30$) useremo il valore di t critico, tenendo conto dai gradi di liberà e dal livello di fiducia prescelto. La formula da utilizzare è la 4.15.

	Valori critici di *t* (ipotesi monodirezionale)				
gl	0,05	0,025	0,01	0,005	0,001
1	6,313	12,706	31,821	63,657	318,309
2	2,920	4,303	6,965	9,925	22,327
3	2,353	3,1824	4,541	5,841	10,215
4	2,132	2,776	3,747	4,604	7,173
5	2,015	2,571	3,365	4,032	5,893
6	1,943	2,447	3,143	3,707	5,208
7	1,895	2,365	2,998	3,499	4,785
8	1,860	2,306	2,896	3,355	4,501
9	1,833	2,262	2,821	3,250	4,297
10	1,812	2,228	2,764	3,169	4,144
11	1,796	2,201	2,718	3,106	4,025
12	1,783	2,179	2,681	3,055	3,930
13	1,771	2,160	2,650	3,012	3,852
14	1,761	2,145	2,624	2,977	3,787
15	1,753	2,131	2,602	2,947	3,733
17	1,740	2,110	2,567	2,898	3,646
19	1,729	2,093	2,539	2,861	3,580
21	1,721	2,080	2,518	2,831	3,527
24	1,711	2,064	2,492	2,796	3,467
27	1,703	2,052	2,473	2,771	3,421
30	1,697	2,042	2,457	2,756	3,385
40	1,684	2,021	2,423	2,704	3,307
75	1,665	1,992	2,377	2,643	3,202
150	1,655	1,976	2,351	2,609	3,145
∞	1,645	1,960	2,326	2,576	3,090

Tabella 4.2 – *Valori critici della distribuzione del t di Student*

4.3. GLI INTERVALLI DI FIDUCIA DELLA VARIANZA

Quando dobbiamo stimare la varianza di una popolazione (parametro σ^2), occorre fare riferimento alla distribuzione χ^2 (chi quadrato). Come sappiamo, la distribuzione χ^2 può essere definita come una sommatoria di punti z elevati al quadrato:

$$\chi^2_\nu = \sum_{i=1}^{\nu} z^2 \tag{4.19}$$

dove ν = gradi di libertà, corrispondenti al numero dei punti *z* al quadrato sommati tra loro.

Con un grado di libertà, e cioè con un solo punto z^2, la forma della distribuzione sarà assolutamente asimmetrica, con la massima frequenza in corrispondenza del valore 0. Ciò trova una sua chiara spiegazione nel fatto che la media della distribuzione *z*, corrispondente anche al valorre di massima frequenza, è appunto 0. All'aumentare dei gradi di libertà la forma della distribuzione diventa sempre più simmetrica e acquista una certa somiglianza con quella della normale (vedi fig. 4.4).

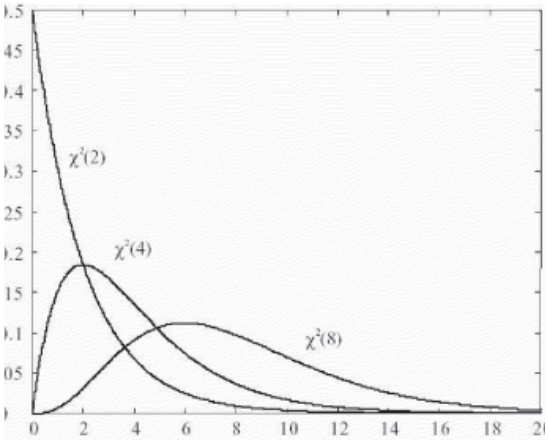

Figura 4.4 – *Distribuzione teorica* χ^2 *con 2, 4 e 8 gradi di libertà*

Sapendo che

$$\chi^2_{n-1} = \frac{(n)s^2}{\sigma^2}, \tag{4.20}$$

o, nel caso si usi la varianza stimata,

$$\chi^2_{n-1} = \frac{(n-1)\hat{s}^2}{\sigma^2}, \tag{4.20bis}$$

analogamente a quanto abbiamo fatto per la (4.7) otteniamo:

$$p\left(\chi^2_{(1-\frac{\alpha}{2})} \leq \frac{(n-1)\hat{s}^2}{\sigma^2} \leq \chi^2_{(\frac{\alpha}{2})} \right) = 1 - \alpha \tag{4.21}$$

Invertiamo la disuguaglianza e isoliamo σ^2, moltiplicando tutti i membri della disequazione per σ^2:

$$p\left(\frac{(n-1)\hat{s}^2}{\chi^2_{(\frac{\alpha}{2})}} \leq \sigma^2 \leq \frac{(n-1)\hat{s}^2}{\chi^2_{(1-\frac{\alpha}{2})}} \right) = 1 - \alpha \tag{4.22}$$

I termini destro e sinistro della disuguaglianzza tra parentesi ci indicano i limiti dell'intervallo di fiducia della varianza dell'universo, essendo noto quello della varianza stimata del campione. Se invece si conoscesse la varianza, non si userebbero i gradi di libertà $n - 1$, ma la numerosità del campione n.

Se fosse nota la varianza dell'universo, e si volessero conoscere i limiti dell'intervallo di fiducia della varianza stimata campionaria, la disuguaglianza diventerebbe, attraverso alcuni elementari passaggi algebrici, la seguente:

$$\frac{\chi^2_{(1-\frac{\alpha}{2})}\sigma^2}{(n-1)} \leq \hat{s}^2 \leq \frac{\chi^2_{(\frac{\alpha}{2})}\sigma^2}{(n-1)} \; . \tag{4.23}$$

Analogamente a quanto fatto per il t di Student, nella tabella 4.3 riportiamo alcuni valori critici del χ^2 per ipotesi monodirezionale. Osserviamo che mentre le distribuzioni z e t sono simmetriche con media uguale a 0, per cui i valori di destra e di sinistra di z e t a parità di area di rifiuto sono in valore assoluto uguali, e differiscono solo per il segno, i valori di χ^2 sono diversi anche in valore assoluto, e sono tutti positivi, trattandosi di sommatorie di quadrati.

Essendo la distribuzione χ^2 asimmetrica, avremo aree di rifiuto corrispondenti alle due code ($\alpha/2$ e $1 - \alpha/2$) di area uguale (la grandezza dell'area *è* la probabilità associata a quel certo valore), ma di forma diversa (vedi fig. 4.5). Con l'aumentare dei gradi di libertà, le due aree tenderanno ad assumere forme uguali, diventando la distribuzione sempre più simmetrica.

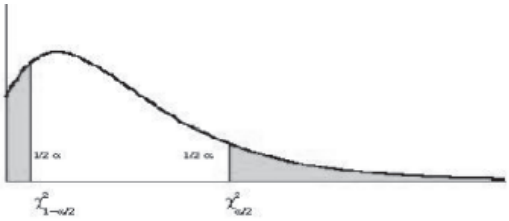

Figura 4.5 – *Aree di rifiuto per una distribuzione* χ^2

	Area sottostante la distribuzione χ^2												
gl	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995
1	7,88	6,63	5,02	3,84	2,71	1,32	0,45	0,10	0,02	0,00	0,00	0,00	0,00
2	10,60	9,21	7,38	5,99	4,61	2,77	1,39	0,58	0,21	0,10	0,05	0,02	0,01
3	12,84	11,34	9,35	7,81	6,25	4,11	2,37	1,21	0,58	0,35	0,22	0,11	0,07
4	14,86	13,28	11,14	9,49	7,78	5,39	3,36	1,92	1,06	0,71	0,48	0,30	0,21
5	16,75	15,09	12,83	11,07	9,24	6,63	4,35	2,67	1,61	1,15	0,83	0,55	0,41
6	18,55	16,81	14,45	12,59	10,64	7,84	5,35	3,45	2,20	1,64	1,24	0,87	0,68
7	20,28	18,48	16,01	14,07	12,02	9,04	6,35	4,25	2,83	2,17	1,69	1,24	0,99
8	21,95	20,09	17,53	15,51	13,36	10,22	7,34	5,07	3,49	2,73	2,18	1,65	1,34
9	23,59	21,67	19,02	16,92	14,68	11,39	8,34	5,90	4,17	3,33	2,70	2,09	1,73
10	25,19	23,21	20,48	18,31	15,99	12,55	9,34	6,74	4,87	3,94	3,25	2,56	2,16
11	26,76	24,72	21,92	19,68	17,28	13,70	10,34	7,58	5,58	4,57	3,82	3,05	2,60
12	28,30	26,22	23,34	21,03	18,55	14,85	11,34	8,44	6,30	5,23	4,40	3,57	3,07
13	29,82	27,69	24,74	22,36	19,81	15,98	12,34	9,30	7,04	5,89	5,01	4,11	3,57
14	31,32	29,14	26,12	23,68	21,06	17,12	13,34	10,17	7,79	6,57	5,63	4,66	4,07
15	32,80	30,58	27,49	25,00	22,31	18,25	14,34	11,04	8,55	7,26	6,26	5,23	4,60
16	34,27	32,00	28,85	26,30	23,54	19,37	15,34	11,91	9,31	7,96	6,91	5,81	5,14
17	35,72	33,41	30,19	27,59	24,77	20,49	16,34	12,79	10,09	8,67	7,56	6,41	5,70
18	37,16	34,81	31,53	28,87	25,99	21,60	17,34	13,68	10,86	9,39	8,23	7,01	6,26
19	38,58	36,19	32,85	30,14	27,20	22,72	18,34	14,56	11,65	10,12	8,91	7,63	6,84
20	40,00	37,57	34,17	31,41	28,41	23,83	19,34	15,45	12,44	10,85	9,59	8,26	7,43

Tabella 4.3 – *Valori critici monodirezionali della distribuzione* χ^2

ESEMPIO 4.4

Facciamo un esempio pratico. Immaginiamo di aver fatto un'indagine su 20 dipendenti di un'amministrazione pubblica e di aver calcolato media e deviazione standard della loro retribuzione mensile in euro: $M \pm s = 1250 \pm 27$. La varianza sarà allora uguale a 27^2, e cioè 729. Voglio ora calcolare i limiti inferiore e superiore dell'intervallo di fiducia della media e della varianza per il possibile universo di appartenenza di questa amministrazione, con $\alpha = 0{,}05$. (La curiosità può non essere peregrina: potrebbe esserci una tipologia delle amministrazioni costruita sulla base della retribuzione ai dipendenti, e voglio accertarmi della probabilità che l'amministrazione in questione appartenga effettivamente alla categoria a cui pretende di ap-

partenere). Per calcolare l'intervallo di fiducia per la media al 95%, supponiamo che la variabile retribuzione mensile si distribuisca normalmente e, dato che si tratta di un piccolo campione con $n < 30$, utilizziamo la (4.15), e cioè la distribuzione *t*. Come possiamo vedere dalla tabella 4.2, trattandosi di ipotesi bidirezionale, ed avendo 19 gradi di libertà, il valore di *t* critico che dobbiamo utilizzare, corrispondente a $p = 0{,}025$, è 2,093. La nostra disuguaglianza diventa dunque:

$$-t_{\left(\frac{\alpha}{2}\right)} \leq \frac{M-\mu}{\frac{s}{\sqrt{n-1}}} \leq t_{\left(\frac{\alpha}{2}\right)}; \quad -2{,}093 \leq \frac{1250-\mu}{\frac{27}{\sqrt{19}}} \leq 2{,}093;$$

$$-2{,}093 \times \frac{27}{4{,}359} \leq 1250-\mu \leq 2{,}093 \times \frac{27}{4{,}359}; \quad -12{,}964 \leq 1250-\mu \leq 12{,}964.$$

Il limite inferiore dell'intervallo di fiducia sarà dunque 1250 – 12,964 = 1237,04. Il limite superiore sarà invece 1262,96.

Vediamo ora di determinare l'intervallo di fiducia per la varianza. Qui dobbiamo impiegare la (4.22), avendo peraltro l'avvertenza di usare la numerosità del campione, anziché i gradi di libertà. Come vediamo dalla tabella 4.3, il valore di χ^2 per $p = 0{,}025$ è 32,85, mentre quello per $p = 0{,}975$ è 10,12.

$$\frac{20 \times 729}{32{,}85} \leq \sigma^2 \leq \frac{20 \times 729}{10{,}12};$$

$$443{,}836 \leq \sigma^2 \leq 1440{,}71.$$

Avremo allora una probabilità del 95% che l'intervallo [443,836; 1440,71] includa la varianza della popolazione da cui è estratto il campione.

4.4. Gli intervalli di fiducia per una proporzione

Quando dobbiamo stimare la proporzione (o frequenza relativa) π con cui una certa caratteristica è presente nella popolazione presa in esame, occorre riferirsi alla proporzione campionaria *p,* facendo riferimento alla distribuzione binomiale. Tale distribuzione è discreta, quindi non permette di costruire intervalli di fiducia continui. Lo stesso ragionamento vale ovviamente per tutte le osservazioni relative alla frequenza di dati appartenenti a universi bernoulliani (mutuamente escludentisi: sesso, ecc.).

Approssimiamo la distribuzione della variabile casuale $\hat{p}$ a una distribuzione normale standardizzata, con media μ_p e scarto quadratico medio σ_p. Per far ciò occorre che l'ampiezza del campione *n* sia sufficientemente grande e che le frequenze sia dei successi *np* sia degli insuccessi *n(1 – p)* siano maggiori di 5.

Il calcolo della media e dell'errore standard sarà dato da

$$\mu_P = np \tag{4.24}$$

$$\sigma_p = \sqrt{\frac{pq}{n}} \tag{4.25}$$

Sostituiamo nell'equazione degli intervalli di fiducia per la media (4.7), tenendo presente che la procedura è analoga, ma che l'errore standard σ è determinato direttamente dal valore di *p*. Otteniamo

$$p(\hat{p} - z_{crit}\sqrt{\frac{pq}{n}}) \leq \pi \leq (\hat{p} + z_{crit}\sqrt{\frac{pq}{n}}) = 1 - \alpha \tag{4.26}$$

dove:

- p = proporzione o frequenza relativa nel campione
- $\hat{p}$ = stimatore puntuale di p ($\hat{p} = \frac{f}{n}$, dove f è il numero dei successi e n il totale delle osservazioni)
- $\sqrt{\frac{pq}{n}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = errore standard della proporzione campionaria σ_p
- π = proporzione ignota della popolazione

Esempio 4.5

Vediamo un esempio pratico. Una ditta tessile vuole stimare la proporzione *p* di operai disposti a lavorare durante il turno di notte. A tale scopo estrae un campione casuale di 200 operai e lo intervista: il 78% si dichiara favorevole al turno di notte, il restante si dichiara contrario.

Dobbiamo trovare l'intervallo di fiducia al 95% per la proporzione π incognita della popolazione di operai.

I dati che abbiamo a disposizione sono:

1. $\hat{p} = \frac{f}{n} = 0,78$
2. $1 - \alpha = 0,95$
3. $\alpha = 0,05$
4. $z_{crit} = \pm 1,96$

Sostituiamo nella disuguaglianza della (4.26), e otteniamo:

$$0,78 - 1,96\sqrt{\frac{0,78(1-0,78)}{200}} \leq \pi \leq 0,78 + 1,96\sqrt{\frac{0,78(1-0,78)}{200}};$$

$$0,72 \leq \pi \leq 0,84.$$

Avremo una probabilità del 95% che l'intervallo [0,72; 0,84] includa la proporzione della popolazione da cui è estratto il campione.

Quando l'ampiezza del campione non è sufficientemente grande, non è possibile approssimare alla distribuzione normale standardizzata, ma occorre procedere con la binomiale. Infatti, se utilizzassimo erroneamente la normale, potremmo ottenere degli intervalli di fiducia negativi o superiori a 1. È evidente che nel caso di una proporzione sarebbe un risultato senza alcun senso.

Se il campione è molto piccolo, con n compreso tra 1 e 10, è opportuno ricorrere al metodo della procedura grafica, proposta da Clipper & Pearson (1934). Se ne ricava che

$$p(p_{i,\frac{\alpha}{2}} < \hat{p} < p_{s,\frac{\alpha}{2}}) \geq 1 - \alpha \tag{4.27}$$

dove

- $p_{i,\frac{\alpha}{2}}$ = limite critico inferiore;
- $p_{s,\frac{\alpha}{2}}$ = limite critico superiore.

Per consentire un rapido calcolo esiste un'apposita tabella costruita con il metodo grafico di Clopper-Pearson, che riporta i limiti di fiducia calcolati per una probabilità compresa tra 0 e 1 a partire da n, α e il numero di successi. L'intervallo di fiducia che si ottiene è simmetrico solo quando $p = 0{,}50$.

Le procedure illustrate fino ad ora si riferiscono ad un campione estratto da una popolazione infinita. Quando invece la popolazione è finita, cioè composta da un numero limitato di elementi, la varianza stimata risulterebbe maggiore di quella effettiva. È necessario perciò applicare la correzione per la popolazione finita:

$$\frac{N-n}{N}. \tag{4.28}$$

Di conseguenza, la formula per il calcolo dell'errore standard diventa

$$\sigma_p = \sqrt{\frac{pq}{n}\left(\frac{N-n}{N}\right)}. \tag{4.29}$$

Rispetto al campione estratto da una popolazione infinita, gli intervalli di fiducia sono minori e vanno stimati apportando una correzione, come proposto da Burstein (1975). Vediamo come si procede.

Si calcola il limite inferiore L_1

$$\frac{X}{X + (n - X + 1)F_{\alpha/2,v1,v2}}. \tag{4.30}$$

dove

- X = numero di elementi, appartenenti al campione n, che presentano la caratteristica in oggetto;
- $F_{\alpha/2,\, v1,v2}$ = F di Fisher.

Successivamente si calcola il limite inferiore corretto (L_1 corretto)

$$\frac{X-0,5}{n}-\left(\frac{X-0,5}{n}-L_1\right)\sqrt{1-\frac{n}{N}}\ . \tag{4.31}$$

Dopo aver calcolato il limite superiore L_2

$$\frac{(X+1)F_{\frac{\alpha}{2},v1,v2}}{n-X+(X+1)F_{\frac{\alpha}{2},v1,v2}}. \tag{4.32}$$

si procede col calcolo del limite superiore corretto (L_2 corretto)

$$\frac{X}{X+(n-X+1)F_{\alpha/2,v1,v2}}. \tag{4.33}$$

dove

$$X' = X+\frac{X}{n}.$$

Come già accennato precedentemente, l'intervallo di fiducia di un campione estratto da una popolazione finita è minore rispetto a quello estratto da una popolazione infinita e si annulla quando l'ampiezza campionaria n coincide con l'ampiezza della popolazione N.

4.5. Gli intervalli di fiducia della grandezza dell'effetto

Nel precedente Cap. 3 abbiamo analizzato a fondo la grandezza dell'effetto, sempre più frequentemente ritenuta un'indicazione necessaria da riportare nell'analisi dei dati delle ricerche. Peraltro, la stima della grandezza dell'effetto, per quanto ci fornisca un'informazione molto più accurata della semplice significatività statistica, non è esente da errori che possono derivare sia dalle modalità di campionamento, sia dalla grandezza stessa del campione. Abbiamo visto in precedenza che il calcolo di questi indici è basato sui dati campionari e questo porta con sé tutte le difficoltà ed i rischi connessi all'utilizzo di questo tipo di dati.

Se la grandezza dell'effetto è calcolata su un campione ragionevolmente ampio la sua accuratezza è certamente più garantita rispetto al caso in cui venga calcolata su un campione di dimensioni ridotte. Per ridurre il margine di errore connesso all'utilizzo di campioni piccoli si ricorre agli intervalli di fiducia della grandezza dell'effetto.

Volendo un intervallo di fiducia del 95%, e avendo ottenuto un certo valore di grandezza dell'effetto, per uno dei tanti indici che abbiamo visto nel capitolo 3 (ad esempio il *d* di Cohen, il rapporto tra la differenze tra le medie e la deviazione standard; vedi paragrafo 3.3.1), ciò che ora vogliamo calcolare è la gamma delle variazioni che potremmo ottenere di questa stima se prendessimo dalla medesima popolazione altri campioni della stessa grandezza. In altre parole, ciò significa che per 100 nuovi campioni di uguale ampiezza presi a caso dalla popolazione, 95 di questi avrebbero una stima della grandezza dell'effetto compresa nel range di valori indicato dall'intervallo di fiducia. Trattandosi di *grandezza* dell'effetto, è ovvio che se l'intervallo di fiducia include il valore 0, questo equivale ad avere un risultato statisticamente non significativo; al contrario, se lo 0 non è incluso nell'intervallo, questo è statisticamente significativo.

Il caso del *d* di Cohen è particolarmente semplice da risolvere. Infatti questo indice è normalmente distribuito ed ha una deviazione standard uguale a:

$$\sigma_d = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}}, \tag{4.34}$$

dove n_1 ed n_2 sono, rispettivamente, le numerosità del gruppo sperimentale e del gruppo di controllo (Hedges e Olkin, 1985). Per calcolare gli intervalli di fiducia per il *d* potremo allora procedere esattamente come abbiamo fatto per calcolare gli intervalli di fiducia della media, determinando cioè i punti *z* di destra e di sinistra corrispondenti al livello α che abbiamo prefissato. A questo punto, se $\alpha = 0{,}05$, gli intervalli di fiducia al 95% per l'indice *d* sono dati da questa disuguaglianza:

$$-1{,}96 < \frac{d}{\sigma_d} \leq 1{,}96. \tag{4.35}$$

Esempio 4.6

Riprendiamo i dati dell'esempio 3.8 del paragrafo 3.3.1 del Cap. 3 (tabella 3.12). Come si ricorderà, si trattava di una ricerca mirante a individuare le differenze di genere relativamente al punteggio ottenuto in un questionario che misurava la "ricerca di sensazioni forti". Si avevano due campioni, rispettivamente di maschi e di femmine, che presentavano medie rispettivamente di 28,9±6,9 e di 14,4±4,65. Il *d* rilevato era stato di 2,47.

Calcoliamo allora la deviazione standard del *d*, in base alla (4.34):

$$\sigma_d = \sqrt{\frac{10 + 10}{10 \times 10} + \frac{2{,}47^2}{2 \times (10 + 10)}} = 0{,}594.$$

A questo punto, passiamo a calcolare i limiti di sinistra e di destra dell'intervallo di fiducia, in base alla (4.35). Il limite di sinistra sarà dato da:

$$d - 1,96 \times \sigma_d = 2,47 - 1,164 = 1,306.$$

E di converso il limite di destra sarà dato da

$$d - 1,96 \times \sigma_d = 2,47 + 1,164 = 3,634.$$

Questi limiti non includono lo 0, pertanto possiamo concludere fiduciosamente che la grandezza dell'effetto è statisticamente significativa.

Come abbiamo detto, l'esempio relativo al d è abbastanza semplice, perché questo indice si distribuisce normalmente. Peraltro, questo non è vero per la maggior parte degli indici che abbiamo passato in rassegna nel Capitolo 3, e ciò comporta la necessità di utilizzare altri metodi per individuare gli intervalli di fiducia. Va detto preliminarmente che questi altri metodi sono troppo complessi per poter essere affrontati in questa sede, e comunque escludono la possibilità di un calcolo manuale. È questo in particolare il caso di due tra gli indici più usati, il g di Hedges (par. 3.3.1) e lo η^2 (par. 3.2.1). In questi casi si può far ricorso a del software facilmente reperibile, e in particolare segnaliamo la possibilità di usare sia SPSS che STATISTICA (vedi rispettivamente Smithson, 2001; Steiger e Fouladi, 1997), nonché ovviamente con un minimo di lavoro personale SAS e R (o S-PLUS). Su questo problema, vedi Fidler e Thompson (2001). In ogni caso, se si ritiene opportuno calcolare in un'analisi intervalli di fiducia e grandezza dell'effetto, sarebbe poi paradossale fermarsi qui e non calcolarne i limiti …

4.6. INTERVALLI DI FIDUCIA PER IL τ (TAU) DI KENDALL

Prima di chiudere il capitolo, vediamo anche come questo discorso degli intervalli di fiducia si possa applicare ai test non parametrici. Non possiamo ovviamente passarli in rassegna tutti, ci limiteremo a titolo esemplificativo al coefficiente τ (tau) di Kendall. Il τ di Kendall è un coefficiente di correlazione per ranghi, quindi di tipo non parametrico. Non è una buona stima di r (al contrario del coefficiente di correlazione per ranghi r_s di Spearman), ha un valore inferiore di r_s, può essere utilizzato nelle stesse condizioni e sugli stessi dati, ma ha il vantaggio di poter essere usato facilmente in presenza di molti ranghi uguali. I risultati tra i due test sono molto simili, ma molti autori preferiscono lo r_s di Spearman principalmente perché più noto e più semplice. Recentemente, peraltro, è stato fatto rilevare (Long e Cliff, 2004) che questo coefficiente, a differenza di r e di r_s, richiedendo meno assunzioni, non viene influenzato da fattori che influiscono sull'attendibilità degli altri coefficienti di correlazione, e soprattutto è

insensibile al fatto che le distribuzioni delle due variabili non siano normali. Ciò porta a preferire senz'altro il τ, in caso di non normalità.

Come sappiamo, il coefficiente τ di Kendall si calcola ordinando una delle due variabili (*x*) secondo l'ordine naturale crescente della graduatoria, assegnando il rango 1 al valore più piccolo fino a N per il valore più grande. I valori della variabile associata (*y*) relativi agli stessi soggetti verranno quindi spostati, l'ordine dei punteggi segue quello della prima variabile. Si contano solo i valori della variabile y e si assegna +1 ogni volta che le coppie di ranghi sono concordanti e –1 ogni volta che sono discordanti dall'ordine naturale.

Infine occorre sommare tutti i confronti (*S*) per confrontarli con il massimo totale possibile

$$\tau = \frac{2S}{n(n-1)}, \tag{4.36}$$

dove *S* è la somma di tutti i valori (+1 e –1), e *1/2 n (n– 1)* è il massimo totale possibile, dato dalla combinazione di *n* oggetti presi a due a due. In questa forma il τ è anche noto come "tau-a" (Cliff e Charlin, 1991)

Per chiarire questo procedimento, più complesso da spiegare che da eseguire praticamente, ricorriamo a un esempio pratico, esempio 4.7. Poniamo che per *n* bambini le due variabili siano livello socio-economico (LSE) e rendimento scolastico (RS). Attribuiti i valori di rango a LSE e RS, ordiniamo i bambini in base all'LSE. Veniamo al primo bambino, quello che ha rango 1 per l'LSE. Tutti gli altri bambini hanno evidentemente valori di rango per l'LSE superiori. Vediamo qual è il valore di rango di questo bambino per quel che riguarda RS. Supponiamo che sia 4. Vi saranno allora tre bambini che per RS avranno un rapporto inverso con il nostro rispetto a LSE; e *n* – 5 che avranno lo stesso rapporto. Assegneremo quindi rispettivamente –3 e +(*n* – 1) punti a questo bambino. Passeremo quindi al bambino successivo. *S* è la sommatoria di tutti questi punteggi.

Con campioni piccoli (tra 4 e 10) sono disponibili delle tavole per trovare i valori critici, mentre quando abbiamo a che fare con campioni grandi la significatività del τ di Kendall può essere verificata con la distribuzione normale, quindi è possibile utilizzare i punti *z*:

$$z = \frac{\tau - \mu_\tau}{\sigma_\tau} \tag{4.37}$$

dove μ_τ è la media della distribuzione campionaria di τ, e σ_τ è la deviazione standard della distribuzione campionaria di τ.

La distribuzione campionaria di τ, che si approssima soddisfacentemente alla normale, ha media $\mu_\tau = 0{,}00$ e varianza

$$\sigma_\tau^2 = \frac{2(2n+5)}{9n(n-1)}, \tag{4.38}$$

dove *n* è il numero delle coppie di dati (Kendall e Gibson, 1990).

La (4.37), tenuto conto del fatto che la media è assunta uguale a 0, passando alla radice quadrata della (4.38) per ottenere la deviazione standard di τ, diventa

$$z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}. \tag{4.39}$$

A questo punto si procede analogamente a quanto abbiamo fatto nel caso della (4.10):

$$-z_{crit} \times \frac{\sqrt{2(2n+5)}}{3\sqrt{n(n-1)}} \leq \tau \leq z_{crit} \times \frac{\sqrt{2(2n+5)}}{3\sqrt{n(n-1)}}. \tag{4.40}$$

ESEMPIO 4.7

Come detto sopra, in una ricerca si vuole verificare il rapporto esistente tra livello socio-economico (LSE) e rendimento scolastico (RS) in un campione di 10 bambini. Vogliamo calcolare il τ di Kendall e determinarne gli intervalli di fiducia per α = 0,05. Nella tabella 4.4, ordinati per valore di rango di LSE, sono riportati i valori di rango di LSE e di RS, e i valori positivi e negativi relativi alla concordanza o discordanza tra i valori di rango di LSE e i valori di rango di RS.

LSE	RS	Diff. +	Diff. –
1	3	7	2
2	2	8	0
3	5	5	2
4	1	6	0
5	4	5	0
6	6	4	0
7	8	2	1
8	9	1	1
9	7	1	0
10	10	0	0

Tabella 4.4 – *Ranghi e differenze di rango per livello socioeconomico e rendimento scolastico*

A questo punto, possiamo calcolare il τ. Dalla tabella vediamo che *S* (sommatoria delle differenze negative e positive) è uguale a 33. Applicando la (4.33), abbiamo che

$$\tau = \frac{2 \times 33}{10 \times 9} = 0{,}73.$$

Applicando la (4.40), determiniamo gli intervalli di fiducia del valore di τ:

$$-1,96 \times \frac{\sqrt{2(2 \times 10 + 5)}}{3\sqrt{10 \times 9}} \leq \tau \leq 1,96 \times \frac{\sqrt{2(2 \times 10 + 5)}}{3\sqrt{10 \times 9}};$$

$$-0,487 \leq \tau \leq 0,487.$$

Il nostro valore di $\tau = 0,73$ ricade al di fuori di questi intervalli. Ciò significa che la relazione che abbiamo trovato tra queste due variabili è statisticamente significativa.

Si osservi che questo modo di procedere, che è quello di usuale riscontro anche a livello manualistico (cfr. Siegel, 1956), assume che gli $n!$ possibili arrangiamenti per rango siano tutti ugualmente probabili. Sembra invece più opportuno trattare il τ "parametricamente", e stimarne la varianza dal campione, anziché in base alla (4.38) (Cliff e Charlin, 1991). Nel caso che si adottino tali stime, occorrerà nella (4.40) sostituire al valore di z_{crit} quello di t_{crit}. Le simulazioni fatte in proposito non ci sembra però che abbiano portato a vantaggi tali da indurre a preferire la più complessa procedura di stima (cfr. Long e Cliff, 2004).

Capitolo 5
L'analisi di Potenza

5.1. La potenza di un test statistico

La potenza è la capacità di individuare un effetto qualora abbia luogo; essa è legata direttamente all'errore di secondo tipo: accettare l'ipotesi nulla quando in realtà vi è un effetto. Questa relazione è data da:

$$\text{Potenza} = 1 - \beta \tag{5.1}$$

dove β = probabilità di commettere un errore di secondo tipo.

Espressa in modo quantitativo, la potenza varia tra 0 ed 1. Ad esempio un valore di potenza di 0,95 (o 95%) significa che $\beta = 0{,}05$, in altre parole, ho il 5% di probabilità di sbagliare e ritenere che non vi sia un effetto quando invece ha avuto luogo.

La potenza è influenzata e determinata da tre fonti:

1. Livello di α (e direzionalità dell'ipotesi nulla)
2. Grandezza del campione
3. Grandezza dell'effetto

In linea generale possiamo dire che quando queste tre componenti aumentano anche il valore della potenza tende ad incrementarsi.

In riferimento al *livello di* α, Cohen (1988) ci fa comprendere molto chiaramente la relazione esistente tra questo valore β e la potenza corrispondente: all'aumentare di α, aumenta anche la probabilità di scoprire delle differenze, quindi β diminuisce e la potenza $(1 - \beta)$ cresce; viceversa la diminuzione di α causa una diminuzione della probabilità di mettere in evidenza un effetto, quindi della potenza, ed un incremento di β.

Come più volte abbiamo rilevato (cfr. soprattutto il Cap. 2), nella pratica della ricerca tradizionalmente viene posta molta più attenzione al controllo dell'errore di primo tipo piuttosto che all'errore di secondo tipo. Questo atteggiamento è molto pericoloso e spesso fuorviante, oltre che sostanzialmente non corretto dal punto di vista metodologico. L'inferenza statistica si deve, quindi, porre l'obiettivo di cercare un compromesso accettabile tra le probabilità connesse ai due tipi di errore.

Simon (1999) ha suggerito l'utilizzo di una regola informale che prevede di fissare α a 0,05 e β a 0,2. Questo significa che, per essere accettabile, la potenza dovrebbe ave-

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*, ISBN 88-8453-226-4,

re un valore di almeno 0,80 (o 80%), mentre la probabilità di commettere un errore di secondo tipo dovrebbe essere del 20%. Se la generalità dei metodologi concorda con Simon, non sembra di contro che la comunità scientifica, di fatto, avvalori l'utilizzo della regola da lui proposta.

Il rapporto esistente tra i valori di α e β definisce anche la gravità attribuita ai due tipi di errore (Cohen, 1988). Per visualizzare questa relazione ci serviamo della formula:

$$q = \frac{\beta}{\alpha} \tag{5.2}$$

Se ad esempio avessimo $\alpha = .001$ e $\beta = .50$, il rapporto q sarebbe uguale a (0,50/0,001) = 500 e questo equivarrebbe ad asserire che rifiutare erroneamente l'ipotesi nulla è considerato 500 volte più grave che accettare l'ipotesi nulla sbagliando. Altresì, seguendo le indicazioni fornite dalla letteratura e ponendo $\alpha = 0,05$ e la potenza a 0,80, da cui avremmo un $\beta = (1 - 0,80) = 0,20$, il rapporto q corrisponde a (0,20/0,05) = 4, da cui deriva che rifiutare l'ipotesi nulla per errore è considerato quattro volte più grave che accettarla erroneamente.

Questo rapporto tra α e β, tale per cui $\beta = 4\alpha$ e, di conseguenza, $(1 - \beta) = 1 - 4\alpha$, può essere utilizzato come criterio per la scelta di β in funzione di α, tenendo però sempre presente che, per quanto si tratti di una indicazione ragionevole, essa non ha nessuna base teorica.

La potenza è influenzata non solo dalla grandezza di α, ma anche dalla *direzionalità dell'ipotesi nulla* (Cohen, 1988). La scelta tra un test monodirezionale o bidirezionale è importante per la determinazione della zona di rifiuto (rispettivamente, ad una coda o a due code) e, conseguentemente, per il valore dell'indice al quale si rifiuta l'ipotesi nulla. In una distribuzione normale, con $\alpha = .05$, z assume come valore critico 1,654 nel caso di un test ad una coda, mentre assume come valore critico 1,96 per un test a due code. Il fatto che il valore critico dell'indice (in valore assoluto) sia sistematicamente inferiore in un test monodirezionale rispetto ad uno bidirezionale determina maggiore potenza nel primo caso rispetto al secondo. Quindi, con un α bidirezionale il test avrà minore potenza nel determinare l'effetto rispetto ad un disegno di ricerca con lo stesso valore di α ma monodirezionale, ammesso che il risultato vada nella direzione predetta. Un test usato con una ipotesi monodirezionale, infatti, non ha alcuna potenza nel determinare un risultato che vada nella direzione opposta rispetto a quella indicata. Sapere che una ipotesi monodirezionale garantisce maggiore potenza potrebbe indurre i ricercatori a cercare di utilizzare sempre questo tipo di disegno, ma la constatazione che questo procedimento preclude la possibilità di ottenere dei risultati che non siano in linea con quanto predetto dovrebbe frenare questa tentazione (Cohen, 1988).

Il legame tra potenza e *grandezza del campione* ci riporta ad un problema connesso direttamente al campionamento: l'affidabilità dei risultati ottenuti. Maggiore è la numerosità del campione, minore è l'errore e, di conseguenza, maggiore sarà l'affidabilità

o precisione dei risultati. Il passaggio successivo è intuitivamente chiaro: maggiore è l'affidabilità dei risultati, maggiore è la probabilità di determinare l'effetto. Un incremento della numerosità del campione comporta un aumento nella potenza del test (Cohen, 1988).

La *grandezza dell'effetto* è già stata ampiamente discussa in precedenza, essa rappresenta il grado in cui l'effetto è presente, in altre parole il grado in cui l'ipotesi nulla è falsa. La grandezza dell'effetto è un parametro che assume valore 0 quando l'ipotesi nulla è vera ed assume qualunque altro valore diverso da 0 quando l'ipotesi nulla è falsa. La grandezza dell'effetto è la determinante più importante della potenza. Maggiore è la grandezza dell'effetto, a parità di livello α e di grandezza del campione, maggiore è la potenza del test (Cohen, 1988). È intuitivo che sia molto più facile rilevare differenze grandi rispetto a differenze piccole.

L'analisi di potenza, inoltre, non è collegata semplicemente alla probabilità di commettere errori nell'identificare o meno un effetto, bensì riguarda da vicino un altro obiettivo fondamentale nella maggior parte delle ricerche: la replicabilità. Ottenbacher (1996) invita gli studiosi a riflettere con maggiore attenzione sul fatto che una potenza bassa può ridurre la probabilità di replicare con successo i risultati di uno studio.

Infine, una ulteriore difficoltà nell'utilizzo dell'analisi di potenza si lega ad un problema molto più diffuso di quanto si creda: le abilità dei ricercatori. Muller & Lavange (1992) hanno posto l'accento sul problema dell'accostamento tra analisi di potenza e tipo di analisi effettuata sui dati. Purtroppo non è insolito che alcuni ricercatori facciano confusione tra disegni di ricerca quali ANOVA, ANCOVA e MANOVA, rischiando poi di condurre una analisi di potenza per un tipo di disegno quando in realtà ne è stato utilizzato un altro.

Il calcolo della potenza può essere effettuato in due modi: utilizzando alcune tabelle fornite in numerosi testi ed articoli (Kraemer e Thiemann, 1987; Cohen, 1988; Lipsey, 1990; Zar, 1996) oppure mediante dei programmi statistici. Il problema dell'efficienza di questi metodi verrà discusso in seguito.

È fondamentale, prima di tutto, introdurre una distinzione tra due tipi di analisi di potenza che possono essere condotte: a priori ed a posteriori.

5.2. L'analisi di potenza *a priori*

L'analisi a priori si utilizza quando l'esperimento non è ancora stato condotto e serve per determinare la grandezza del campione necessaria a garantire un certa potenza del test.

Come è stato già detto, la numerosità del campione è una caratteristica fondamentale del disegno della ricerca. Un campione troppo ristretto comporta, quando anche si ottenessero dei risultati significativi, il rischio che le risposte fornite ai quesiti della

ricerca non siano affidabili. Se di contro il campione è troppo ampio si corre il rischio di sprecare inutilmente tempo e risorse, ma anche di ottenere un risultato significativo dovuto alla grandezza stessa del campione e non alla presenza di un effetto.

La grandezza ottimale del campione va calcolata in funzione di (i) grandezza dell'effetto; (ii) livello di α; e (iii) potenza del test.

5.3. L'ANALISI DI POTENZA *A POSTERIORI*

Si utilizza dopo aver già condotto l'esperimento e serve per conoscere la potenza del test che è stato utilizzato. In questo caso, la potenza è ricavata in funzione dei tre parametri discussi nell'introduzione:

1. livello di α;
2. grandezza del campione;
3. grandezza dell'effetto.

L'analisi di potenza a posteriori non è una buona procedura per il controllo della bontà della propria ricerca. L'analisi dovrebbe essere utilizzata piuttosto prospetticamente per meglio costruire l'esperimento prima che sia condotto. Come abbiamo visto, l'analisi di potenza a priori permette prima di tutto di determinare quale dovrà essere la grandezza ottimale del campione.

5.4. METODI PER L'ANALISI DI POTENZA

Come abbiamo annunciato precedentemente, è possibile procedere ad una analisi di potenza, sia essa a priori o a posteriori, avvalendosi dell'uso di tabelle predisposte, o di specifici software.

Vedremo ora entrambe queste possibilità.

5.4.1 Tavole per l'analisi di potenza

L'utilizzo delle tavole per l'analisi di potenza è piuttosto semplice ed intuitivo. Nel caso di una analisi di potenza a priori il ricercatore dovrà avere come valori noti: α, la potenza e la grandezza dell'effetto. Tali valori dipenderanno sia dalle norme date in letteratura per problemi di ricerca analoghi, sia dagli scopi del particolare tipo di indagine condotta. Stabiliti i parametri, basterà andare a cercare sulla tavola adatta l'incrocio di questi valori per ricavare la grandezza del campione necessaria.

Nel caso di una analisi di potenza a posteriori il procedimento è sostanzialmente identico, ma questa volta dovremo conoscere α, la grandezza del campione e la

grandezza dell'effetto. Tutti e tre questi parametri sono definiti in anticipo: le scelte del livello di α e della grandezza del campione sono compiute al momento della definizione del disegno della ricerca, mentre la grandezza dell'effetto sarà calcolata sui dati raccolti. Sarà sufficiente andare a controllare l'incrocio dei valori noti sulla tavola rilevante e troveremo la potenza del test.

Cohen (1988) ha fornito una serie di tavole che permettono di condurre analisi di potenza, sia a priori che a posteriori, per numerosi test statistici, prendendo in considerazione tre valori di α (0,01, 0,05 e 0,10) e, laddove possibile, le direzionalità ad una e a due code, ponendo talvolta ulteriori varianti in base al tipo di test utilizzato.

Nei due paragrafi seguenti riportiamo modificate alcune delle tavole di Cohen (1988) e le indicazioni sul loro utilizzo suddividendole in base al tipo di analisi di potenza per il quale devono essere utilizzate.

5.4.1.1. Tavole di Cohen per l'analisi di potenza a priori

La tabella 5.1 è, modificata, una delle cinque che Cohen (1988) propone per determinare la grandezza ottimale del campione, nel caso in cui venga utilizzato il *t* di Student per campioni indipendenti.

Come visto (Cap. 3), *d* è l'indice di grandezza dell'effetto calcolato per questo test.

	d										
Potenza	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	1,00	1,20	1,40
0,25	332	84	38	22	14	10	8	6	5	4	3
0,50	769	193	86	49	32	22	17	13	9	7	5
0,60	981	246	110	62	40	28	21	16	11	8	6
0,67	1144	287	128	73	47	33	24	19	12	9	7
0,70	1235	310	138	78	50	35	26	20	13	10	7
0,75	1389	348	155	88	57	40	29	23	15	11	8
0,80	1571	393	175	99	64	45	33	26	17	12	9
0,85	1797	450	201	113	73	51	38	29	19	14	10
0,90	2102	526	234	132	85	59	44	34	22	16	12
0,95	2600	651	290	163	105	73	54	42	27	19	14
0,99	3675	920	409	231	148	103	76	58	38	27	20

Tabella 5.1 – *Grandezza del campione in funzione della potenza e dell'indice d (test t) per* $\alpha = 0{,}05$ *bidirezionale*

ESEMPIO 5.1

Poniamo di stabilire una ipotesi bidirezionale con $\alpha = 0,05$, di volere una potenza di 0,80 ed un effetto di media entità, $d = 0,50$. I parametri di partenza sono:

$\alpha = 0,05$ $d = 0,50$ potenza = 0,80

Nella tabella 5.1 all'incrocio tra la colonna corrispondente a $d = 0,50$ e la riga corrispondente a potenza = 0,80, troviamo un valore di 64. Questo significa che, se desideriamo avere una probabilità di 0,80 di trovare un effetto di 0,50 utilizzando il test *t* con una ipotesi bidirezionale e $\alpha = 0,05$, il campione dovrà avere 64 partecipanti.

Lasciando invariate α e la potenza, ma riducendo la grandezza attesa dell'effetto, ad esempio $d = 0,20$, cioè presupponendo che l'effetto sia piuttosto piccolo, vediamo come varia la grandezza del campione. Riassumiamo i parametri di partenza:

$\alpha = 0,05$ $d = 0,20$ potenza = 0,80

Nella tabella 5.1, in corrispondenza di questi valori, troviamo $n = 393$. Il campione necessario è molto più grande del precedente (di circa sei volte) e questo per una ragione molto semplice: l'effetto è piccolo, quindi avrà minor probabilità di essere messo in evidenza, a meno che non incrementi notevolmente la grandezza del campione. Più l'effetto è ridotto e più avrò bisogno di campioni grandi perché esso emerga, soprattutto se desideriamo avere una alta probabilità (.80) di trovarlo qualora ci fosse.

Prendiamo adesso in considerazione, modificata (tabella 5.2) una delle cinque tavole che Cohen (1988) ha messo a punto per la grandezza ottimale del campione nel caso in cui si voglia utilizzare il coefficiente di correlazione *r*.

	r								
Potenza	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
0,25	99	24	12	8	6	4	4	3	3
0,50	277	69	30	17	11	8	6	5	4
0,60	368	92	40	22	14	10	7	5	4
0,67	430	107	47	26	16	11	8	6	4
0,70	470	117	51	28	18	12	8	6	4
0,75	537	133	58	32	20	13	9	7	5
0,80	618	153	68	37	22	15	10	7	5
0,85	727	180	78	43	26	17	12	8	6
0,90	864	213	93	50	31	20	13	9	6
0,95	1105	272	118	64	39	25	16	11	7
0,99	1585	389	168	91	55	35	23	15	10

Tabella 5.2 – *Grandezza del campione in funzione della potenza e del coefficiente r per $\alpha = 0,05$ monodirezionale e $\alpha = 0,10$ bidirezionale*

Esempio 5.2

Poniamo di partire da una ipotesi monodirezionale con $\alpha = 0,05$, di ipotizzare un effetto forte $r = 0,50$ ed una potenza di 0,80. I parametri di partenza sono:

$\alpha = 0,05$ $r = 0,50$ potenza = 0,80

Incrociando i parametri nella tabella 5.2 troviamo che la numerosità del campione dovrà essere di 22 partecipanti, quindi un numero relativamente ridotto.

Poniamo, adesso, lasciando invariati i valori di α e di r, di ipotizzare per il test una potenza superiore, 0,95, poiché per gli scopi della mia ricerca è assolutamente fondamentale evitare l'errore di secondo tipo. I parametri sono:

$\alpha = 0,05$ $r = 0,50$ potenza = 0,95

In base alla tabella 5.2 il campione dovrà essere di 39 partecipanti. La numerosità richiesta non è molto aumentata: presupporre una forte grandezza dell'effetto ($r = 0,50$) è comunque un fattore che di per sé incrementa la probabilità di mettere in evidenza un effetto. Ciò ci riporta alla questione del rapporto tra potenza e grandezza dell'effetto: questa è, senza dubbio, la determinante più importante della prima.

Prendiamo adesso in considerazione, modificata, una delle 42 tavole di Cohen (1988) per la grandezza del campione nel caso si utilizzi il test χ^2 (tabella 5.3). Come abbiamo visto (Cap. 3), la misura di grandezza dell'effetto che viene utilizzata nel caso di questo test è l'indice *w* (Cohen, 1988).

	w								
Potenza	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
0,25	713	178	79	45	29	20	15	11	9
0,50	1175	294	131	73	47	33	24	18	15
0,60	1374	343	153	86	55	38	28	21	17
0,67	1521	380	169	95	61	42	31	24	19
0,70	1601	400	178	100	64	44	33	25	20
0,75	1734	434	193	108	69	48	35	27	21
0,80	1887	472	210	118	75	52	39	29	23
0,85	2073	518	230	130	83	58	42	32	26
0,90	2318	580	258	145	93	64	47	36	29
0,95	2704	676	300	169	108	75	55	42	33
0,99	3502	876	389	219	140	97	71	55	43

Tabella 5.3 – *Grandezza del campione in funzione della potenza e dell'indice w (test χ^2) per $\alpha = 0,01$ con 6 gradi di libertà*

ESEMPIO 5.3

Abbiamo una tabella 4×3, e sappiamo che i gradi di libertà corrispondono a *gl* = (numero di righe-1) × (numero di colonne-1) = (4-1) × (3-1) = 3×2 = 6; poniamo poi di stabilire $\alpha = 0{,}01$, di volere una potenza accettabile, cioè 0,80, e di aspettarci una grandezza dell'effetto di media entità, $w = 0{,}30$. Riassumendo i parametri, avremo:

$\alpha = 0{,}01$ $gl = 6$ $w = 0{,}30$ potenza = 0,80

Controllando l'incrocio dei parametri noti nella tabella 5.3, troviamo che il campione dovrà avere una numerosità pari a 210 partecipanti.

Supponiamo che la letteratura in merito al nostro oggetto di indagine ci suggerisca di poter contare su un effetto di maggiore intensità e di decidere, quindi, di porre $w = 0{,}50$, lasciando invariati i valori degli altri parametri. Avremo:

$\alpha = 0{,}01$ $gl = 6$ $w = 0{,}50$ potenza = 0,80

Dalla tabella 5.3, emerge che, date queste condizioni, il numero di partecipanti necessari è di 75, un notevole risparmio trattandosi praticamente di un terzo del campione precedente. Purtroppo è raro poter contare su grandezze dell'effetto così forti ed ancor meno spesso si trovano in letteratura indicazioni che ci orientino verso questi valori. Ecco perché, nella pratica di una analisi di potenza a priori, capiterà più frequentemente di dover mantenere un atteggiamento improntato alla prudenza per quanto riguarda il livello di grandezza dell'effetto da ipotizzare.

Proponiamo, infine, modificata, una delle 33 tavole che Cohen (1988) fornisce per determinare *n* nel caso venga utilizzato il test *F* (tabella 5.4).

	f											
Potenza	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,50	0,60	0,70	0,80
0,10	194	49	23	13	9	6	5	4	3	3	2	2
0,50	580	146	65	37	24	17	13	10	7	5	4	3
0,70	785	197	88	50	32	23	17	13	9	7	5	4
0,80	918	230	103	58	38	27	20	15	10	8	6	5
0,90	1122	281	126	71	46	32	24	19	12	9	7	6
0,95	1303	327	146	83	53	37	28	22	14	10	8	6
0,99	1676	420	187	106	68	48	36	27	18	13	10	8

Tabella 5.4 – *Grandezza del campione in funzione della potenza e dell'indice f (test F) per $\alpha = 0{,}01$ con 8 gradi di libertà*

L'indice di grandezza dell'effetto impiegato in questo caso è f (cfr. Cap. 3). I gradi di libertà indicati nella tabella si riferiscono al numeratore del rapporto F e corrispondono a $(k - 1)$, dove k sono i livelli della variabile indipendente. Nel caso di ANOVA fattoriali, con più di una variabile indipendente, i gradi di libertà delle interazioni sono dati ad esempio da $(k - 1) \times (r - 1)$, dove k ed r sono i livelli delle due variabili indipendenti, oppure da $(k - 1) \times (r - 1) \times (p - 1)$, nel caso di tre variabili indipendenti, e così via.

Esempio 5.4.1

ANOVA ad una via: Poniamo di avere una variabile indipendente su 9 livelli (ad esempio tutta la popolazione italiana da 0 a 90 anni suddivisa in decadi di età), di conseguenza $gl = (9\text{-}1) = 8$, e stabiliamo il livello di α a 0,01. Vogliamo assicurarci una potenza accettabile, diciamo 0,80, ed ipotizziamo una grandezza dell'effetto di media entità, $f = 0{,}25$. I parametri di partenza sono:

$\alpha = 0{,}01$ $gl = 8$ $f = 0{,}25$ potenza = 0,80

Dalla tabella 5.4, all'incrocio tra la colonna $f = 0{,}25$ e la riga potenza = 0,80, troviamo il valore 38, ma questo non significa certo che il campione totale deve essere di 38, bensì vuol dire che ogni gruppo (derivato da ciascun livello della variabile indipendente) deve essere composto da 38 partecipanti. Quindi il campione totale è dato da $N = (38 \times 9) = 342$.

Cosa accade se, a parità di tutte le altre condizioni, ipotizzassimo una grandezza dell'effetto lievemente inferiore, ad esempio $f = 0{,}20$? I parametri diventano:

$\alpha = 0{,}01$ $gl = 8$ $f = 0{,}20$ potenza = 0,80

In questo caso, dalla tabella 5.4, estraiamo un numero di 58 partecipanti per ogni gruppo, che diventa di $(58 \times 9) = 522$ partecipanti per il campione totale. Come sempre, una diminuzione nella grandezza dell'effetto comporta un incremento notevole nella numerosità del campione.

Esempio 5.4.2

ANOVA fattoriale: In questo esempio verranno utilizzate delle tavole di Cohen (1988) che non sono state riportate nel testo per esigenze di spazio; è importante comunque che il lettore, in questa fase, comprenda il procedimento specifico per questo tipo di ANOVA, a prescindere dai valori estrapolati dalle tavole.

Poniamo di avere un disegno della ricerca con tre variabili indipendenti (A, B e C) aventi rispettivamente 2, 3 e 4 livelli. Questo disegno fattoriale avrà, quindi, $(2 \times 3 \times 4) = 24$ gruppi o celle. Poniamo il livello di α a 0,05. Per ogni effetto principale (A, B e C) e per ciascuna interazione (A×B, A×C, B×C e A×B×C) dovremo stabilire a priori sia la potenza sia la grandezza dell'effetto. Poniamo che la potenza sia uguale per tutti i test, cioè 0,80, mentre la grandezza dell'effetto sarà 0,10 per A, 0,25 per B, 0,40 per C e 0,30 per tutte le interazioni. Procediamo alla determinazione della grandezza del campione partendo dagli effetti principali.

I parametri per A sono:

$\alpha = 0{,}05$ $gl = 2 - 1 = 1$ $f = 0{,}10$ potenza = 0,80

Dalla tavola di Cohen (1988) specifica per questi parametri, ricaviamo il valore 394. Questo valore, in realtà, non fa parte di un disegno con due gruppi (tali sono i livelli della variabile indipendente considerata) bensì di un disegno fattoriale con 24 gruppi. Quindi per ricavare da esso la numerosità di ciascuno dei 24 gruppi (n_c) dovremo utilizzare la seguente formula per disegni fattoriali:

$$n_c = \frac{(n' - 1) \times (gl + 1)}{numero\ di\ celle} + 1, \tag{5.3}$$

dove n' è il numero ricavato dalla tavola, gl sono i gradi di libertà della variabile presa in considerazione e *numero di celle* corrisponde precisamente al numero di gruppi o celle del disegno fattoriale. Sostituendo i valori ottenuti per la variabile A nella formula 5.3 otteniamo:

$$n_c = \frac{(394 - 1) \times (1 + 1)}{24} + 1 = \frac{393 \times 2}{24} + 1 = \frac{786}{24} + 1 = 32,75 + 1 = 33,75 \approx 34.$$

Ora sarà sufficiente moltiplicare la numerosità di ciascuna cella per il numero di celle ed avremo la numerosità del campione totale, quindi $N = (24 \times 34) = 816$.

Prendiamo adesso in considerazione la seconda variabile indipendente B; i suoi parametri sono:

$\alpha = 0,05$ $gl = 3 - 1 = 2$ $f = 0,25$ potenza = 0,80

La relativa tavola di Cohen (1988) ci dà il valore 52. Sostituendo nella (5.3) troviamo:

$$n_c = \frac{(52 - 1) \times (2 + 1)}{24} + 1 = \frac{51 \times 3}{24} + 1 = \frac{153}{24} + 1 = 6,375 + 1 = 7,375 \approx 8.$$

Infine ricaviamo facilmente il campione totale: $N = (8 \times 24) = 192$. La numerosità richiesta da questa variabile è molto inferiore rispetto alla prima e questo è da attribuirsi principalmente all'incremento della grandezza dell'effetto.

Vediamo cosa accade con la terza variabile indipendente (C). I parametri sono:

$\alpha = 0,05$ $gl = 4 - 1 = 3$ $f = 0,40$ potenza = 0,80

La relativa tavola di Cohen (1988) ci dà il valore 18. Dalla (5.3) troviamo:

$$n_c = \frac{(18 - 1) \times (3 + 1)}{24} + 1 = \frac{17 \times 4}{24} + 1 = \frac{68}{24} + 1 = 2,83 + 1 = 3,83 \approx 4.$$

Quindi N sarà dato da $(4 \times 24) = 96$. Un risultato atteso dato che abbiamo ulteriormente incrementato la grandezza dell'effetto.

I tre effetti principali di questo esempio hanno dato come esito grandezze del campione molto diverse tra loro: 816 partecipanti per A, 192 per B e 96 per C. Vediamo, rapidamente, cosa accade andando ad utilizzare le interazioni.

Per l'interazione A×B i parametri sono:

$\alpha = 0,05$ $gl = (2 - 1)(3 - 1) = 2$ $f = 0,30$ potenza = 0,80

La relativa tavola di Cohen (1988) ci dà il valore 36. Dalla (5.3) troviamo:

$$n_c = \frac{(36 - 1) \times (2 + 1)}{24} + 1 = \frac{35 \times 3}{24} + 1 = \frac{105}{24} + 1 = 4,375 + 1 = 5,375 \approx 6.$$

Quindi $N = (6 \times 24) = 144$.
Per l'interazione A×C i parametri sono:

$\alpha = 0{,}05$ $gl = (2 - 1)(4 - 1) = 3$ $f = 0{,}30$ potenza = 0,80

La relativa tavola di Cohen (1988) ci dà ora il valore 31. Dalla (5.3) troviamo:

$$n_c = \frac{(31 - 1)\,x\,(3 + 1)}{24} + 1 = \frac{30\,x\,4}{24} + 1 = \frac{120}{24} + 1 = 5 + 1 = 6.$$

Quindi $N = (6 \times 24) = 144$.
Per l'interazione B×C i parametri sono:

$\alpha = 0{,}05$ $gl = (3 - 1)(4 - 1) = 6$ $f = 0{,}30$ potenza = 0,80

La tavola di Cohen (1988) ci dà qui il valore 22. Dalla (5.3) troviamo:

$$n_c = \frac{(22 - 1) \times (6 + 1)}{24} + 1 = \frac{21 \times 7}{24} + 1 = \frac{147}{24} + 1 = 6{,}125 + 1 = 7{,}125 \approx 8.$$

Quindi $N = (8 \times 24) = 192$.
Infine per l'interazione A×B×C i parametri sono:

$\alpha = 0{,}05$ $gl = (2 - 1)(3 - 1)(4 - 1) = 6$ $f = 0{,}30$ potenza = 0,80

La relativa tavola di Cohen (1988) ci dà il valore 22. Dalla (5.3) troviamo:

$$n_c = \frac{(22 - 1) \times (6 + 1)}{24} + 1 = \frac{21 \times 7}{24} + 1 = \frac{147}{24} + 1 = 6{,}125 + 1 = 7{,}125 \approx 8.$$

Quindi, $N = (8 \times 24) = 192$.

Riassumiamo nella tabella 5.5 tutti i valori di N ricavati dagli effetti principali e dalle interazioni del disegno fattoriale analizzato.

Effetto	Parametri				n_c	N
	α	*gl*	*f*	potenza		
A	0,05	1	0,10	0,80	34	816
B	0,05	2	0,25	0,80	8	192
C	0,05	3	0,40	0,80	4	96
A×B	0,05	2	0,30	0,80	6	144
A×C	0,05	3	0,30	0,80	6	144
B×C	0,05	6	0,30	0,80	8	192
A×B×C	0,05	6	0,30	0,80	8	192

Tabella 5.5 – *Grandezza del campione (principali ed interazioni) di una ANOVA*

Come si procedere per decidere quale grandezza del campione utilizzare di fronte a risultati così diversi? Una prima considerazione da farsi è che nei disegni fattoriali la questione centrale sono le interazioni, mentre gli effetti principali risultano secondari. Di conseguenza sarebbe superfluo preoccuparsi della numerosità enorme (816) richiesta dalla variabile A. Strategia migliore è andare a considerare le indicazioni fornite dalle interazioni: un campione di 192 partecipanti sarebbe ottimale, ed è quindi auspicabile attenersi a questa numerosità.

5.4.1.2. Tavole di Cohen per l'analisi di potenza a posteriori

Riportiamo, modificata, (tabella 5.6) una delle sei tavole che Cohen (1988) utilizza per determinare a posteriori la potenza nel caso si utilizzi il test *t* per campioni indipendenti. Qui l'indice di grandezza dell'effetto è il *d*.

ESEMPIO 5.5

Poniamo di avere una ipotesi monodirezionale con $\alpha = 0{,}01$, di aver utilizzato un campione di 50 partecipanti e che l'indice di grandezza dell'effetto, calcolato sui dati raccolti, sia $d = 0{,}40$. Riassumiamo i parametri:

$\alpha = 0{,}01$ $\qquad$ $d = 0{,}40$ $\qquad$ $n = 50$

Nella tabella 5.6 all'incrocio tra la colonna $d = 0{,}40$ e la riga $n = 50$ otteniamo un valore di potenza di 0,36.

Il test utilizzato in queste condizioni risulta avere una potenza piuttosto bassa, la probabilità di rilevare un effetto, qualora esista, è del 36%, mentre la probabilità di commettere un errore di secondo tipo è $(1 - 0{,}36) = 0{,}64$, cioè del 64%.

Vediamo di quanto migliora la potenza del test portando la numerosità del campione a 100 e lasciando invariate le altre condizioni. I parametri saranno quindi:

$\alpha = 0{,}01$ $\qquad$ $d = 0{,}40$ $\qquad$ $n = 100$

In questo caso, dalla tabella 5.6, estraggo un valore di potenza di 0,69. La probabilità di rilevare un effetto esistente è salita al 69%, praticamente raddoppiata. Questo ci serve per comprendere chiaramente quale influenza diretta ha la grandezza del campione sulla potenza del test e quindi l'importanza di effettuare una analisi di potenza a priori per determinare *n* ed effettuare una ricerca con presupposti ragionevolmente solidi.

Prendiamo adesso in considerazione, modificata, una delle sei tavole che Cohen (1988) ha messo a punto per ricavare la potenza quando si utilizza il coefficiente di correlazione *r* (tabella 5.7).

						d					
n	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	1,00	1,20	1,40
8	0,02	0,03	0,04	0,05	0,08	0,12	0,14	0,19	0,30	0,43	0,57
10	0,02	0,03	0,04	0,07	0,10	0,14	0,18	0,25	0,40	0,55	0,70
12	0,02	0,03	0,05	0,08	0,12	0,17	0,23	0,31	0,49	0,66	0,81
14	0,02	0,03	0,06	0,09	0,14	0,20	0,28	0,38	0,57	0,75	0,88
16	0,02	0,04	0,06	0,10	0,16	0,24	0,34	0,44	0,64	0,82	0,92
18	0,02	0,04	0,07	0,12	0,19	0,27	0,38	0,49	0,71	0,87	0,95
20	0,02	0,04	0,08	0,13	0,21	0,30	0,42	0,54	0,76	0,91	0,97
22	0,02	0,05	0,08	0,15	0,23	0,34	0,46	0,59	0,81	0,94	0,98
24	0,02	0,05	0,09	0,16	0,25	0,37	0,50	0,64	0,85	0,95	0,99
26	0,02	0,05	0,10	0,17	0,28	0,41	0,55	0,68	0,89	0,97	0,99
28	0,02	0,05	0,11	0,19	0,30	0,44	0,59	0,72	0,91	0,98	*
30	0,03	0,06	0,11	0,20	0,32	0,48	0,62	0,75	0,93	0,99	
34	0,03	0,06	0,13	0,23	0,37	0,53	0,69	0,81	0,95	0,99	
38	0,03	0,07	0,15	0,26	0,42	0,60	0,75	0,86	0,97	*	
40	0,03	0,07	0,15	0,28	0,45	0,62	0,78	0,88	0,98		
44	0,03	0,08	0,17	0,31	0,49	0,67	0,82	0,91	0,99		
48	0,03	0,08	0,19	0,34	0,53	0,71	0,85	0,94	0,99		
50	0,03	0,09	0,20	0,36	0,55	0,73	0,87	0,95	0,99		
54	0,04	0,10	0,21	0,39	0,59	0,77	0,90	0,96	*		
58	0,05	0,10	0,23	0,41	0,62	0,81	0,92	0,97			
64	0,05	0,11	0,26	0,46	0,68	0,85	0,94	0,98			
72	0,05	0,12	0,29	0,52	0,74	0,89	0,97	0,99			
80	0,05	0,14	0,33	0,57	0,78	0,92	0,98	*			
88	0,06	0,16	0,36	0,62	0,83	0,95	0,99				
100	0,06	0,18	0,41	0,69	0,88	0,97	*				
140	0,07	0,25	0,57	0,84	0,96	*					
180	0,08	0,33	0,69	0,93	0,99						
250	0,11	0,46	0,84	0,98	*						
300	0,13	0,55	0,91	0,99							
350	0,16	0,61	0,95	*							
400	0,18	0,69	0,97								
500	0,22	0,80	0,99								
600	0,27	0,87	*								
700	0,32	0,92									
800	0,37	0,95									
900	0,42	0,97									
1000	0,46	0,98									

* Al di sotto di questo punto la potenza è superiore a 0,995

Tabella 5.6 – *Potenza del test t con* $\alpha = 0{,}01$ *monodirezionale*

					r				
n	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
8	0,08	0,12	0,18	0,26	0,37	0,52	0,68	0,85	0,97
9	0,08	0,13	0,20	0,29	0,42	0,57	0,74	0,90	0,99
10	0,08	0,14	0,22	0,32	0,46	0,62	0,79	0,93	0,99
12	0,09	0,15	0,25	0,38	0,54	0,71	0,87	0,97	*
14	0,10	0,17	0,28	0,43	0,60	0,78	0,91	0,98	
16	0,10	0,19	0,31	0,48	0,66	0,83	0,95	0,99	
18	0,11	0,20	0,34	0,52	0,71	0,87	0,97	*	
20	0,11	0,22	0,37	0,56	0,75	0,90	0,98		
23	0,12	0,24	0,41	0,62	0,81	0,94	0,99		
26	0,12	0,26	0,45	0,67	0,85	0,96	*		
29	0,13	0,28	0,49	0,71	0,89	0,97			
33	0,14	0,30	0,53	0,76	0,92	0,99			
37	0,15	0,33	0,57	0,80	0,95	0,99			
42	0,16	0,36	0,62	0,85	0,97	*			
48	0,17	0,39	0,67	0,89	0,98				
54	0,18	0,43	0,72	0,92	0,99				
60	0,19	0,46	0,76	0,94	0,99				
68	0,20	0,50	0,81	0,96	*				
76	0,22	0,54	0,85	0,98					
84	0,23	0,58	0,88	0,99					
92	0,24	0,61	0,90	0,99					
100	0,26	0,64	0,92	0,99					
120	0,29	0,71	0,96	*					
140	0,32	0,77	0,98						
160	0,35	0,82	0,99						
180	0,38	0,86	0,99						
200	0,41	0,89	*						
250	0,47	0,94							
300	0,54	0,97							
400	0,64	0,99							
500	0,72	*							
600	0,79								
700	0,84								
800	0,88								
900	0,91								
1000	0,94								

* Al di sotto di questo punto la potenza è superiore a 0,995

Tabella 5.7 – *Potenza del coefficiente r con* $\alpha = 0{,}05$ *monodirezionale*

Esempio 5.6

Poniamo di aver formulato una ipotesi monodirezionale con $\alpha = 0{,}05$, di aver utilizzato un campione composto da 160 partecipanti e di aver ottenuto un coefficiente *r* di 0,10 dai dati raccolti. I parametri a disposizione sono:

$\alpha = 0{,}05$ $r = 0{,}10$ $n = 160$

Dalla tabella 5.7, incrociando i parametri noti, otteniamo un valore di potenza pari a 0,35, in altre parole, avevamo una probabilità del 35% di rilevare un effetto, posto che questo ci fosse. In questo caso la numerosità del campione sembrerebbe più che soddisfacente ma la potenza ottenuta è comunque troppo bassa, perché? Perché la grandezza dell'effetto è comunque troppo ridotta: con un effetto così minimale la probabilità di rilevarlo è scarsa anche utilizzando moltissimi soggetti.

Lasciando invariati gli altri parametri, proviamo ad incrementare, seppure non di molto, la grandezza dell'effetto: supponiamo che il coefficiente *r*, calcolato sui dati raccolti, fosse stato 0,20. I parametri sarebbero stati:

$\alpha = 0{,}05$ $r = 0{,}20$ $n = 160$

Utilizzando la tabella 5.7, avremmo trovato un valore di potenza pari a 0,82, praticamente il valore indicato dalla letteratura come ottimale. La probabilità di mettere in evidenza un effetto esistente è passata dal 35% all'82%. Non è necessario insistere sulla rilevanza della grandezza dell'effetto nel determinare la potenza, essa ci pare evidente.

Vediamo ora la tabella 5.8, che si riferisce al χ^2. Modificata, è una delle 46 tavole che Cohen (1988) ha presentato per questo test.

Per la grandezza dell'effetto, come è noto, si fa riferimento all'indice *w*.

Esempio 5.7

Poniamo di avere una tabella 4×2, quindi con 3 gradi di libertà. Avendo posto $\alpha = 0{,}05$, con un campione di 50 partecipanti, abbiamo trovato una grandezza dell'effetto pari a $w = 0{,}30$. I parametri di partenza sono:

$\alpha = 0{,}05$ $gl = 3$ $w = 0{,}30$ $n = 50$

Dalla tabella 5.8, intersecando i dati ottenuti, troviamo una potenza di 0,40, corrispondente al 40% di probabilità di far emergere un effetto qualora esista. Come è chiaro, la potenza ottenuta non può essere considerata soddisfacente.

Osservando la tabella 5.8 e scorrendo la colonna corrispondente alla grandezza dell'effetto ottenuta dai dati, $w = 0{,}30$, è possibile notare che, per ottenere una potenza accettabile (ad esempio di 0,80) avrei dovuto utilizzare un campione di almeno 120 partecipanti, grande cioè più del doppio di quello che invece è stato impiegato.

					w				
n	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
25	0,07	0,12	0,21	0,36	0,54	0,71	0,85	0,93	0,98
30	0,07	0,13	0,25	0,42	0,62	0,80	0,90	0,97	0,99
35	0,07	0,15	0,29	0,49	0,70	0,86	0,95	0,99	*
40	0,07	0,16	0,32	0,55	0,76	0,90	0,97	0,99	
50	0,08	0,19	0,40	0,65	0,86	0,96	0,99	*	
60	0,09	0,22	0,47	0,74	0,92	0,98	*		
70	0,09	0,26	0,54	0,81	0,95	0,99			
80	0,10	0,29	0,60	0,86	0,98	*			
100	0,12	0,36	0,71	0,93	0,99				
120	0,13	0,42	0,80	0,97	*				
160	0,16	0,55	0,90	0,99					
200	0,19	0,65	0,96	*					
250	0,23	0,76	0,99						
300	0,27	0,84	*						
350	0,32	0,90							
400	0,36	0,93							
500	0,44	0,98							
600	0,52	0,99							
700	0,59	*							
800	0,65								
900	0,71								
1000	0,76								

* Al di sotto di questo punto la potenza è superiore a 0,995

Tabella 5.8 – *Potenza del test* χ^2 *con* $\alpha = 0{,}05$ *e 3 gradi di libertà*

Vediamo invece ora la tabella 5.9., ancora modificata, che è una delle 33 tavole di Cohen (1988) per l'ANOVA, in cui si fa quindi riferimento è *f*. La tavola contiene, inoltre, i gradi di libertà del numeratore del rapporto *F*, quindi in generale $gl = (k - 1)$, dove k sono i livelli della variabile indipendente.

Come abbiamo detto per l'analisi di potenza a priori, è importante ricordare che, nel caso di ANOVA fattoriali, i gradi di libertà delle interazioni corrispondono, ad esempio, a $(k - 1) \times (r - 1)$, dove k ed r sono i livelli delle due variabili indipendenti, oppure a $(k - 1) \times (r - 1) \times (p - 1)$, nel caso di tre variabili indipendenti, e così via.

Per quanto concerne il valore di n indicato nella tabella, esso non si riferisce al campione totale bensì alla numerosità di ciascun sottogruppo, costruito sulla base del numero dei livelli della variabile indipendente. Se si ha a disposizione solo il numero totale dei partecipanti, per avere n sarà sufficiente suddividere il campione totale per il numero di livelli della variabile indipendente, anche se i gruppi hanno numerosità diversa:

$$n = \frac{N}{k}. \tag{5.4}$$

	f											
n	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,50	0,60	0,70	0,80
5	0,05	0,06	0,07	0,09	0,11	0,14	0,17	0,22	0,32	0,44	0,56	0,69
6	0,05	0,06	0,07	0,10	0,13	0,16	0,21	0,26	0,39	0,53	0,67	0,79
7	0,05	0,06	0,08	0,11	0,14	0,19	0,25	0,31	0,46	0,62	0,76	0,87
8	0,05	0,06	0,08	0,12	0,16	0,22	0,28	0,36	0,53	0,69	0,83	0,92
9	0,05	0,07	0,09	0,13	0,18	0,24	0,32	0,40	0,59	0,75	0,88	0,95
10	0,05	0,07	0,10	0,14	0,20	0,27	0,35	0,45	0,64	0,81	0,91	0,97
12	0,06	0,07	0,11	0,16	0,23	0,32	0,42	0,53	0,74	0,88	0,96	0,99
14	0,06	0,08	0,12	0,18	0,27	0,38	0,49	0,61	0,81	0,93	0,98	*
16	0,06	0,08	0,13	0,21	0,31	0,43	0,55	0,67	0,86	0,96	0,9	
18	0,06	0,09	0,14	0,23	0,34	0,48	0,61	0,73	0,90	0,98	*	
20	0,06	0,09	0,16	0,26	0,38	0,52	0,66	0,78	0,93	0,99		
23	0,06	0,10	0,18	0,29	0,43	0,59	0,73	0,84	0,96	*		
26	0,06	0,11	0,20	0,33	0,48	0,65	0,79	0,89	0,98			
29	0,06	0,12	0,22	0,36	0,53	0,70	0,83	0,92	0,99			
32	0,07	0,12	0,24	0,40	0,58	0,75	0,87	0,94	0,99			
36	0,07	0,13	0,26	0,44	0,63	0,80	0,91	0,97	*			
40	0,07	0,15	0,29	0,48	0,68	0,84	0,94	0,98				
44	0,07	0,16	0,32	0,53	0,73	0,88	0,96	0,99				
48	0,08	0,17	0,34	0,57	0,77	0,90	0,97	0,99				
54	0,08	0,19	0,38	0,62	0,82	0,94	0,98	*				
60	0,08	0,21	0,42	0,67	0,86	0,96	0,99					
68	0,09	0,23	0,47	0,73	0,90	0,98	*					
76	0,09	0,25	0,52	0,78	0,93	0,99						
84	0,10	0,28	0,56	0,82	0,95	0,99						
92	0,10	0,30	0,60	0,85	0,97	*						
100	0,11	0,32	0,64	0,88	0,98							
120	0,12	0,38	0,73	0,94	0,99							
140	0,14	0,44	0,79	0,97	*							
180	0,16	0,54	0,89	0,99								
250	0,22	0,69	0,97	*								
350	0,29	0,84	*									
450	0,36	0,92										
600	0,47	0,98										
700	0,53	0,99										
900	0,65	*										
1000	0,70											

* Al di sotto di questo punto la potenza è superiore a 0,995

Tabella 5.9 – *Potenza del test F con* $\alpha = 0{,}05$ *e 2 gradi di libertà*

ESEMPIO 5.8.1

ANOVA ad una via: Poniamo di avere una variabile indipendente su tre livelli, quindi $gl = (3 - 1) = 2$, di aver posto $\alpha = 0{,}05$, di aver impiegato un campione totale (N) composto da 180 partecipanti (quindi, come abbiamo visto, $n = 180/3 = 60$), e di aver calcolato sui dati raccolti l'indice di grandezza dell'effetto, $f = 0{,}15$. I parametri di partenza sono:

$\alpha = 0{,}05$ $gl = 2$ $f = 0{,}15$ $n = 60$

Incrociando questi dati nella tabella 5.9, troviamo che la potenza del test è pari a 0,42, decisamente non soddisfacente.

Poniamo però il caso che la grandezza dell'effetto calcolata sui dati fosse stata $f = 0{,}18$, come procediamo per ricavare la potenza dato che questa colonna nella tabella non esiste? I parametri sono:

$\alpha = 0{,}05$ $gl = 2$ $f = 0{,}18$ $n = 60$

Una interpolazione lineare ci permette di ottenere un valore approssimato della potenza partendo dai valori più prossimi presenti nella tabella. Il valore di f più vicino per difetto è 0,15 e, come abbiamo visto, per la colonna corrispondente a questo valore la potenza è 0,42; altresì il valore più prossimo per eccesso è $f = 0{,}20$ ed in questo caso la potenza corrispondente è 0,67 (vedi tabella 5.9). Per effettuare l'interpolazione utilizziamo la seguente formula:

$$x_{\text{inf}} + \frac{\left(par - par_{\text{inf}}\right)}{\left(par_{\text{sup}} - par_{\text{inf}}\right)} \times \left(x_{\text{sup}} - x_{\text{inf}}\right), \tag{5.5}$$

dove: x_{inf} e x_{sup} sono i valori della variabile da ricavare corrispondenti ai parametri prossimi per difetto e per eccesso; par_{inf} ed par_{sup} corrispondono ai valori del parametro più vicini, rispettivamente per difetto e per eccesso; e par è il valore del parametro trovato sul campione. È importante sapere e ricordare che questa formula per l'interpolazione può essere utilizzata per tutti i test presentati e tutti i rispettivi parametri.

Tornando al nostro esempio: x_{inf} e x_{sup} corrisponderanno ai valori di potenza approssimati per difetto e per eccesso, par_{inf} ed par_{sup} saranno i valori dell'indice f più vicini e par sarà il valore dell'indice f ottenuto dal campione. A questo punto, sostituendo questi valori nella formula 5.5, otteniamo:

$$0{,}42 + \frac{\left(0{,}18 - 0{,}15\right)}{\left(0{,}20 - 0{,}15\right)} \times \left(0{,}67 - 0{,}42\right) = 0{,}42 + \frac{\left(0{,}03\right)}{\left(0{,}05\right)} \times \left(0{,}25\right) = 0{,}57.$$

Per i parametri proposti, la potenza del test è 0,57, corrispondente al 57% della probabilità di rilevare un effetto esistente.

È utile familiarizzarsi con la procedura di interpolazione poiché nella pratica della ricerca potrà capitare spesso di incontrare valori degli indici di grandezza dell'effetto che non saranno inclusi nelle tabelle. Invitiamo, pertanto, il lettore ad esercitarsi in questo procedimento per i vari test ed indici proposti nel presente capitolo.

Esempio 5.8.2

ANOVA fattoriale: Anche in questo esempio, come per l'analisi di potenza a priori per ANOVA fattoriali, verranno utilizzate delle tavole di Cohen (1988) che esigenze di spazio ci hanno indotto a non riportare nel testo; quello che qui ci interessa è che il lettore comprenda semplicemente la procedura per questo tipo di ANOVA.

Poniamo di avere un disegno fattoriale con due variabili indipendenti (A e B) rispettivamente su 3 e 4 livelli: avremo un numero di celle pari a (3×4) = 12. Supponiamo di aver posto $\alpha = 0{,}05$ e di aver calcolato la grandezza dell'effetto per ciascuna variabile e per l'interazione: $f = 0{,}25$ per A, $f = 0{,}40$ per B e f = 0,30 per A×B. Il campione totale è di 72 partecipanti, per cui ciascuna cella avrà un valore $n_c = (72/12) = 6$.

In realtà è necessario effettuare una correzione sul valore n_c a causa di una discrepanza nei gradi di libertà: le tavole di Cohen (1988) sono state costruite basandosi su una ANOVA ad una via, quindi presupponendo un solo effetto ed un solo valore dei gradi libertà, mentre noi stiamo valutando un disegno in cui entrano in gioco più effetti contemporaneamente e, con essi, diversi valori di gradi di libertà. Senza entrare più in profondità nella questione, basti sapere che è sufficiente calcolare n' utilizzando la seguente formula:

$$n' = \frac{N - numero\ di\ celle}{df_{effetto} + 1} + 1. \tag{5.6}$$

Questo significa che nel caso dell'effetto A, utilizzando la formula 5.6, avremo:

$$n' = \frac{72 - 12}{(3 - 1) + 1} + 1 = \frac{60}{3} + 1 = 21.$$

Per l'effetto B, avremo:

$$n' = \frac{72 - 12}{(4 - 1) + 1} + 1 = \frac{60}{4} + 1 = 16.$$

Per l'interazione A×B, avremo:

$$n' = \frac{72 - 12}{(3 - 1)(4 - 1) + 1} + 1 = \frac{60}{6 + 1} + 1 = \frac{60}{7} + 1 = 8{,}6 + 1 = \mathbf{9{,}6}.$$

A questo punto possiamo procedere a ricavare la potenza per i singoli effetti principali e per la loro interazione.

I parametri dell'effetto A sono:

$\alpha = 0{,}05$ $\quad gl = 2$ $\quad f = 0{,}25$ $\quad n' = 21$

Utilizzando la tavola di Cohen (1988) adatta a questi parametri, troviamo che la potenza ha valore 0,40, quindi piuttosto bassa.

I parametri dell'effetto B sono:

$\alpha = 0{,}05$ $\quad gl = 3$ $\quad f = 0{,}40$ $\quad n' = 16$

Mediante la corrispondente tavola di Cohen (1988), troviamo un valore di potenza pari a 0,75, quindi molto migliore della precedente.

I parametri dell'interazione A×B sono:

$\alpha = 0{,}05$ $gl = 6$ $f = 0{,}30$ $n' = 9{,}6$

In questo caso, avendo un valore di *n* che non compare sulla tavola, dovremo ricorrere nuovamente ad una interpolazione lineare. Il valore di *n* prossimo per difetto è 9 ed a questo valore, sulla tavola corrispondente (Cohen, 1988), troviamo una potenza di 0,35, mentre il valore prossimo per eccesso è 10 ed a questo corrisponde una potenza di 0,39. Sostituendo i valori ottenuti nella formula 5.5, troviamo:

$$0,35 + \frac{(9,6 - 9)}{(10 - 9)} \times (0,39 - 0,35) = 0,35 + 0,6 \times 0,04 = 0,35 + 0,024 = 0,374.$$

La potenza dell'interazione A×B è 0,374, particolarmente bassa ed inferiore ad entrambe quelle trovate in precedenza per gli effetti principali.

Questo risultato in verità non ci dovrebbe stupire particolarmente: dato che i gradi di libertà delle interazioni sono dati dal prodotto dei gradi di libertà dei singoli effetti, le interazioni, nei disegni fattoriali, avranno sempre più gradi di libertà rispetto agli effetti e, considerando la formula 5.6, risulta chiaro che il loro *n'* sarà generalmente inferiore di quello degli effetti. Questo comporta che, a parità di α ed *f*, la potenza delle interazioni sarà sempre minore rispetto alla potenza dei singoli effetti principali.

Una possibile strategia indicata da Cohen (1988) per ovviare, almeno in parte, al problema della scarsa potenza delle interazioni potrebbe essere quella di alzare leggermente il livello di α (ad esempio 0,10) per le interazioni, lasciandolo adeguatamente basso (0,05) per gli effetti principali. Chiaramente questo comporta una perdita di credibilità qualora l'ipotesi nulla per l'interazione venga rifiutata: laddove per $\alpha = 0{,}05$ la probabilità di rifiutare l'ipotesi nulla quando è vera è del 5%, ponendo $\alpha = 0{,}10$ questa probabilità sale al 10%. D'altra parte questa perdita in affidabilità dei risultati potrebbe essere considerata un prezzo equo da pagare in favore di un incremento della potenza. Come suggerisce Cohen (1988), una decisione del genere deve, però, essere presa considerando non soltanto il disegno della ricerca e la grandezza dell'effetto, ma anche le questioni teoriche sulle quali è fondata la ricerca che viene condotta.

5.4.2. Programmi per l'analisi di potenza

Attualmente si trovano in commercio una varietà di programmi che permettono di effettuare analisi di potenza. Passeremo velocemente in rassegna i più diffusi cercando di metterne in luce le caratteristiche salienti.

1. **POWER AND PRECISION** (Borestein, Cohen e Rothstein, 1997), questo prodotto è commercializzato anche dalla SPSS con il nome di **SAMPLE POWER**. È piuttosto semplice da utilizzare, ma non prevede la possibilità di calcolare la potenza per disegni a misure ripetute e per la MANOVA.

2. **PASS** (NCSS Statistical Software, 1999), effettua anche analisi di potenza per disegni a misure ripetute. Non prevede la possibilità di calcolare la potenza per la MANOVA.

3. **G POWER** (2000), ha il pregio indiscutibile di essere gratuito e ne esistono due versioni: una DOS ed una per Mac. Molto semplice da utilizzare, a differenza dei programmi citati in precedenza, tutte le informazioni devono essere inserite in una finestra unica e questo rende il procedimento molto intuitivo e rapido. Questo programma fornisce, inoltre, un grafico che permette di visualizzare la relazione tra la grandezza del campione e la potenza.

4. **SAS MACRO**, Friendly (1991) ha scritto una Macro di SAS denominata **MPOWER** che può effettuare soltanto analisi di potenza a posteriori, ma include l'analisi per la MANOVA.

Infine, è importante sapere che anche due programmi ampiamente utilizzati nella ricerca psicologica forniscono delle analisi di potenza:

• **SPSS**, fornisce soltanto una valutazione della potenza a posteriori; infatti, permette di richiedere, durante l'ANOVA, una analisi di potenza (la procedura è: *Options* → *Display* → *Observed power*) che viene quindi riprodotta nell'output.

• **SYSTAT 10**, permette di effettuare sia analisi di potenza *a priori* che a posteriori per la maggior parte dei test statistici utilizzati e fornisce in output un grafico che visualizza la potenza in funzione della numerosità del campione.

Capitolo 6
Simulazione e ricampionamento

6.1. Il metodo Monte Carlo

In questo capitolo vedremo una serie di metodi, tutti o quasi curiosamente dotati di nomi altamente suggestivi, che certamente ne hanno favorito in qualche misura l'affermazione, che consentono di affrontare il problema della VeSN in modi decisamente alternativi. Tutti questi metodi si basano su tecniche di simulazione, e vanno in genere sotto il nome di metodi di ricampionamento (*resampling*). Prima però di entrare nel merito dell'uso delle tecniche di ricampionamento (randomizzazione, *cross-validation*, *jackknife* e *bootstrapping*), è allora opportuno spendere qualche parola sul metodo Monte Carlo, che è in qualche misura alla base di tutti questi metodi (salvo i metodi cosiddetti di probabilità esatta).

Esempio 6.1

Prima di entrare nello specifico, può valere la pena di presentare un esempio classico, che consentirà al lettore di entrare nello spirito della proposta simulativa. Si tratta di un esempio, e ciò potrà apparire curioso, apparentemente assai lontano dai problemi che c'interessano, e per di più almeno a prima vista ben lontano dalla possibilità di soluzione con tecniche probabilistiche, come quelle alla base della simulazione: il calcolo del pi greco, il famoso 3,14, così importante per il calcolo delle circonferenze e dell'area dei cerchi, al cui valore ognuno di noi penserebbe di dover arrivare solo con metodi analitici.

Si consideri la figura 6.1. In questa figura noi vediamo una circonferenza inscritta in un quadrato. Supponiamo di essere un po' monelli, di aver visto questa figura su un manifesto incollato ad un muro, di avere accanto a noi un cesto d'uova abbandonato da una massaia distratta, e di esserci venuta la voglia di colpire l'intera figura (non solo il cerchio, quindi, ma anche gli spicchi del quadrato non coperti dal cerchio) con lanci di uova. Immaginando che non intervengano altre variabili, ad esempio una particolare infermità al braccio che fa sì che le uova vadano a colpire prevalentemente una parte della figura, noi possiamo tranquillamente assumere che ogni parte di questa abbia le stesse probabilità di essere colpita da un uovo.

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*,
ISBN 88-8453-226-4,

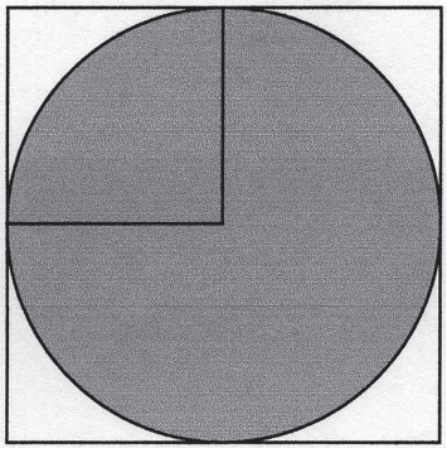

Figura 6.1 – *Un cerchio inscritto in un quadrato*

Concentriamo ora la nostra attenzione sull'angolo in alto a sinistra della figura, dove è evidenziato un quarto del quadrato con relativo spicchio di cerchio inscritto. Ora, sino dalle elementari noi sappiamo che l'area del cerchio è πr^2, dove r è il raggio del cerchio, ma è anche metà del lato del quadrato. Il quarto di quadrato in alto a sinistra avrà allora una superficie pari a $1/4\ r^2$, e lo spicchio di cerchio al suo interno avrà una superficie pari a $1/4\ \pi r^2$. Ora, ci sembra ragionevole assumere che il rapporto tra la quantità di uova che finiranno sullo spicchio di cerchio e la quantità di uova che finiranno sulla parte del quarto del quadrato non occupata dal cerchio, sia pari al rapporto tra le due aree. In altri termini, detto n_s il numero di uova che finisce sullo spicchio e n_c il numero di uova che finisce sul quadratino,

$$\frac{n_s}{n_c} = \frac{\frac{1}{4}\pi r^2}{r^2} = \frac{1}{4}\pi. \tag{6.1}$$

Ma ciò significa che π è dato da quattro volte il rapporto tra n_s e n_c. Per calcolare il suo valore, noi potremmo quindi semplicemente contare quante uova finiscono sullo spicchio di cerchio, e quante invece vanno a imbrattare la parte di quadratino non occupata. Questo metodo sarebbe sicuramente molto più divertente per gli scolari, ma sarebbe però estremamente impreciso, perché per raggiungere valori di π accettabili dovremmo effettuare un numero di lanci impressionantemente elevato, a livello di qualche migliaio. Ed è qui che interviene la simulazione.

Noi sappiamo che tutti i punti che sono all'interno dello spicchio di cerchio distano tra 0 e r dal centro della figura. Ora il nostro quadratino può essere considerato un quadrante di uno spazio delimitato da due assi cartesiani, e i suoi limiti sono dati dai punti di coordinate (0, 0) e $(-r, r)$. Noi possiamo allora far generare al calcolatore quante coppie di numeri casuali vogliamo, ponendo come vincolo che siano compresi il primo tra $-r$ e 0 e il secondo tra 0 e r, e considereremo il primo numero un valore di coordinata di ascissa x e il secondo un valore di

coordinata di ordinata *y*. In questo modo, ogni coppia di valori corrisponde alle coordinate di un punto interno al quadratino. Ora, noi sappiamo che per il teorema di Pitagora la distanza di un punto dall'origine è data dalla radice quadrata di $x^2 + y^2$. È allora evidente che tutti i punti così generati che hanno una distanza uguale o inferiore dall'origine appartengono allo spicchio di cerchio, mentre quelli che hanno una distanza superiore apparterranno alla parte rimanente del quadratino. Il rapporto tra il numero dei primi e il numero dei secondi, moltiplicato in base alla (6.1) per 4, ci darà allora il valore di π.

Noi abbiamo così fatto generare al calcolatore iterativamente 100.000 coppie di valori casuali di coppie *x* e *y*, e i risultati delle 100000 iterazioni sono riportati nella tabella 6.1, da cui si vede che con l'aumentare del numero di iterazioni il valore di π si approssima sempre più al fatidico 3,14159265… .

N. iterazioni	Valore π
1000	3,152000
10000	3,127200
19000	3,134105
28000	3,135286
37000	3,132973
46000	3,138696
55000	3,134545
64000	3,134562
73000	3,134685
82000	3,136976
91000	3,136396
100000	3,13772

Tabella 6.1 – *Valori simulati di* π

Questi valori sono stati ottenuti con un semplice programma in *basic*, che molti lettori saranno in grado di implementare, forse con qualche piccola modifica, nei linguaggi *basic* disponibili sul loro calcolatore, e che proponiamo nel box 6.1. Abbiamo numerato le righe, per rendere le cose più chiare, anche se in molti *basic* questo non si fa più, e facciamo presente che le parti che iniziano con # non sono istruzioni, ma commenti al programma.

Facciamo infine presente che il risultato da noi ottenuto non è molto brillante, ma la cosa è probabilmente dovuta a difetti del generatore di numeri casuali (o meglio, pseudo-casuali) usato. Questo problema lo discuteremo meglio più avanti.

```
10 # Programma per calcolare il valore di pi greco con 100000 iterazioni
20 # Nella riga 100 vengono dimensionati i vettori che conterranno rispettivamente
   i valori di ascissa (x), di ordinata (y) e di pi nelle successive iterazioni
100 dim x(100000); dim y(100000); dim p(100000)
150 # nella riga 200 vengono posti i valori iniziali del numero delle iterazioni
   (n) e delle prove in cui si rimane all'interno del cerchio (r)
200 n = 0; r=0
250 # nelle righe 300-600 si svolge il loop delle 100000 iterazioni, che vengono
   conteggiate aumentando uno alla volta il valore di n
300 for i = 1 to 100000 step 1; n=n+1
350 # nella riga 400 viene assegnato un numero casuale compreso tra 0 e 1 ad ascis-
   sa ed ordinata all'i-esima interazione e viene calcolata la distanza a di tale
   punto dall'origine
400 x(i)=(rnd+1); y(i)=(rnd+1); a=sqr(x(i)^2+y(i)^2)
450 # nella riga 500 se la distanza a è inferiore a 1, e quindi il punto cade nel
   cerchio, si aumenta il valore di r di 1 e si calcola il valore di pi alla i-esima
   iterazione, dato da 4 volte il rapporto tra r e n
500 if a < 1 then r=r+1; p(i)=4*r/n
600 next i
650 nella righe 700-800 si crea un loop che consente di presentare come output i
   valori calcolati di pi a intervalli di 3000 iterazioni
700 for i = 1000 to 100000 step 3000; print p(i)
800 next i
9999 end
```

Box 6.1 – *Un semplice programma in basic per calcolare il valore di* π

6.1.1. Da Buffon a Metropolis e al metodo Monte Carlo

Si osservi che il problema del calcolo del π con la simulazione è tutt'altro che nuovo, ma in forma diversa era stato posto già nel corso del XVIII secolo da uno dei padri della scienza moderna, il naturalista francese Georges Louis Leclerc Conte di Buffon (1707-1788). Il problema venne posto da Buffon nel corso di una conferenza tenuta nel 1733 alla Académie Royale des Sciences, e quindi pubblicato, con la sua soluzione, solo 44 anni più tardi (Buffon, 1777), anche se aveva fatto discutere a lungo i matematici dell'epoca. (Sul problema di Buffon esiste una letteratura copiosissima; cfr. Schuster, 1974).

ESEMPIO 6.2

Il problema, nella sua forma più semplice, era il seguente: poste due linee parallele a distanza d una dall'altra, e posto un ago di lunghezza l, inferiore a d, qual è la probabilità che lanciando questo ago a caso in modo che una sua estremità cada sempre nello spazio compreso tra le due linee, l'ago ne intersechi una delle due? (vedi figura 6.2).

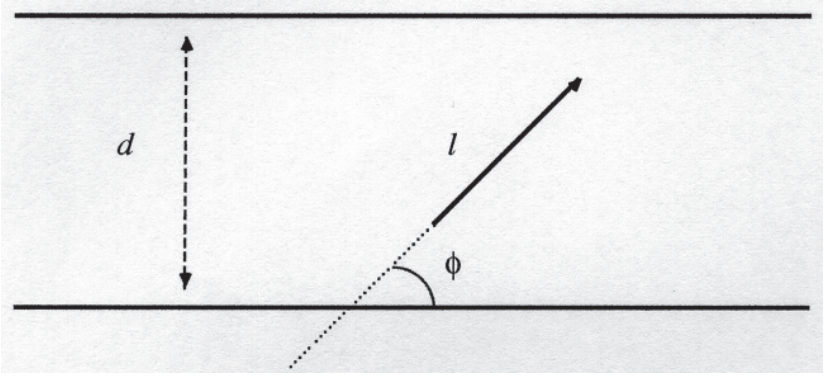

Figura 6.2 – *L'ago di Buffon*

La soluzione del problema non è semplicissima, e se qui la forniamo, avvertiamo che il lettore non si deve sentire obbligato a seguirla, quella che conta è la sua conclusione. Il principio è analogo a quello che abbiamo stabilito per l'esempio precedente: la probabilità è pari al rapporto tra la superficie su cui l'ago può cadere e che interseca una retta, e l'intera superficie su cui può cadere, che interseca e che non interseca una retta.

Ora, fatta una serie di n lanci, posto uguale a s il numero di volte che l'ago interseca una parallela, la probabilità che ciò avvenga è

$$p(s,n) = \frac{s}{n}. \tag{6.2}$$

Posto uguale a ϕ l'angolo che l'ago forma con le parallele, per il ragionamento sopra fatto relativo al rapporto tra le aree, tenuto conto che per il calcolo dell'area dovremo ricorrere all'integrale definito del coseno dell'angolo ϕ tra 0 e 2π (cioè, i valori di ϕ corrispondenti all'intera circonferenza), la nostra probabilità sarà uguale a

$$\frac{s}{n} = \int_0^{2\pi} \frac{l\left|\cos\phi\right|}{d} \frac{\mathrm{d}\phi}{2\pi} = \frac{2l}{\pi d} \int_0^{\pi/2} \cos\phi \cdot \mathrm{d}\phi = \frac{2l}{\pi d}. \tag{6.3}$$

Ne segue che

$$\pi = \frac{2ln}{ds}. \tag{6.4}$$

È questo allora un nuovo modo di calcolare il valore di π. Quindi armiamoci di coraggio, tracciamo su un foglio le nostre parallele, e cominciamo a lanciare il nostro ago. In capo ad alcune migliaia di lanci saremo giunti a un valore soddisfacentemente approssimato. Ma evidentemente non è questo il modo migliore per calcolare il valore del π. Il problema dell'ago di Buffon godette peraltro di grande popolarità per tutto il XIX secolo, attirò l'interesse di studiosi come Laplace (1812-1887), che per primo pensò di utilizzarlo per il calcolo empirico del π; e non mancò chi si dedicò intensamente al lancio di aghi e di freccette (che portano a conse-

guenze meno disturbanti del lancio di uova). Famoso rimase un certo Capitano O.C. Fox, che a quanto si tramanda nel 1864 trascorse il tempo libero che gli lasciava la necessità di guarire da ferite riportate nella Guerra Civile a lanciare per centinaia di volte aghi (cfr. Beckman, 1971), e ottenne un π variante tra 3,178 e 3,1416. Ma anche l'Italia ha la sua piccola gloria: il matematico Mario Lazzerini (1901), che ottenne un valore approssimato al sesto decimale (355/113, e cioè 3,1415929) dopo 3408 lanci dell'ago. Noi invece, astuti come volpi, come abbiamo fatto per l'esempio 6.1, ci rivolgeremo al calcolatore.

Il problema, ancora una volta, è quello di determinare le condizioni per cui l'ago tocca una parallela. Poniamo per semplicità $d = 1{,}5$ e $l = 1$. In questo caso, la (6.4) si riduce a $\pi = 4n/3s$. Ora si osservi la figura 6.3. Se in tale figura l'ago corrisponderà al segmento AB, da essa risulta evidente che l'ago toccherà una parallela ogni qual volta il segmento CD sarà più lungo di 0,75. Si osservi che l'ago è l'ipotenusa del triangolo rettangolo ABC, i cui cateti sono costituiti da BC e da AC.

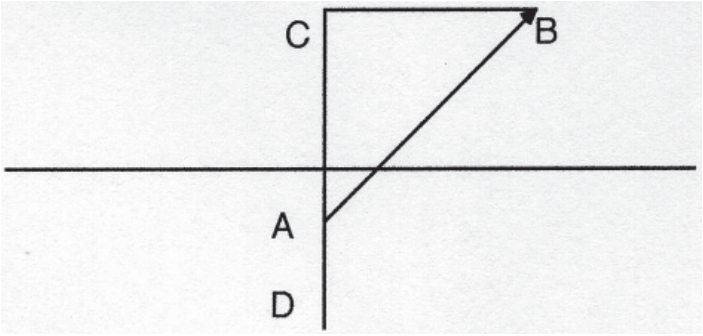

Figura 6.3 – *vedi testo*

Ora, essendo l'ipotenusa uguale a 1, BC non è altro che cosϕ, e avrà quindi un campo di variazione tra 0 (quando l'ago cade in corrispondenza di AC, e AB e AC coincidono) e 1 (quando l'ago cade parallelamente alle rette, e AB e BC coincidono). CD, a sua volta, è dato dalla somma di AD e del cateto AB, e cioè,

$$CD = AD + \sqrt{1 - BC^2} \quad . \tag{6.5}$$

Noi allora procederemo come per l'esempio precedente, considerando però un esempio di s ogni caso in cui CD risulta maggiore di 0,75, assegneremo a AD e BC dei valori casuali compresi tra 0 e 1, ripeteremo 100.000 volte questa assegnazione, e calcoleremo i valori di π ad ogni interazione.

I risultati sono presentati nella tabella 6.2, posti a confronto, nell'ultima colonna, con quelli ottenuti con la simulazione precedente. Lasciamo come esercizio per il lettore la costruzione del semplice programma in *basic*. I dati, come si può vedere, sono ben peggiori di quelli di Lazzerini (1901). Peraltro, la strabiliante accuratezza del nostro campione ha fatto sorgere diversi dubbi negli studiosi (cfr. Badger, 1994).

Iterazioni	Ago di Buffon	Spicchio
1000	3,23102	3,15200
10000	3,12207	3,12720
19000	3,13661	3,13410
28000	3,13515	3,13528
37000	3,13905	3,13297
46000	3,14788	3,13870
55000	3,14402	3,13455
64000	3,15240	3,13456
73000	3,14242	3,13468
82000	3,14834	3,13698
91000	3,14525	3,13640
100000	3,14688	3,1377

Tabella 6.2 – *Valori di* π *ottenuti nelle due simulazioni*

Ma, si badi, ben prima che nascesse il metodo Monte Carlo, di cui stiamo per occuparci, non mancarono scienziati di grande valore che capirono come, attraverso la ripetizione intensiva anche simulata di procedimenti riconducibili al calcolo numerico si poteva giungere alla soluzione di problemi matematici anche complessi. È il caso, nel 1899, di Lord Raleigh, che individuò in queste tecniche il modo di risolvere equazioni differenziali paraboliche; e di Kolmogorov, che nel 1931 mostrò che attraverso i processi stocastici si poteva giungere alla soluzione di certe equazioni integro-differenziali. Ma un significato analogo ebbero i campioni studiati all'inizio del '900 da Gosset, lo statistico più noto con lo pseudonimo di Student (1908a, 1908b). Questi, per studiare le caratteristiche delle distribuzioni, ricorse all'estrazione a caso delle lunghezze delle dita di 3000 criminali, non per studiare queste caratteristiche psicometriche, ma servendosi delle loro mani come di generatori di numeri casuali.

Il metodo Monte Carlo è nato alla fine della II Guerra Mondiale. Il nome, di grande successo, gli venne dato dal fisico e matematico americano di origine greca Nicholas Constantine Metropolis (1915-1999), che ne fu il fondamentale creatore, e a cui si deve anche una vivace ricostruzione delle origini del metodo (cfr. Metropolis, 1987). La storia del nome, che evoca la capitale del Principato di Monaco, e, intenzionalmente, il gioco d'azzardo, è curiosa. Stanislaw Marcin Ulam (oggi forse ricordato per una delle sue scoperte meno significative, la cosiddetta "spirale di Ulam" che consente di individuare nelle serie numeriche i numeri primi), fisico e matematico di origine polacca, era appassionato ai giochi di carte, e quando le prime idee su quello

che sarebbe stato il metodo cominciarono a germogliare, pensò che avrebbero potuto essere applicate alla soluzione di solitari (cfr. Eckhardt, 1987). Poiché le riunioni si tenevano a casa di Ulam, questa venne chiamata da Metropolis Monte Carlo, e il nome passò poi al metodo.

Metropolis, come Ulam, fu uno degli scienziati reclutati nel 1943 da Robert J. Oppenheimer al Los Alamos National Laboratory per il progetto Manhattan, da cui sarebbe poi nata la bomba atomica. Qui iniziò una stretta collaborazione con Enrico Fermi, a cui volle attribuire poi il merito dell'intuizione primitiva da cui sarebbe nato il metodo Monte Carlo. Un altro straordinario studioso che ebbe un ruolo determinante nello sviluppo del metodo Monte Carlo assieme a Metropolis e Ulam fu John von Neumann, le cui intuizioni e i cui suggerimenti costituirono uno stimolo costante per Metropolis e il gruppo di studiosi raccolto attorno a lui. Finita la guerra, Metropolis tornò a Chicago, dove aveva studiato, per poi riprendere il suo posto a Los Alamos nel 1948, diventando il capo di quella che sarebbe stata la famosa Divisione T, che avrebbe realizzato nel 1952 un tipo completamente diverso di computer digitale, il MANIAC, rispetto all'ormai sorpassato ENIAC (il primo calcolatore elettronico mai realizzato), su cui era stata realizzata la prima simulazione Monte Carlo nel 1948. Il nome, acronimo di "mathematical and numerical integrator and computer", fu coniato scherzosamente da Metropolis stesso, stanco dell'inflazione di acronimi che stava invadendo la nuova scienza dei calcolatori (e che procede tuttora in modo diluviale).

Nel 1953, assieme a A.W. e M.N. Rosenbluth e a A.H. e E. Teller pubblicò quello che sarebbe rimasto come il famoso "algoritmo di Metropolis", in un articolo tuttora citatissimo, in cui si mostrava nella sua pienezza la potenza del metodo Monte Carlo per lo sviluppo della fisica. L'algoritmo consentiva di trovare, per mezzo di integrali multidimensionali, le media di grandezze fisiche date con la formula di Gibbs della meccanica statistica. Con tale metodo, si potevano così risolvere integrali di dimensione superiore a dieci, che avevano sino ad allora posto problemi di calcolo ritenuti insolubili.

Nel 1957 Metropolis tornò a Chicago, dove fondò lo Institute for Computer Science, per poi tornare definitivamente a Los Alamos nel 1965. I suoi contributi allo sviluppo del calcolo elettronico hanno costituito una delle pagine più significative dello sviluppo della scienza nel XX secolo. Ma il suo nome rimane indissolubilmente legato al metodo Monte Carlo, che ha indubbiamente costituito il più poderoso progresso avutosi da sempre nel calcolo.

6.1.2. I numeri casuali

Il metodo Monte Carlo si articola in diverse tecniche, a seconda dei problemi specifici a cui va applicato, aventi tutti alla base la generazione di numeri casuali, e la loro elaborazione sulla base di vincoli dati dalla forma delle funzioni e dei parametri di controllo di queste, che regolano i fenomeni che si devono studiare, e che vengono così simulati al calcolatore. Poniamo che io mi ponga il problema di come evolve il

rapporto tra predatori e prede in un certo ambito naturale — è tra l'altro questo uno dei più classici problemi di dinamica non lineare, che trovò la sua prima soluzione negli anni '20 con il famoso modello di Lotka (1925) e Volterra (1926) — per una discussione del problema nel suo significato odierno, vedi Bertuglia e Vaio (2003, p. 199 e sgg.). Si assume che, se i predatori si nutrono esclusivamente di una certa specie di prede, con un numero basso di predatori il numero delle prede aumenti; ma ciò, a sua volta, dovrebbe comportare un aumento del numero dei predatori, che trovano più cibo a disposizione. L'aumento del numero dei predatori comporterà però una diminuzione del numero delle prede, che a sua volta si rifletterà in una diminuzione del numero dei predatori. Ciò comporterà un nuovo aumento del numero delle prede, e così via, in un ciclo che si rinnova continuamente.

Il modello può essere così espresso:

$$\begin{cases} \dfrac{\mathrm{d}n_1}{\mathrm{d}t} = an_1 - bn_1n_2 \\ \dfrac{\mathrm{d}n_2}{\mathrm{d}t} = cn_1n_2 - dn_2 \end{cases}. \tag{6.6}$$

dove n_1 è il numero dei predatori, n_2 il numero delle prede, a il tasso di natalità delle prede, b il tasso di mortalità delle prede per mano dei predatori, c il tasso di natalità dei predatori per predazione, e d il loro tasso di mortalità. In altri termini, la variazione del numero dei predatori e del numero delle prede in funzione del tempo è data da queste due equazioni, regolate dai quattro parametri *a, b, c* e *d*.

Quale può essere l'utilizzo del metodo Monte Carlo in questo caso? Nella normale ricerca scientifica, il modello viene verificato osservando il numero dei predatori e il numero delle prede a scadenze regolari di tempo in un determinato habitat, e rilevando se i dati così raccolti si adattano al modello espresso nel (6.2). Evidentemente, i difetti che questo modo di procedere presenta sono molteplici. Una ricerca di questo genere si basa necessariamente su una quantità di dati limitata. Il modello (come tutti o quasi i modelli) costituisce una semplificazione della realtà, che non tiene conto di altre variabili interagenti non controllate, come ad esempio l'arrivo di nuove specie di predatori o di prede in quel determinato habitat, la modificazione ambientale che può prodursi per eventi naturali o per inquinamento, l'intervento di epizoozie che alterano le popolazioni studiate, e così via. E gli esempi potrebbero continuare. La simulazione consente di superare molti di questi inconvenienti. Il calcolatore ci consente di generare in misura pressoché illimitata numeri casuali, e sostituendo alle osservazioni reali sul numero di predatori e prede i numeri casuali noi possiamo così "simulare" quel che accadrebbe in un numero grande a volontà di ambienti diversi e popolazioni diverse, e verificare come si comporta il modello, a cui possiamo apportare tutte le modifiche che vogliamo, confrontandolo poi con quanto realmente osservato.

I numeri casuali sono il cuore della simulazione, ed è quindi opportuno spendere qualche parola per definirli. Noi diciamo che all'interno di certi limiti (ad esempio, tra 0 e 1000) sono casuali i numeri che vengono generati in base a una funzione tale

che rende equiprobabile la generazione di ciascun numero. Detto così, la definizione sembra semplice e di facile applicabilità, ma in realtà le cose sono tanto più complesse che nella pratica si ricorre quasi sempre alla generazione di numeri pseudo-casuali: numeri cioè generati in base a qualche funzione tale per cui non viene raggiunta l'equiprobabilità, ma qualcosa che molto le assomiglia, tanto da far passare alla serie generata una serie di test di casualità. Ma vi è anche un altro motivo per preferire i numeri pseudocasuali ai numeri casuali: la generazione dei numeri pseudo-casuali si effettua utilizzando una funzione deterministica; in altri termini, dati gli stessi valori di partenza (detti in genere *seeds*, letteralmente *semi*), verranno generati sempre gli stessi numeri in sequenza. Ciò consente il controllo e la replicabilità dei dati che si ottengono, che non potrebbe ottenersi con numeri genuinamente casuali, che condurrebbero sempre a sequenze differenti. Ma ciò comporta anche una conseguenza meno gradevole: ogni funzione deterministica (o quasi: l'affermazione non va presa alla lettera) dopo un certo numero di applicazioni iterate (la cosiddetta *lunghezza del ciclo*) tornerà al punto di partenza, e comincerà a generare sequenze già generate in precedenza. È quindi opportuno che le sequenze siano sufficientemente lunghe da evitare ripetizioni.

I generatori di numeri casuali di più comune impiego sono i Generatori Congruenziali Lineari (GCL). Essi funzionano in base alla seguente funzione ricorsiva

$$x_{i+1} + (ax_i + c) \bmod m, \qquad i = 1, 2, 3 \ldots n. \tag{6.7}$$

Si parte da un *seed* x_0, e la funzione viene iterata n volte. I parametri a, c e m sono detti rispettivamente *costante moltiplicativa*, *incremento* e *modulo*, e vengono fissati dal ricercatore. La scelta di questi parametri, e del *seed*, è di estrema importanza, perché da essi dipendono le proprietà statistiche della serie generata, e soprattutto la lunghezza del ciclo. Il lettore che sia interessato a costruirsi propri generatori di numeri casuali (i GCL ne costituiscono solo una della numerose famiglie) può far riferimento a una bibliografia oggi vastissima, e può utilmente iniziare con Knuth (1981).

Una volta generata la sequenza, occorre però che essa sia effettivamente utilizzabile come se si trattasse di numeri effettivamente casuali. La prima preoccupazione, l'abbiamo detto, è quella di evitare di generare cicli troppo brevi. La seconda è relativa alla forma della distribuzione.

Di massima, il generatore sviluppa delle sequenze distribuite uniformemente. In altri termini, la probabilità che i numeri vengano generati in punti diversi della serie sarà uguale per tutti i punti della serie. Potrà però servire avere serie che abbiano distribuzioni diverse della uniforme. Non stupirà sapere che le distribuzioni più usate siano la normale e la poissoniana, ma anche altre (per esempio, quella di Weibull, o, in econometria, quella di Lorenz) sono di frequente utilizzo. Queste distribuzioni si possono ottenere o direttamente con generatori appositi, o applicando algoritmi specifici alla serie uniforme generata.

Occorrerà a questo punto fare dei test statistici sulla serie per determinare se i numeri pseudo-casuali rispondono alle nostre esigenze. In altri termini, se sono

equiprobabili relativamente alla loro distribuzione. I cinque test statistici che vengono più frequentemente usati sono riepilogati nella seguente tabella 6.3.

Test	Uso
χ^2	Suddivisa la sequenza in una serie di intervalli ugualmente intervallati, confronta la frequenza in ogni intervallo con la frequenza attesa.
Kolmogorov-Smirnov	Confronta la distribuzione teorica dei numeri generati (uniforme, normale, ecc.) con la distribuzione osservata.
Gaps	Consente di rilevare la probabilità del ripresentarsi degli stessi intervalli (*gaps*) tra le presentazioni dello stesso numero.
Runs	Si esamina la presenza di sequenze regolari (*runs*) nella successione: numeri pari o dispari, numeri in serie ascendente o discendente, ecc. Non sono accettabili sequenze con troppi o troppo pochi *runs*.
Autocorrelazione	Si vede se esiste una correlazione tra numeri divisi da un intervallo di lunghezza *m* (detto *lag*). Permette di individuare regolarità non immediatamente visibili tra numeri non contigui.

Tabella 6.3 – *I test di randomness di uso più frequente*

Dando per scontata la conoscenza del χ^2, e rinviando a un manuale di statistica per *gaps, runs* e *autocorrelazione* (che, ricordiamo, è di largo impiego nello studio delle serie temporali), spendiamo qualche parola sul test di Kolmogorov-Smirnov, anche perché ne daremo una dimostrazione pratica dell'uso nell'esempio 6.3.

L'effettuazione del test di Kolmogorov-Smirnov è di estrema semplicità. Innanzi tutto si ipotizza che i dati osservati, nel nostro caso la serie dei numeri simulati, appartengano a una qualche distribuzione di cui sia nota la legge di probabilità. La maggior parte dei programmi di statistica oggi generano numeri casuali, e si chiede all'utente di specificare in base a quale distribuzione si vuole che siano generati. Per esempio, si assume che essi siano distribuiti in modo uniforme. Si costruisce quindi la distribuzione cumulativa di tali numeri: in altri termini, si vede qual è la frequenza del numero che ha valore più basso, e questa frequenza costituirà il primo valore cumulato. Si vede poi qual è la frequenza del numero che ha valore immediatamente superiore, e questa frequenza verrà sommata alla precedente. La somma costituirà il secondo valore. E così via, sino al valore più alto, sommando sempre la frequenza trovata a quella dei valori precedenti.

Questa distribuzione cumulata verrà allora confrontata con la distribuzione *teorica* cumulata, le due curve verranno confrontate, e valore per valore si metteranno a confronto i valori della curva osservata e della curva teorica. Si individuerà qual è, in valore assoluto, la massima distanza tra le due curve, D, e il valore della statistica Z sarà dato da

$$Z = \sqrt{n} \cdot D, \tag{6.8}$$

dove n è il numero dei valori di cui si sono calcolate le frequenze. Per $n \geq 30$ (e cioè, per la quasi totalità dei casi che ci interessano — per gli altri occorre rintracciare le tabelle, peraltro non molto comuni), una buona approssimazione del valore di probabilità è data da

$$p = 1 - \frac{1}{e^{\left(2D^2 + \frac{2D}{3\sqrt{n}} + \frac{1}{18n}\right)}}. \tag{6.9}$$

ESEMPIO 6.3

Negli esempi sopra riportati, abbiamo fatto ricorso alla generazione di numeri casuali per simulare i lanci rispettivamente di uova e dell'ago. In realtà, si trattava di numeri pseudocasuali, generati da un GCL che abbiamo costruito a questo scopo. Pur essendo orgogliosi del nostro generatore, è però opportuno procedere all'effettuazione di test di *randomness*, per vedere se si trattava effettivamente di numeri accettabilmente casuali. Come abbiamo visto, i test di *randomness* sono numerosi, ma ci limitiamo qui per brevità a quello che è forse il più usato test, quello di Kolmogorov-Smirnov, le cui basi razionali abbiamo esposto sopra.

Per effettuare il test, preliminarmente abbiamo generato 100.000 numeri pseudo-casuali con il nostro GCL. Si trattava di numeri compresi tra 0,000001 e 1. Abbiamo quindi suddiviso i nostri numeri tra 40 *bins* (urne: vengono così detti gli intervalli spaziati tra cui si suddividono i numeri generati). Così, nel primo bin abbiamo posto i numeri più piccoli di 0,025, nel secondo quelli compresi tra 0,025 e 0,050, e così via, fino al 40° bin, che conteneva i numeri compresi tra 0,975 e 1. I numeri generati secondo questa procedura sono nella tabella 6.4.

La nostra ipotesi nulla assume che i dati siano stati generati secondo una distribuzione uniforme. In altri termini, non dovrebbero esistere differenze significative nelle frequenze dei diversi *bins*. Ora, il test principe per verificare un effetto di questo tipo è il χ^2.

Noi peraltro sappiamo benissimo che, come abbiamo già rilevato nel 2° Cap., questo test è molto sensibile alla frequenza del campione che viene esaminato, e in questo caso la frequenza totale è uguale a 100.000. Con un valore di N così elevato sarebbe pressoché impossibile trovare un χ^2 non significativo. E si badi: qui noi vogliamo accettare l'ipotesi nulla, non rifiutarla!

Ricorriamo allora al test di Kolmogorov-Smirnov. Rileviamo preliminarmente che la gamma delle frequenze nei diversi *bins* è compresa tra 2385 e 2609. La differenza massima positiva è +0,115 e negativa −0,142. A questo corrisponde uno Z di 0,898, e la probabilità associata (bidirezionale) è $p = 0{,}396$.

Bin	Freq.	Bin	Freq.	Bin	Freq.	Bin	Freq.
1	2489	11	2531	21	2526	31	2499
2	2476	12	2567	22	2551	32	2469
3	2510	13	2428	23	2539	33	2544
4	2572	14	2517	24	2508	34	2484
5	2485	15	2467	25	2465	35	2518
6	2463	16	2485	26	2431	36	2521
7	2445	17	2490	27	2454	37	2562
8	2494	18	2456	28	2420	38	2431
9	2609	19	2537	29	2605	39	2519
10	2385	20	2557	30	2550	40	2441

Tabella 6.4 – *Frequenze dei numeri generati nei 40 bins*

Ora, un valore per noi accettabile di probabilità è compreso tra 0,10 e 0,90. Pertanto, possiamo accettare tranquillamente l'ipotesi che il nostro artigianale GLC sia un buon generatore di numeri pseudo-casuali. Ovviamente, per dare un giudizio definitivo dovrebbero essere compiuti anche altri test (*runs, gap*, e così via), ma a livello illustrativo l'esempio può ritenersi sufficiente.

Un'analisi un po' più raffinata di quella che abbiamo condotto potrebbe anche essere questa. Per ogni *bin*, noi possiamo calcolare il valore di χ^2 relativo. Al di là del loro valore, noi avremmo così un campione di 40 χ^2. Possiamo assumere che questi χ^2 si distribuiscano a loro volta secondo la distribuzione χ^2. Allora potremmo effettuare un test di Kolmogorov-Smirnov confrontando la distribuzione dei χ^2 che abbiamo calcolato empiricamente con la distribuzione teorica del χ^2.

6.1.3. Tipi di simulazione Monte Carlo

Ovviamente non è qui possibile dar conto di tutti i problemi che sono dietro al metodo Monte Carlo. Tra l'altro, il più grande numero di applicazioni si è avuto in settori che sono abbastanza lontani dai nostri interessi, dalla fisica all'economia. Ci limiteremo quindi a qualche indicazione puramente sommaria delle principali linee verso cui si indirizzano teoria ed applicazione di questi metodi, rimandando il lettore interessato alla sterminata bibliografia che può trovare in proposito, a partire da quello che è considerato la "Bibbia" del Monte Carlo, e cioè il Fishman (1997).

I problemi che si possono affrontare con il metodo Monte Carlo appartengono soprattutto a due categorie: problemi *probabilistici* e problemi *deterministici*. Il nome stesso è sufficientemente esplicativo per indicare di cosa si tratta. I primi, che sono quelli che in questa sede forse ci interessano più direttamente, consistono nel generare serie di

numeri casuali secondo regole (funzioni di densità di probabilità) che riteniamo rispecchino l'andamento del problema che stiamo studiando, osservare il comportamento di questi numeri, e concludere che la soluzione del problema consiste nel loro comportamento. In questo caso parliamo di *simulazione Monte Carlo* in senso proprio.

Se il problema che abbiamo di fronte è invece deterministico, noi ci troviamo di fronte a un sistema il cui stato e il cui comportamento sono già definiti. Il nostro interesse non è allora quello di trovare una soluzione, che già possediamo, ma di poter moltiplicare artificialmente il numero di osservazioni, per risolvere ad esempio complessi problemi di calcolo, o per trovarci di fronte a dati che in natura sarebbe eccessivamente arduo, se non impossibile, rilevare. Di fatto, ed è il caso ad esempio dei problemi della meccanica statistica, è difficile che ci si trovi di fronte a problemi interamente deterministici. Più spesso, alcuni parametri vengono trasformati in variabili aleatorie, e la soluzione che si cerca è allora di tipo probabilistico.

Le componenti fondamentali di un algoritmo Monte Carlo sono le seguenti:

1. La/le *funzione/i di densità di probabilità* che descrivono il sistema. Le più comuni sono l'uniforme, la normale (comprese l'uniforme normale e la log-normale), la poissoniana, la gamma, la Weibull, e così via; ma il loro numero è pressoché illimitato.
2. Il *generatore di numeri casuali*: anche qui ve ne sono di molti tipi, il più comune è il congruenziale lineare (GCL) che già abbiamo descritto.
3. La *regola di campionamento,* e cioè una regola per campionare numeri casuali (disponibili nelle unità in cui sono suddivisi) dalla funzione di densità di probabilità scelta.
4. Il *binning*, e cioè la collocazione degli output all'interno di intervalli a distanza prefissata (*bins*). In luogo del *binning*, si possono assegnare dei *punteggi*, o stabilire delle *categorie* di *output*.
6. La *riduzione della varianza*: una serie di metodi che consentono di stimare la varianza in funzione soprattutto del numero delle iterazioni, e di ridurla per poter ridurre il tempo di simulazione.

Il nostro interesse, per quel che riguarda l'uso della simulazione nelle tecniche di *resampling*, oggetto di questo capitolo, si appunta comunque su un obiettivo molto limitato. A noi interesserà generare numeri casuali da associare ai valori dei campioni che osserviamo, in modo da poter costruire nuovi campioni i cui componenti siano scelti casualmente dai campioni originali. Come questo potrà essere fatto, e quali potranno essere i test statistici da utilizzare per determinare la reale casualità (*randomness*) dei numeri generati, è quanto vedremo nei prossimi paragrafi.

6.2. IL RICAMPIONAMENTO

Abbiamo già accennato al problema del ricampionamento nel secondo capitolo. È però venuto il momento di affrontare il problema con una maggiore ampiezza, anche perché si tratta di una sostanziale alternativa al tradizionale modo di verifica delle ipotesi.

Per ricampionamento (*resampling*) sostanzialmente s'intende un modo di operare sul campione già estratto, che, attraverso operazioni di duplicazione o sostituzione dei valori del campione a meno di uno o più elementi alla volta, consente di costruire nuovi campioni da quello preesistente, su cui fare nuove stime dei parametri. Poniamo che, ad esempio, il parametro d'interesse sia la media: il ricercatore troverà così non più una media (quella del campione originale), ma una popolazione di medie, una per ogni campione generato, e verificherà dove la media originale si colloca all'interno di questa popolazione. In altri termini, ora la variabile casuale non è più il valore della popolazione osservata (la statura, il quoziente intellettivo, il locus of control esterno, o quel che si vuole), ma il parametro (la media, la varianza, la statistica utilizzata).

Le tecniche di ricampionamento furono così chiamate da Julian Simon (1969), che creò nel 1966, a fini soprattutto didattici, un metodo che sarebbe poi stato riscoperto indipendentemente da Bradley Efron (1979), che gli diede il nome, diventato poi popolarissimo, di "*bootstrap*" (stringhe degli scarponi).

In realtà, le tecniche di ricampionamento risalgono a diversi decenni prima, e come in molti altri settori della statistica del Novecento, possono essere fatte risalire, come senso generale, a Fisher (1935; che peraltro non ne fu poi uno sponsor convinto). Fisher ideò i cosiddetti test esatti. Vediamone il senso. Prima di procedere, rileviamo che non è qui per ovvi motivi possibile presentare passo passo i calcoli eseguiti, trattandosi di procedure che richiedono calcoli iterati anche per migliaia di volte. Ci limitiamo a dire che tutte le simulazioni sono state condotte con R 1.9.1. Peraltro, i migliori *packages* di statistica hanno o sottoprogrammi specifici per il *resampling*, ed è il caso di R (che di fatto è la versione *open* di S-PLUS, e perciò particolarmente raccomandabile), di SAS e di Systat; o, come MacAnova, consentono con estrema facilità di operare il *resampling*.

6.3. Testi di randomizzazione

6.3.1. Test della probabilità esatta

Tutti conoscono il cosiddetto test della probabilità esatta di Fisher per tabelle di contingenza, che viene abitualmente calcolato ricorrendo al coefficiente ipergeometrico (cfr. Luccio, 1996). In realtà, il ragionamento che è alla base di questo test, che parte dall'analisi di un solo campione, può essere così schematizzato.

Supponiamo di avere un campione $\mathbf{x}$: $\{x_1, x_2, \ldots, x_n\}$, che sia costituito da valori che possono essere collocati a livello di scala d'intervallo o di rapporto. Il problema che poniamo è quello dell'esistenza o meno di un numero di valori superiori alla media maggiore (rispettivamente minore) del numero di valori inferiori alla media. A noi non interessa conoscere l'entità dello scarto dalla media, interessa solo il numero (la frequenza) dei valori superiori o rispettivamente inferiori. Oppure supponiamo

d'avere misure ripetute sugli stessi soggetti dopo un certo trattamento, e vediamo se tra prima e dopo il trattamento si hanno scarti positivi o negativi. È la situazione che nei manuali di statistica viene usualmente presentata come *test dei segni*. In realtà, non è altro che una possibile estensione dell'uso del cosiddetto *test della binomiale*.

In questo senso, ci troviamo di fronte a un universo bernoulliano, costituito da eventi mutuamente escludentesi. Degli n eventi, k saranno positivi e $n - k$ negativi. Ci è ben noto che sotto ipotesi nulla, la probabilità di avere un evento positivo $p(\mathrm{P})$ è uguale alla probabilità di avere un evento negativo, $p(\mathrm{N})$, e, trattandosi di un universo bernoulliano, sono entrambe uguali a 0,5. Ora, come sa ogni studente, la probabilità associata al presentarsi di k eventi positivi e $n - k$ negativi sarà data da

$$\binom{n}{k} k^{0,5} (n-k)^{0,5}. \tag{6.10}$$

Lo studente comprende abbastanza facilmente il senso dell'espansione binomiale qui presentata, mentre gli è poi però difficile capire quanto gli viene detto in seguito: a questa probabilità "istantanea" bisogna poi aggiungere la "coda" data dalla somma delle probabilità degli eventi meno favorevoli.

Secondo Fisher (1935), ciò che il ricercatore deve semplicemente fare è calcolare *tutti* i valori di probabilità associati a tutti i possibili esiti di un esperimento che abbia come risultato una serie di valori positivi contrapposta a una serie di valori negativi su un unico campione. Noi poi possiamo ordinare queste probabilità in ordine di attribuzione ad esiti progressivamente sempre più lontani dall'ipotesi nulla (solo in una direzione, per ipotesi monodirezionali; verso destra *e* verso sinistra, per ipotesi bidirezionali). Vedremo poi in che punto si colloca la probabilità associata all'evento che abbiamo effettivamente osservato. Questo punto costituisce una sorta di asse di decisione: lo spazio delle probabilità di tutti gli esiti realizzabili con n eventi viene diviso in due: da una parte la probabilità osservata *più* tutte le probabilità associate a possibili esiti ancora più sfavorevoli per l'ipotesi nulla (somma di probabilità che va raddoppiata nel caso di ipotesi bidirezionali), dall'altra il resto delle probabilità che sommate a questa devono dare 1.

ESEMPIO 6.4

Un esempio chiarirà meglio quanto detto. Supponiamo di avere sottoposto 10 persone a una dieta dimagrante. La tabella 6.5 ne presenta i risultati in chilogrammi. Nell'ultima colonna è presentato il segno dello scarto tra prima e dopo la dieta. Ovviamente, la nostra ipotesi nulla prevede l'inefficacia della dieta.

Qui, n è uguale a 10 e k è uguale a 8. Noi, però, per eseguire un test *esatto* dobbiamo calcolare tutte le probabilità corrispondenti a tutti i valori possibili di k, e vedere dove si colloca la probabilità corrispondente al k osservato, appunto 8. Questi valori sono presentati nella tabella 6.5 e nella figura 6.6.

Soggetti	Prima	Dopo	Segno
1	75	72	+
2	88	80	+
3	79	81	-
4	68	62	+
5	74	69	+
6	76	73	+
7	87	88	-
8	65	60	+
9	92	78	+
10	99	89	+

Tabella 6.5 – *Risultati di una dieta*

k	Probabilità
0	$\binom{10}{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{10} = 0,0001$
1	$\binom{10}{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{9} = 0,0098$
2	$\binom{10}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{8} = 0,0439$
3	$\binom{10}{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{7} = 0,1172$
4	$\binom{10}{6}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{6} = 0,2051$
5	$\binom{10}{5}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{5} = 0,2461$
6	$\binom{10}{6}\left(\frac{1}{2}\right)^{6}\left(\frac{1}{2}\right)^{4} = 0,2051$
7	$\binom{10}{7}\left(\frac{1}{2}\right)^{7}\left(\frac{1}{2}\right)^{3} = 0,1172$
8	$\binom{10}{8}\left(\frac{1}{2}\right)^{8}\left(\frac{1}{2}\right)^{2} = 0,0439$
9	$\binom{10}{9}\left(\frac{1}{2}\right)^{9}\left(\frac{1}{2}\right)^{1} = 0,0098$
10	$\binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{0} = 0,0001$

Tabella 6.6 – *Probabilità corrispondenti a tutti i possibili valori di k*

Il valore osservato, con *k* uguale a 8, evidenziato in grassetto nella tabella, corrisponde a 0,0439, e sarebbe cioè inferiore al fatidico 0,05. Ma ad esso vanno sommati i valori relativi a k uguale a 9 e a 10, entrambi più sfavorevoli alla realtà dell'ipotesi nulla. Pertanto, il valore complessivo di probabilità associato all'evento osservato è 0,0538, non significativo. Una rappresentazione chiara di quanto detto si ha nel grafico della figura 6.4.

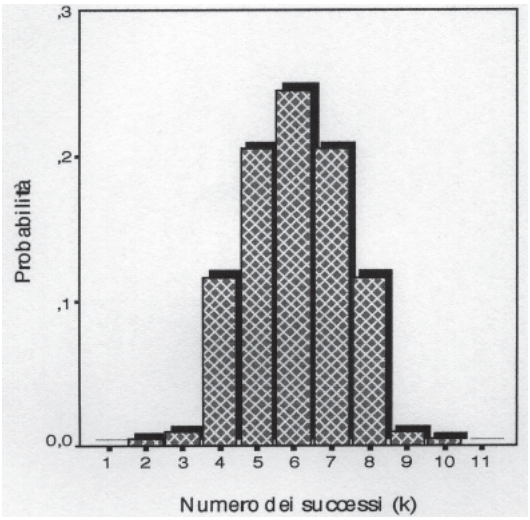

Figura 6.4 – *Probabilità in funzione del numero dei successi*

Qui si vede con chiarezza come si distribuiscono le probabilità. All'estrema destra, abbiamo la probabilità osservata (*k* uguale a 8) e al di là le probabilità meno favorevoli. La loro somma sarà, sull'ammontare complessivo delle probabilità (e cioè, 1), la probabilità assegnata a questo esito, posto che H_0 sia vera.

Veniamo allora al test della probabilità esatta con cui abbiamo aperto il paragrafo, e per cui il ragionamento è assolutamente analogo. La situazione da cui noi partiamo è quella classica di una tabella di contingenza. Abbiamo due variabili ortogonali A e B, ciascuna su due livelli, con frequenze che possiamo disporre come nella tabella 6.7.

	B1	**B2**	**Totale**
A1	*a*	*b*	*a+b*
A2	*c*	*d*	*c+d*
Totale	*a+c*	*b+d*	*N*

Tabella 6.7 – *Tabella di contingenza 2 × 2*

Noi conosciamo le straordinarie proprietà simmetriche di una tale tabella, e sappiamo ad esempio che ci basta conoscere un totale di riga (per esempio, $a+b$), un totale di colonna (per esempio, $b+d$), la frequenza di una qualsiasi cella (per esempio, c) ed N, per conoscere l'intera tabella. Per questo motivo, il ragionamento che ora faremo, riferito a una singola cella, vale in realtà per l'intera tabella.

Come sappiamo, il problema che qui ci poniamo è quello dell'indipendenza tra le due variabili. In altre parole, noi vogliamo determinare se la probabilità condizionale di appartenere a un determinato livello di una variabile, posto che si appartenga a un determinato livello dell'altra variabile, sia o meno diversa dalla probabilità condizionale di appartenere allo stesso livello della prima variabile, posto che si appartenga all'altro livello dell'altra variabile. In altri termini, le nostre ipotesi saranno quelle della (6.11). (In realtà, trattandosi di un approccio tipicamente fisheriano, questo modo di esporre il problema in termini di ipotesi nulla e di ipotesi sostantiva non sarebbe corretto).

Come detto, quanto vale per l'incrocio tra B_1 e A vale poi per l'intera tabella. Come sempre il nostro problema è quello di determinare la probabilità associata ai dati a nostra disposizione, posto che H_0 sia vera. Noi possiamo qui determinare questa probabilità in modo "istantaneo", e cioè escludendo la "coda" dei possibili eventi più sfavorevoli per l'ipotesi nulla. Tale probabilità è data dal coefficiente ipergeometrico, che può essere così calcolato.

$$\begin{aligned} H_0 &: p(B_1|A_1) = p(B_1|A_2); \\ H_1 &: p(B_1|A_1) \neq p(B_1|A_2). \end{aligned} \tag{6.11}$$

La probabilità è qui calcolata in modo frequentistico. Ciò significa che il denominatore è dato da tutti i modi in cui, su N osservazioni, si possono avere $a+c$ eventi B_1: e cioè, dalle combinazioni di N elementi di classe $a+c$. Il numeratore poi deve tener conto dei modi in cui e gli elementi A_1 (che sono $a+b$) e A_2 (che sono $c+d$) possono essere anche B_1: e cioè, del prodotto delle combinazioni di $a+b$ elementi di classe a per le combinazioni di $c+d$ elementi di classe c:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}. \tag{6.12}$$

A questa probabilità "istantanea", associata a questo pattern di frequenze, va poi assommata la "coda", che viene ottenuta diminuendo un valore alla volta la frequenza della cella considerata con la frequenza più bassa, e aumentando le frequenze delle altre celle, in modo da mantenere inalterati i totali di riga e di colonna, sinché non si giungerà ad avere frequenza pari a 0 in una cella.

Nei manuali di statistica, il discorso in generale si ferma qui. In realtà, questo è solo un pezzo del ragionamento di Fisher, che potremmo articolare nel suo complesso

in questo modo. Facciamo una simulazione, e proviamo a vedere in quanti modi diversi potremmo articolare queste frequenze, salvando sempre i totali di riga e di colonna. Noi abbiamo visto, nei coefficienti ipergeometrici di cui sopra, l'istantaneo e la coda, solo una parte delle probabilità che potevamo calcolare, quella della coda di sinistra. Ma noi possiamo proseguire nel nostro calcolo verso destra, fino al centro della distribuzione (di fatto, per l'intera distribuzione). Il valore di probabilità da noi ottenuto occuperà allora una posizione ben precisa all'interno della distribuzione. L'esempio che segue chiarirà con valori numerici concreti quanto abbiamo esposto.

ESEMPIO 6.5

Immaginiamo di avere due variabili ortogonali S e P, ciascuna su due livelli (per esempio, sesso e professione), con frequenze che possiamo disporre in una tabella di contingenza (tabella 6.8).

Per il calcolo del test della probabilità esatta di Fisher ricorriamo, come abbiamo detto, alla probabilità ipergeometrica. A questo scopo, individuiamo una cella (di solito, quella a frequenza più bassa, in questo caso 2), e determiniamo la probabilità "istantanea" associata a questo pattern di frequenze, più la "coda" (ottenuta diminuendo un valore alla volta la frequenza della cella individuata, e aumentando le frequenze delle altre celle, in modo da mantenere inalterati i totali di riga e di colonna).

	P1	**P2**	**Totale**
S1	3	7	10
S2	8	2	10
Totale	11	9	20

Tabella 6.8 – *Tabella di contingenza 2 × 2*

Facciamo ora una simulazione, e proviamo a vedere in quanti modi diversi potremmo articolare queste frequenze, salvando sempre i totali di riga e di colonna. Noi abbiamo visto, nei tre coefficienti ipergeometrici di cui sopra, solo una parte delle probabilità che potevamo calcolare, quella della coda di sinistra. Ma noi possiamo proseguire nel nostro calcolo verso destra, fino al centro della distribuzione. I dati relativi sono presentati nella tabella 6.9 (i valori relativi al pattern osservato sono in grassetto):

$$\frac{\binom{10}{2}\binom{10}{7}}{\binom{20}{9}} + \frac{\binom{10}{1}\binom{10}{8}}{\binom{20}{9}} + \frac{\binom{10}{0}\binom{10}{9}}{\binom{20}{9}} = 0,0322 + 0,0027 + 0,0001 = 0,0350.$$

Frequenze cella S2P2	Probabilità istantanea	Probabilità cumulata
0	0,00006	0,00006
1	0,00268	0,00274
2	**0,03215**	**0,03489**
3	0,15004	0,18493
4	0,31507	0,50000

Tabella 6.9 – *Probabilità associate ai diversi pattern di frequenza*

Come si può vedere, il valore di probabilità cumulata corrispondente al pattern osservato (tutti gli altri valori non sono osservati, ma simulati; l'espressione "simulazione" peraltro si comincerà ad usare sistematicamente solo nel dopoguerra) è di 0,034. Se il limite α è stato posto a 0,05, possiamo senz'altro escludere che vi sia indipendenza tra le variabili A e B.

6.3.2. Confronto tra medie

Evidentemente questo caso è molto semplice, ma secondo Fisher si poteva estendere a situazioni ben più complesse. Poniamo di avere due campioni indipendenti, sottoposti a due diversi livelli di trattamento, che chiameremo S e C. Ricorrendone gli estremi (proprietà metriche della variabile dipendente) potremo effettuare un confronto tra medie, utilizzando come statistica il *t* di Student.

Ora, i nostri campioni sono sempre di dimensioni ridotte, e salvo effetti di notevole grandezza e potenza molto elevata del test statistico, anche un valore significativo del *t* ci lascia, come abbiamo detto nel corso di tutto il volume, molti dubbi sull'effettiva possibilità di trarre delle conclusioni decisive dai nostri dati. La proposta di Fisher (da lui poi di fatto non molto perseguita, per l'onerosità dei calcoli che richiedeva, ben superiori a quelli da fare con una tabella di contingenza) si ispirava direttamente a quanto abbiamo visto relativamente al test della probabilità esatta.

I nostri dubbi sono di due ordini: il primo è che ben poche garanzie abbiamo sulla normalità della distribuzione da cui sono stati tratti i campioni che confrontiamo, data la loro scarsa consistenza numerica; il secondo è che, in queste condizioni, è sempre verosimimile che le differenze siano attribuibili al caso.

Si tratta allora di rovesciare il ragionamento, e trattare come variabile casuale non già la popolazione da cui i campioni sono estratti, ma la statistica usata (nel nostro caso, il *t*). Noi ora abbiamo confrontato le medie dei due campioni, e abbiamo ottenuto un *t*.

Ma cosa avverrebbe se potessimo moltiplicare, servendoci degli stessi dati, che sono gli unici che abbiamo a disposizione, le coppie dei campioni da confrontare? Creeremmo una pluralità di *t*, e potremmo studiarne la distribuzione, e vedere dove si colloca all'interno di questa distribuzione proprio il *t* che abbiamo osservato all'inizio.

La soluzione proposta allora da Fisher è questa. Prendiamo il primo campione, e invertiamone un valore con un valore del secondo campione. Avremmo così due nuovi campioni, di cui potremo calcolare il *t*. Sostituiamo poi sistematicamente a questo primo valore tutti gli altri valori del secondo campione, uno alla volta, mettendo quello al loro posto, e ogni volta calcoliamo il *t* relativo. Facciamo poi la stessa cosa con il secondo valore, quindi con il terzo, e così via, fino all'ultimo. In questo modo, se i campioni hanno ognuno *n* soggetti, noi otterremo una popolazione di n^2 statistiche *t*. Ne potremmo studiare la distribuzione, e vedere al suo interno quali sono gli intervalli di fiducia che possiamo porre, e dove si colloca il nostro *t* originale.

Quello che abbiamo fatto invertendo due valori, uno per campione, possiamo ora farlo invertendo due valori per campione, ponendo cioè due valori del primo campione nel secondo campione, e prendendo due valori del secondo campione e ponendoli nel primo. Proseguiremo poi sino ad esaurire tutte le possibile coppie. A questo punto, avremo la possibilità di calcolare un altro buon numero di statistiche, ed esattamente le combinazioni di *n* elementi di classe 2 (tante sono le possibili coppie di un campione di *n* valori) al quadrato. Per renderci conto dell'onerosità del calcolo, per due campioni di soli 10 valori ciascuno si tratta già di 2025 coppie, che vanno aggiunte alle 100 ottenute con il passaggio precedente, per un totale provvisorio di 2125 statistiche.

Totale provvisorio, perché dobbiamo ora calcolare le triple. Qui ci troviamo di fronte ovviamente al quadrato delle combinazioni di *n* elementi di classe 3; nel caso sia il nostro solito *n* di 10, abbiamo altre 14400 statistiche, da sommare alle 2125 precedenti, per un totale ancora provvisorio di 16525.

Ma ovviamente non è finita. Le quadruple ci portano altre 44100 statistiche, che sommate alle precedenti ci dà un totale di 60625. E con le quintuple abbiamo ancora 63504 statistiche per un totale di 124129. Questo numero va poi raddoppiato, per tener conto di sestuple, etc, sino alla sostituzione totale dei due campioni. In totale, il test esatto richiederebbe il calcolo di 248258 statistiche *t*. Non deve sorprendere se questo modo di procedere non abbia suscitato entusiasmi, ed anche oggi, che la potenza di calcolo disponibile rende la cosa affrontabile, si preferiscono vie ben meno onerose.

Anche qui, un esempio renderà chiaro quanto stiamo dicendo.

ESEMPIO 6.6

Anche questo esempio è semplificato, perché ci fermeremo al primo passo, e cioè alla sostituzione di un solo elemento.

I dati originali sono presentati nella tabella 6.10. Come si vede, si hanno due campioni (S e C), ciascuno di 8 soggetti. Ora, le medie dei due campioni sono rispettivamente 5,36250 e 10,88587. Calcoliamo il *t* per campioni indipendenti, e otteniamo un valore di –2,2931 con

14 gradi libertà, che corrisponde a una probabilità a una coda di 0,02213. Con il tradizionale modo di interpretare i dati, questo valore sarebbe significativo, se α è stato posto a 0,05.

Gruppo S	Gruppo C
2.07	1,02
6.19	3,05
6.21	3,06
8.49	4,18
12.22	6,02
14.90	7,34
18.23	8,98
18.78	9,25

Tabella 6.10 – *Dati fittizi relativi a due campioni indipendenti*

Operiamo allora ordinatamente le sostituzioni richieste, e cioè invertiamo il 2,07 del gruppo S con lo 1,02 del gruppo C, e calcoliamo di nuovo il t; quindi con il 3,05, e di nuovo calcoliamo il *t*. E così via, sinché non abbiamo eseguito tutte e 64 le sostituzioni richieste, e non abbiamo i nostri nuovi 64 *t*.

Di questi *t* possiamo allora studiare la distribuzione. L'istogramma della figura 6.5 ce ne mostra la forma, approssimativamente a campana. Possiamo dire inoltre che la loro media è di –1,678, e la varianza di 0,607. La distribuzione appare ragionevolmente normale, e gli intervalli di fiducia (di destra, dato il valore di *t* ottenuto) sono di –3,325 (allo 0,025), e di –2,90 (allo 0,05). Contrariamente alle nostre aspettative, dunque, basate sulla semplice ispezione dei dati osservati, il valore rientra pienamente all'interno della distribuzione, e non ci autorizza a respingere l'ipotesi nulla. Evidentemente, una conclusione definitiva sarebbe possibile solo portando a termine l'onerosissimo compito di calcolo di tutte le statistiche in gioco (in questo caso, *solo* 17768!). Peraltro, si badi che con l'aumentare delle sostituzioni, i *t* che si ottengono dovrebbero essere sempre più bassi, e quindi è facile che quello da noi calcolato vada a finire nell'area di rifiuto.

L'eccessivo onere richiesto dai test esatti ha portato a riflettere seriamente sulla possibilità di sfruttare concretamente l'intuizione di Fisher (trattare la statistica usata come variabile casuale, e utilizzare i dati richiesti come universo), campionando peraltro i valori in modo da ridurre l'onere del calcolo.

Va ricordato che già negli anni '20 era stato il grande von Mises (1928; ma vedi anche von Mises, 1964) a lanciare questa idea. Si sarebbe però dovuto attendere il dopoguerra, e le straordinarie possibilità di calcolo offerte dai calcolatori elettronici,

per vedere l'idea marciare con le proprie gambe. In ogni caso, il problema del numero di campioni ha portato ai test non esatti di randomizzazione.

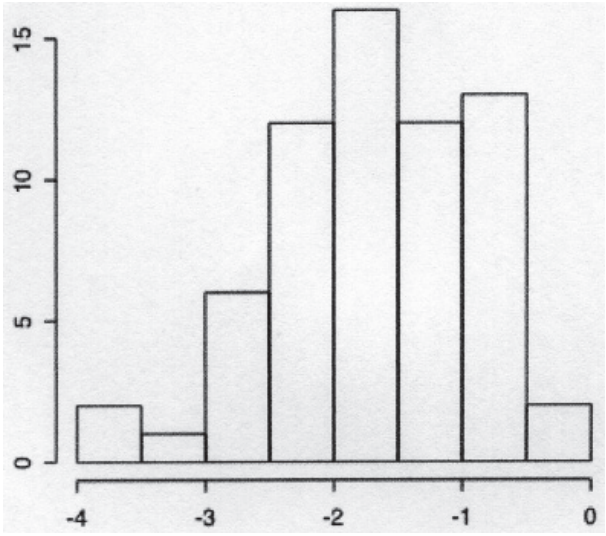

Figura 6.5 – *Distribuzione dei t simulati*

6.3.3. Test non esatti di randomizzazione

I test non esatti di randomizzazione, o test di randomizzazione *tout court*, sfruttano sì l'intuizione di Fisher, ma limitano il numero di campioni utilizzati per studiare la distribuzione della statistica usata. In linea di massima, si ritiene che siano sufficienti 1000 campioni (ovviamente presi a caso) se si vogliono porre gli intervalli di fiducia a livello 0,05, e almeno 5000 campioni se si vuole andare a livello 0,01 (Manly, 1997).

ESEMPIO 6.7

Per semplicità, limitiamo il nostro esempio ai dati che abbiamo già presentato nella tabella 6.10. Supponiamo peraltro di trovarci davanti a due campioni dipendenti. Ad esempio, potrebbe trattarsi dello stesso soggetto su cui vengono rilevati dei dati in condizioni basali (C), e successivamente dopo essere stato sottoposto a un certo trattamento (condizione S). Si ricorderà che in questo caso viene costruito il vettore delle differenze, soggetto per soggetto, tra S e C, e questo verrà trattato come se fosse un unico campione. Su questi dati, il relativo valore di *t* per dati appaiati è 4,957, corrispondente a una probabilità, a una coda, di 0,0009.

Il campione, peraltro, è molto piccolo, e non staremo quindi a ripetere quanto più volte detto a proposito delle cautele con cui questa significatività statistica va considerata. Vediamo allora di procedere a un test di randomizzazione. Per fare questo, costruiremo prima, per valutare la significatività a livello del 5%, 1000 coppie di campioni, in cui permuteremo l'ordine dei dati presenti nei gruppi S e C. Ogni volta permuteremo l'ordine nei due campioni, per cui gli accoppiamenti non saranno più tra agli stessi soggetti. Evidentemente non lavoreremo su tutte le possibili permutazioni, ma su una quota molto ridotta. Le permutazioni sono infatti 8!, e cioè 40320 per ogni campione, e il numero di coppie completo, che sarebbe richiesto da un test esatto, è di oltre un 1 miliardo e seicento milioni! Per valutare poi il livello dell'1% giungeremo a 5000 coppie.

Abbiamo ritenuto utile calcolare oltre al valore di *t* medio per le1000 e le 5000 iterazioni, gli intervalli di fiducia a livello 0,025 (test bidirezionale) e 0,05 (test monodirezionale) per le 1000 iterazioni, e 0,005 (test bidirezionale) e 0,01 (test monodirezionale) per le 5000 iterazioni. Presentiamo infine per le due simulazioni i coefficienti di variazione CV, dati dal rapporto tra deviazione standard e *t* medio.

I dati sono presentati nella tabella 6.11. Si tratta di dati indubbiamente soddisfacenti, perché da un lato mostrano per entrambe le simulazioni un coefficiente di variazione molto basso, il che dimostra una scarsa variabilità dei dati, e quindi una sostanziale omogeneità ed affidabilità del processo. Dall'altro mostrano che il *t* calcolato sui dati originali si pone decisamente fuori degli intervalli di fiducia così calcolati, e ciò dimostra attendibilmente l'esistenza dell'effetto del trattamento.

Iterazioni	***t* medio**	**CV**	**Intervalli di fiducia a una coda**	**Intervalli di fiducia a due code**
1000	2,33996	0,18038	$1{,}824 < t < 3{,}150$	$1{,}778 \leq t \leq 3{,}445$
5000	2,33874	0,18203	$1{,}754 < t < 3{,}768$	$1{,}731 \leq t \leq 4{,}006$

Tabella 6.11 – *Dati relativi al test di randomizzazione con 1000 e 5000 iterazioni*

6.4. La *cross-validation*

La prima realizzazione pratica di quanto abbiamo detto è rappresentata dalla *cross-validation*. L'idea della *cross validation*, sviluppata a partire dagli anni '40 sui disegni correlazionali, trova la sua origine nella teoria dei test, e in particolare negli studi sull'attendibilità (*reliability*). Ricordiamo che una delle caratteristiche dell'attendibilità di un test è data dalla proprietà di misurare la stessa caratteristica psicologica in ogni sua parte, e gli psicometristi hanno sviluppato diverse tecniche per misurare questa proprietà.

Una prima sistematizzazione dell'uso della *cross-validation* nelle ricerche correlazionali è stata data da Mosier (1951). Immaginiamo di condurre una ricerca in cui studiamo la possibile influenza della variabile indipendente su una variabile dipendente. Calcolando la correlazione tra le due variabili, otteniamo un valore di r che per un dato numero n di osservazioni può, per la gioia nostra e della redazione della rivista a cui manderemo il nostro lavoro, associarsi a un valore di p inferiore al famoso numero magico di 0,05. Peraltro, come abbiamo ben visto nel Cap. 2, il senso di questo risultato è indipendente dalla significatività statistica. Come interpretarlo?

La *cross-validation* ci aiuta in questo compito. Noi possiamo per esempio fare, in analogia con il cosiddetto *split-half* che si usa in psicometria, una partizione casuale del campione in due sottocampioni (nel caso della *cross-validation* semplice o doppia), o in più sottocampioni con ripetute operazioni indipendenti di campionamento (nel caso della *cross-validation* multipla).

A questo punto utilizziamo non più i coefficienti di correlazioni, ma le equazioni di regressione. Sappiamo infatti che, ammessa una relazione lineare, il rapporto tra la variabile indipendente x e il valore teorico, o predetto, della variabile dipendente y', deve essere

$$y' = a + bx, \tag{6.13}$$

dove, applicando il principio dei minimi quadrati, l'intercetta a è uguale a

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n}, \tag{6.14}$$

e il coefficiente angolare b è uguale a

$$b = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}. \tag{6.15}$$

(Vale la pena di ricordare che il coefficiente di correlazione r di Bravais Pearson non è altro che b, il coefficiente angolare di questa equazione di regressione, quando i valori delle variabili sono trasformati in punti z, e dove l'intercetta è uguale a 0).

Nella *cross-validation semplice* si procede quindi così. Si divide, come si è detto, il campione originale in due sottocampioni di uguale grandezza, con valori scelti a caso. Del primo sottocampione si calcolano i parametri della retta di regressione della (6.12). Attraverso questi, si calcolano quindi i valori predetti del secondo sottocampione, e viene qui calcolato il coefficiente di correlazione tra valori della variabile indipendente e valori predetti. Questo coefficiente prende il nome di *coefficiente di cross-validation*.

Esempio 6.8.1

Viene condotta una ricerca in una scuola media superiore su 20 studenti sui rapporti tra reddito familiare (variabile indipendente, RF), espresso in migliaia di euro l'anno, e rendimento scolastico (variabile dipendente, RS), espresso come media ponderata in test di profitto in italiano, latino, matematica, fisica e lingua straniera. I dati sono riportati nella tabella 6.12.

Ss	**RS**	**RF**	**Ss**	**RS**	**RF**
1	36	33	11	48,4	102
2	36,2	34	12	50	65
3	37,2	37	13	51,2	40
4	38,4	35	14	51,2	22
5	41	36	15	52	109
6	42,6	35	16	53,6	154
7	42,8	42	17	53,8	43
8	44	48	18	55	55
9	45,6	47	19	60	68
10	48	35	20	60	72

Tabella 6.12 – *Rendimento scolastico (RS) e reddito familiare (RF) di 20 studenti*

Se noi calcoliamo il coefficiente di correlazione r per questi dati, otteniamo un valore di 0,481, a cui corrisponde un t di 2,326, con $p = 0{,}032$, quindi significativo. Per fare una *cross-validation semplice*, dobbiamo allora estrarre 10 coppie di valori a caso, e ottenere così i due sottocampioni che ci servono. I dati relativi ai due sottocampioni sono ora nella tabella 6.12.

Troviamo la retta di regressione del primo campione, che è:

$$y' = 41{,}95 + 0{,}086x.$$

Otteniamo così i valori predetti di RS (colonna RS' della tabella 6.13). Calcoliamo ora il coefficiente di cross-validazione, che corrisponde al coefficiente di correlazione tra i valori osservati RS e i valori predetti RS' del secondo campione. Ora, abbiamo così un r di 0,502, che, seppur più alto dello 0,481 trovato sull'intero campione, non è significativo, corrispondendogli un t di solo 1,641, essendosi ridotto n a 10. Il risultato della *cross-validation semplice* ci induce così a concludere per uno scarso significato della correlazione trovata: anche se significativo, l'r è stato ottenuto su un campione disomogeneo, e presumibilmente i risultati non sono replicabili.

I sottocampione				II sottocampione			
Ss	**RS**	**RS'**	**RF**	**Ss**	**RS**	**RS'**	**RF**
1	37,2	45,85	37	1	41	45,99	47
2	38,4	46,02	35	2	45,6	47,54	65
3	42,6	46,63	35	3	50	45,39	40
4	44	46,84	48	4	51,2	44,87	34
5	48	47,42	35	5	36,2	44,79	33
6	48,4	47,47	102	6	36	45,56	42
7	51,2	47,88	22	7	42,8	45,65	43
8	52	48,00	109	8	53,8	46,68	55
9	53,6	48,23	154	9	55	47,80	68
10	60	49,16	72	10	60	45,05	36

Tabella 6.13 – *I due sottocampioni estratti dalla tabella 6.11*

Peraltro, la *cross-validation semplice* non appare un metodo molto affidabile, proprio perché le conclusioni finiscono con il riferirsi a un campione molto ridotto. È stata allora proposta la *cross-validation doppia*. Con questa vengono calcolate la retta di regressione e i valori predetti in entrambi i sottocampioni, evidentemente i valori predetti nel primo sottocampione mediante la retta di regressione ottenuta sul secondo, e viceversa per quest'ultimo. Viene quindi calcolato, come coefficiente di *cross-validation*, il coefficiente di correlazione tra valori osservati e valori predetti per l'intero campione.

ESEMPIO 6.8.2

Rimanendo sempre sugli stessi dati, la retta di regressione per il secondo sottocampione è la seguente:

$$y' = 40,46 + 0,145x.$$

In base a questa, calcoliamo i valori predetti per il primo sottocampione, che sono riportati nella terza colonna della tabella 6.13. Ora il coefficiente di *cross-validation* è di 0,484, a cui corrisponde un t di 2,345, con p = 0,031. È un risultato pressoché identico a quello ottenuto in origine, e dovrebbe essere considerato abbastanza rassicurante.

Un progresso sostanziale si ottiene però attraverso la *cross-validation multipla*. In questo caso, il metodo che raccomandiamo consiste nel suddividere sempre il campione in due sottocampioni casuali, e procedere come nel caso della *cross-validation* doppia. Il procedimento, però, verrà poi ripetuto un numero elevato di volte, e il coefficiente di *cross-validation* sarà la media dei coefficienti di correlazione ottenuti. Di più, come nel caso dei test esatti, potremo studiare la distribuzione dei coefficienti ottenuti.

Esempio 6.9

I dati sono sempre quelli degli esempi precedenti, e quindi la correlazione originale è, come abbiamo visto, di 0,481, corrispondente a un *p* significativo. Peraltro, la simulazione ci fa cambiare rapidamente idea sul significato di questa correlazione statisticamente significativa.

Noi abbiamo eseguito 1500 simulazioni, creando quindi 1500 coppie di sottocampioni, su cui abbiamo calcolato, come nel caso della cross-validation doppia, i coefficienti. Il coefficiente di cross-validation medio è risultato essere di 0,096 ≠ 0,05. La distribuzione è risultata soddisfacentemente normale, come può vedersi dall'istogramma della figura 6.6. Gli intervalli di fiducia di sinistra sono rispettivamente 0,753 e 0,679.

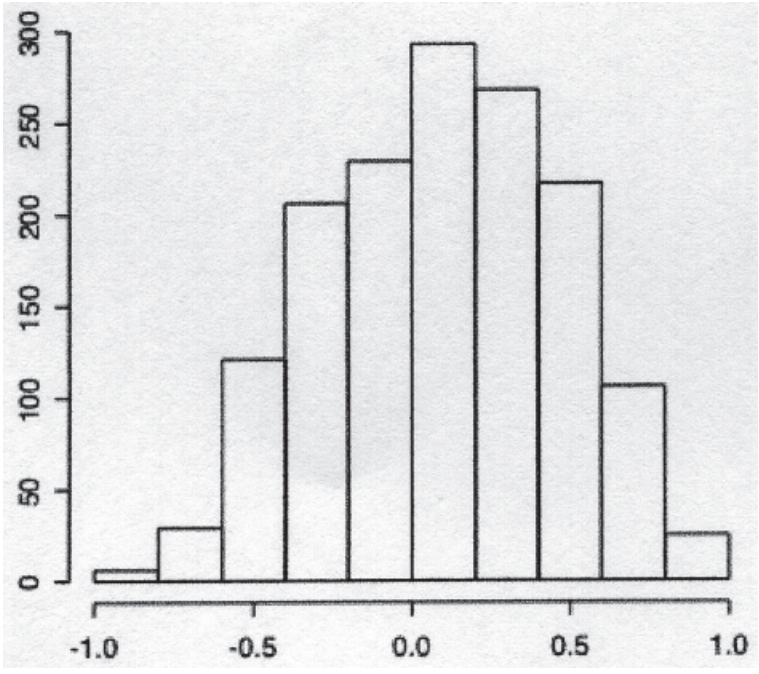

Figura 6.6 – *Istogramma della distribuzione dei coefficienti di cross-validation dopo 1500 simulazioni.*

Quali conclusioni trarre? L'omogeneità del campione è modesta, i risultati appaiono difficilmente replicabili, e la simulazione mostra che la correlazione trovata, seppur significativa statisticamente, ha uno scarso senso.

La tecnica della *cross-validation*, specie multipla, trova ancora oggi dei sostenitori, ad esempio Thompson (1993), ma la possibilità di artefatti nel calcolo dei parametri di regressione dovuta alla piccolezza dei campioni impiegati ha reso molti ricercatori particolarmente cauti, raccomandando piuttosto l'uso di tecniche di *jackknife* (per esempio, Ang, 1998). E sono queste le tecniche che ora esamineremo.

6.5. IL *JACKKNIFE*

La tecnica di *jackknife* è stata proposta da Quenouille nel 1949 e da Tukey nel 1958. Il senso di questa tecnica sarà ora perfettamente chiaro, in base a quanto abbiamo sin qui detto.

Noi abbiamo il nostro campione **x**, e su questo calcoliamo la nostra statistica, che chiamiamo genericamente θ. Successivamente, se **x** è costituito di n valori, noi costituiamo n sottocampioni di **x**, ognuno di $n-1$ valori, togliendo in altre parole un valore alla volta — come se tagliassimo con un "coltellino" (è questo il significato di *jackknife*) un valore alla volta dal campione originale. Questi sottocampioni sono più o meno delle dimensioni del campione originale, e il rischio di artefatti dovuti alla piccolezza del campione viene così minimizzata. A questo punto su questi campioni viene calcolata la statistica θ', e per ogni campione si ottiene il cosiddetto pseudo-valore (*PV*):

$$PV = n \cdot \theta - (n-1)\theta'. \tag{6.16}$$

Avremo così un campione di n pseudovalori, la cui media prende il nome di *jackknifed coefficient* (*jc*). Dividendo il *jc* per l'errore standard stimato, otteniamo il t di Student associato, di cui potremo valutare la significatività.

Si osservi che questo metodo è in grado di minimizzare l'effetto della presenza di *outliers* tra i valori osservati. Di massima, si considera *outlier* un valore che sia di oltre due unità standard diverso dalla media. Si pensi a un esperimento sui tempi di reazione: il soggetto potrebbe rispondere troppo velocemente, perché ha risposto con un automatismo, prima di esaminare il significato dello stimolo; o potrebbe essersi distratto, e rispondere con abnorme lentezza. Ma queste sono interpretazioni che si dà lo sperimentatore, e che nulla garantisce. E la discussione sull'opportunità o meno di eliminare gli *outliers* dai dati è sempre molto vivace, anche perché la distorsione prodotta è strettamente legata alla grandezza del campione (cfr. Miller, 1991). Il *jackknife* ci risolve in modo indolore questo problema, perché l'effetto degli *outliers* viene di fatto a scomparire nel gran numero di campioni utilizzati. E proprio il problema degli *outliers* fu alla base della proposta del metodo.

Esempio 6.10

Per il nostro primo esempio, possiamo riprendere i dati della tabella 6.12, e vedere se il *jackknife* è in grado di fornirci il modo di interpretare in maniera più sensata il nostro problema. Avendo $n = 20$, ricaviamo quindi per prima cosa 20 campioni di 19 osservazioni ciascuno, da cui otteniamo mediante la (6.15) i nostri 20 *PV*, la cui media è il coefficiente di *jackknife jc*. Calcoliamo di questo il *t*, e la probabilità ad esso associata. I dati sono presentati nella tabella 6.14.

Sottocamp.	**PV**	**Sottocamp.**	**PV**
1	0,908	11	0,135
2	0,867	12	0,536
3	0,757	13	0,088
4	0,830	14	-0,459
5	0,754	15	0,785
6	0,707	14	0,217
7	0,613	17	-0,117
8	0,527	18	0,181
9	0,515	19	0,357
10	0,316	20	0,571
jc = 0,454	*t* =23,881	*p* = 0,000	

Tabella 6.14 – *Pseudo-valori, coefficiente di jackknife e t per i dati della tabella 6.11*

Il *jc* è molto simile al coefficiente di correlazione trovato in origine (che era di 0,481). A questo *jc* corrisponde un *t* significativo, il che ci consente di asserire che la correlazione trovata è improbabilmente dovuta a fluttuazioni casuali dei valori.

Esempio 6.11

A questo primo esempio, ne facciamo seguire un secondo, che renderà chiaro quanto abbiamo detto a proposito degli *outliers*. I dati si riferiscono a un esperimento sui tempi di reazione, in cui si hanno due gruppi indipendenti, ciascuno di 12 soggetti, sottoposti il primo (K) a un compito di tempo di reazione in cui c'è compatibilità spaziale tra stimolo e risposta; il secondo (I) a un compito di tempo di reazione in cui non c'è compatibilità spaziale. In altri termini, nel compito K lo stimolo può provenire da destra o da sinistra, e il pulsante che il soggetto deve premere per dare la sua risposta è situato nella stessa direzione spaziale da cui proviene lo stimolo. Per il gruppo I, direzione dello stimolo e posizione del pulsante sono indipendenti. I tempi di reazione, in millisecondi, sono una variabile che si pone su una scala a rapporti (esiste

uno 0 non arbitrario), e può quindi essere usata la statistica *t*. L'ipotesi è a una coda, poiché ci aspettiamo che i tempi K siano inferiori ai tempi I. I dati sono presentati nella tabella 6.15.

Campione K		**Campione I**	
Sogg	**TR**	**Sogg**	**TR**
1	442	1	632
2	383	2	555
3	530	3	1052
4	388	4	554
5	186	5	746
6	602	6	488
7	407	7	198
8	478	8	1006
9	514	9	558
10	369	10	521
11	436	11	710
12	1010	12	633

Tabella 6.15 – *Tempi di reazione mediani per compiti compatibili (K) e incompatibili (I)*

Si osservi che in questo caso la situazione è più complessa che nel caso precedente, perché per ogni elemento che eliminiamo nel primo campione, dobbiamo effettuare un confronto con i venti sottocampioni ottenuti dal secondo campione, eliminando un elemento alla volta. Quindi non abbiamo a che fare con 12 pseudovalori, ma con 144 (12 sottocampioni del primo gruppo per 12 campioni del secondo gruppo).

I dati sono presentati nella tabella 6.16. Come si può vedere, malgrado la notevole differenza tra le medie, il *t* non è significativo. Peraltro, è possibile che questo sia dovuto alla presenza di *outliers*, due nel gruppo K (soggetti 5 e 12) e 3 nel gruppo I (soggetti 3, 7 e 8). Eliminando questi soggetti, infatti, pur diminuendo lievemente la differenza tra le due medie (da 159 a 144,77) e il numero dei soggetti (da 12 + 12 a 10 + 9; e abbiamo detto più volte che più grandi sono i campioni, più è facile ottenere un *t* significativo), il *t* sale a 3,683, e diventa largamente significativo.

Se ciò conferma l'effetto distorcente degli *outliers*, ci mostra anche come questi costituiscano un problema di assai difficile soluzione per le analisi tradizionali. Eliminandoli, i dati diventano più facilmente statisticamente significativi, ma quali sono le basi reali che giustificano questo comportamento? È chiaro che si eliminano informazioni, e l'avere un *t* significativo non può costituire l'unica base per questo sacrificio.

Con il *jackknife*, il problema si risolve. I 144 campioni che otteniamo ci mostrano un coefficiente *jc* ben più alto di entrambi i valori trovati, e statisticamente significativo, come dimostra l'elevatissimo valore di *t* (45,033) a lui associato.

Campioni	Statistica		gl	*p*
Originali	Media K	M_K = 478,75		
	Media I	M_I =637,75		
	confronto	t =1,750	22	0,081
Senza *outliers*	Media K	M_K = 454,90		
	Media I	M_I =599,67		
	confronto	t = 3,863	17	0,001
Jackknife	coefficiente	jc = 6,553		
	t	t = 45,033	143	0,000

Tabella 6.16 – *Analisi dei dati della tabella 6.14, con e senza outliers e con jackknife*

6.6. Il *bootstrap*

La tecnica di *bootstrap* rappresenta un ulteriore passo in avanti rispetto al *jackknife*. Il *bootstrap* nasce con il nome di "ricampionamento" (*resampling*) a metà degli anni '60, per merito di Julian Simon (cfr. Simon, 1969; Simon e Holmes, 1969) come sviluppo delle tecniche di simulazione Monte Carlo, applicate allora all'insegnamento della statistica (Simon, Atkinson e Shevokas, 1976). Rapidamente le tecniche divennero uno strumento di importanza fondamentale per la statistica inferenziale, sino ad essere definite l'"unico grande progresso" compiuto da allora dalla statistica (Kotz e Johnson, 1993). Nel 1979 Bradley Efron sviluppò la stessa tecnica, indipendentemente da Simon, e gli diede il nome fortunatissimo di *bootstrap* (letteralmente, stringa degli scarponi). Il concetto è ispirato al racconto, molto popolare nel mondo anglosassone, e diventato proverbiale, del ragazzo che si tirò su dalla palude in cui è caduto afferrandosi alle stringhe delle scarpe — o, come fece il Barone di Münchausen, al codino. Da allora la tecnica è sempre stata associata soprattutto al nome di Efron (cfr. Efron, 1979, 1981, 1982, nonché Efron e Tibshirani, 1993) anche se Simon non si è mai stancato di rivendicarne la priorità.

La differenza sostanziale tra *bootstrap* e *jackknife* consiste nel fatto che in quest'ultimo si ricorre a un campionamento senza reimmissione, mentre nel *bootstrap* si ha campionamento con reimmissione. In altri termini, nel *jackknife* ogni volta che si estrae un elemento dall'insieme originale di n elementi per formare il nuovo sottocampione di $n - 1$ elementi, l'elemento ogni volta estratto non viene reimmesso nell'insieme originale, per cui gli $n - 1$ elementi saranno tutti diversi gli uni dagli altri.

Il numero di campioni che si può così creare può essere eccessivamente limitato. Nel *bootstrap*, di contro, ogni elemento che viene estratto viene reimmesso subito nell'insieme originale, per cui potrebbe poi essere di nuovo estratto. Ciò aumenta in misura considerevole l'universo di campioni creabili.

Il vantaggio di questo metodo appare allora evidente. Noi possiamo così costruire quanti campioni vogliamo, mentre nel *jackknife* siamo vincolati a un numero di campioni che può essere eccessivamente ridotto.

Vediamo allora di presentare un esempio di *bootstrap*. I dati che qui analizziamo sono ancora quelli della tabella 6.10. Ciò ci consentirà di abbozzare un primo confronto tra tecniche di *bootstrap* e tecniche di randomizzazione, almeno per disegni per valori appaiati.

ESEMPIO 6.12

Si ricorderà che i dati della tabella 6.10 si riferivano a un confronto tra medie di due campioni appaiati, e che dai dati originali si era ottenuto un t significativo di 4,956. Nella tabella 6.16 vediamo i risultati dell'analisi, per un *bootstrap* con 5000 iterazioni (terza riga), che sono confrontati con i risultati ottenuti con quelli ottenuti con la randomizzazione, sempre per 5000 iterazioni, e già visti in precedenza (seconda riga).

Iterazioni	***t* medio**	**CV**	**Intervalli di fiducia a una coda**	**Intervalli di fiducia a due code**
Random.	2,33874	0,18203	$1{,}754 < t < 3{,}768$	$1{,}731 < t < 4{,}006$
Bootstrap	2,61090	0,49900	$0{,}350 < t < 6{,}634$	$0{,}136 < t < 7{,}714$

Tabella 6.16 – *Dati relativi al test di randomizzazione e al bootstrap con e 5000 iterazioni per t per campioni dipendenti con intervalli di fiducia a 0,01 e 0,005*

Di questi risultati vanno rilevate alcune cose. Innanzitutto, il CV nel *bootstrap* è molto più alto che nella randomizzazione, ciò che mostra una maggiore dispersione dei dati nel primo. In secondo luogo, questa maggiore dispersione comporta un forte allargamento degli intervalli

di fiducia, per cui il *t* originale, che in base al test di randomizzazione era significativo all'1 %, qui risulta non significativo. Risulterebbe invece significativo al 5% monodirezionale, essendo il limite superiore qui posto a 4,419.

Il *bootstrap* risulterebbe, in base a questa simulazione, più conservativo e meno affidabile della randomizzazione. Noi sospettiamo che sia proprio così, e ci spieghiamo difficilmente l'entusiasmo che questa tecnica ha sollevato, mentre i test di randomizzazione appaiono al confronto decisamente trascurati. Peraltro, quanto detto vale per piccoli campioni (che sono peraltro solitamente quelli di maggior interesse per lo psicologo), mentre per campioni di maggiori dimensioni il problema probabilmente va visto in modo diverso.

Capitolo 7
Conclusioni

7.1. Uno sguardo retrospettivo

Siamo così giunti alla fine del nostro lavoro, ma prima di scrivere la parola fine sarà opportuno rivedere in estrema sintesi come siamo giunti a questo punto, offrendo al lettore come ciambella di salvataggio un quadro riassuntivo che lo aiuti a non naufragare nel mare di formule e di tabelle di cui il testo è intessuto.

Nel Cap. 1 abbiamo visto come sono nati i due approcci che hanno dominato le analisi inferenziali dei dati in psicologia, e in poche altre discipline, dalla sociologia alla medicina, il PVA di Fisher (1925) e il FAA di Neyman e Pearson (1933). Abbiamo visto che nell'approccio fisheriano non si ponevano a confronto un'ipotesi nulla H_0 e un'ipotesi sostantiva H_1, ma si ipotizzava esclusivamente che i dati osservati appartenessero a una data distribuzione — che la stima di un dato parametro tratto dai dati osservati corrispondesse a quello di una data popolazione. Se la probabilità che questo fosse vero era troppo bassa (tipicamente, inferiore al valore che avrebbe poi acquistato un significato mitico dello 0,05) secondo Fisher andava escluso che questa corrispondenza ci fosse.

Nei termini di Neyman e Pearson, la corrispondenza equivaleva all'accettazione dell'ipotesi nulla, a cui però essi opponevano un'ipotesi sostantiva, e cioè che il parametro fosse effettivamente diverso. Essi allora ritenevano che il processo dovesse svolgersi in modo più complesso di quanto ipotizzato da Fisher. In primo luogo, andava fissato il valore di α, e cioè della probabilità di commettere l'errore di I tipo (decidere per H_1 quando è vera H_0), e qui tornava in primo piano il fatidico 0,05. Il discorso però si faceva più complesso. Innanzitutto si chiedeva di stabilire il coefficiente di massima verosimiglianza L per i casi in cui fosse rispettivamente vera H_0 (L_0) e H_1 (L_1) e di determinare il loro rapporto. α a questo punto era la probabilità condizionale che il rapporto tra L_0 e L_1 fosse inferiore a un k prefissato, posto che H_0 fosse vera. Questo rapporto (*lemma di Neyman e Pearson*, cfr. paragrafo 1.3.3.3) consentiva allora di suddividere lo spazio dei parametri in due aree, una critica, tale per cui il rapporto tra L_0 e L_1 era inferiore a k, che portava a respingere H_0 e accettare H_1, e una di accettazione, tale per cui il rapporto tra L_0 e L_1 era superiore a k, che portava a respingere H_1 e accettare H_0.

Alcune cose vanno sottolineate, anche perché altrimenti quanto detto risulterebbe alquanto oscuro. In primo luogo, ed è questa la differenza più importante tra i due ap-

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*, ISBN 88-8453-226-4,

procci, lo FAA prevede una presa in considerazione di entrambe le ipotesi, e non solo dell'ipotesi nulla. In questo senso, come è stato sottolineato soprattutto da Huberty (1993), lo FAA è un "test di ipotesi", e non un test di significatività tout-court, come il PVA. Così, alla specularità delle due ipotesi corrisponde anche una specularità dei possibili errori, di primo tipo (con probabilità α), ma anche di secondo tipo (con probabilità β), legato questo all'importantissimo concetto di potenza del test statistico (cfr. Cap. 5). Ciò comporta una strategia completamente differente da parte del ricercatore che sceglie l'uno o l'altro approccio. In un caso ci si muove in termini di tutto o nulla: o il valore di p è statisticamente significativo, e in questo caso si respinge l'ipotesi nulla. Nell'altro caso, la decisione è meno netta, ma probabilmente più sensata sul piano della fondatezza della ricerca: le due ipotesi sono sempre entrambe in gioco, la decisione per l'una o per l'altro è relativa, e suscettibile di essere cambiata alla luce di nuovi dati di altre ricerche, che qui sono auspicate, nel PVA non contemplate.

Abbiamo peraltro visto che, curiosamente, questi due approcci decisamente alternativi sono stati poi fusi in un ibrido, che si è imposto paradigmaticamente nel mondo della ricerca. L'ibrido ha accolto la terminologia del FAA (ipotesi nulla e sostantiva, errori di primo e secondo tipo), ma ha limitato il processo decisionale a quanto dettato dal PVA: il ricercatore si limita a verificare se la statistica da lui utilizzata ha associato un valore di p inferiore all'α prefissato: se è così, respinge l'ipotesi nulla e accetta la sostantiva, altrimenti accetta l'ipotesi nulla.

Sin dagli anni '60 si sono levate le prime voci di dissenso verso questo modo di procedere, ma è stato soprattutto in questi ultimi venti anni che le critiche si sono fatte insistenti e largamente motivate. Nel Cap. 2 abbiamo proposto un largo catalogo di inconvenienti a cui la VeSN, per come oggi impiegata nel mondo della ricerca, dà luogo. In particolare, abbiamo visto come, secondo la maggior parte degli autori che hanno affrontato il problema, il semplice valore di probabilità (e conseguente decisione di significatività o meno dei risultati) venga ritenuto insufficiente a prendere una decisione sensata sui dati di una ricerca. In particolare, è stata da più parti sottolineata l'opportunità di accompagnare l'esposizione di ogni analisi inferenziale con una misura della grandezza dell'effetto; con l'esposizione degli intervalli di fiducia (o confidenza), sia per quel che riguarda i valori stimati dei parametri, sia per le misure della grandezza dell'effetto; con un'analisi di potenza, preferibilmente a priori, comunque almeno a posteriori. A questi aspetti specifici sono stati dedicati rispettivamente i Cap. 3 (grandezza dell'effetto), 4 (intervalli di fiducia), 5 (analisi di potenza).

Nel Cap. 6, poi, abbiamo voluto presentare dei modi alternativi di usare le analisi inferenziali, impiegando le tecniche di resampling, e in particolare le tecniche di *randomizzazione*, di *jackknife* e di *bootstrapping*. Si tratta di tecniche di simulazione che, servendosi solo dei dati raccolti nell'esperimento che viene analizzato, consentono però di trattare le statistiche utilizzate come variabili casuali, e di analizzarle quindi in base alle caratteristiche delle loro distribuzioni. Sono tecniche oggi alla portata di qualsiasi ricercatore, da sfruttare quindi a pieno per le straordinarie possibilità che offrono.

Potremmo concludere il nostro lavoro quindi in questi termini: nelle analisi inferenziali si tenga conto del quadro che si offre al ricercatore in tutta la sua ampiezza, e complessità, e (in linea, tra l'altro, con quanto raccomanda l'*American Psychological*

Association) i valori di significatività vengano sempre accompagnati dagli indici di cui si è detto (grandezza, intervalli di fiducia e potenza). A questo aggiungiamo la raccomandazione di accompagnare l'analisi con qualche appropriata misura di *resampling*.

Potremmo fermarci qua. In realtà, qualche considerazione ulteriore va pur fatta, e riguarda, sia pur succintamente, tre ordini di problemi: (i) quanto questo discorso, nonostante la gran mole di lavori che si è accumulata in proposito in questi anni, sia stato fatto proprio, oltre che dai metodologi, dai ricercatori; (ii) come viene insegnata l'inferenza statistica ai futuri ricercatori; (iii) se è proprio necessario, almeno in questa misura soverchiante, utilizzare pressoché solo queste tecniche per analizzare i dati della ricerca psicologica.

7.2. La VeSN nelle ricerche psicologiche pubblicate

Le analisi condotte (Cohen, 1962; Sedlmeier e Gigerenzer, 1989) sulle principali riviste psicologiche hanno dimostrato che i ricercatori sono del tutto insensibili al problema, ma soprattutto che negli ultimi anni, malgrado l'ampio dibattito apertosi in proposito, non si è registrato nessun miglioramento. L'attenzione dei ricercatori si è particolarmente concentrata sulla potenza dei test statistici impiegati, ma questa è solitamente molto bassa, di media inferiore a 0,5, e cioè, come sottolinea Hunter (1997), il livello del caso. La significatività, di più, è di norma assunta *sic et sempliciter* come conferma dell'ipotesi sostantiva, non vi è nessuna riflessione sulla potenza (i valori suaccennati sono i risultati di analisi post hoc condotte da ricercatori successivi, e non dagli autori delle ricerche), assolutamente eccezionale è poi la presenza di altri indici, da quelli sulla grandezza dell'effetto agli intervalli di fiducia.

Ma ciò che è forse più sconfortante è che a quasi trent'anni dalla prima ricerca pionieristica di Cohen (1962) sul *Journal of Personality and Social Psychology*, la successiva analisi di Sedlmeier e Gigerenzer (1989) ha prodotto risultati pressoché identici: in trent'anni ricercatori (e redazioni delle riviste scientifiche) sembrano non essersi accorti minimamente del problema, e han seguitato a operare esattamente allo stesso modo. E oggi, malgrado il lavoro ampiamente pubblicizzato della Task Force dell'*American Psychological Association*, e l'uscita da tre anni del più volte citato nuovo manuale di pubblicazione (American Psychological Association, 2001), anche in assenza di dati precisi tutto fa sembrare che nulla sia cambiato.

E in Italia? Anche qui il dibattito sulla VeSN ha avuto scarsa eco. Vi è stato (ma riferito alla situazione in sociologia) un buon contributo di Pisati (2002). Al congresso dell'Associazione Italiana di Psicologia di Bari una relazione di Agnoli (2002) ha suscitato un buon interesse. Nel giugno del 2003 si è tenuta una giornata di studio sul problema a Firenze, che ha visto la partecipazione di pressoché tutti i docenti italiani del raggruppamento disciplinare M.PSI/03, quello cioè che raggruppa la psicometria e le materie affini. Ma è stata appunto una discussione tutta interna agli studiosi delle applicazioni della statistica alla psicologia, che ha visto estranei i ricercatori.

Noi abbiamo allora voluto esaminare l'analisi dei dati eseguita in tutti i lavori di ricerca pubblicati da quella che è oggi unanimemente reputata la più stimata rivista italiana di psicologia, l'unica forse che si ritiene soddisfi agli standard internazionali più esigenti, il *Giornale Italiano di Psicologia* (GIP), che oggi compie il trentesimo anno di vita, essendo stata fondata nel 1974.

Sin dall'inizio, la rivista era articolata in più sezioni, in parte di intervento e discussione, di rassegne della letteratura, a cui più recentemente si sono aggiunti recensioni, "strumenti", e un "articolo bersaglio" con *peer commentaries*. Ma il "cuore" della rivista è sempre stato costituito dalla sezione "Studi e ricerche", che pubblica lavori di ricerca, prevalentemente sperimentale, ma anche *field research* e studi teorici. Sono 611 gli articoli pubblicati in questa sezione nei 30 anni di vita della rivista (sono qui assenti i dati relativi all'ultimo numero del 2003, che stiamo analizzando), ed è su questi che abbiamo condotto la nostra analisi. La ricerca è ancora in corso, per cui i dati che presentiamo riguardano un primo esame sommario, mentre stiamo ancora calcolando nuovi indici statistici, per quanto è possibile farlo a posteriori.

Quel che volevamo vedere è se era stata condotta un'analisi statistica dei dati. Se questa era stata condotta, ci interessava vedere se si trattava di una VeSN. Nel caso che si trattasse di una VeSN, ci interessava allora analizzare i punti nodali emersi, e in particolare: (i) se fosse stata condotta un'analisi a priori della potenza; (ii) nel caso non fosse stata condotta, quali erano i valori a posteriori della potenza delle statistiche utilizzate; (iii) se venivano forniti indici aggiuntivi, come le stime della grandezza dell'effetto e gli intervalli di fiducia.

Come si può vedere dalla Tabella 7.1, si tratta di 377 lavori, corrispondenti al 61,7% del totale. Si osservi che molte ricerche presentavano più tipi di analisi dei dati. Ad esempio, spesso a un'analisi fattoriale faceva seguito un'ANOVA, per vedere le eventuali differenze medie nei punteggi fattoriali tra i sottogruppi in cui poteva essere suddiviso il campione. Il criterio di classificazione scelto considera solo il tipo di analisi prevalente, con tutti i rischi di arbitrarietà che questa decisione comporta. La ricerca dell'esempio precedente sarebbe così stata compresa nella categoria terza, analisi multivariata.

Tipo di analisi	**frequenza**	**%**
VESN	377	61,7
Descrittiva	86	14,08
Multivariata	76	12,44
Teorica	34	5,56
Errori	22	3,6
Altre analisi	16	2,62
Totale	611	100

Tabella 7.1 – *Riepilogo dei lavori presentati sul GIP dal 1974 ad oggi, suddivisi per tipo di analisi*

Come si vede dalla tabella 7.1, 86 lavori erano di tipo descrittivo, senza nessun tentativo di andare in alcuni casi al di là delle percentuali dei comportamenti osservati. In questa categoria abbiamo incluso anche le ricerche fenomenologiche (16), non infrequenti nei primi dieci anni di vita della rivista, e poi pressoché scomparse. 76 ricerche erano invece multivariate, quasi esclusivamente analisi fattoriali fino all'inizio degli anni '90. In questo periodo cominciano a comparire modelli di equazioni strutturali (Analisi Confermative, LISREL). Vi sono poi ricerche teoriche (34), e ricerche compiute con analisi di dati che non assumono la VeSN (modelli matematici, ricerche psicofisiche, ricerche verificazioniste, come i modelli loglineari che sono 7, il primo comparso nel 1994).

Le analisi statistiche richiedenti la VeSN di questi lavori sono presentati nella Tabella 7.2.

Analisi	**freq**	**%**
ANOVA	549	59,67
χ^2	108	11,74
t	89	9,67
r	49	5,33
Wilcoxon	20	2,17
Mann e Withney	13	1,41
Manova	10	1,09
binomiale	10	1,09
z	9	0,98
rho	8	0,87
ANCOVA	6	0,65
mcr	6	0,65
Kendall	5	0,54
McNemar	5	0,54
Fisher	4	0,43
Friedman	3	0,33
k	3	0,33
Kolmogorov Sm	3	0,33
Kruskall	3	0,33
Friedman	3	0,33
L	2	0,22
altri	12	1,13
Totale	920	100

Tabella 7.2 – *Analisi statistiche richiedenti la VeSN*

Il totale superiore a 377 è spiegato dal fatto che in quasi tutti i lavori sono presenti più analisi statistiche. Come si può vedere la parte del leone è fatta dal confronto tra medie: quasi il 60% delle analisi sono ANOVA, e quasi il 10% *t* di Student. Il χ^2 riguarda circa il 12% delle analisi, e quasi esclusivamente si riferisce a tabelle di contingenza. Le analisi correlazionali sono poco di più del 5% del totale. Il resto è ampiamente disperso tra 29 altri modelli di analisi.

Di queste 920 analisi, solo 3 sono accompagnate da analisi di potenza a priori, e nessuna a posteriori. Non vi sono mai misure della grandezza dell'effetto, e assolutamente eccezionale (12 casi) è la presentazione di intervalli di fiducia. La cosa più grave è che in 211 casi i dati sono presentati in modo tale da rendere impossibile il calcolo di questi indici anche a posteriori. Tipicamente in molte ANOVA ci si limita a presentare a fianco dei valori di *F* le medie dei gruppi. Solo in questi ultimi anni (probabilmente più per merito del software utilizzato, che fornisce certi elementi per *default*, che dei ricercatori) sono cominciati a comparire nei grafici le barrette che indicano gli intervalli di fiducia, ma è rarissimo che i valori numerici esatti siano poi riportati nel testo degli articoli, e l'informazione resta quindi inutilizzabile. Ma accade anche che i ricercatori si limitino a dire, nel confronto tra le medie di più gruppi, quale sia la media maggiore e quale la minore, ritenendo però superfluo darne il valore numerico.

Noi allora abbiamo calcolato a posteriori la potenza delle statistiche utilizzate in queste 920 analisi. Dove era possibile, si è calcolata la grandezza dell'effetto, utilizzando uno degli indici esposti nel Cap. 3, e la grandezza dell'effetto è stata poi categorizzata in una delle tre grandezze convenzionali (piccola, media o grande) indicate da Cohen (1988). Dove non era possibile, sono stati calcolati tutti e tre i valori di potenza per le tre grandezze convenzionali. È stata quindi calcolata per i sei quinquenni in cui abbiamo suddiviso i trent'anni di vita del GIP la media della potenza per i tre valori convenzionali di grandezza dell'effetto. I risultati sono presentati nella Tabella 7.3. È stato usato il software G*Power (2000).

quinquennio	piccola	media	grande
1974-78	0,1408	0,5167	0,7796
1979-83	0,1248	0,3964	0,6543
1984-88	0,1521	0,4784	0,7112
1989-93	0,1358	0,4618	0,7147
1994-98	0,1727	0,5266	0,7773
1999-2003	0,2365	0,6260	0,8320

Tabella 7.3 – *La potenza media per quinquennio per grandezze convenzionali piccole, medie e grandi degli effetti negli articoli pubblicati sul GIP (1974-2003)*

I dati sono abbastanza sconfortanti, anche se nell'ultimo quinquennio vi è stato un certo miglioramento, per tutte e tre le grandezze convenzionali. Si badi che sino al 1998 la potenza per grandezza media dell'effetto (quella di più frequente riscontro nelle ricerche analizzate) si mantiene intorno al 50%, quindi a livello del caso. Ed anche nell'ultimo quinquennio la potenza sale al 62%, quindi ben al disotto di quello 0,80 considerato consigliabile. Si può tranquillamente concluderne che nella generalità dei casi l'ipotesi sostantiva viene accettata senza nessuna reale garanzia della sensatezza di tale accettazione.

Le discussioni che si svolgono nelle stanze dei metodologi non sembra dunque che turbino più che tanto i sonni dei ricercatori. Il paradigma della VeSN, come lo abbiamo esposto all'inizio di questo volume, seguita ad imperare. Ed è anche difficile attribuire a una maggiore consapevolezza da parte dei ricercatori il miglioramento della potenza verificatosi negli ultimi anni, dato che non si accompagna alla presentazione degli indici raccomandati dall'APA.

Evidentemente, il problema non è avvertito dai ricercatori. È probabile che ciò dipenda in primo luogo dal modo in cui, nel corso della propria formazione, essi hanno appreso ad analizzare i dati delle loro ricerche. È quanto cercheremo ora di capire.

7.3. L'insegnamento della statistica ai futuri psicologi e la VeSN

Come notano Haller e Krauss (2002, p. 2), "l'insegnamento della VeSN potrebbe essere giustificato solo se gli studenti sono capaci di afferrare il *significato* di quel che stanno facendo", ma disgraziatamente gli studi sin qui condotti (e la dolorosa nostra esperienza autobiografica) sembrano dimostrare che dopo un corso di statistica di norma uno studente medio non ha la più lontana idea di cosa la VeSN effettivamente significhi (Falk e Greenbaum, 1995; Gigerenzer e Krauss, 2001). Al meglio, per usare la famosa espressione di Gigerenzer (1998), ha appreso un "rituale di calcolo".

A dimostrazione, Gigerenzer e Krauss (2001) hanno mostrato che, alla domanda, "Cosa significa dire che una misura statistica è significativa al livello del 5%?", praticamente nessuno studente al termine di un corso di statistica per psicologia era in grado di dare la risposta esatta, e cioè, "La probabilità dei dati osservati (o di dati meno probabili) è minore di 0,05, posto che l'ipotesi nulla sia vera". Quasi tutte le risposte cadevano in una di queste tre categorie: (i) risposte prive di senso, del tipo: "Significa che la misura è del 5% al di sopra della percentuale del caso"; (ii) risposte che indicavano la credenza che il 5% indicasse la probabilità dell'ipotesi, e non dei dati; (iii) altre risposte di varia insensatezza, tra cui la credenza che la significatività riguarda la replicabilità dei risultati.

Ma quel che deve far particolarmente riflettere è che, come ha dimostrato Oakes (1986), la maggior parte degli psicologi accademici non hanno idee molto più chiare de-

gli studenti. Haller e Kraus (2002), giustamente preoccupati di questi dati, hanno allora voluto vedere se per caso certe concezioni erronee non venissero specificamente trasmesse nell'insegnamento, attraverso errori o confusioni presenti nei manuali, o concezioni erronee proprie addirittura degli insegnanti di statistica agli studenti di psicologia.

Per ciò che riguarda i manuali, i citati Haller e Kraus, ma anche Sedlmeier e Gigerenzer (1989), ci propongono un vero e proprio museo degli orrori, che non risparmia neppure i prodotti di qualche mostro sacro della statistica, come Nunally (1975), che in tre pagine (194-196) riesce a darci la bellezza di otto definizioni diverse di significatività, tutte errate, da "l'improbabilità che i risultati osservati siano dovuti a errore" a "il pericolo di accettare un risultato sperimentale come reale quando di fatto è dovuto solo a errore".

È questa una situazione tipica dei soli manuali nordamericani, o assimilati? Può essere interessante vedere cosa accade in Italia, pur senza pretese di essere esaustivi — anzi, è opportuno sottolineare che le considerazioni che seguono vogliono solo essere un invito a condurre una ricerca approfondita su manuali, e in genere forme di insegnamento della statistica applicata alla psicologia, che al momento manca. Ora, dobbiamo dire che il problema della VeSN è presentato in forma sostanzialmente corretta nei tre manuali di statistica per psicologi di forse più ampia diffusione nei corsi di laurea in psicologia italiani (Caudek e Luccio, 2001, cap. 8; Ercolani e Areni, 1983, pp. 139-146; Vidotto, Xausa e Pedon, 1996, pp. 274-284). Peraltro, la riforma degli studi universitari italiani, con il famoso 3+2, ha portato a livello di laurea junior a un largo uso in molte sedi di succinti manualetti, poco più che prontuari di formule, pressoché privi di ogni riflessione teorica, e a un radicale ridimensionamento della presenza dei manuali di più ampio respiro, come i tre succitati. Su questi testi andrebbe probabilmente condotta un'analisi più approfondita.

Un altro testo però largamente usato (più in passato che oggi) è il Blalock (1960), che per molti psicologi (e sociologi) è stato e in parte è ancora per molti anni la "Bibbia" della statistica per le scienze sociali. Ora, qui viceversa c'è da rimanere perplessi. Il problema della VeSN è infatti affrontato nel Cap. 8, con le modalità espositive tipiche di questo manuale: poco trattamento matematico, stile conversativo accattivante, sul piano del buon senso. Blalock introduce correttamente i concetti di errore di primo e secondo tipo, solo che a p. 146 della traduzione italiana afferma testualmente "Il secondo tipo (errore beta) deriva dall'errore sul piano logico che consiste nell'affermare ciò che si deduce; il primo tipo (errore alfa) insorge quando introduciamo affermazioni probabilistiche nella teoria". Ad aumentare la confusione, a p. 147 si sostiene che è solo l'ipotesi nulla "che viene in effetti sottoposta a verifica". E poco sotto si confonde l'accettazione dell'ipotesi sostantiva (evidentemente per rifiuto dell'ipotesi nulla) con la grandezza dell'effetto.

La pluralità dei corsi di laurea diffusi sul territorio nazionale, la loro disomogeneità, il fatto che spesso l'insegnamento sia affidato (specie a contratto) a giovani magari brillanti per altre competenze, ma che non hanno la statistica al centro dei loro interessi scientifici, produce conseguenze molto meno desiderabili. Non consideriamo i testi a livello di dispensa, spesso poco più che *samizdat*, che talvolta si presentano come plagi di altri più noti testi, arricchiti però spesso di strafalcioni; purtroppo anche a livello

di traduzioni abbiamo a volte a che fare con prodotti che lasciano senz'altro perplessi. Prendiamo così un manuale di "Statistica per psicologi", largamente usato nei corsi di laurea di psicologia delle Università italiane, e cioè il Greene e D'Oliveira (1999). Il problema della significatività statistica viene trattato in questo libro tra le pp. 25 e 30, e in cinque pagine gli autori riescono ad accumulare tante di quelle inesattezze e franche sciocchezze da rendere probabilmente necessario l'invio per alcuni anni in un campo di rieducazione degli sventurati giovanetti che ne dovessero venire a contatto.

La prima perla (p. 25) è l'affermazione per cui "*meno* probabilmente le differenze sono dovute a variabili casuali, *maggiormente* possiamo essere fiduciosi che esista una differenza reale significativa". Qui la significatività non c'entra, siamo ai fondamentali: hanno idea gli autori di cosa siano le variabili casuali? Questa curiosa concezione di "variabile casuale" ritorna poi sempre, per cui non ci soffermeremo oltre.

Veniamo a questioni più specifiche. A pag. 28 si dice: "è convenzione accettare quote dell'1% o del 5% come base per rifiutare l'ipotesi nulla e quindi accettare che l'ipotesi sperimentale sia confermata". Poco sotto, "succede spesso che i risultati possano essere più significativi". E a pag. 27, "se una percentuale di probabilità rimane sotto a un certo livello … lo sperimentatore può rifiutare l'ipotesi nulla e accettare dunque una conferma dell'ipotesi sperimentale".

Questo manuale viene propagandato come un testo che "riduce l'incubo" della matematica nell'apprendimento della statistica. Lo slogan proposto è: "Più psicologia, più metodologia e meno matematica". Gli autori si sono dimenticati di aggiungere: "e niente statistica".

Per ciò che riguarda gli insegnanti di statistica per psicologi, specie gli esercitatori, i dati che ci presentano Haller e Krause per la Germania e Oakes per gli Stati Uniti sono semmai ancor più sconfortanti. Noi non abbiamo dati per quel che riguarda la situazione italiana, ma qualche considerazione può ben farsi. Le discipline che riguardano l'analisi dei dati e le applicazioni della statistica alla psicologia sono in Italia raggruppate nel gruppo disciplinare M.PSI/03. Ora, per molti anni questo gruppo, anche per una sua debolezza intrinseca dovuta allo scarso numero dei suoi appartenenti, è stato utilizzato come terra di conquista da parte degli altri raggruppamenti disciplinari, per cui finivano ad occupare qui delle posizioni persone anche valide, ma prive di preparazione specifica, che trovavano difficoltà a ottenere delle posizioni in raggruppamenti più consoni ai loro reali interessi (e competenze).

La situazione, a livello di docenti di ruolo, è in questi ultimi anni molto migliorata sotto questo punto di vista, ed oggi gli appartenenti al raggruppamento sono senz'altro molto competenti in materia. L'esplosione e la disseminazione così ampia dei corsi di laurea in psicologia su tutto il territorio nazionale ha però dilatato enormemente il bisogno di insegnamenti, ad un livello che per il loro numero ancora esiguo i docenti strutturati sono lontanissimi dal poter soddisfare. Si sono così in molti casi dovuti affidare questi insegnamenti a giovani ricercatori di altri settori della psicologia, forse anche brillanti, ma provvisti per questo problema della formazione appunto dei ricercatori, che come abbiamo visto lascia molto a desiderare.

Né la situazione si è fatta migliore quando si è fatto ricorso a giovani statistici. È raro che gli statistici italiani, di solito più vicini a problematiche econometriche o

demografiche, abbiano competenza specifica nella statistica inferenziale, o almeno in quella che più è utilizzata nella ricerca psicologica. E il risultato è che gli studenti non acquisiscono nessuna consapevolezza critica, e seguitano ad apprendere senza capirli dei meri rituali di calcolo.

7.4. LA VESN È UNA VIA OBBLIGATA?

Sembrerebbe, giunti a questo punto, che si possa anche così concludere: la VeSN, nella sua versione largamente praticata, è una via che appare quasi obbligata, poco suscettibile di essere modificata, perché la sua logica (erronea) appare invece cogente, lineare, e un approccio più complesso, che tenga conto anche di grandezza dell'effetto, intervalli di fiducia e potenza del test non è altrettanto facilmente afferrabile da parte di un ricercatore la cui preparazione di base è già di per sé abbastanza carente.

Se questo è vero, *oleum et operam perdimus* nel cercare di modificare le cose. Ai ricercatori la VeSN posta in questi termini piace, propone un rituale semplice e rassicurante, che non richiede neppure di essere capito. C'è un numero magico, lo 0,05, se vado sotto il mio esperimento è riuscito, se resto al di sopra pazienza, ricomincio da capo. Di più, questo modo di procedere riceve continui rinforzi: un articolo così confezionato viene più facilmente accettato da un giornale scientifico, nel presentare i miei risultati a un congresso non suscito obiezioni, anzi, vengo complimentato dai colleghi.

Ma è indispensabile questo modo di procedere? Abbiamo già detto che nella sua generalità la scienza trova questo modo di elaborare i dati un po' inconsueto, e non sono molte le discipline che gli assegnano un ruolo di così assoluto privilegio: la psicologia, appunto, la sociologia, la medicina, l'agraria. Anche la psicologia, peraltro, ha sviluppato sin dall'inizio del '900 un modo alternativo di elaborare i suoi dati: parliamo dell'analisi fattoriale, che è specificamente una creazione di uno psicologo, Spearman, e che grazie ad altri due psicologi, Thompson e Thurstone, ha avuto lo straordinario sviluppo che conosciamo.

Ma l'analisi fattoriale, intesa come analisi esplorativa, è la madre di tutte le analisi multivariate, e in genere la madre di tutte le analisi fondate su strutture di covarianza, dalle analisi confermative ai modelli in genere di equazioni strutturali. Si dirà: queste analisi richiedono grandi campioni, a differenza della VeSN, che può lavorare su piccoli campioni (anche se non tanto piccoli come quelli che troppo spesso si vedono pubblicati). Vero. Ma c'è un aspetto comune a queste analisi, e ad altre che stanno diventando di impiego sempre più frequente, come i modelli log-lineari, su cui val la pena di fermare l'attenzione: la verifica dei modelli.

La verifica dei modelli, salvo in settori molto particolari come la psicofisica (ma di fatto è spesso scarsa la consapevolezza da parte dello psicologo del fatto che il lavoro di analisi dei dati che sta effettuando *è* una verifica di modelli) è l'attività principe

nell'analisi dei dati della gran parte delle scienze della natura, a cominciare dalla fisica. Lo scienziato, o sulla base di quanto è già consegnato alla letteratura, o sulla base di sue ipotesi innovative, e infine a posteriori all'ispezione dei dati, ipotizza che la natura abbia arrangiato le cose in modo tale che queste seguano un modello, o, in altri termini, che il loro presentarsi sia formulabile sulla base di una funzione matematica, o un insieme di funzioni, a cui si dà sovente il nome di legge. Il modello può essere la legge di Fechner; ma anche l'equazione di specificazione di un'analisi fattoriale; ma anche l'insieme di equazioni di regressione multipla che in un modello di equazioni strutturali legano tra loro variabili esogene e endogene e errori.

Che la verifica dei modelli abbia scarso peso in psicologia è dimostrato dal modestissimo rilievo che questa metodologia ha nei manuali di analisi dei dati psicologici. Anche l'analisi fattoriale non viene rappresentata in questi termini. E di solito lo studente riceve solo delle informazioni molto sommarie sulla stima della cosiddetta *goodness of fit* dei dati alla funzione ipotizzata, limitata di massima a una delle statistiche che invece, seppur diffusa, andrebbe analizzata con la massima attenzione, e cioè il χ^2.

Non è evidentemente il caso di esporre in qualche dettaglio cosa la verifica dei modelli sia. Ci limitiamo a segnalare come questa disattenzione porti spesso ricercatori anche di gran peso a commettere errori grossolani nella formulazione dei modelli da sottoporre a verifica, il più comune dei quali, come denunciavano Forster e Sober (1994) è la sovraparametrizzazione, e cioè la proposizione di funzioni con un numero di parametri sovrabbondante, che le rende infalsificabili, ma prive di contenuto informativo. E vorremmo sottolineare la straordinaria importanza che hanno nuovi approcci alla verifica dei modelli, come quello proposto una trentina di anni or sono da Akaike (1973; cfr. Bozdogan, 1987), che consentono di verificare i modelli in termini di informazione (secondo la definizione di Kullback e Leibler, 1951), come *distanza* tra i dati osservati e il modello.

Tutto ciò nel mondo della psicologia stenta ad entrare. È molto più facile concepire le ricerche in termini di gruppo sperimentale e gruppo di controllo, o disegni analoghi, e interpretare i dati secondo le procedure rassicuranti della VeSN. Vantaggio non trascurabile, non si è neppure obbligati a capire quel che si sta facendo.

Bibliografia

Aaron, B., Kromrey, J.D. & Ferron, J.M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula.* Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED 433 353).

Abelson, R.P. (1995). *Statistics as Principled Argument.* Hillsdale, NJ: Erlbaum.

Abelson, R.P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*, 12-15.

Agnoli, F. (2002). *Come presentare i risultati di ricerche sperimentali: ambiguità, errori, controversie*. Congresso AIP, Sez. Psicologia Sperimentale, Bari.

Akaike, H. (1973). Information Theory and an extension of the Maximum Likelihood Principle. In B.N. Petrov & F. Csaki (eds), *2nd International Symposium on Information Theory* (267-81). Budapest: Akademiai Kiado.

American Psychological Association (1994*). Publication Manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association (2001*). Publication Manual of the American Psychological Association.* Washington, DC: Author.

Ang, R.P. (1998). Use of Jackknife statistic to evaluate result replicability. *Journal of General Psychology, 125*, 218-228.

Badger, L. (1994). Lazzarini's Lucky Approximation of Pi, *Mathematics Magazine*, *67,* 83-91.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Beckmann, P. (1971). *The History of p*. New York, NY: St. Martin's Press.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-542.

Bertuglia, G.S. & Vaio, F. (2003). *Non linearità, caos, complessità*. Torino: Bollati Boringhieri.

Blalock, H.M. (1960). *Social Statistics.* New York: McGraw-Hill. [tr. it. *Statistica per la ricerca sociale.* Bologna: Il Mulino, 1960]

Bland, J.M. (2000). *An Introduction to Medical Statistics.* New York: Oxford University Press.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni R. Istituto Superiore Scienze Economiche e Commerciali di Firenze*, *8*, 3-62.

Borestein, M., Cohen, J. & Rothstein, H. (1997). *Power and precision*. Dataxiom, Inc., [Online] Available URL: http://www.dataxiom.com.

Box Fisher, J. (1978). *R.A.Fisher: The Life of a Scientist*. New York: Wiley.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Brandstätter, E. (1999). Confidence Intervals as an Alternative to Significance Testing. *Methods of Psychological Research Online, 4, No. 2*: http://www.ipn.uni-kiel.de/mpr/.

Bachmann Christina, Luccio Riccardo, Salvadori Emilia, *La verifica della significatività dell'ipotesi nulla in psicologia*, ISBN 88-8453-226-4,

Buffon, G.L.L. comte de (1777). Sur le jeu de franc carreau. Essai d'arithmétique morale. *Histoire Naturelle, Générale et Particulière, Suppl. 4*, 46-123.

Burstein, H. (1975). Finite population correction for binomial confidence limits. *Journal of American Statistic Association*, *70*, 67-69.

Cahan, S. (2000). Statistical significance is not a "kosher certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results. *Educational Researcher*, *29*, 31-34.

Cannon, A., Howse, J., Hush, D. & Scovel, C. (2002). *Learning with the Neyman/Pearson and min/max criteria. LANL Technical Report: LA/UR/02-2951.* Los Alamos, NM: Los Alamos National Laboratory.

Carver, R.P. (1978). The case against significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.

Caudek, C. & Luccio, R. (2001). *Statistica per psicologi*. Bari, Roma: Laterza.

Chow, S. (1996). *Statistical Significance: Rationale, Validity, and Utility*. London: Sage.

Chow, S. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21*, 169-240.

Chow, S. (1999). In defense of significance tests. Commentary on Krüeger on social bias. *Psycoloqui, 10* (006). http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy? 10.006.

Clipper, C.J. & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of binomial. *Biometrika*, *26*, 404-413.

Cliff, N. & Charlin, V. (1991). Variances and covariances of Kendall's tau and their estimation. *Multivariate Behavioural Research, 26,* 693-707.

Cohen, J. (1962). The statistical power of abnormal social psychology research. *Journal of Abnormal and Social Psychology*, *65*(3), 145-153.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised edition). New York: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*, 155- 159.

Cohen, J. (1992b). Statistical Power Analysis. *Current Directions in Psychological Science, 1,* 98-105.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.

Cooper, H. & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8*, 168-173.

Cooper, H. & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8*, 168-173.

Daniel, L.G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the School*, *5(2)*, 23-32.

Dempster, A.P. (1973). The direct use of likelihood for signicance testing. *Memoirs No. 1, Proceedings of Conference on Foundational Questions in Statistical Inference*, 335-54. Aarhus, Denmark: Barndorff-Nielsen, Blaesild & Schou. [Reprinted in *Statistics and Computing*, 1997, *7,* 247-252].

Dunlap, W.P., Cortina, J.M., Vaslow, J.B. & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170-177.

Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science*, *15*, 131-137.

Edgington, E.S. (1995). *Randomization tests*. New York, NY: M. Dekker.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1-26.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 63*, 589-599.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

Efron, B. (1983). Computer-intensive methods in statistics, *Scientific American*, May, 116-130.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36-48.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Efron, B. & Tibshirani, R.J. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals and Other Measures of Statistical Accuracy, *Statistical Science*, vol. 1, pp. 54- 77.

Ercolani, A.P. & Areni, A. (1983). *Statistica per la ricerca in psicologia*. Bologna: Il Mulino.

Eves, H. (1969). *In Mathematical Circles*. Boston: Prindle, Weber & Schmidt.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83-96.

Falk, R. & Greenbaum, C. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75-98.

Fidler, F. & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61,* 575-604.

Fisher, R.A. (1927). Statistical methods for research workers. *Science Progress, 21,* 340-341.

Fisher, R.A. (1921a). On the mathematical foundations of theoretical statistics. *Philosophical Transactions, A, 222*, 309-368.

Fisher, R.A. (1921b). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3-32.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R.A. (1928[2]). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R.A. (1929). Letter. *Nature,* August 17th, 266-267.

Fisher, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of Royal Statistic Society, B, 17,* 69-78.

Fisher, R.A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Fishman, G.S. (1997). *Monte Carlo: Concepts, Algorithms and Applications*. New York, NY: Springer Verlag,

Fleiss, J.L. (1994). Measures of effect size for categorical data. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York: Sage.

Forster, M. & Sober, E. (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*, 1-35.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70,* 245-251.

Friendly, M. (1991). *SAS macro programs: mpower*. [On-line] Available URL: http://www.math.yorku.ca/SCS/sasmac/mpower.html.

G*Power (2000). [Online]. Available URL: http://www.psychologie.uni-trier.de:8000/projects/gpower.html.

Gauquelin, M. (1955). *L'influence des astres*. Paris: Editions du Dauphin.

Gauquelin, M., Gauquelin, F. & Eysenk, H.J. (1979). Personality and position of planets at birth. *British Journal of Social and Clinical Psychology, 18,* 71-75.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*, 199-200.

Gigerenzer, G. & Krauss, S. (2001). Statistisches Denken oder statistische Rituale? Was sollte man unterrichten? In: M. Borovcnik, J. Engel & D. Wickmann (Hrsg.), *Anregungen zum Stochastikunterricht: Die NCTM-Standards 2000, Klassische und Bayessche Sichtweise im Vergleich*. Franzbecker: Hildesheim, 53-62.

Gigerenzer, G. & Murray, D.J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5***,** 3-8.

Greene, J. & D'Oliveira, M. (1999). *Learning to Use Statistical Tests in Psychology.* New York: McGraw-Hill. [tr. it. *Statistica per psicologi.* Milano: McGraw-Hill Libri Italia, 2000]

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Haase, R.F., Waechter, D.M. & Solomon, G.S. (1982). How significant is a significant difference? Average effect size of research in Counseling Psychology. *Journal of Counseling Psychology, 29,* 58-65.

Haller, H. & Krauss, S. (2002). Misinterpretation of significance: a problem students share with their teachers? *Methods of Psychological Research Online, 17(1),* 1-20.

Hardy, G.H. (1908). Mendelian proportions in a mixed population. *Science, 28,* 49-50

Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds) (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hawking, I. (1990). *The Taming of Chance*. Cambridge: Cambridge University Press

Hays, W.L. (1963). *Statistics for psychologists*. New York: Holt, Rinehart & Winston.

Hays, W.L. (1981). *Statistics for Psychologists.* New York: Holt, Rinehart and Winston.

Hedges, L.V. (1981). Distributional theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6***,** 107-128.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego; Academic Press.

Hinkle, D.E., Wiersma, W. & Jurs, S.G. (1994). *Applied statistics for the behavioral sciences* (3rd ed.). Boston: Houghton Mifflin Company.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-803.

Hubbard, R., & Armstrong, J. S. (1992). Are null results becoming an endangered species in marketing? *Marketing Letters*, *3*, 127-136.

Hubbard, R. & Armstrong, J.S. (1994). Replicationss and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, *11*, 233-248.

Hubbard, R. & Armstrong, J.S. (1997). Publication bias against null results. *Psychological Reports, 80*, 337-338.

Hubbard, R. & Vetter, D.E. (1996). An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. *Journal of Business Research, 35*, 153-164.

Huberty, C.J. (1987). On statistical testing. *Educational Researcher, 16(8),* 4-9.

Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317-333.

Hunter, J.E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*, 3-7.

Hunter, W. (2000). About some misconceptions ant the discontent with statistical tests in psychology. *Methods of Psychological Research Online, 5 (1).*

Kelley, T.L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences, 21***,** 554-559.

Kendall, M. & Gibbons, J.D. (1990). *Rank Correlation Methods*. New York, NY: Oxford University Press.

Kerlinger, F.N. (1979). *Behavioral Research: A conceptual approach.* New York, NY: Holt, Rinehart and Winston.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56(5),* 746-759.

Kotz, S. & Johnson, N.L. (1993). *Process Capability Indexes.* London: Chapman & Hall.

Knuth, D.E. (1981). *Seminumerical Algorithms: The Art of Computer Programming.* Reading, MA: Addison-Wesley.

Kraemer, H.C. & Thiemann S. (1987). *How Many Subjects?* London, UK: Sage Publications.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions,* Chicago: Chicago University Press [trad. it. *La struttura delle rivoluzioni scientifiche.* Torino, Einaudi, 1969].

Kullback, S. & Leibler R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics 22,* 79-86.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*, 635-642.

Lane, D.M. & Quinones, M.A. (1997). Toward resolving the significance testing debate: Electronic publishing and editorial decision making. *12th Annual Conference of the Society for Industrial and Organizational Psychology*, April, 1997. St. Louis, MO.

Laplace, P.S. (1812-1995). *Théorie Analytique des Probabilités*. Reprint in: P.S. Laplace. *Œuvres*, *Tome VII*. Paris: Jacques Gabais.

Lazzerini, M. (1901). Un'applicazione del calcolo della probabilità alla ricerca sperimentale di un valore approsimato di pi. *Periodico di Matematica*, *4*, 140-143

Levin, J.R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

Levin, J.R. & Robinson, D.H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review, 11*, 143-155.

Levin, J.R. & Robinson, D.H. (2000). Rejoinder: Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher, 29*, 34-36

Lundquist, E.F. (1940). *Satatistical Analysis in Educational Research.* Boston: Houghton Mifflin.

Lindsay, R.M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research, 11*, 33-57.

Lipsey, M.W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, California: Sage.

Loftus, G.R. (1996). Psychology will be much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161-170.

Long, J.S. (1983). *Covariance Structure Models*. Newbury Park, CA: Sage.

Long, J.S. & Cliff, N. (2004). Confidence intervals of Kendall's tau. *British Journal of Mathematical and Statistical Psychology, 57,* 31-41.

Lotka, A.J. (1925). *Elements of Physical Biology*. Baltimore: Williams & Wilkins.

Luccio, R. (1982). A prova di leprecauno. Fenomeni paranormali e razionalità scientifica. *Giornale Italiano di Psicologia, 9,* 205-222.

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

Luccio, R. & Salvadori, E. (2002). L'effetto placebo. *Ricerche di Psicologia, 25 (4)*, 165-197.

Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology.* New York, NY: Chapman & Hall.

Maxwell, S.E. & Delaney, H.D. (1990). *Designing experiments and analyzing data. A model comparison perspective*. Belmont, CA: Wadsworth.

McLean, A.L. (2001). Statistics on the catwalk — the importance of models in training researchers in statistics. In C. Batanero (Ed.), *Training Researchers in the Use of* Statistics. Granada, Spain. International Association for Statistical Education.

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist, 15*, 295-300.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In: L. L. Harlow, S.A. Mulaik & J.H. Steiger (Eds), *What if there were no significance tests?,* 393-426. Mahwah, NJ: Erlbaum.

Melton, A. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

Metropolis, N. (1987). The beginning of the Monte Carlo Method. *Los Alamos Science,* No. 15, 125-130.

Metropolis, N. & Ulam, S. (1949). The Monte Carlo method, *Journal of the American Statistical Association*, *44*, 335-341.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.V. & Teller, E. (1953). Equation of state calculations by fast computing machines.*Journal of Chemical Physics, 21*, 1087.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Miller, J. (1991). Reaction time analysis with outliers exclusion: Bias varies with sample size. *Quarterly Journal of Experimental Psychology, 43A*, 907-912.

Milosevic, V.M. (1995). *Teorijska Statistika*. Beograd: Naucna Knjiga.

Mosier, C.I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11*, 5-11.

Mulaik, S. A., Raju, N.S. & Harshman, R.A. (1997). There is a time and a place for significance testing. in: L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds). *What if there were no significance tests?* 65-115. Mahwah, NJ: Lawrence Erlbaum.

Muller, K.E. & Lavange, L.M. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association, 87,* 1209-1216.

NCSS Statistical Software (1999). *PASS.* [On-line] Available: http://www.ncss.com.

Neyman, J. (1950). *First Course in Probability and Statistics*. New York: Holt

Neyman, J. (1981). Egon S Pearson (August 11, 1895-June 12, 1980): An appreciation. *Annals of Statistics, 9,* 1-2.

Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London, A 231*, 289-337.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

Nix, T.W. & Barnette J.J. (1998). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5(2),* 3-14.

Nunnally, J.C. (1960). The place of statistics in psychology. *Education and Psychological Measurement*, *20*, 641-650.

Nunnally, J.C. (1975). *Introduction to Statistics for Psychology and Education.* New York: McGraw-Hill.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* New York: Wiley.

Orsi, R. (1985). *Probabilità e inferenza statistica.* Bologna: Il Mulino.

Ottenbacher, K.J. (1996). The power of replications and replications of power. *American Statistician, 50*, 271-275.

Pearson, E.S. (1926). Review of *Statistical Methods for Research Workers* (R.A. Fisher). *Science Progress*, *20,* 733-734.

Pearson, E.S. [Unsigned] (1929a). Statistics in Biological Research. *Nature,* June 8th, 866-867.

Pearson, E.S. (1929a). Letter. *Nature,* October 19th, 615.

Pearson, E.S. (1966). *The selected papers of E S Pearson.* Berkeley, Calif.

Pearson, E.S. (1990). *'Student', A Statistical Biography of William Sealy Gosset* [Edited and Augmented by R.L. Plackett with the Assistance of G.A. Barnard]. Oxford: University Press.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments? *British Medical Journal, 316*, 1236-1238.

Pisati, M. (2002). Nelle stime non c'è certezza. Uso, abuso e non uso dell'inferenza statistica nella ricerca sociale. *Rassegna Italiana di Sociologia, 63 (1),* 115-141.

Pollard, P. & Richardson, J. (1987). On the probability of making Type I errors. *Psychological Bullettin, 102,* 159-163.

Primi, C. (2002). *Indici di bontà di adattamento nei modelli di equazioni strutturali*. Firenze: Loggia de' Lanzi.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B, 11*, 18-84.

Quételet, L.A.J. (1835). *Sur l'homme et le dévelopment de ses facultés, essai d'une physique sociale.* Paris: Huber.

Reid, C. (1997). *Neyman.* New York, NY.

Rindskopf, D.M. (1997). Testing „small," not null, hypotheses: classical and Bayesian approaches. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds). *What if there were no significance tests?* 319-332. Mahwah, NJ: Lawrence Erlbaum.

Robinson, D. & Levin, J.R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26(5),* 21-26.

Rosenthal, R. & Rosnow, R.L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw Hill.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD and alternative indices. *American Psychologist*, *46*, 1086-1087.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds), *A hand-book for data analysis in the behavioral sciences: Methodological issues,* 519-559. Hillsdale, NJ: Erlbaum.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds), *The handbook of research synthesis,* 231-244. New York, NY: Sage.

Rosenthal, R. & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166-169.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Rosnow, R.L. & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1*, 331-340.

Schafer, J.P. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education*, *61*(4), 383-387.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105* (2), 309-316.

Shaughnessy, J.M. (1983). The psychology of inference and the teaching of probability and statistics: two sides of the same coin? In R.W. Scholz (Ed.), *Decision making under certainty: cognitive decision research, social interaction, development and epistemology.* Amsterdam: North-Holland.

Schuster, E.F. (1974). Buffon's needle experiment. *American Mathematical Monthly, 81,* 26-29

Siegel, S. (1956). *Nonparametric Methods for Behavioral Sciences.* New York, NY: McGarw-Hill.

Simon, J.L. (1969). *Basic Research Methods in Social Science.* New York, NY: Random House. [third edition, with Paul Burstein, 1985].

Simon, J.L. & Holmes, A. (1969). A really new way to teach probability and statistics. *The Mathematics Teacher*, *62,* 283-288.

Simon, J.L., Atkinson, D.T. & Shevokas, C. (1976). Probability and Statistics: Experimental Results of a Radically Different Teaching Method, *The American Mathematical Monthly*, *83*, 733-739.

Simon, S. (1999, January 7). *Re: Type I and Type II error. Educational Statistics Discussion List (EDSTAT-L).* [Online]. Available E-mail: edstat-l@jse.stat.ncsu.edu [1999, January 7].

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605-632.

Snyder, P. & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, *61*(4), 334-349.

Steiger, J.H. & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds), *What if there were no significance tests?* (pp. 221-257). Mahwah, NJ: Erlbaum.

Stegmüller, W. (1973). *"Jenseits von Popper un Carnap": Die logischen Grundlagen des statistchen Schliessens.* Berlin: Springer Verlag.

Sterling, R. (1959). Publications decisions and their possible effects on inferences drawn from tests of significance — or vice versa. *Journal of the American Statistical Association, 54*, 30-34

Sterling, T.D., Rosenbaum, W.L. & Weinkam, J.J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108-112.

Student (1908a). The probable error of a mean. *Biometrika*, *6*, 1-25.

Student (1908b). Probable error of a correlation coefficient. *Biometrika*, *6*, 302-310.

Student [William Sealy Gosset] (1929). Letter. *Nature,* July 20th, 866-867.

Sullivan, J.R. (2000). A Review of Post-1994 Literature on Whether Statistical Significance Tests Should be Banned. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX, January 29, 2000.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25(2),* 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26,* 29-32.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: strong arguments move the field. *Journal of Experimental Education, 70,* 80-93.

Thompson, B. (2002). What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher, 31(3),* 24-31.

Thompson, B. (2003). "Statistica", "pratica", "clinica": quanti tipi di significatività deve considerare chi opera nel counselling? *Bollettino di Psicologia applicata, 240,* 3-13.

Tyler, R.W. (1931). What is statistical significance? *Educational Research Bulletin, 10,* 115-118.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.

Tukey, J.W. (1991). The philosophy of multiple comparison. *Statistical Science, 6,* 100-116.

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

Volterra, V. (1926). Variazioni e fluttuazioni del numero di individui in specie animali conviventi. *Memorie della Regia Accademia Nazionale dei Lincei, Serie VI, 2,* 31-113.

von Mises, R. (1928-1957). *Probability, statistics, and truth.* London: Macmillan.

von Mises, R. (1964). *Mathematical theory of probability and statistics.* New York, NY: Academic Press.

Weinberg, W. (1908). Über den Nachweis der Verebung beim Menschen. *Natur in Wüttemberg, 64,* 368-382.

Welkowitz, J., Ewen, R.B. & Cohen, J. (1982). *Introductory statistics for the behavioral sciences.* San Diego, CA: Harcourt Brace Jovanovich, Publishers.

Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Williams, R.H., Zumbo, B.D., Ross, D. & Zimmerman, D.W. (2003). On the Intellectual Versatility of Karl Pearson. *Human Nature Review, 3,* 296-301.

Wilson, F.D., Smoke, G.L. & Martin, J.D. (1973). The replication problem in sociology: A report and a suggestion. *Sociological Inquiry*, *43*, 141-149.

Wilson, S.A., Becker, L.A. & Tinker, R.H. (1995). Eye movement desensitization and reprocessing (EMDR) treatment for psychologically traumatized individuals. *Journal of Consulting and Clinical Psychology, 63,* 928-937.

Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *Journal of the American Statistical Association*, *46*, 19-34.

Yu, C.H. (2003). Resampling methods: Concepts, applications, and Justifications. *Practical Assessment*, *Research & Evaluation*, *8* (19). [On-line] Available URL: http://PAREonline.net/getvn.asp?v=8&n=19.

Zar, J.H. (1996). *Biostatistical analysis* (3rd Ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Ziliak, S.T. & McCloskey, D.N. (2004). *Size Matters: The Standard Error of Regressions in the "American Economic Review"*. Paper presented at ASSA Meetings, San Diego, California.

Il volume affronta il problema della verifica dell'ipotesi nulla, nel corso degli ultimi anni animatamente dibattuto nel mondo dell'analisi dei dati della ricerca psicologica. Negli anni si è consolidato un paradigma di interpretazione inferenziale che è purtroppo frutto di un ibrido tra due approcci parzialmente incompatibili, che fanno capo rispettivamente da un lato a R. A. Fisher, dall'altro a J. Neyman e E. Pearson. Il volume esamina il costituirsi storico di questo paradigma, gli inconvenienti a cui seguita a dar luogo e indica le principali vie per superare tali inconvenienti. Prende infine in esame il problema dell'insegnamento della statistica ai futuri psicologi.

Christina Bachmann, psicologa, insegna Psicometria presso l'Università degli Studi di Firenze e conduce ricerche sulla probabilità in psicologia.

Riccardo Luccio è ordinario di Tecniche di ricerca e analisi dei dati presso l'Università degli Studi di Firenze.

Emilia Salvadori, dottoranda di Psicologia, conduce ricerche su modelli matematici del rischio.

€ 18,00

ISBN 88-8453-227-2

9 788884 532275