UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

RELAZIONE FINALE

Indice di accordo tra valutazioni: la Kappa di Cohen

Relatore: Prof. Fortunato Pesarin

Firma……

Laureando: Gianluca Toffolo

ANNO ACCADEMICO 2008/2009

INDICE

INTRODUZIONE

Il modo in cui viene effettuata la misurazione qualifica nel complesso l'attività di valutazione: se si appoggia ad un apprezzamento intuitivo delle prestazioni, viene detta intuitiva; se invece opera per ridurre le variabili soggettive si può allora parlare di valutazione oggettiva.

Comunque è impossibile parlare di valutazione oggettiva in assoluto, perché le contaminazioni soggettive esistono se non altro a livello di determinazione dei criteri di misura.

Per misurare occorre uno strumento adatto allo scopo e quindi bisogna intendersi sul concetto di misura.

Perché una misura sia tale, bisogna che sia il risultato del confronto di un dato osservato con una posizione identificabile su una scala; pertanto l'operazione preliminare alla misurazione vera e propria è l'esplicitazione della scala di cui ci si serve.

Il mio studio è atto all'analisi di dati utilizzando un indice di accordo tra valutazioni: *Kappa di Cohen*, che si applica a variabili nominali, volendo a ordinali, (qualitativa e quantitativa), per le quali perde per intero la nozione di distanza, quindi sottopesa le distanze grandi e sovrapesa quelle piccole. Una misura più pesa, più pesa nel disagreement.

Nel primo capitolo ho descritto l'analisi delle frequenze, specificando il test χ^2 che risulta particolarmente utile nella fase iniziale dell'analisi statistica, quando si ricercano le variabili più significative e le relazioni di associazione tra esse e la validità del test χ^2.

Nel secondo capitolo le misure di associazione fondate sul valore del χ^2, ricavato da una tabella di contingenza di dimensioni minime 2 x 2 oppure di dimensioni generiche M x N, i confronto di una proporzione osservata con una attesa, il metodo esatto di Fisher per piccoli campioni e analisi per grandi campioni; utili per comprendere gli sviluppi inferenziali del *Kappa di Cohen*.

Nel terzo capitolo darò una definizione dell'indice *Kappa di Cohen* trattando i vari contenuti che generano questo tipo di indice di accordo tra valutazioni, generalmente ritenuto una misura di concordanza per dati dicotomici. Esso è stato originariamente concepito come una misura tra raters accordo, per la valutazione delle scale psicometriche, ma serve anche per la presenza/assenza di dati microbiologici per l'esame di potabilità ecc. Una volta acquisiti dei dati di natura campionaria può sorgere il problema

di procedere alla verifica d’ipotesi per valutare se i due valutatori sono in accordo statisticamente significativo o meno.
Di seguito ho illustrato alcuni esercizi per il calcolo dell’indice Kappa di Cohen e il significato delle concordanza casuale.

1. ANALISI DELLE FREQUENZE

1.1 CONFRONTO TRA DISTRIBUZIONE OSSERVATE E DISTRIBUZIONI ATTESE

Nella pratica sperimentale, è frequente la necessità di verificare se esiste accordo tra una distribuzione osservata e la corrispondente distribuzione attesa o teorica. Il test viene definito **test per la bontà dell'adattamento.** Sia per dati qualitativi che possono essere classificati in categorie nominali, sia per dati quantitativi distribuiti in classi di frequenza.

È lo scopo per il quale è stato proposto il **test** χ^2 (chi-quadro o chi- quadrato). È uno dei **metodi non parametrici**, con i quali è possibile stabilire se una serie di dati, raccolti in natura od in laboratorio, è in accordo con una specifica ipotesi sulla loro distribuzione o sulla loro frequenza relativa per classi.

Il test χ^2 serve anche per il confronto tra 2 o più distribuzioni osservate. Il suo uso più frequente è per la verifica dell'associazione tra le varie modalità di due o più caratteri qualitativi. Risulta particolarmente utile nella fase iniziale dell'analisi statistica, quando si ricercano le variabili più significative e le relazioni di associazione tra esse.

La prima asserzione, quella della casualità dell'evento, è chiamata **ipotesi nulla** e viene indicata con H_0.

La seconda, quella dell'esistenza di una differenza reale anche se le cause sono ignote, è chiamata **ipotesi alternativa** e viene indicata con H_1.

La scelta tra le due ipotesi avviene sulla base della probabilità stimata con il test. Essa è la probabilità di trovare per caso la distribuzione osservata o una distribuzione che si allontani ancor più da quella attesa, nella condizione che l'ipotesi nulla sia vera. Se la probabilità calcolata è piccola, la logica dell'inferenza statistica rifiuta l'ipotesi nulla, accettando implicitamente l'ipotesi alternativa.

Per affrontare questo problema di inferenza statistica, è possibile ricorrere al test $\chi^2_{(g.d.l.)}$ (chi-quadrato), proposto da **Pearson** nel 1900**.**

Con questo test, le ipotesi sono sulla distribuzione di tassi e proporzioni, ma per la stima della probabilità utilizza le **frequenze assolute**, secondo la formula

$$\chi^2_{(g.d.l.)} = \sum_{i=1}^{n} \frac{(f_i^{oss} - f_i^{att})^2}{f_i^{att}}$$

dove:

- f_i^{oss} = frequenza osservata i-esima;

- f_i^{att} = frequenza attesa i-esima;

- *g.d.l.* = numero di gruppi (n) meno uno (gdl = n-1);

- e la sommatoria Σ è estesa a tutti gli **n** gruppi.

La **distribuzione della densità di probabilità** del χ^2 (*g. d. l.*) dipende dai suoi gradi di libertà, abbreviati in *g.d.l.*. Conteggiati nel calcolo delle frequenze attese, per definizione i gradi di libertà sono il numero di classi che restano indipendenti, conoscendo il numero totale dei dati.

Il numero di *g.d.l.* corrisponde al numero di osservazioni indipendenti e al numero di gruppi meno uno.

Ma quando tra *n* variabili casuali sussistono *k* vincoli lineari, cioè relazioni che riducono il numero di osservazioni indipendenti, i gradi di libertà del corrispondente χ^2 diminuiscono di un numero pari a *k*.

Secondo uno schema valido per tutti i test statistici, il **procedimento** logico che deve essere seguito nell'applicazione del χ^2 comprende diverse fasi, che possono essere riassunte in 7 passaggi:

1 - stabilire l'**ipotesi nulla** (H_0) e l'eventuale **ipotesi alternativa** (H_1);

2 - scegliere **il test** più appropriato per saggiare l'ipotesi nulla H_0, secondo le finalità della ricerca e le caratteristiche statistiche dei dati;

3 - specificare il **livello di significatività** (indicato con α), l'**ampiezza del campione** e i **gradi di libertà**;

4 - trovare la **distribuzione di campionamento** del test statistico nell'ipotesi nulla H_0, di norma fornita da tabelle;

5 - stabilire la zona di rifiuto (che negli esercizi di norma sarà prefissata al 5% indicato con la simbologia α = 0.05);

6 - calcolare il **valore del test statistico** sulla base dei dati sperimentali, stimando la **probabilità** P ad esso associata;

7 - sulla base della probabilità, **trarre le conclusioni**:

- se la probabilità P calcolata risulta superiore a quella α prefissata, concludere che non è possibile rifiutare l'ipotesi nulla H_0;

- se la probabilità P calcolata risulta inferiore a quella α prefissata, rifiutare l'ipotesi nulla e quindi implicitamente accettare l'ipotesi alternativa H_1.

Per la comprensione dell'inferenza statistica con il test chi quadrato, è utile ricordare che quanto più le differenze tra osservato ed atteso sono grandi, tanto più il valore del χ^2 sarà elevato.

- Quindi, la probabilità che tali differenze siano dovute solo al caso sarà bassa e si rifiuterà l'ipotesi nulla, accettando implicitamente l'ipotesi alternativa H_1.

Al contrario, quando le differenze tra osservato ed atteso sono ridotte, ugualmente basso sarà il valore del χ^2;

- Pertanto, sarà elevata la probabilità che esse siano imputabili esclusivamente al caso e si accetterà l'ipotesi nulla H_0.

1.2. CONDIZIONI DI VALIDITA' DEL χ^2

Fissata la probabilità, il valore critico del chi quadrato è totalmente determinato dai suoi gradi di libertà e quindi dal numero di gruppi.

Appare logico pensare che il risultato sia tanto più attendibile quanto più elevato è il numero di osservazioni nell'esperimento.

Nel test χ^2 il numero di osservazioni, sia in totale che entro ogni classe, determina la condizione essenziale di validità.

Il χ^2 è valido solamente quando è applicato a grandi campioni.

Definito il principio, sotto l'aspetto pratico esiste scarsa concordanza su quando un campione possa essere universalmente ritenuto di grandi dimensioni.

Si possono formare 2 classi di "credibilità" o **validità** del test.

1- Il test è valido quando il numero totale di osservazioni è superiore a 100*;*
2- Il test perde ogni attendibilità quando il numero di osservazioni è inferiore a 30. Il motivo è che, con così pochi dati, le variazioni casuali diventano così ampie da non poter mai rifiutare l'ipotesi nulla con una probabilità ragionevolmente bassa, per quanto distanti possano essere le frequenze osservate e quelle attese.

A questa condizione sul numero totale di dati è necessario aggiungerne una seconda:

- il numero di frequenze attese entro ogni classe non deve essere minore di 5.

È quindi utile ricordare che, quando ha un numero abbastanza alto di gradi di libertà, il chi quadrato è meno sensibile agli errori determinati da frequenze attese piccole.

2. LE TABELLE DI CONTINGENZA

2.1. LE TABELLE DI CONTINGENZA 2 X 2

Quando si confrontano le frequenze di risposte binarie in due campioni indipendenti, è utile costruire una tabella a doppia entrata, chiamata **tabella di contingenza.** Per ognuno dei due gruppi, deve essere riportato il conteggio di risposte binarie, quali il numero di successi e quello di insuccessi oppure di quelli che presentano la caratteristica X e di quella alternativa Y.

Il test chi quadrato permette di verificare se le proporzioni di successi e di insuccessi nei due gruppi sono indipendenti dal trattamento al quale sono sottoposti oppure se esiste associazione tra essi.

Per esempio, si supponga di voler verificare se vivere in una zona ad alto inquinamento atmosferico incide sulla frequenza di malattie polmonari. A questo scopo, in una zona con tassi elevati d'inquinamento e in una con livelli molto bassi, sono stati analizzati alcune decine d'individui residenti da alcuni anni, contando quanti sono coloro che presentano malattie polmonari.

DISTRIBUZIONE **OSSERVATA** IN TABELLA 2 X 2

	Con malattie	Senza malattie	Totale
Alto inquinamento	**32 a**	**48 b**	80 n_1
Basso inquinamento	**13 c**	**57 d**	70 n_2
Totale	45 n_3	105 n_4	150 **N**

Nei testi di statistica, non esiste uniformità su come costruire la tabella. La **convenzione** qui seguita è quella proposta da H. **Zeisel,** che riporta

- le due modalità della variabile casuale sulle righe;
- le due modalità della variabile effetto sulle colonne.

Il test chi quadrato utilizza **i casi effettivamente contati, non le frequenze relative o percentuali,** anche se su di esse vengono formulate le ipotesi.

Un'altra convezione, in questo caso generalmente seguita, suggerisce di indicare le frequenze riportate in ognuna delle 4 celle con le lettere minuscole **a**, **b**, **c**, **d**, (con la disposizione utilizzata nella tabella precedente). Il totale generale dei dati è indicato con la lettera maiuscola **N**.

Per comprendere la procedura del chi quadrato in tabelle 2 x 2, è bene seguire alcuni passaggi logici.

1- Se fosse vera l'ipotesi nulla (H_0: vivere in una zona ad alto inquinamento atmosferico non cambia la frequenza di malattie polmonari, rispetto ad una zona a basso inquinamento), la frequenza relativa di persone con malattie polmonari nei 2 gruppi a confronto sarebbe uguale; le differenze riscontrate sarebbero da interpretare come variazioni casuali.

2- La stima migliore di questa frequenza relativa o incidenza percentuale, valida nella condizione che l'ipotesi nulla sia vera, è data dalla somma delle persone con malattie polmonari nei 2 gruppi (a + c cioè 32 + 13 = 45) rapportate al numero totale di persone osservate:

(a + c)/N cioè 45 / 150 = 0,3.

3- Considerando che i due campioni a confronto hanno un numero differente di osservazioni, sempre nel caso che l'ipotesi nulla sia vera, - nel primo campione (che è composto da 80 individui) dovremmo aspettarci di trovare 24 persone (0,3 x 80 = 24) con malattie polmonari e - nel secondo campione (composto da 70 individui) di trovarne 21 (0,3 x 70 = 21).

I quattro valori attesi possono essere presentati in una tabella 2 x 2, come i valori osservati.

Per la sua costruzione, è utile riportare dapprima i 4 totali marginali ed il totale generale.

Successivamente, si calcola ognuno dei 4 valori attesi, moltiplicando il totale di riga per il totale di colonna, diviso per il totale generale:

a = n1 x n3 / N;

b = n1 x n4 / N;

c = n2 x n3 / N

d = n2 x n4 / N

DISTRIBUZIONE **ATTESA** IN TABELLA 2 X 2

	Con malattie	Senza malattie	Totale
Alto inquinamento	**24 a**	**56 b**	80 n_1
Basso inquinamento	**21 c**	**49 d**	70 n_2
Totale	45 n_3	105 n_4	150 **N**

Per stimare l'atteso di ogni casella, noi abbiamo bisogno di 3 informazioni:

- il totale di riga,
- il totale di colonna,
- il totale generale (N).

Poiché i dati sono 4, ne deriva che i gradi di libertà è uno solo (gdl = 4 – 3 = 1).

Colui che propose questo metodo per primo, **Karl Pearson**, attribuì erroneamente un numero maggiore di gradi di libertà. Fu **R. A. Fisher** che mostrò il procedimento esatto.

Stimata la distribuzione attesa nell'ipotesi che sia vera l'ipotesi nulla, dalle differenze tra osservato ed atteso si calcola il valore del chi quadrato, mediante la formula generale già presentata:

$$\chi^2_{(g.d.l.)} = \sum_{i=1}^{n} \frac{(f_i^{oss} - f_i^{att})^2}{f_i^{att}}$$

dove:

- f_i^{oss} = frequenza osservata i-esima
- f_i^{att} = frequenza attesa i-esima

ed estendendo la sommatoria (Σ) ai dati di tutte quattro le caselle.

Con i dati dell'esempio

$$\chi^2_{(1)} = (32-24)^2/24 + (48-56)^2/56 + (13-21)^2/21 + (57-49)^2/49 =$$

$$= 2{,}666 + 1{,}143 + 3{,}048 + 1{,}306 = 8{,}163$$

si ottiene un valore del chi quadrato, con 1 gdl, uguale a 8,163

La tavola sinottica del $\chi^2_{(1)}$ riporta

- il valore critico di **3,84** alla probabilità α **= 0.05** e
- il valore critico di **6,64** alla probabilità α **= 0.01**.

Il valore calcolato (8,163) è superiore sia a quello della probabilità 0.05 che di quella 0.01; di conseguenza, si rifiuta l'ipotesi nulla ed implicitamente si accetta l'ipotesi alternativa.

Questa procedura è utile per capire il reale significato del test χ^2 in tabelle di contingenza 2 x 2. Inoltre, il confronto tra distribuzione osservata e distribuzione attesa mostra in quali caselle si trovano le differenze più importanti. Nell'esempio, tale confronto mostra che le persone con malattie polmonari (riportate nella tabella delle frequenze osservate) sono più frequenti nella zona con maggior inquinamento e sono meno frequenti nella zona senza inquinamento atmosferico, rispetto all'ipotesi nulla che esse abbiano la stessa frequenza percentuale (riportate nella tabella delle frequenze attese).

Si può ottenere lo stesso risultato ed evitare il lungo calcolo delle frequenze attese, con il ricorso alla formula per il calcolo rapido del chi quadrato per le tabelle di contingenza 2 x 2

$$\chi^2_{(1)} = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{n_1 \cdot n_2 \cdot n_3 \cdot n_4}$$

dove, con la simbologia e i valori riportati nella **tabella osservata**

	Con malattie	Senza malattie	Totale
Alto inquinamento	**32 a**	**48 b**	80 n_1
Basso inquinamento	**13 c**	**57 d**	70 n_2
Totale	45 n_3	105 n_4	150 **N**

- a, b, c, d sono le frequenze osservate nei due campioni a confronto;

- n_1**,** n_2**,** n_3**,** n_4 sono i totali marginali;

- N **è il totale generale di osservazioni.**

Il calcolo, con i dati sperimentali dell'esempio precedentemente utilizzato, fornisce

$$\chi^2_{(1)} = \left[(32 \cdot 57 - 48 \cdot 13)^2 \cdot 150\right] / (80 \cdot 70 \cdot 45 \cdot 105) = 8{,}163$$

è un **valore identico** a quello calcolato in precedenza, con la **formula estesa**.

L'**equivalenza tra le due formule** potrebbe essere dimostrata con una serie di passaggi matematici; ma per l'utente della statistica applicata è sufficiente ricordare le due formule, da usare nelle differenti condizioni.

2.2. CONFRONTI TRA FREQUENZE RELATIVE CON LA DISTRIBUZIONE NORMALE

Per il **teorema del limite centrale,** in campioni abbastanza numerosi

- la distribuzione della frequenza relativa π di una popolazione è approssimativamente normale;

- con media campionaria p e deviazione standard della popolazione σ_π (dove $\sigma^2 = p \cdot q$).

L'assunzione rimane valida anche per le percentuali, che tuttavia devono essere trasformate in frequenze relative, per utilizzare le formule proposte.

Questa **approssimazione** della distribuzione chi quadrato alla distribuzione normale non è ritenuta corretta, quando il numero totale di osservazioni **N** è piccolo.

Si ha un uso corretto della distribuzione normale nel confronto tra rapporti, quando N_p e N_q sono entrambi maggiori di 5.

In grandi campioni, se p_1 e p_2 sono le proporzioni osservate di casi con la caratteristica in esame in due campioni indipendenti, è possibile verificare la significatività della loro differenza con **un test Z**:

$$Z = \frac{p_1 - p_2}{\sqrt{p^*(1-p^*) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

dove

- p^* è la proporzione media ponderata dei 2 gruppi a confronto, ottenuta con

$$p^* = \frac{m_1 + m_2}{n_1 + n_2}$$

in cui

- m_1 e m_2 sono i casi positivi nei gruppi **1** e **2** a confronto,

- composti rispettivamente da n_1 e n_2 casi.

Si pone il problema di verificare se le due proporzioni differiscono di una quantità predeterminata π.

La tabella del χ^2 fornisce la probabilità per un **test a due code o bilaterale**. In altri termini, è possibile formulare solo una ipotesi alternativa: le due proporzioni a confronto appartengono a popolazioni differenti. Con i simboli, si scrive

$$H_1 : \pi_1 \neq \pi_2$$

Nel caso di tabelle 2 x 2, con il test chi quadrato è solo possibile dimostrare che le 2 percentuali a confronto sono differenti, quando si è in grado di rifiutare l'ipotesi nulla.

Con la distribuzione normale applicata alle proporzioni o percentuali, sono possibili due diverse impostazioni dell'ipotesi alternativa H_1. E' possibile verificare:

1 - se esiste una differenza nelle frequenze relative tra i due gruppi, senza predeterminare quale dei due debba essere il maggiore (o il minore): si tratta di un **test bilaterale o a due code,** come già per il test χ^2:

$$H_1 : \pi_1 \neq \pi_2$$

2 - se un gruppo ha una frequenza relativa significativamente maggiore (oppure minore): è un **test unilaterale o a una coda:** si confrontano

$$H_0 : \pi_1 \leq \pi_2 \quad \text{contro } H_1 : \pi_1 > \pi_2$$

In ognuno di questi ultimi 2 casi ad una coda, viene a priori rifiutata come non accettabile od illogica la possibilità alternativa a quella proposta.

La distinzione tra test a due code e test a una coda non è solamente una questione di logica. Ha effetti pratici importanti: da essa dipende la distribuzione delle probabilità ed il valore critico per rifiutare l'ipotesi nulla, come chiarisce il grafico.

Scegliendo la probabilità del 5%,

- in un test a due code, si hanno due zone di rifiuto collocate ai due estremi, ognuna con un'area di 2,5%
- in un test a una coda, si ha una sola zona di rifiuto, con un'area di 5 %.

Esistono maggiori probabilità di rifiutare l'ipotesi nulla quando si effettua un test ad una coda, che quando si effettua un test a due code. Anche nella rappresentazione grafica, risulta evidente in modo visivo che, alla stessa probabilità totale, in un test unilaterale il valore critico è minore di quello in un test bilaterale. Come verrà più ampiamente discusso nel capitolo 4, **il test unilaterale è più potente del test bilaterale** (definizione: la potenza di un test è la capacità di rifiutare l'ipotesi nulla quando essa è falsa).

2.3. CONFRONTO DI UNA PROPORZIONE OSSERVATA CON UNA ATTESA: IL TEST Z PER GRANDI CAMPIONI E LA DISTRIBUZIONE BINOMIALE PER PICCOLI CAMPIONI

La distribuzione **Z** permette il confronto tra la proporzione osservata in un singolo esperimento e la corrispondente proporzione attesa o teorica.

La formula può essere derivata da quella già utilizzata per la distribuzione di una osservazione campionaria x rispetto alla media della popolazione μ, quando sia nota la varianza σ^2 della popolazione, attraverso la relazione

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2}}$$

poiché la varianza di una proporzione è totalmente definita dal suo valore medio p e dal numero totale di osservazioni essendo $\sigma^2 = n \cdot p \cdot (1 - p)$.

Nel caso di una proporzione, il **test Z** diventa

$$Z = \frac{x - n \cdot p}{\sqrt{n \cdot p \cdot (1 - p)}}$$

ricordando che:

- p = proporzione attesa o teorica;
- n = numero totale di osservazioni o dati dell'esperimento;
- x = numero di individui osservati con la caratteristica in esame;
- $n \cdot p$ = numero atteso di individui con la caratteristica in esame.

Nel test Z, la distribuzione delle probabilità è simmetrica ed il risultato evidenzia se la differenza è positiva oppure negativa. L'ipotesi alternativa H_1 può essere non solo bilaterale ma anche unilaterale.

2.4. TABELLE DI CONTINGENZA 2 X 2 IN PICCOLI CAMPIONI: IL METODO ESATTO DI FISHER

Il χ^2 è valido solo per grandi campioni. Se il numero di frequenze attese è piccolo, nel caso di tabelle 2 x 2 si deve ricorrere al metodo esatto di Fisher, derivato dalla **distribuzione ipergeometrica**. E' lo stesso principio per cui, nel caso di una sola proporzione e un campione piccolo, si ricorre alla distribuzione binomiale.

Per passare da indicazioni di principio a raccomandazioni pratiche, per la scelta appropriata del test è consigliato utilizzare il metodo esatto di Fisher in sostituzione del chi quadrato quando

- il campione ha un numero totale di osservazioni inferiore a circa 30;
- e/o almeno una frequenza attesa è inferiore a 5.

Sono criteri identici alle raccomandazioni precedenti, che consigliavano di evitare l'uso del χ^2 quando il valore di $n \cdot p$ oppure quello di $n \cdot (1 - p)$ sono inferiori a 5.

Il metodo delle probabilità esatte di Fisher, è di estrema utilità sotto l'aspetto didattico, perché spiega con chiarezza la logica dell'inferenza statistica.

L'uso di questo metodo richiede l'impiego dei fattoriali; di conseguenza, è di semplice e rapida applicazione solo quando il numero di osservazioni è molto piccolo. Il metodo potrebbe essere applicato anche nel caso di campioni di dimensioni medie; ma con un numero più alto di dati, diviene possibile stimare la probabilità solamente con l'uso di calcolatori.

Il metodo permette di stimare la specifica probabilità (P_i) di ottenere una tabella 2 x 2 uguale a quella osservata.

Usando la medesima simbologia dei precedenti paragrafi, riportata nella tabella seguente

	Risposta X	Risposta x	Totale
Campione Y	**a**	**b**	$n_1 = a + b$
Campione y	**c**	**d**	$n_2 = c + d$
Totale	$n_3 = a + c$	$n_4 = a + d$	$N = a + b + c + d$

con la **distribuzione ipergeometrica** la probabilità P_i è calcolata con la formula

$$P_i = \frac{n_1! n_2! n_3! n_4!}{a! b! c! d! N!}$$

Con questa **formula abbreviata**, (abbrevia i tempi richiesti dal calcolo manuale) la probabilità (P_i) di trovare quel particolare insieme dei dati osservati è determinata dal rapporto tra il prodotto dei fattoriali dei quattro totali marginali ed il prodotto dei fattoriali delle quattro frequenze osservate moltiplicato il numero totale di osservazioni.

Il metodo di Fisher si fonda sul concetto che, tenendo fissi i totali, i numeri riportati nelle 4 caselle possano assumere per caso qualsiasi valore. Sulla base di questo presupposto, si può calcolare la probabilità di ottenere ognuna delle risposte possibili.

Per stabilire se esiste una differenza significativa tra le due distribuzioni osservate dei campioni Y e y, non è sufficiente calcolare la probabilità della distribuzione osservata. Come con la precedente distribuzione binomiale, nel caso di metodi esatti si deve stimare la probabilità totale di osservare una combinazione di dati così estrema oppure più estrema.

A questo fine, si riduce di 1 il numero di osservazioni nella casella con il numero minore, modificando i valori delle altre caselle per mantenere uguali i totali marginali; successivamente, si calcola la probabilità di ottenere ognuna di queste risposte. E' necessario elencare tutte le possibili combinazioni delle osservazioni più estreme e quindi calcolare le probabilità esatte associate ad ognuna di queste possibili combinazione dei dati.

Per poter decidere tra le due ipotesi, la probabilità che occorre stimare è data dalla somma della probabilità della distribuzione osservata e di quelle delle risposte più estreme nella stessa direzione.

La probabilità così stimata corrisponde ad un test ad una coda; per un test a due code, si deve moltiplicare per due questa probabilità.
In modo più dettagliato, i passaggi per calcolare la probabilità che **permette di rifiutare l'ipotesi nulla** sono:
1 - calcolare la probabilità associata ai dati osservati;
2 - individuare la casella con il numero minore; se è zero, è sufficiente questa probabilità, perché la risposta osservata è quella più estrema;
3 - se è diverso da zero, ridurre il valore di 1, modificando le frequenze nelle altre tre caselle, in modo che i totali marginali (e quindi quello totale) restino immutati;
4 - calcolare la probabilità associata alla nuova tabella;
5 - ripetere le operazioni 3 e 4, finché il valore minore diventa zero;
6 - per un test ad una coda, sommare tutte queste probabilità;
7 - per un test a due code, moltiplicare per 2 il risultato della precedente operazione 6;
8 - se la probabilità totale calcolata è inferiore al valore di probabilità prefissato come limite critico (di solito 0,05), si rifiuta l'ipotesi nulla H0 ed implicitamente si accetta l'ipotesi alternativa H1, che può essere sia bilaterale che unilaterale.

2.5. LE TABELLE 2 x N CON LA FORMULA GENERALE E QUELLA DI BRANDT-SNEDECOR. LE TABELLE M x N

Il metodo del χ^2 per tabelle 2 x 2, con 1 grado di libertà, può essere esteso al caso generale di tabelle a due entrate, ognuna con classificazioni multiple anziché dicotomiche, con più gradi di libertà. Con l'applicazione dei medesimi concetti ed il ricorso a formule analoghe, è possibile il confronto tra M popolazioni indipendenti, per verificare l'ipotesi nulla che tutte le N percentuali o proporzioni a confronto siano uguali.
Sono le tabelle M x N in cui l'ipotesi nulla è

$$H_0 : \pi_1 = \pi_2 = \pi_3 = ... = \pi_M$$

e l'ipotesi alternativa è

H_1 = almeno una delle π è diversa dalle altre.

Il caso più semplice di tabelle M x N è la tabella di contingenza **2 x N**, per risposte dicotomiche di N gruppi a confronto. Essa ha $N-1$ gradi di libertà, derivati dalla formula generale

$$(N-1)\cdot(2-1)$$

Anche in queste tabelle, è bene evitare di avere caselle con **frequenze teoriche od attese inferiori a 5**, per non avere una eccessiva perdita di potenza del test. Tuttavia, la tolleranza in merito a queste condizioni di validità diviene maggiore: si accettano frequenze attese di 1 o 2, oppure un numero più alto di frequenze uguali a 4-5, poiché le variazioni casuali tendono a compensarsi.

Il χ^2 con parecchi gradi di libertà è meno sensibile agli errori determinati da frequenze attese piccole.

Anche per il calcolo del χ^2 in tabelle 2 x N sono stati proposti procedimenti abbreviati. Una formula frequentemente proposta nei testi di statistica applicata è quella di **Brandt e Snedecor**

$$\chi^2_{(g.d.l.)} = \frac{C \cdot 100}{\overline{p} \cdot (1 - \overline{p})}$$

con C uguale a

$$C = \sum_{i=1}^{k} p_i \cdot n_i - \overline{p} \cdot \sum_{i=1}^{k} n_i$$

e dove

- k = numeri di gruppi a confronto;
- p_i = frequenza percentuale del carattere in esame nel gruppo i;
- n_i = frequenza assoluta del carattere in esame nel gruppo o campione i;
- N = numero totale di osservazioni;
- p = frequenza percentuale media di tutti i gruppi per il carattere in esame.

Nel caso più generale di una tabella di contingenza M x N, il χ^2 è più frequentemente utilizzato come test per l'indipendenza tra i caratteri riportati in riga madre (di norma, i Trattamenti) e quelli riportati nella prima colonna (le Categorie). L'ipotesi nulla è che vi sia indipendenza tra tali variabili, mentre l'ipotesi alternativa bilaterale è che esista associazione.

In molti test di statistica applicata è sconsigliato avere caselle con frequenze attese inferiori a 5. In altri testi, si sostiene che la maggiore robustezza del chi quadrato con più gradi di libertà permette risultati attendibili anche quando si dispone di frequenze minori. Tuttavia, qualora si avessero alcune frequenze molto basse, è bene riunire questi gruppi in un numero inferiore di categorie, aggregando ovviamente in modo logico le variabili che sono tra loro più simili.

In una tabella di contingenza M x N, i gradi di libertà sono:

$$(M-1)\times(N-1)$$

dove **M** è il numero di colonne e **N** è il numero di righe.

Il valore del chi quadrato può essere ottenuto con la formula generale, fondata sullo scarto tra frequenze osservate e frequenze attese.

Anche per le tabelle M x N sono state proposte formule rapide. In realtà, sono metodi più complessi di quelli già illustrati e non presentano vantaggi apprezzabili nel tempo richiesto e nelle approssimazioni dei calcoli, rispetto alla formula generale. Inoltre, nell'interpretazione dei risultati hanno lo svantaggio di evidenziare la differenza complessiva, ma non ogni singola differenza tra la distribuzione attesa e quella osservata.

Quando si analizzano e si interpretano i risultati in tabelle M x N dopo il calcolo del χ^2, se **si è rifiutata l'ipotesi nulla** non è semplice individuare con precisione a quali caselle, a quali associazioni positive o negative, sia imputabile in prevalenza il risultato complessivo. A questo scopo elenco due metodi.

Il più semplice consiste nel riportare in una tabella M x N il contributo al valore del chi quadrato fornito da ogni casella; ma è utile solo per la descrizione. Il secondo si fonda sulla scomposizione e sull'analisi dei singoli gradi di libertà.

Il contributo al valore totale dato da ogni casella è evidenziato riportando per ognuna di essa, in una tabella M x N, il valore del rapporto

$$\left(\frac{f_{i,j}^{oss} - f_{i,j}^{att}}{f_{i,j}^{att}} \right)^2 .$$

La scomposizione dei gradi di libertà di queste tabelle complesse è un altro modo che permette di avere informazioni più dettagliate, sugli effetti di ogni particolare gruppo di dati.

La proprietà additiva del χ^2 e dei relativi gradi di libertà consente la scomposizione di una tabella M x N in tanti test 2 x 2 , ognuno con 1 *g.d.l.*, quanti sono i gradi di libertà totali della matrice.

Quando si è interessati ad individuare la causa di una significativa deviazione dall'ipotesi nulla, è possibile costruire i test che ne spiegano le quote maggiori.

Prendendo come schema di riferimento una teorica tabella **3 x 3** con la relativa simbologia

	TRATT. 1	TRATT. 2	TRATT. 3	Totale
Blocco A	a_1	a_2	a_3	n_1
Blocco B	b_1	b_2	b_3	n_2
Blocco C	c_1	c_2	c_3	n_3
Totale	n_4	n_5	n_6	N

con 9 dati si ottiene un χ^2 che ha 4 gradi di libertà. Se risulta significativo, è utile scomporre questa valutazione globale, per conoscere quali confronti singoli **2 x 2** siano la causa di questa differenza tra frequenze osservate e frequenze attese.

Con 4 gradi di libertà è possibile fare solamente 4 confronti. Se impostati correttamente, la somma dei valori di questi 4 $\chi^2_{(1)}$ con 1 *g.d.l*. deve essere uguale al valore complessivo del $\chi^2_{(4)}$ con 4 *g.d.l*. calcolato su tutti i dati.

La ripartizione deve essere eseguita in modo gerarchico, stabilita una prima suddivisione, le ripartizioni successive devono essere attuate sempre all'interno della precedente. È il modo per rendere i confronti ortogonali, la conclusione precedente non deve dare informazioni sul test successivo.

Con la tabella 3 x 3 presentata, una possibile partizione dei 4 gradi di libertà è quella di seguito riportata:

1)

a_1	a_2
b_1	b_2

2)

$a_1 + a_2$	a_3
$b_1 + b_2$	b_3

3)

$a_1 + b_1$	$a_2 + b_2$
c_1	c_2

4)

$a_1 + a_2 + b_1 + b_2$	$a_3 + b_3$
$c_1 + c_2$	c_3

Anche dalla semplice osservazione risulta evidente che esistono molte possibilità differenti di suddivisione della medesima tabella.

La scelta dipende dal ricercatore, che è totalmente libero di scegliere i raggruppamenti di caselle che gli sembrano più logici ed utili per spiegare la significatività ottenuta; ma tale scelta deve essere fatta "a priori" non "a posteriori", per non alterare la probabilità di scegliere una distribuzione casualmente significativa. Scelta a priori significa che essa

deve essere fatta in modo totalmente indipendente dai dati rilevati; non è corretto individuare quali gruppi hanno le frequenze maggiori e quali le frequenze minori e successivamente pianificare la suddivisione, sulla base delle differenze osservate, scegliendo quelle che danno valori del chi quadrato maggiori.

2.6. CLASSIFICAZIONE DEI COEFFICIENTI D'ASSOCIAZIONE O D'INDIPENDENZA

Quando i dati sono classificati sulla base di due variabili **categoriali** o **qualitative**, le **frequenze** sono riportate in una **tabella di contingenza.**

Di solito si utilizzano **frequenze assolute**, sia per facilitare i calcoli, sia perché le dimensioni del campione hanno un effetto rilevante sulla significatività del test e quindi è conveniente conoscerle esattamente. Ma è possibile utilizzare anche le **frequenze relative**, in particolare quando si vuole facilitare il confronto tra due o più rilevazioni, che ovviamente solo di rado hanno campioni con lo stesso numero di osservazioni.

Le tabelle hanno dimensioni minime **2 x 2**; ma possono essere molto più ampie, indicate genericamente con **M x N** (**M** righe x **N** colonne).

I valori che quantificano le relazioni tra le due variabili qualitative sono chiamati **coefficienti di associazione**; si parla di **correlazione,** quando le variabili sono quantitative.

Il test del χ^2 serve per verificare **le ipotesi sulla indipendenza** (corrispondente a una **associazione nulla**),

- tra le modalità della variabile riportata nelle **righe;**
- e le modalità della variabile riportata nelle **colonne**.

È prassi che la dimensione delle righe, per analogia con l'asse delle ascisse nella regressione, corrisponda alla variabile classificatoria che dovrebbe essere **esplicativa** (come la dose di un farmaco oppure la località nella quale si è raccolto un campione di alcune specie animali o vegetali) e l'altra dimensione, quella delle colonne, sia una **risposta o variabile dipendente** (come l'effetto del farmaco che può essere nullo, moderato o forte oppure le varie specie raccolte), analogamente all'asse delle ordinate.

Per le due variabili, i gruppi possono essere formati sulla base di **dati misurati su scale differenti**:

1 - **qualitativi o nominali**, come l'elenco delle località e quello delle specie;

2 - **ordinali o di rango**, come l'intensità della risposta al farmaco (nulla, moderata, forte) o la classificazione delle specie in classi d'età (giovani, adulti, vecchi) o livelli di sviluppo;

3 - **di intervalli e/o di rapporti** (come l'età o le dimensioni) raggruppati in classi, con intervalli differenti oppure costanti (nelle tabelle di contingenza, di solito non sono fatte distinzioni tra questi due tipi di scala, per i quali possono essere applicati i test parametrici).

Da queste tre classificazioni del tipo delle due variabili, derivano **tabelle a due entrate** che utilizzano scale differenti, quali

- nominale per ambedue le variabili;
- nominale per una e ordinale per l'altra;
- ordinale per ambedue le variabili;
- nominale per una e intervallare per l'altra;
- in tutte le combinazioni di scala possibili, fino a intervallare per entrambe.

Non esiste una misura ideale dell'associazione o concordanza tra le due variabili**,** che sia valida per tutte le situazioni.

Una classificazione utile per ordinare la presentazione degli indici più frequentemente utilizzati, propone

- una suddivisione per misure **nominali**, **ordinali** e in **classi d'intervalli**;
- abbinata a quelle delle dimensioni in **tabelle 2 x 2** e in **tabelle M x N.**
- per vari indici non esiste una differenza determinata dalle dimensioni della tabella, in quanto l'indice valido per tabelle M x N molto spesso è solo una generalizzazione dell'indice proposto per la tabella **2 x 2**.

3. IL KAPPA DI COHEN:

3.1. STIMA DELL'ACCORDO (AGREEMENT) TRA DUE VALUTAZIONI CON SCALA NOMINALE.

Le misure del **grado di associazione**, la cui significatività è ottenuta con il test χ^2, fa riferimento a **due variabili**. Ad esempio, nelle tabelle 2 x 2 col χ^2 si è valutato il grado di associazione tra livello di inquinamento (alto o basso) di un'area e la presenza di persone residenti con malattie polmonari (si o no).

In altre situazioni, si utilizza **una sola variabile** per valutare il **grado di accordo** tra **due valutatori**. Ad esempio, in medicina può essere interessante verificare se due chirurghi che decidono sulla necessità di operare forniscono risposte concordanti; nella ricerca ambientale, se due commissioni che agiscono in modo indipendente approvano o respingono gli stessi progetti.

Un problema identico si pone anche per lo **stesso valutatore**, quando agisce in **due momenti differenti**. Ad esempio, se lo stesso chirurgo fornisce o meno la medesima risposta sulla necessità di un intervento chirurgico prima e dopo aver preso visione di una nuova analisi clinica; se un ricercatore, di fronte agli stessi soggetti in due momenti differenti, fornisce la stessa classificazione.

In una visione più generale, il problema è importante tutte le volte in cui si confrontano due o più distribuzioni di frequenza. L'appartenenza degli esperti a scuole con impostazioni culturali differenti e la diversa esperienza dei ricercatori possono determinare classificazioni anche notevolmente discordanti, per effettuare correttamente test sulla similarità della distribuzione. Ad esempio, con una tabella 2 x 2 oppure a più dimensioni (M x N) spesso si vuole valutare se M specie hanno la stessa distribuzione nelle N aree campionate. Ma tale analisi come condizione di validità richiede necessariamente che la classificazione delle specie abbia seguito gli stessi criteri. In altri termini, che la classificazione sia riproducibile, che i criteri utilizzati siano affidabili.

Il problema non è valutare quale delle due classificazioni sia quella corretta o la migliore; è una domanda alla quale è possibile rispondere con una impostazione logica e con metodi differenti.

Il kappa di Cohen è una misura dell'**accordo** (*coefficient of agreement)* tra le **risposte qualitative o categoriali** di due persone (*inter-observer variation*) oppure della medesima persona in momenti differenti (*intra-observer variation*), valutando gli **stessi oggetti**.

La metodologia è stata presentata da Jacob **Cohen** (nel 1960).

Prendendo in considerazione una situazione caratteristica della ricerca psicologica, si supponga che due medici abbiano analizzato separatamente e in modo indipendente il comportamento delle stesse 200 persone, classificandole in tre differenti tipologie nominali (A = disordini della personalità, B = neurosi, C = psicosi), con i seguenti risultati:

		Medico 1			
	Categorie	A	B	C	Totale
Medico 2	A	50	26	24	100
	B	24	4	32	60
	C	6	30	4	40
	Totale	80	60	60	200

Si tratta di valutare se i giudizi forniti dai **due esperti** sono **riproducibili**, **affidabili**; in altri termini, si chiede di determinare il **grado**, la **significatività** e la **stabilità** campionaria del loro **accordo.**

Per il **coefficiente di concordanza,** devono essere realizzate le seguenti **condizioni di validità**:

1 - le unità (in questo caso i 200 soggetti analizzati) sono indipendenti;

2 -le categorie della scala nominale sono indipendenti, mutuamente esclusive e esaustive;

3 - i giudici operano in modo indipendente.

Queste assunzioni ne implicano altre due:

4 - i due giudici hanno lo stesso livello di competenza;

5 - non esistono restrizioni nell'attribuzione alle categorie.

Per entrare nella logica del coefficiente, è importante comprendere che se la classificazione dei pazienti fosse effettuata su criteri indipendenti, cioè se le due serie di attribuzioni fossero realizzate in modo puramente casuale, si avrebbe ugualmente un certo numero di giudizi coincidenti: un paziente potrebbe essere attribuito alla stessa categoria, per solo effetto del caso. Per meglio illustrare il concetto di concordanza e evidenziare la logica che porta a ricavare l'indice **k** proposto da **Cohen**, è vantaggioso utilizzare le proporzioni riportate nella tabella successiva. Esse sono semplicemente la trasformazione in frequenze relative (con totale uguale a 1,0) delle frequenze assolute precedenti (con totale uguale a 200)

	Categorie	Medico 1			
		A	B	C	Totale
Medico 2	A	**0,25** *(0,20)*	**0,13** *(0,15)*	**0,12** *(0,15)*	0,50
	B	**0,12** *(0,12)*	**0,02** *(0,09)*	**0,16** *(0,09)*	0,30
	C	**0,03** *(0,08)*	**0,15** *(0,06)*	**0,02** *(0,06)*	0,20
	Totale	0,40	0,30	0,30	1,00

Entro ogni casella,

- in grassetto sono riportate le **proporzioni osservate** (p_o da **o**bserved); ad esempio, nella casella 1,1 si ha **0,25 = 50/200** (presi dalla tabella precedente con le frequenze assolute);

- in corsivo quelle **attese** (p_e da **e**xpected), nella condizione che l'ipotesi nulla sia vera, cioè che l'attribuzione dell'individuo alla categoria sia stata casuale; ad esempio sempre nella 1,1 si ha *0,20* **= 0,4 x 0,5** (totali marginali presi da questa ultima tabella di frequenze relative)**.**

Come nelle tabelle del chi quadrato, le **proporzioni attese** entro ogni casella sono date dai prodotti delle proporzioni marginali.

Si tratta di valutare quanto differiscono le classificazioni effettuate dai due medici.

Prima di Jacob Cohen, era seguita la procedura proposta nel 1950 da J. P. **Guilford.** In esso si ricorre al χ^2, per saggiare la significatività, e al coefficiente di contingenza **C** di Pearson, per ricavare una misura dell'accordo che sia più facilmente valutabile, cioè indipendente dalle dimensioni del campione. Con i dati dell'esempio:

- per ottenere il χ^2 mediante la formula applicata alle proporzioni

Si calcolava

$$\chi^2 = \sum \frac{(p_o - p_e)^2}{p_e} \cdot N = 64{,}59$$

con 4 gdl

- per C di Pearson si calcolava

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{64{,}59}{64{,}59 + 200}} = \sqrt{0{,}244} = 0{,}494$$

Jacob **Cohen** contesta questo metodo.

Il risultato del χ^2 è altamente significativo (infatti il valore critico del χ^2 con **4** gdl e α **= 0.001** è **18,467**), quindi si allontana dall'ipotesi di distribuzione casuale.

In realtà, egli scrive, è semplice dimostrare che l'uso del χ^2 e quindi del **C** fondato su di esso sono logicamente indifendibili, come misura dell'accordo.

Quando è applicato a una tabella di contingenza, il test χ^2 serve per

- verificare l'ipotesi nulla rispetto all'associazione, non alla concordanza (anche se la distribuzione dell'ipotesi nulla è calcolata nello stesso modo). Infatti, come nel caso dell'esempio, sul valore totale $\chi^2_{(4)} = 64{,}59$ il contributo maggiore è dato dalla casella **3,2** con:

$$\frac{(0{,}15 - 0{,}06)^2}{0{,}06} \cdot 200 = 27{,}00$$

un χ^2 parziale uguale a **27,00**.

Questo valore così alto **non dipende dall'accordo** tra i due medici, ma dal fatto opposto: essi hanno fornito una classificazione differente degli stessi pazienti (cioè la malattia B per il medico 1 e la malattia C per il medico 2) e in misura maggiore dell'atteso, cioè delle frequenze fondate sull'ipotesi nulla di casualità. Quindi il valore ottenuto risulta elevato, non perché i due medici concordano, ma perché essi non concordano. Più in generale, il valore del χ^2 **misura se due distribuzioni qualitative sono associate** (non importa se in modo positivo o negativo, trattandosi di valori elevati al quadrato), ma senza fornire la direzione dell'accordo, che è l'aspetto fondamentale e specifico di questa valutazione della concordanza.

Come conclusione dei concetti precedenti, si deduce che una misura dell'accordo tra le due distribuzioni può essere ricavata:

- dalla differenza tra la proporzione osservata dei giudizi che sono effettivamente coincidenti e la proporzione di quelli attesi nell'ipotesi di totale casualità dei giudizi (H_0 vera);
- rapportata a quella della non associazione attesa.

La formula proposta da **Cohen standardizza la differenza** tra proporzione totale osservata e proporzione totale attesa**,** dividendola per la massima differenza possibile non casuale.

Nelle ultime due tabelle dei dati, l'informazione utile è fornita dalle frequenze collocate lungo la diagonale principale (nella tabella 3 x 3, le caselle 1,1; 2,2; 3,3).

Nel caso dell'esempio, con **le proporzioni** la somma della diagonale principale

- 0,25 + 0,02 + 0,02 = 0,29 è la **proporzione** totale **osservata** p_o = 0,29

- 0,20 + 0,09 + 0,06 = 0,35 è la **proporzione** totale **attesa** p_e = 0,35.

L'**indice k** proposto da Cohen è:

$$k = \frac{p_o - p_e}{1 - p_e} = \frac{0{,}29 - 0{,}35}{1 - 0{,}35} = \frac{-0{,}06}{0{,}65} = -0{,}0923$$

Con **le frequenze assolute**, sovente è possibile una stima più semplice e rapida.

Dopo aver calcolato

- le frequenze osservate f_o = 50 + 4 + 4 = 58 (nella prima tabella)

- e quelle attese f_e = 40 + 18 + 12 = 70 (nella tabella sottostante)

	Categorie	Medico 1			
		A	B	C	Totale
Medico 2	A	***40***	***30***	***30***	100
	B	***24***	***18***	***18***	60
	C	***16***	***12***	***12***	40
	Totale	80	60	60	200

utilizzando appunto solo i valori collocati sulla diagonale principale, il calcolo dell'**indice k** diventa:

$$k = \frac{f_o - f_e}{N - f_e} = \frac{58 - 70}{200 - 70} = \frac{-12}{130} = -0{,}0923$$

Con entrambe le formule, il valore dell'accordo risulta **k = -0,09**. In questo caso, è un valore **negativo.** Esso indica che i due medici si trovano d'accordo su una proporzione di casi che è minore di quella che si sarebbe ottenuta con una attribuzione casuale dei pazienti alle varie categorie. In conclusione, i due medici forniscono valutazioni tendenzialmente discordanti (anche se per una piccola quantità).

Il valore di **k** teoricamente può variare tra **– 1** e **+ 1**. In realtà **l'indice k** ha **significato** solo quando è **positivo**.

Da questa osservazione derivano due conseguenze:

1 - la sua **significatività** deve essere verificata mediante **il test unilaterale:**

H_0 **: k ≤ 0** contro H_1 **: k > 0**

2 - il valore **massimo teorico** è **k** = **+1,0.**

Questa ultima affermazione è vera, cioè si può ottenere **k = +1,** solamente quando sono realizzate contemporaneamente le seguenti due condizioni:

1 - tutte le frequenze osservate non collocate sulla diagonale, cioè quelle che indicano il disaccordo (*disagreement*), sono 0.

2 - i totali marginali dei due valutatori (cioè i totali delle righe e quelli delle colonne) sono identici.

Infatti essi indicano che i due valutatori hanno trovato le stesse proporzioni delle categorie utilizzate. Nella tabella con le proporzioni fino ad ora utilizzata, le frequenze marginali dei due medici sono differenti, esattamente quelle riportate nella tabella sottostante (per il medico 1 esse sono 0,40, 0,30, 0,30; per il medico 2 sono 0,50, 0,30, 0,20)

Medico	Categorie		
	A	B	C
1	0,40	0,30	0,30
2	0,50	0,30	0,20
Minimi	***0,40***	***0,30***	***0,20***

A causa di questa differenza nei totali marginali, il **k massimo** (k_M) ottenibile con la formula precedente non potrà mai essere **k = +1,00** ma un valore inferiore. Tale valore massimo possibile può essere ricavato con alcuni passaggi:

1) confrontare i **singoli totali marginali** (prime due righe della tabella) e per ogni categoria **scegliere il valore minore** (terza riga in grassetto e corsivo),

2) calcolare p_{oM}, la **proporzione osservata massima**, utilizzando la somma di queste proporzioni minime:

$$p_{oM} = 0{,}40 + 0{,}30 + 0{,}20 = 0{,}90$$

3) calcolare il **k massimo** (k_M) con

$$k_M = \frac{p_{oM} - p_e}{1 - p_e}$$

con i dati dell'esempio, dove

- p_{oM} = 0,90
- p_e = 0,35

mediante

$$k_M = \frac{p_{oM} - p_e}{1 - p_e} = \frac{0{,}90 - 0{,}35}{1 - 0{,}35} = \frac{0{,}55}{0{,}65} = 0{,}846$$

si ricava che il valore massimo possibile di **k**, é k_M **= 0,846**.

E' una conseguenza del fatto che i due valutatori forniscono una classificazione differente degli stessi soggetti, poiché per le categorie in oggetto essi "vedono" frequenze differenti nella stessa popolazione.

Da questa prima analisi sul k_M può derivare un primo effetto.

Per ottenere ricerche più attendibili, dove k_M sia **1**, sarebbe vantaggioso fornire indicazioni più vincolanti ai due valutatori, con una preparazione preliminare più accurata e precisa tramite anche la frequenza ad appositi corsi. Dopo il corso, valutare nello stesso modo se il k_M è migliorato.

Una seconda conseguenza potrebbe essere quella di calcolare un valore di **k corretto** (k_C), attraverso la relazione

$$k_C = \frac{k}{k_M}$$

in modo che il valore massimo raggiungibile sia sempre 1 e quindi sia la scala di valutazione sia i confronti siano omogenei.

Ma Cohen sconsiglia tale trasformazione, che nel ragionamento precedente appariva logica e razionale, con la motivazione che se i totali marginali sono differenti è perché i due valutatori hanno fornito effettivamente risposte differenti. Quindi esiste un reale **non-accordo nella valutazione**, che giustamente è **compreso nell'indice k calcolato** senza la correzione.

Nella presentazione di questo metodo, dopo la illustrazione

a) del **significato di k**,

b) del calcolo del valore k

c) e di quello **massimo possibile** (k_M),

si pongono altri tre problemi:

d) stimare l'**intervallo di confidenza di k**,

e) valutare la **significatività statistica e il significato disciplinare del risultato**, cioè del valore di **k** ottenuto,

f) testare la **significatività della differenza** tra **due valori** di **k.**

Nel caso di **grandi campioni** (N ≥100), per calcolare **l'intervallo di confidenza di k** secondo Cohen è possibile il ricorso alla distribuzione normale standardizzata,

$$k \pm Z_{a/2} \cdot \sigma_k$$

dove σ_k è un **errore standard** (pure essendo indicato come una deviazione standard) in quanto **k è una media**.

Il valore di σ_k può esser calcolato utilizzando

- sia le **frequenze relative** o **proporzioni**

$$\sigma_k = \sqrt{\frac{p_o \cdot (1 - p_o)}{N \cdot (1 - p_e)^2}}$$

- sia le **frequenze assolute**

$$\sigma_k = \sqrt{\frac{f_o \cdot (N - f_o)}{N \cdot (N - f_e)^2}} = \frac{\sqrt{f_o \cdot \left(1 - \frac{f_o}{N}\right)}}{N - f_e}$$

I **limiti di confidenza di kappa** sono compresi

- con probabilità del 95% tra

$$k \pm 1{,}96 \cdot \sigma_k$$

- con probabilità del 99% tra

$$k \pm 2{,}58 \cdot \sigma_k$$

Utilizzando i dati dell'esempio,

- sia mediante la tabella delle frequenze relative o proporzioni, dove p_o = 0,29 e p_e = 0,35 e **N = 200,**

$$\sigma_k = \sqrt{\frac{0{,}29 \cdot (1 - 0{,}29)}{200 \cdot (1 - 0{,}35)^2}} = \sqrt{\frac{0{,}2059}{84{,}5}} = 0{,}0494$$

- sia mediante la tabella delle frequenze assolute, dove f_o = 58 e f_e = 70 e **N = 200,**

$$\sigma_k = \sqrt{\frac{58 \cdot (200 - 58)}{200 \cdot (200 - 70)^2}} = \sqrt{\frac{8{,}236}{3.380.000}} = 0{,}0494$$

si ottiene σ_k = 0,0494.

Poiché il valore sperimentale ricavato è k = - 0,09, alla probabilità del 95% **il valore reale di k** è compreso

$$-0{,}09 \pm 1{,}96 \cdot 0{,}0494$$

tra il valore minimo = - 0,138 (–0,09 – 0,048)

e il valore massimo = - 0,042 (–0,09 + 0,0489).

Per la **significatività statistica di k**, teoricamente per valutare l'ipotesi nulla H_0**: k = 0** che è ottenibile quando p_o **=** p_e, la formula dell'errore standard σ_{k0}:

- con le **frequenze relative** diventa

$$\sigma_{k0} = \sqrt{\frac{p_e}{N \cdot (1 - p_e)}}$$

- con le **frequenze assolute** diventa

$$\sigma_{k0} = \sqrt{\frac{f_e}{N \cdot (1 - f_e)}}$$

Con i dati dell'esempio,

- sia mediante la tabella delle **frequenze relative** o **proporzioni**, dove p_e = 0,35 e **N = 200,**

$$\sigma_{k0} = \sqrt{\frac{0{,}35}{200 \cdot (1 - 0{,}35)}} = \sqrt{\frac{0{,}35}{130}} = 0{,}0519$$

- sia mediante la tabella delle **frequenze assolute**, dove f_e = 70 e **N = 200,**

$$\sigma_{k0} = \sqrt{\frac{70}{200 \cdot (200 - 70)}} = \sqrt{\frac{70}{26.000}} = 0{,}0519$$

si ottiene σ_{k0} = 0,0519.

Nella significatività di un **k** sperimentale, per la sua rilevanza pratica ai fini della potenza del test e un approccio teoricamente più corretto, è importante ricordare un concetto già evidenziato. Benché, in un esperimento reale, il valore di **k** possa variare tra **–1** e **+1,** quasi sempre nella ricerca si vuole valutare **se esiste un accordo** significativo.

Pertanto **in realtà il test è unilaterale** con ipotesi

H_0**: k ≤ 0** contro H_1**: k > 0**

Sempre Cohen, per il test di significatività con **grandi campioni** (**N ≥100**) e come quasi sempre avviene quando si utilizzano tabelle di dimensioni superiori a 2 x 2, propone il ricorso alla distribuzione normale standardizzata

$$Z = \frac{k}{\sigma_{k0}}$$

Nel caso dell'esempio, il valore di **k** è risultato negativo (**k = -0,09**).

Di conseguenza, non ha senso verificare se è maggiore di zero (cioè H_1**: k > 0)**, cioè se esiste un accordo che sia contemporaneamente **positivo** e **significativo**, tra i due medici nella classificazione da essi effettuata per gli stessi pazienti.

Nelle due formule dell'**errore standard** (σ_k e σ_{k0}), utili

- il primo (σ_k) per l'**intervallo di confidenza**

- il secondo (σ_{k0}) per la **significatività di k**,

si evidenzia che il numero totale di osservazioni (N), ha un ruolo importante. Ne deriva che, come in quasi tutti i test, con grandi campioni anche un valore di **k** piccolo può risultare significativo, mentre con un campione piccolo anche un valore grande di **k** può non essere statisticamente significativo.

Per ottenere una **interpretazione univoca e adimensionale** di **k** come stima di **Agreement** o **Reproducibility**, sono state proposte griglie di valutazione.
Nella tabella successiva, sono riportate le due più frequentemente utilizzate.

Kappa	**Agreement**
< 0.00	Nessun accordo
0.00-0.20	Lieve accordo
0.21-0.40	Accordo equo
0.41-0.60	Moderato accordo
0.61-0.80	Sostanziale accordo
0.81-1.00	Quasi perfetto accordo

Kappa	**Reproducibility**
> 0.75	Excellent
$0.40 \leq k \leq 0.75$	Good
$0.00 \leq k < 0.40$	Marginal

La prima, a sinistra e più dettagliata, è stata proposta da J. Richard **Landis** e Gary G. **Koch** del 1977.
La seconda, riportata a destra, è stata proposta da Joseph L. **Fleiss** nel suo testo del 1981.

3.2. ESEMPIO 1. Valutare il grado di accordo tra due giudici nella seguente tabella 3 x 3 (tra parentesi e in grassetto sono evidenziate le frequenze attese e quelle osservate limitatamente alla diagonale, in quanto sono le uniche informazioni utili).

		Giudice A			
		1	2	3	Totale
Giudice B	1	**88 *(60)***	14	18	120
	2	10	**40 *(18)***	10	60
	3	2	6	**12 *(4)***	20
	Totale	100	60	40	N=200

Risposta. Dopo aver ricavato

- le frequenze osservate f_o = 88 + 40 + 12 = 140

- le frequenze attese f_e = 60 + 18 + 4 = 82

è semplice osservare che in questo caso esiste un accordo maggiore di quello possibile per solo effetto del caso.

Dalle frequenze si ricava il valore di **k**

$$k = \frac{f_o - f_e}{N - f_e} = \frac{140 - 82}{200 - 82} = 0{,}492$$

che risulta **k = 0,492**.

Per il calcolo dell'**intervallo di confidenza** si calcola

$$\sigma_k = \sqrt{\frac{f_o \cdot (N - f_o)}{N \cdot (N - f_e)^2}} = \sqrt{\frac{140 \cdot (200 - 140)}{200 \cdot (200 - 82)^2}} = \sqrt{\frac{8.400}{2.784.800}} = 0{,}0549$$

l'**errore standard σ_k = 0,0549**.

Per il test che verifica la **significatività dell'accordo** si calcola

$$\sigma_k = \sqrt{\frac{f_e}{N \cdot (N - f_e)}} = \sqrt{\frac{82}{200 \cdot (200 - 82)}} = \sqrt{\frac{82}{23.600}} = 0{,}0589$$

l'**errore standard σ_k = 0,0589**.

Questi stessi risultati possono essere ottenuti con la tabella delle frequenze relative o proporzioni

		Giudice A			
		1	2	3	Totale
Giudice B	1	**0,44 *(0,30)***	0,07	0,09	0,60
	2	0,05	**0,20 *(0,09)***	0,05	0,30
	3	0,01	0,03	**0,06 *(0,02)***	0,10
	Totale	0,50	0,30	0,20	1,00

sempre ricordando che N = 200.

Dopo aver ricavato

- le frequenze relative osservate p_o = 0,44 + 0,20 + 0,06 = 0,70

- le frequenze relative attese p_e = 0,30 + 0,09 + 0,02 = 0,41

si calcola il valore di **k**

$$k = \frac{p_o - p_e}{1 - p_e} = \frac{0{,}70 - 0{,}41}{1 - 0{,}41} = \frac{0{,}29}{0{,}59} = 0{,}492$$

che risulta **k = 0,492**.

Con le frequenze relative, può essere utile calcolare il valore k_M

$$k_M = \frac{p_{oM} - p_e}{1 - p_e} = \frac{(0{,}50 + 0{,}30 + 0{,}10) - 0{,}41}{1 - 0{,}41} = \frac{0{,}49}{0{,}59} = 0{,}831$$

Per il calcolo dell'**intervallo di confidenza** si calcola

$$\sigma_k = \sqrt{\frac{p_o \cdot (1 - p_o)}{N \cdot (1 - p_e)^2}} = \sqrt{\frac{0{,}70 \cdot (1 - 0{,}70)}{200 \cdot (1 - 0{,}41)^2}} = \sqrt{\frac{0{,}21}{69{,}62}} = 0{,}0549$$

l'**errore standard σ_k = 0,0549.**

Per il test che verifica la **significatività dell'accordo** si calcola

$$\sigma_{k0} = \sqrt{\frac{p_e}{N \cdot (1 - p_e)^2}} = \sqrt{\frac{0{,}41}{200 \cdot (1 - 0{,}41)}} = \sqrt{\frac{0{,}41}{118}} = 0{,}0589$$

l'**errore standard σ_{k0} = 0,0589.**

Con k = 0,492 e σ_k = 0,0549 si ottiene l'**intervallo di confidenza**.

Alla probabilità del 95% esso è compreso

$$k \pm Z_{a/2} \cdot \sigma_k = 0{,}492 \pm 1{,}96 \cdot 0{,}0549$$

- tra il valore minimo = 0,384 (0,492 - 0,108)

- e il valore massimo = 0,600 (0,492 + 0,108).

La **significatività statistica** del valore **k = 0,492** cioè la verifica dell'ipotesi

$$H_0 \textbf{: k} \leq \textbf{0} \text{ contro } H_1 \textbf{: k} > \textbf{0}$$

con

$$Z = \frac{k}{\sigma_{k0}} = \frac{0{,}492}{0{,}0589} = 8{,}35$$

determina Z = 8,35

Nella distribuzione normale **unilaterale**, a **Z = 8,35** corrisponde una probabilità **P < 0.0001.**

L'interpretazione conclusiva è che esiste un accordo statisticamente significativo, ma oggettivamente non alto. Infatti ha un livello o una intensità

- *moderate* secondo una classificazione,

- *good* secondo l'altra.

In queste condizioni, ai fini dell'interpretazione appare più utile l'**intervallo di confidenza**: il valore **reale** di **kappa** è compreso in una scala molto ampia, essendo incluso con probabilità del 95% tra

- un livello ***fair,*** nel limite inferiore (k = 0,384) e

- un livello ***moderate,*** nel limite superiore (k = 0,600).

Per la significatività della **differenza tra due k indipendenti** ($k_1 - k_2$), dove l'ipotesi alternativa ovviamente può essere sia unilaterale sia bilaterale, **Cohen** propone

$$Z = \frac{k_1 - k_2}{\sqrt{\sigma_{k1}^2 + \sigma_{k2}^2}}$$

dove

$$\sigma_k = \sqrt{\frac{p_o \cdot (1 - p_o)}{N \cdot (1 - p_e)^2}}$$

per ognuno dei due campioni in modo indipendente

Per il calcolo dell'**errore standard** di **k,** necessario alla verifica dell'ipotesi nulla H_0**: k = 0,** è stata proposta una nuova formula asintotica, quindi per **grandi campioni** e con l'uso della distribuzione Z, indicata con $se(k)$ essa è:

$$se(k) = \frac{\sqrt{p_e + p_e^2 - \sum p_{i+} p_{+i} (p_{i+} + p_{+i})}}{(1 - p_e)\sqrt{N}}$$

Può essere utile il confronto con quella originaria di Cohen, dalla quale differisce per il numeratore, come svolto nell'esempio successivo.

3.3. ESEMPIO 2. Un dentista ha registrato sulle cartelle dei pazienti la sua opinione, cioè la necessità di estrarre il dente cariato, prima e dopo la radiografia.
Il conteggio delle valutazioni ha dato i seguenti risultati

		Dopo		
	Estrazione	SI	NO	Totale
Prima	SI	40	5	45
	NO	25	30	55
	Totale	65	35	N=100

Fornire una **misura quantitativa della variazione di giudizio** o inversamente della **riproducibilità del giudizio** nei due diversi esami.
Risposta. Benché i calcoli possano essere effettuati indifferentemente con le frequenze assolute e con quelle relative, per una visione più chiara dei risultati è vantaggioso utilizzare quelle relative.
Dopo trasformazione, i dati diventano

		Dopo		
	Estrazione	SI	NO	Totale
Prima	SI	**0,40** *(0,2925)*	**0,05** *(0,1575)*	0,45
	NO	**0,25** *(0,3575)*	**0,30** *(0,1925)*	0,55
	Totale	0,65	0,35	1,00

ricordando che
- in grassetto sono riportate le **proporzioni osservate**,
- in corsivo e tra parentesi **quelle attese** e che
- il **numero totale** di osservazioni è N = 100.

Dopo aver ottenuto p_o = 0,40 + 0,30 = **0,70** e p_e = 0,2925 + 0,1925 = **0,485** si ricavano
- il valore di **k**

$$k = \frac{0,70 - 0,485}{1 - 0,485} = \frac{0,215}{0,515} = 0,417$$

- il suo errore standard $se(k)$

$$es(k) = \frac{\sqrt{0{,}485 + 0{,}485^2 - (0{,}45 \cdot 0{,}65 \cdot (0{,}45 + 0{,}65) + 0{,}55 \cdot 0{,}35 \cdot (0{,}55 + 0{,}35))}}{(1 - 0{,}485)\sqrt{100}}$$

$$es(k) = \frac{\sqrt{0{,}485 + 0{,}2352 - 0{,}3218 - 0{,}1733}}{515} = \frac{0{,}474}{515} = 0{,}092$$

La **significatività di k** per la verifica di

H_0 **: k ≤ 0** contro H_1 **: k > 0**

fornisce un valore

$$Z = \frac{0{,}417}{0{,}092} = 4{,}53$$

Il risultato (Z = 4,53) è così grande che, nella tabella della distribuzione normale standardizzata unilaterale, corrisponde a un probabilità P < 0,0001.

Se ne deve dedurre che il valore di **k è altamente significativo**, quindi statisticamente maggiore di zero.

Tuttavia, poiché k = 0,417 non è molto alto, il grado di accordo tra le due distribuzioni è

- ***moderate*** secondo la scala di **Landis** e **Koch**
- ***good*** secondo quella di **Fleiss**.

Con la formula di **Cohen**

$$\sigma_{k0} = \sqrt{\frac{p_e}{N \cdot (1 - p_e)}} = \sqrt{\frac{0{,}485}{100 \cdot (1 - 0{,}485)}} = \sqrt{\frac{0{,}485}{51{,}5}} = 0{,}097$$

l'errore standard ha come risultato σ_{k0} = 0,097.

E' un valore più grande e quindi fornisce una stima di Z più prudenziale (più bassa) ai fini del rifiuto dell'ipotesi nulla k = 0; ma la differenza con il risultato precedente è ridotta.

Con questo valore dell'**errore standard**, il risultato del **test per la significatività**

$$Z = \frac{0{,}417}{0{,}097} = 4{,}30$$

sarebbe stato Z = 4,30.

Non avrebbe modificato sostanzialmente l'interpretazione del risultato ottenuto con l'errore standard precedente.

CONCLUSIONI

In alcuni testi di statistica applicata presentano solo la nuova formula, altri testi evidenziano che per essa la condizione di normalità è più vincolante e che pertanto in esperimenti standard, con campioni inferiori alle 100 unità, sia preferibile utilizzare sempre quella proposta da **Cohen.**

Anche per l'intervallo di confidenza più recentemente è stata proposta una formula asintotica dell'errore standard di k, che con grandi campioni appare più precisa. È stata presentata da J. L. **Fleiss**. Secondo altri autori di testi divulgativi, fondamentalmente non è migliore e ha gli stessi limiti dell'altra già proposta per il test di significatività: fornisce risultati non molto diversi da quella di **Cohen**, è più vantaggiosa per la significatività, ma è meno valida per i campioni che sono inferiori a 100 unità.

Concludo dicendo che questa tesi ha rappresentato per me l'apprendimento di una metodologia di calcolo utile per valutazioni tra raters o giudici, necessaria per valutare se i giudizi forniti dai due esperti sono riproducibili e affidabili.

TAVOLA DEI FATTORIALI

Fattoriali dei numeri fino a 65

n.	Fattoriale	n.	Fattoriale	n.	Fattoriale
0	1	**22**	1.12×10^{21}	**44**	2.66×10^{54}
1	1	**23**	2.59×10^{22}	**45**	1.2×10^{56}
2	2	**24**	6.2×10^{23}	**46**	5.5×10^{57}
3	6	**25**	1.55×10^{25}	**47**	2.59×10^{59}
4	24	**26**	4×10^{26}	**48**	1.24×10^{61}
5	120	**27**	1.09×10^{28}	**49**	6.1×10^{62}
6	720	**28**	3.05×10^{29}	**50**	3.04×10^{64}
7	5040	**29**	8.84×10^{30}	**51**	1.55×10^{66}
8	40320	**30**	2.65×10^{32}	**52**	8.06×10^{67}
9	362880	**31**	8.2×10^{33}	**53**	4.2×10^{69}
10	3628800	**32**	2.6×10^{35}	**54**	2.3×10^{71}
11	39916800	**33**	8.6×10^{36}	**55**	1.27×10^{73}
12	4.78×10^{8}	**34**	2.95×10^{38}	**56**	7.1×10^{74}
13	6.23×10^{9}	**35**	1.03×10^{40}	**57**	4.05×10^{76}
14	8.72×10^{10}	**36**	3.7×10^{41}	**58**	2.35×10^{78}
15	1.3×10^{12}	**37**	1.37×10^{43}	**59**	1.39×10^{80}
16	2.1×10^{13}	**38**	5.23×10^{44}	**60**	8.3×10^{81}
17	3.56×10^{14}	**39**	2.04×10^{46}	**61**	5.07×10^{83}
18	6.4×10^{15}	**40**	8.16×10^{47}	**62**	3.14×10^{85}
19	1.22×10^{17}	**41**	3.34×10^{49}	**63**	1.98×10^{87}
20	2.43×10^{18}	**42**	1.4×10^{51}	**64**	1.27×10^{89}
21	5.1×10^{19}	**43**	6.04×10^{52}	**65**	9.25×10^{90}

VALORI DI CHI QUADRATO (FINO A 20 GRADI DI LIBERTÀ)

G.L.	P=0.10	P=0.05	P=0.01	P=0.005
1	2.705	3.841	6.635	7.879
2	4.605	5.991	9.210	10.597
3	6.251	7.815	11.345	12.838
4	7.779	9.488	11.277	14.860
5	9.236	11.07	15.086	16.749
6	10.645	12.592	16.812	18.547
7	12.017	14.067	18.475	20.278
8	13.362	15.507	20.090	21.955
9	14.684	16.919	21.666	23.589
10	15.987	18.307	23.209	25.188
11	17.275	19.675	24.725	26.757
12	18.549	21.026	26.217	28.299
13	19.812	22.362	27.688	29.819
14	21.064	23.685	29.141	31.319
15	22.307	24.996	30.578	32.801
16	23.542	26.296	31.999	34.267
17	24.769	27.587	33.409	35.718
18	25.989	28.869	34.805	37.156
19	27.204	30.143	36.191	38.582
20	28.412	31.410	37.566	39.997
30	40.264	43.776	50.893	
40	51.81	55.76	63.69	
50	63.17	67.50	76.15	

BIBLIOGRAFIA

Paul W. Miele, Jr., Kenneth J. Berry, *Permutation methods: A distance function approach*, New York, Springer, 2001.

H. Zeisel nel 1947 (nel volume *Say it with figures*, Harper & Row, New York; tradotto in italiano nel 1968, in *Ditelo coi numeri*, Marsilio, Padova).

P. Sprent e N. C. Smeeton del 2001 *Applied nonparametric statistical methods*, 3rd ed. Chapman & Hall/CRC, London, XII + 461 p..

Jacob Cohen nel 1960, *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement, Vol. XX, No. 1, pp. 37-46)

J. P. Guilford nel 1950, *Fundamental Statistics in Psychology and Education* (2nd ed., New York, McGraw-Hill).

J. L. Fleiss, J. C. M. Lee e J. R. Landis nel 1979, *The large sample variance of kappa in the case of different sets of raters*, pubblicato su Psychological Bulletin Vol. 86, pp. 974-977).

J. L. Fleiss nel 1981, nel volume *Statistical Methods for Rates and Proportions* (2nd ed. New York, John Wilwy & Sons).

J. Richard Landis e Gary G. Koch nel 1977 (*The measurement of observer agreement for categorial data* pubblicato da Biometrics, Vol. 33, pp. 159-174).

L. Fleiss nel 1981 *Statistical Methods for Rates and Proportions* (John Wiley & Sons).

Bernard Rosner del 2000 *Fundamentals of Biostatistics* (5^{th} ed. Duxbury, Australia, XVII + 792 p.).

RINGRAZIAMENTI

Un ringraziamento particolare ai miei genitori che mi hanno sempre sostenuto, sia moralmente che economicamente, durante tutti i miei anni di studio.