CAPITOLO V

PROPORZIONI E PERCENTUALI RISCHI, ODDS E TASSI

5.1. Termini tecnici in epidemiologia: misure del rischio 1

5.2. Altri termini tecnici: sensibilita', specificita', valore predittivo e efficienza di un test o di una classificazione 8

5.3. Perche' la varianza di p e' pq e sue conseguenze; varianza e errore standard di una frequenza relativa o assoluta, in una popolazione infinita e finita 19

5.4. Intervallo di confidenza di una frequenza relativa o assoluta con la normale, in una popolazione infinita o finita; metodi grafici per l'intervallo fiduciale e la stima del numero di dati. 28

5.5. Intervallo di confidenza di una proporzione, mediante la distribuzione F 41

5.6. Calcolo del campione minimo necessario, per la stima di una proporzione campionaria con un errore massimo prefissato 47

5.7. Il confronto tra una proporzione campionaria e una proporzione attesa con il test z; dimensione minima del campione, per l'uso della distribuzione normale 52

5.8. La potenza a posteriori e a priori di un test sulla proporzione per un campione, con l'uso della normale 57

5.9. Test per una proporzione: la binomiale per campioni piccoli e l'intervallo di confidenza con F per campioni grandi 64

5.10. La potenza di un test per una proporzione, con l'uso della distribuzione binomiale 68

5.11. Test per la bonta' dell'adattamento di una distribuzione osservata e la distribuzione binomiale, costruita con una proporzione nota e con una proporzione ignota 71

5.12. Test sulla differenza tra due proporzioni, con il metodo di Feldman e Kluger, per abbreviare il metodo esatto di Fisher 78

5.13. Significativita' e intervallo di confidenza della differenza tra due proporzioni, con la distribuzione normale 82

5.14. Potenza a posteriori (1-β) e a priori (n) dei test sulla differenza tra due proporzioni; bilanciamento di due campioni 86

5.15. Il rapporto tra due proporzioni (r): intervallo di confidenza e significativita'; formula test-based di Miettinen per r 101

5.16. Il rapporto tra due odds (or): intervallo di confidenza e significativita'; formula test-based di Miettinen per or 111

5.17. Il rapporto tra due tassi (rr): intervallo di confidenza e significativita'; formula test-based di Miettinen 120

5.18. Dimensioni dei campioni e potenza, per test sulla differenza e sull'odds ratio delle proporzioni di due campioni indipendenti 129

CAPITOLO V

PROPORZIONI E PERCENTUALI
RISCHI, ODDS E TASSI

5.1. TERMINI TECNICI IN EPIDEMIOLOGIA: MISURE DEL RISCHIO

Nel capitolo precedente, sono state presentati alcuni metodi per l'inferenza sulla media e sulla varianza. Essi possono essere utilizzati per misure continue, quando i dati sono misurati con scale a intervalli o di rapporti. Altre volte, **il fenomeno è di tipo nominale o categoriale**, come gli individui guariti o non guariti con la somministrazione di un farmaco. La sua **misura è un conteggio**, quindi una **frequenza assoluta**.

La **prima elaborazione** di questa informazione

- è **il rapporto tra il numero di individui con la caratteristica analizzata e il numero totale di individui che formano la popolazione** analizzata.

Si ottiene una **proporzione,** spesso espressa come **percentuale:**

- la proporzione è quasi sempre richiesta dalle formule per i calcoli e quindi nell'inferenza;
- la percentuale è la misura corrispondente che spesso viene fornita nella comunicazione dei dati e nella statistica descrittiva, poiché nel linguaggio comune è di più immediata comprensione.

Con **la trasformazione della frequenza assoluta in proporzione**, diventa possibile il **confronto delle frequenze rilevate in campioni diversi**, nei quali spesso il numero di individui che formano la popolazione è differente. L'uso delle proporzioni ha essenzialmente tre scopi:

- descrivere la diffusione del fenomeno,
- confrontare situazioni differenti,
- prevedere il risultato, al variare dei fattori di rischio.

Proporzioni e percentuali sono utilizzate con frequenza **in epidemiologia**. Servono per indici e misure differenti, identificati in modo corretto e univoco solo mediante l'uso di **termini tecnici specifici,** se non si vuole ricorrere ogni volta a spiegazioni lunghe e dettagliate. La conoscenza esatta di tali termini è utile anche in molte discipline biologiche, farmacologiche e ambientali, che sono interessate allo studio di patologie di origine diversa, per la ricerca delle cause e dei fattori di rischio.

I termini specifici di uso più frequente sono:

- **prevalenza** e **incidenza**;
- **morbilità**, **letalità, mortalità**;
- **rischio relativo (RR), riduzione del rischio assoluto (RRA), riduzione del rischio relativo (RRR)**.

Quando si misura la **presenza di una malattia** in una popolazione, si utilizza l'**indice di prevalenza (*prevalence*)**. Si parla anche di **tasso di prevalenza**, definito come

- il **rapporto tra il numero di persone ammalate e quello delle persone che formano la popolazione** complessiva, espresso in percentuale:

$$\text{Pr}\,evalenza = \frac{malati}{popolazione} x100$$

La prevalenza può essere misurata in due modi:

- in un **momento preciso**, come avviene in un censimento: è la ***point prevalence***;
- in un **periodo di tempo**, come le persone influenzate nel mese X: è la ***period prevalence.***

In demografia e in epidemiologia, si distingue tra **statistiche di stato** e **statistiche di flusso**:

- si hanno **statistiche di stato,** quando si vuole sapere quante sono le persone, le famiglie, o una categoria qualsiasi di individui, presenti oppure residenti in una zona stabilita, **in un momento preciso**, identificato da una data esatta;
- si hanno statistiche di flusso quando si contano le nascite, i matrimoni o i decessi, in una zona prestabilita, ma in **un periodo di tempo**, che di solito è l'anno, ma che può essere il mese, la settimana o il decennio.

In varie situazioni, oltre alla prevalenza si è interessati a sapere quanti sono i **casi nuovi che compaiono in un intervallo di tempo prefissato**. La **durata del periodo** di osservazione varia in funzione del ciclo o della durata della patologia. Si parla di **incidenza** (***incidence***) o meglio di **tasso di incidenza,** definito come

- il **rapporto tra il numero di persone che si sono ammalate nell'intervallo di tempo e quello delle persone che formano la popolazione** complessiva;

espresso in percentuale è

$$Incidenza = \frac{malati \quad nuovi}{popolazione} x100$$

Quando si calcolano **i tassi** come nella formula precedente,

- gli eventi (riportati al numeratore) avvengono in un intervallo di tempo (**x**) abbastanza lungo,
- durante il quale la popolazione (riportata al denominatore) cambia tra l'inizio (P_t) e la fine (P_{t+x}) del periodo di osservazione.

Di conseguenza, al denominatore deve essere riportata

la **popolazione media (P_m) del periodo**

$$P_m = \frac{P_t + P_{t+x}}{2}$$

In malattie croniche o di lunga durata,

- **il tasso di prevalenza e il tasso di incidenza sono tra loro correlati attraverso la durata**,

per la relazione:

Tasso di Prevalenza = Tasso di Incidenza x Durata

Ovviamente, conoscendo due parametri, si ricava il terzo.

ESEMPIO 1. Per lo studio della frequenza di un'allergia, su un campione di 568 individui 126 hanno presentato i sintomi evidenti della patologia. Secondo le cartelle del medico presso il quale i pazienti sono in cura, tra i 126 ammalati 38 hanno iniziato a presentare la patologia nell'ultimo anno.
Calcolare il tasso di prevalenza, il tasso di incidenza e stimare la durata della patologia.

Risposte. Il **tasso di prevalenza**

$$\Pr evalenza = \frac{126}{568} \cdot 100 = 22{,}18\%$$

è uguale al 22,18 per cento.

Il **tasso di incidenza**

$$Incidenza = \frac{38}{568} \cdot 100 = 6{,}69\%$$

è uguale al 6,69 per cento.

La **durata** dei sintomi di questa malattia

$$Durata = \frac{22{,}18}{6{,}69} oppure \frac{0{,}2218}{0{,}069} = 3{,}32$$

è uguale a 3,32 **anni**.

(Perché l'incidenza è stata valutata in anni).

Altri indicatori usati con frequenza, nello studio di malattie che

- comportano giorni di degenza
- e/o possono causare la morte del paziente,

sono il **tasso di morbilità**, il **tasso di letalità** e il **tasso di mortalità.**

Il **tasso di morbilità**:

$$morbilità = \frac{giorni \quad di \quad malattia}{popolazione \quad a\, rischio} \cdot 100$$

Il **tasso di letalità**:

$$letalità = \frac{morti \quad da \quad causa \quad specifica}{casi \quad diagnosticati} \cdot 100$$

Il **tasso di mortalità**:

$$mortalità = \frac{morti \quad da \quad causa \quad specifica}{popolazione} \cdot 100$$

Per calcolare e **comunicare correttamente il tasso di mortalità**, sono necessarie due avvertenze:

- il periodo di analisi spesso è lungo; quindi, nelle formule in cui al denominatore c'è la popolazione, si deve intendere la popolazione media, tra quella censita all'inizio e quella censita alla fine del periodo di osservazione;
- quando gli eventi sono rari, come nel caso di molte patologie, il rapporto non è moltiplicato per 100 ma per potenze di 10 nettamente superiori. Si utilizzano valori da mille (10^3) a un milione (10^6), in modo che il risultato sia almeno qualche unità o alcune decine.

Ad esempio,

- non si riporta 0,31 per cento (0,31%),
- ma 3 su mille oppure ancora meglio, per una stima più dettagliata, 31 su diecimila.

Nelle misure che riguardano le popolazioni, si usano due termini: **tassi** e **probabilità**.

In comune hanno il numeratore, cioè il numero di individui morti in un intervallo di tempo. Ma

- mentre i **tassi** sono calcolati **in rapporto alla popolazione media**,
- le **probabilità** sono calcolate **in rapporto alla popolazione iniziale**.

Queste ultime in particolare servono quando si costruiscono le **tavole di mortalità** (***life table***) o **analisi della sopravvivenza** (in un paragrafo successivo sono illustrate le differenze tra tassi e probabilità, evidenziando graficamente i concetti con il **diagramma di Lexis)**.

Molto spesso **l'ambiente** e lo **stile di vita** sono la causa principale della comparsa di una malattia. Ogni persona che vive questi fattori è sottoposto a un **rischio (*risk*)** di contrarre la malattia**.** Si parla allora di **rischio relativo (*relative risk*),** in rapporto alla situazione normale o di controllo. La causa della malattia è il **fattore di rischio** (***risk factor***), come può essere lo stile di vita del fumo attivo o passivo per malattie polmonari e l'esposizione a sostanze nocive in ambienti di lavoro non totalmente protetti. Gli individui che vivono la situazione di rischio sono detti **esposti.**

Tale **rischio relativo** (**RR**; anche in inglese da ***Relative Risk***) è misurato mediante **il rapporto**

- **tra l'incidenza** in persone **esposte** al fattore di rischio (I_{esp})
- e **l'incidenza** in persone che non sono **mai state esposte** ($I_{non-esp}$)

$$RR = \frac{I_{esp}}{I_{non-esp}}$$

Se l'esposizione a un fattore ambientale determina un rischio maggiore, ci si può chiedere di quanto diminuirebbe il rischio, eliminando l'esposizione a quel fattore. Tale stima può essere fornita in due modi: dalla

1) **Riduzione del Rischio Assoluto** (**RRA**), misurato mediante **la differenza**

- **tra l'incidenza** in persone **esposte** al fattore di rischio (I_{esp})
- e **l'incidenza** in persone **non esposte** ($I_{non-esp}$)

$$RRA = I_{esp} - I_{non-esp}$$

2) **Riduzione del Rischio Relativo** (**RRR**), misurato mediante **il rapporto**

- **tra la differenza dell'incidenza** in persone **esposte** e in persone **non esposte (** $I_{esp} - I_{non-esp}$ **)**
- e **l'incidenza** in persone **esposte** (I_{esp})

$$RRR = \frac{I_{esp} - I_{non-esp}}{I_{esp}}$$

ESEMPIO 2. In un'analisi sugli effetti dell'amianto nelle malattie polmonari, in cinque anni di osservazione e controlli costanti, su 800 lavoratori esposti 55 hanno presentato patologie specifiche. Nel campione di controllo, formato da 1550 persone non esposte e con età e lavori simili, 38 hanno presentato le stesse patologie.
Determinare il Rischio Relativo, la Riduzione del Rischio Assoluto e la Riduzione del Rischio Relativo.

Risposta. Prima di passare ai calcoli è sempre bene presentare i dati in tabelle. Nel caso specifico, quella più adeguata è

	Patologia		
	Si	No	Totali
Individui **Esposti**	**55**	745	**800**
Individui **Non-Esposti**	**38**	1512	**1550**

Poiché i casi sono quelli nuovi comparsi in un intervallo di tempo, anche se è lungo, è possibile **calcolare le due incidenze**

- per gli **individui esposti,** l'incidenza è $I_{esp} = 55/800 = 0{,}0687$ o 68,7 per mille;
- per gli **individui non-esposti,** l'incidenza è $I_{non-esp} = 38/1550 = 0{,}0245$ o 24,5 per mille.

Il **rischio relativo** (delle persone esposte)

$$RR = \frac{I_{esp}}{I_{non-esp}} = \frac{0{,}0687}{0{,}0245} = 2{,}80$$

è 2,80 volte quello della situazione normale o di confronto, vale a dire quello delle persone non esposte.

La **riduzione del rischio assoluto**

$$RRA = I_{esp} - I_{non-esp} = 0{,}0687 - 0{,}0245 = 0{,}0442$$

sarebbe uguale a 0,0442 o 44,2 per mille.

La **riduzione del rischio relativo**

$$RRR = \frac{I_{esp} - I_{non-esp}}{I_{esp}} = \frac{0{,}0687 - 0{,}0245}{0{,}0687} = \frac{0{,}0442}{0{,}0687} = 0{,}64$$

sarebbe pari al 0,64 o 64 per cento.

In altre discipline, come in tossicologia, è prassi valutare **non la <u>riduzione del rischio</u>** ma **<u>il rischio aggiuntivo</u>**, causato da una sostanza che si reputa nociva e della quale si vogliono misurare gli effetti tossici. La sua analisi statistica è meno intuitiva di quella della riduzione del rischio.

In queste ricerche, spesso si devono valutare gli effetti letali o sulla crescita, mediante la proporzione di decessi o di individui immaturi, contati su un campione di **n** cavie, dopo un determinato periodo di esposizione. Poiché normalmente si hanno decessi di cavie anche quando si somministra il placebo, **l'effetto reale del tossico non può essere misurato come semplice differenza con il placebo.**

Per esempio,

- se somministrando il placebo la proporzione di decessi è 0,15 e quella con il tossico è 0,27
- l'effetto aggiuntivo del tossico **<u>non è</u>** 0,12 (cioè 0,27 - 0,15).

Il calcolo del rischio aggiuntivo deve considerare solo la proporzione che sarebbe sopravvissuta senza il tossico. In modo più formale,

- se con il placebo nel tempo **t** muore una proporzione $\mathbf{p_1}$ degli individui che formano il controllo
- e ovviamente la proporzione di quelli che sopravvivono è $\mathbf{1-p_1}$,

con il tossico la proporzione di decessi sarà

$$\mathbf{p_2 = p_1 + f\,(1-p_1)}$$

ESEMPIO 3. Per valutare l'effetto di una sostanza tossica diluita nell'acqua, in ecotossicologia tra i vari indicatori si utilizza anche la percentuale di Dafnie che non arrivano a maturità sessuale. Si supponga che con il controllo la proporzione di femmine che non hanno prodotto uova sia $p =$ **0,3** e sia stato ripetutamente dimostrato che il tossico determina un rischio aggiuntivo di p = **0,2**.

Quale sarà la proporzione di animali sottoposti all'effetto del tossico che non produrranno uova?

Risposta. Con $\mathbf{p_1 = 0,3}$ e che **f = 0,2**

la proporzione $\mathbf{p_2}$ di individui che non produrranno uova

$$\mathbf{p_2 = 0,3 + 0,2\cdot(1 - 0,3) = 0,44}$$

risulterà uguale a **0,44** (non 0,50 come si sarebbe ottenuto con la somma 0,3 + 0,2 cioè $\mathbf{p_1}$ + **f**).

Occorre sottolineare che **quando la mortalità del controllo è alta** (per es.: $\mathbf{p_1 = 0,70}$), lo stesso effetto della sostanza tossica (**f = 0,2**) determina nel campione esposto una proporzione $\mathbf{p_2}$ di decessi che, in valore assoluto, risulterà minore.

Con $\mathbf{p_1 = 0,70}$ e ancora **f = 0,2**

$$\mathbf{p_2 = 0,70 + 0,2\,(1 - 0,7) = 0,76}$$

la proporzione totale di decessi $\mathbf{p_2}$ risulta uguale a **0,76**.

E' un incremento in valore assoluto di **0,06** (0,76 – 0,70) che potrebbe apparire determinato da un effetto aggiuntivo minore, rispetto all'incremento di **0,14** (0,44 – 0,30) stimato nel caso precedente.

In realtà, il tossico determina una aumento della mortalità o rischio aggiuntivo (**f**) esattamente identico e pari a 0,20.

Per il **confronto tra gli effetti di due o più sostanze tossiche** è quindi importante calcolare correttamente **f**, il **fattore di rischio aggiuntivo** o **la differenza relativa** di $\mathbf{p_2}$ rispetto a $\mathbf{p_1}$.

Il rischio aggiuntivo **f** è dato da

$$\mathbf{f} = \frac{p_2 - p_1}{1 - p_1}$$

ESEMPIO 4. Nell'analisi di un ambiente inquinato, nell'intervallo di tempo $\mathbf{t_1 - t_2}$, la quota di decessi è risultata pari a **0,28**. Con un intervento di risanamento, si suppone di aver eliminato uno dei fattori responsabili dell'inquinamento. Nuove analisi stimano che nello stesso tempo **t** la quota di decessi è scesa a **0,19**. Valutare la differenza relativa o il fattore di rischio eliminato.

Risposta. La **differenza relativa o il fattore di rischio (f)** eliminato
con $\mathbf{p_2 = 0,28}$ e $\mathbf{p_1 = 0,19}$ è

$$f = \frac{0,28 - 0,19}{1 - 0,19} = \frac{0,09}{0,81} = 0,111$$

uguale a 0,111.

5.2. ALTRI TERMINI TECNICI: SENSIBILITA', SPECIFICITA', VALORE PREDITTIVO E EFFICIENZA DI UN TEST O DI UNA CLASSIFICAZIONE.

Anche per valutare la **prestazione** o il **rendimento** (***performance***) di un **test diagnostico** e/o di una **analisi qualitativa** si utilizzano concetti e metodi che sono collegati all'uso di proporzioni. A questo proposito, nella letteratura medica, ambientale, chimica e industriale, ricorrono con frequenza alcuni termini tecnici, che è utile conoscere per le applicazioni generali della statistica:

- **sensibilità** (***sensitivity***),
- **specificità** (***specificity***),
- **valore predittivo** (***predictive value***), che può essere distinto in **valore predittivo positivo** (***positive predictive value***) e **valore predittivo negativo** (***negative predictive value***),
- **efficienza** (***efficiency***).

Quando si applica un test biologico-chimico o si utilizza una procedura classificatoria per identificare la **presenza-assenza di una sostanza specifica oppure di un attributo in un campione di più individui o oggetti**, spesso è richiesto di fornire una valutazione quantitativa della capacità discriminante o selettiva del metodo.
Nella sua forma più semplice e ricorrente, la riposta è espressa in termini qualitativi: il test è dichiarato positivo se la sostanza cercata è presente, negativo se è assente.
Con un campione formato da più unità, la misura è una scala discreta di conteggio della **presenza-assenza**, tradotta poi in una proporzione sul numero totale.

Per valutare il metodo, la **presenza effettiva della sostanza** deve essere indicata da **un'altra analisi diagnostica**, condotta con criteri differenti e che viene ritenuta priva di errore. La misura della

correttezza del test è fornita dalla coincidenza tra il risultato ottenuto con l'analisi di laboratorio e la realtà. L'errore che è possibile commettere è duplice:

- **non trovare una sostanza quando è effettivamente presente**;
- **trovarla quando in realtà è assente**.

Ad esempio, per indicare la presenza di una malattia quando i sintomi non sono ancora evidenti, in medicina si ricercano precursori certi. Un indicatore è corretto quando in tutti gli ammalati è possibile ritrovare quella sostanza o attributo, che invece è sempre assente in tutti gli individui non affetti da quella malattia specifica.

Per facilitare l'esposizione didattica e la comprensione di questi concetti, si ricorre a una impostazione grafica tabellare, che permette il confronto tra la realtà e il risultato campionario di ogni singolo test.

		REALTA' O MALATTIA	
		POSITIVO	NEGATIVO
RISULTATO DEL TEST	POSITIVO	**SENSIBILITA'** Vero Positivo	---------- Falso Positivo
	NEGATIVO	---------- Falso Negativo	**SPECIFICITA'** Vero Negativo

Dalla comparazione, risulta con evidenza che sono possibili quattro esiti.

1 - Se il risultato del **test è positivo** e l'individuo **è affetto** dalla malattia, si ha un **Vero Positivo** (***True Positive***) e si parla di **Sensibilità** (***Sensitivity***) del test;

2 – Se il risultato del **test è positivo** mentre l'individuo **non è affetto** dalla malattia, si ha un **Falso Positivo** (***False Positive***);

3 – Se il risultato del **test è negativo** e l'individuo **è affetto** dalla malattia, si ha un **Falso Negativo** (***False Negative***);

4 – Se il risultato del **test è negativo** mentre l'individuo **non è affetto** dalla malattia, si ha un **Vero Negativo** (***True Negative***) e si parla di **Specificità** (***Specificity***) del test.

Con un campione formato da più individui, le frequenze dei quattro risultati possibili vengono riportati in un tabella di contingenza 2 x 2, che permette di quantificare i concetti illustrati.

Ricorrendo alla simbologia ormai abituale per indicare le frequenze assolute

TABELLA DI CONTINGENZA 2 X 2

		CONDIZIONI DI SALUTE		
		MALATO	NON MALATO	Totale
RISULTATO DEL TEST	POSITIVO	a **Vero Positivo**	b ***Falso Positivo***	$n_1 = a + b$
	NEGATIVO	c ***Falso Negativo***	d **Vero Negativo**	$n_2 = c + d$
	Totale	$n_3 = a + c$	$n_4 = b + d$	$N = a + b + c + d$

si hanno possono ricavare i quattro indici seguenti.

1 - La **sensibilità (*sensitivity*)** di un test o una prova è

- la proporzione di risultati positivi **(il test indica la presenza della malattia) quando il** soggetto è effettivamente ammalato**:**

$$Sensibilità = \frac{a}{a+c} = \frac{Veri \quad Positivi}{Totale \quad Malati}$$

2 – La **proporzione di falsi positivi** è

$$Proporzione \quad Falsi\ positivi = 1 - Sensibilità = \frac{b}{b+d}$$

3 - La **proporzione di falsi negativi** è

$$Proporzione \quad Falsi\,negativi = 1 - Specificità = \frac{c}{a+c}$$

4 - La **specificità** (***specificity***) di un test o una prova è

- la proporzione di risultati negativi **(il test non trova la malattia) quando il** soggetto è effettivamente sano:

$$Specificità = \frac{d}{d+b} = \frac{Veri \quad Negativi}{Totale \quad Sani}$$

5 – La **efficienza** (***efficiency***) del test o della prova è ricavata **sommando la sensibilità e la specificità in modo ponderato** (cioè sia il numeratore che il denominatore):

$$Efficienza = \frac{a+d}{a+b+c+d} = \frac{Veri \quad Positivi + Veri \quad Negativi}{Totale \quad Popolazione}$$

Per l'uso corretto di questi indicatori e una loro valutazione corretta è importante rimarcare che
- la **sensibilità** dipende solamente dalla frequenza di risultati positivi e negativi **entro la popolazione di ammalati**;
- la **specificità** dipende solamente dalla distribuzione dei risultati **entro la popolazione dei non ammalati**.

Ne deriva che questi due indicatori
- **non dipendono dal rapporto tra il numero di ammalati e quello dei non ammalati**
- e quindi sono da considerarsi **indipendenti dalla prevalenza della malattia**.

Sensibilità e specificità non dipendono dalla popolazione testata: sono indipendenti dalla popolazione o dal campione ai quali sono applicati e sono determinati esclusivamente **dalla capacità discriminanti del test rispetto alla realtà di ogni singolo individuo**.

Spesso è richiesto di valutare anche il **tasso di errore**, determinato dalla frequenza dei falsi positivi e dei falsi negativi.
Le **funzioni dei valori predittivi** dei falsi positivi e dei falsi negativi, dai quali deriva la misura dell'**efficienza,** sono stimate mediante rapporti che considerano la popolazione complessiva, cioè l'insieme degli individui ammalati e di quelli non ammalati. Ne consegue che sono **dipendenti dalla**

prevalenza della malattia e quindi variano da caso a caso, come la diffusione della malattia in una popolazione.

La **proporzione di falsi positivi**, la **proporzione di falsi negativi** e l'**efficienza** del test sono indicatore della capacità del test di **scoprire la malattia nella popolazione effettivamente analizzata**.

ESEMPIO 1 (tratto dal testo di James E. De **Muth** del 1999 ***Basic Statistical and Pharmaceutical Statistical Application*** (Marcel Dekker, Inc. New York, XXI + 596 p.)

Si assuma di aver sviluppato una procedura semplice, per **identificare gli individui con anticorpi HIV**. Ovviamente il test dovrebbe dare un risultato positivo con una probabilità molto alta, ma solo quando la persona è realmente infettata dal virus HIV (sensibilità). Una risposta errata, un falso positivo, potrebbe avere conseguenze molto gravi per l'individuo analizzato, determinando non raramente fortissimi attacchi d'ansia, in grado di condurre al suicidio.

Per la verifica della sensibilità e specificità, questo test diagnostico è stato effettuato su 500 volontari, dei quali 100 indiscutibilmente affetti dalla malattia e 400 sicuramente sani.

Il risultato complessivo del test è stato

		VOLONTARI		
		MALATI HIV	SANI	Totale
TEST DIAGNOSTICO	POSITIVO	90	8	98
	NEGATIVO	10	392	402
	Totale	100	400	500

Calcolare: sensibilità, specificità, proporzione di falsi positivi, proporzione di falsi negativi, efficienza.

Risposta.

1 - La **sensibilità (*sensitivity*)** è

$$Sensibilità = \frac{a}{a+c} = \frac{Veri \quad Positivi}{Totale \quad Malati} = \frac{90}{100} = 0.90$$

2 - La **specificità** (***specificity***) è

$$Specificità = \frac{d}{d+b} = \frac{Veri \quad Negativi}{Totale \quad Sani} = \frac{392}{400} = 0{,}98$$

3 – La **proporzione di falsi positivi** è

$$Proporzione \quad Falsi\ positivi = 1 - Sensibilità = \frac{b}{b+d} = 1 - 0{,}90 = \frac{10}{100} = 0{,}10$$

4 - La **proporzione di falsi negativi** è

$$Proporzione \quad Falsi\ negativi = 1 - Specificità = \frac{c}{a+c} = 1 - 0{,}98 = \frac{8}{400} = 0{,}02$$

5 – La **efficienza** (***efficiency***) del test è

$$Efficienza = \frac{a+d}{a+b+c+d} = \frac{Veri \quad Positivi + Veri \quad Negativi}{Totale \quad Popolazione} = \frac{90+8}{500} = 0{,}196$$

Sempre con gli stessi dati dell'esempio, è possibile **ricavare altre informazioni**, per la quali a volte è conveniente utilizzare non le frequenze assolute ma le loro frequenze relative:

		VOLONTARI		
		MALATI HIV	SANI	Totale
TEST DIAGNOSTICO	POSITIVO	0,180	0,016	0,196
	NEGATIVO	0,020	0,784	0,804
	Totale	0,200	0,800	1,000

1 – Con una sensibilità del 90% e una specificità del 98% come già stima per i dati dell'esempio, quale è la probabilità che una persona che ha gli anticorpi HIV risulti positivo al test?

La stima cercata è chiamata **valore positivo predetto** (***predicted value positive*** o **PVP**), per il quale serve conoscere la prevalenza della malattia.

Assumendo che nella popolazione la malattia abbia la stessa frequenza di quella presente nel campione di 500 volontari, quindi con una prevalenza pari a 0,20

è

$$PVP = \frac{Sensibilità \quad x \quad Prevalenza}{\left(Sensibilità \quad x \quad Prevalenza\right) + \left((1 - Specificità) \quad x \quad (1 - Prevalenza)\right)}$$

$$PVP = \frac{0{,}90x0{,}20}{(0{,}90x0{,}20) + (0{,}02x0{,}80)} = \frac{0{,}180}{0{,}180 + 0{,}016} = \frac{0{,}180}{0{,}196} = 0{,}918$$

Ma la proporzione 0,20 è la frequenza degli ammalati di HIV nel campione di 500 volontari.

E' la prevalenza della malattia, che in una popolazione reale quasi sempre è molto minore.

Ad esempio, negli studenti dei college americani, la malattia HIV per alcuni anni ha avuto una prevalenza del 0,2% (due ogni mille studenti).

In questo caso, sempre con una **sensibilità del 90%** e una **specificità del 98%**, quale è la probabilità che una persona effettivamente ammalata risulti positivo al test e quindi sia identificato come tale?

Il **valore positivo predetto** (***predicted value positive*** o **PVP**, detto anche ***positive predictive value*** o **PPV**)

$$PVP = \frac{Sensibilità \quad x \quad Prevalenza}{\left(Sensibilità \quad x \quad Prevalenza\right) + \left((1 - Specificità) \quad x \quad (1 - Prevalenza)\right)}$$

$$PVP = \frac{0{,}90x0{,}002}{(0{,}90x0{,}002) + (0{,}02x0{,}998)} = \frac{0{,}0018}{0{,}0018 + 0{,}01996} = \frac{0{,}0018}{0{,}02176} = 0{,}00827$$

risulta uguale a 0,0827.

Significa che, sebbene specificità e sensibilità appaiano elevate, vi è solamente una probabilità leggermente superiore a 8% che un individuo con gli anticorpi HIV possa essere identificato come tale con il test.

Negli altri 92 casi su cento l'individuo risultato positivo al test in realtà è sano. Questo errore avviene con frequenza rilevante poiché, anche se la probabilità di un singolo errore è bassa, il numero di sani (1 - Prevalenza) è molto alto.

Secondo alcuni testi di epidemiologia, per il test HIV la specificità e la sensibilità sarebbero molto alte, pari per entrambe al 99%; ma con una prevalenza della malattia del 2 su mille si può stimare che il ***positive predictive value*** è

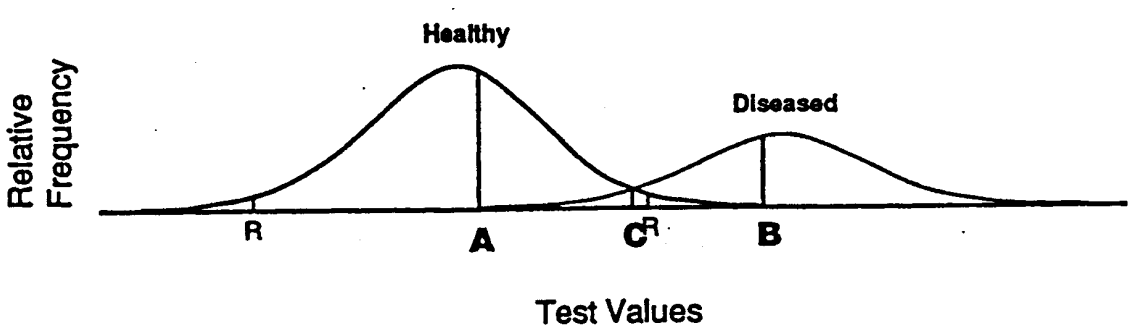
$$PVP = \frac{0{,}99x0{,}002}{(0{,}99x0{,}002)+(0{,}01x0{,}998)} = \frac{0{,}00198}{0{,}00198+0{,}00998} = \frac{0{,}00198}{0{,}01196} = 0{,}1656$$

solamente del 16,5%.

Questo fenomeno, collegato ai **costi umani di un falso positivo nel caso del virus HIV**, è la ragione fondamentale per cui **uno screening su tutta la popolazione non è mai apparsa una idea buona.** *Note that now the positive predictive value is only 16,5%, meaning 5 out of very 6 positive are false positive. This phenomenon is a major reason why screening the general population for HIV infections in not a very good idea.*

L'approccio statistico presentato è per **analisi o test qualitativi**; ma è sempre più diffuso l'uso di analisi di laboratorio che forniscono **risposte quantitative**.

Abitualmente si è in presenza di una condizione patologica dell'individuo o dell'ambiente, quando i valori sono alti. Nella ricerca ambientale è il caso dei livelli di inquinamento; in medicina, di parametri biologici quali colesterolo, trigliceridi, glucosio e globuli bianchi nel sangue.



L'**approccio qualitativo** descritto può essere applicato anche a **risultati numerici**.
Molti test spesso sono caratterizzati da distribuzioni di valori quantitativi che per i sani e per gli ammalati sono simili a quelli riportati nella figura. Da essi si passa a una classificazione qualitativa o binaria, mediante la selezione di un valore soglia, ritenuto biologicamente rilevante.
Questo valore o punto (chiamato spesso con il termine tecnico di ***cutoff***), è tale che **tutti i valori superiori** (come nel grafico) sono considerati **indicazioni positive della presenza della malattia**.

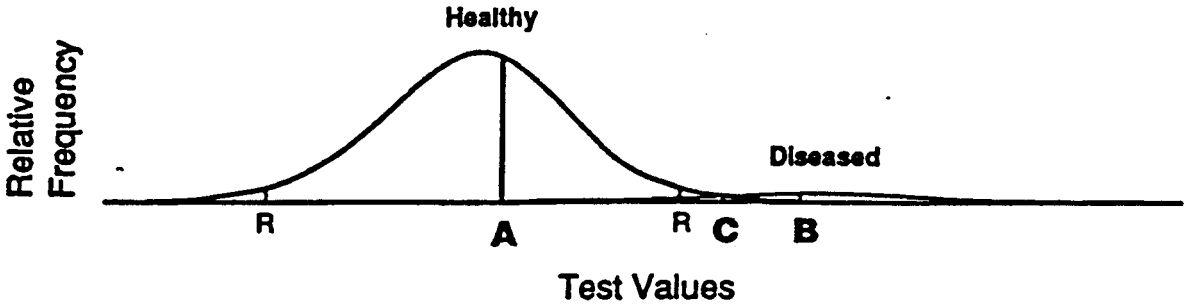
La scelta del valore di cutoff modifica la misura della sensibilità e della specificità del test.
Ad esempio, nella figura precedente in cui la prevalenza della malattia è alta, pari al 33% della popolazione totale (come indica la proporzione dell'area occupata dalle due distribuzioni normali e quindi 66,7% sani e 33,3% ammalati), è facile osservare come scegliendo come valore discriminante o ***cutoff*** diverso si modifichino.

Con un valore di cutoff corrispondente al punteggio indicato
- dal **punto A**, che identifica il livello sotto il quale la frequenza degli ammalati è zero, la **sensibilità del test si approssima al 100%** e **la specificità al 60%**;
- scegliendo invece il **punto B**, sopra il quale la frequenza dei sani è nulla, la **sensibilità è quasi del 60%** e la **specificità del 100%**;
- all'**intersezione tra le due curve**, punto C, **la sensibilità è pari al 90% e la specificità al 95%**;
- mentre il **punto R** indica il **punteggio di riferimento standard**, in rapporto alla popolazione sana: solamente **il 5% degli individui sani ha un valore superiore**.

La scelta del valore di sensibilità e di specificità per test di screening della malattia dipende dai **costi economici dell'analisi** e dal **costo etico o psicologico degli errori** (inevitabili):
- per malattie in cui la **cura è molto efficace** e il **costo dell'analisi è basso**, il punto di *cutoff* deve **massimizzare la sensibilità**;
- per malattie a **mortalità** e/o **morbidità alte** e per le quali **non esiste una cura efficace**, si deve **massimizzare la specificità**;
- in situazioni più sfumate, con **malattie non gravi e una efficienza media delle cure**, si deve **massimizzare l'efficienza**, che è identificata dall'**intersezione delle due curve**.

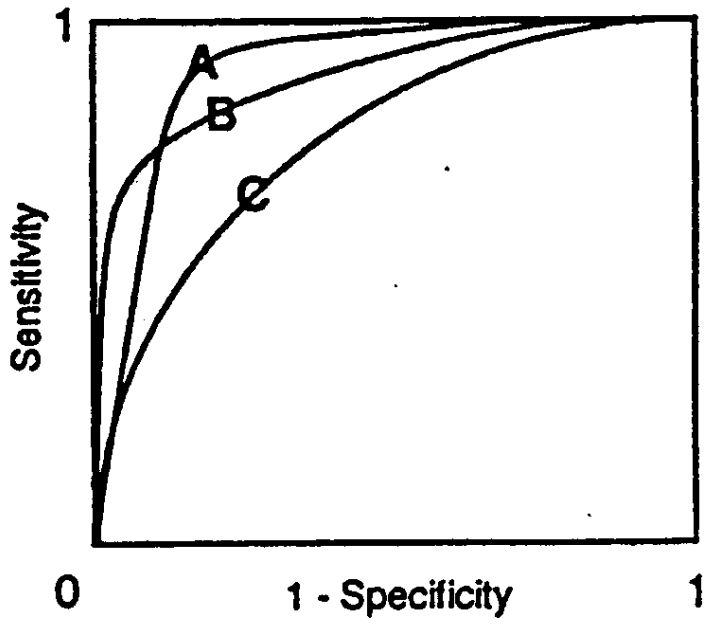
Per molte malattie, la prevalenza nella popolazione reale è bassa.
Ciò non ha effetti rilevanti sulla valutazione del test.

Nella figura successiva è del 5%. Anche in questo caso, porre attenzione al fatto che le due aree devono rispettare le proporzioni tra sani e ammalati presenti nella popolazione totale:



La **sensibilità e la specificità del test non cambiano**, per un dato cutoff.
Ne risulta invece fortemente **influenzata l'efficienza**, il punto di intersezione delle due curve, come pure il **positive predictive value** (PPV) e il **negative predictive value** (NPV).

Per confrontare i livelli di sensibilità e di specificità di due o più metodi quantitativi differenti, che classificano la presenza e l'assenza della stessa malattia, si utilizzano le curve ROC. Tale termine strano nella ricerca biologica e medica (ROC = Receiver Operating Characteristic) è stato utilizzato per la prima volta per descrivere l'abilità di radio riceventi a scoprire il segnale (cioè veri positivi = sensibilità), eliminando il rumore (fasi positivi = 1- specificità).
Il grafico bidimensionale riporta



il confronto tra tre curve di test, con i quali si cerca di ottimizzare la capacità di separare la popolazione in ammalati e non-ammalati.
L'area maggiore sottostante alla curva ROC indica il rapporto migliore tra sensibilità e specificità. Nella figura, il test A e il test B sono entrambi chiaramente migliori del test C poiché hanno una sensibilità migliore per una data specificità. Invece la scelta tra A e B dipende dall'uso che se ne intende fare.
Per ulteriori approfondimenti si rimanda alla letteratura specifica**.**

A conclusione dei vari concetti e metodi già illustrati e come indicazione di quelli che saranno sviluppati in capitoli successivi, è utile un elenco di statistiche derivate da una tabella 2 x 2 che sono state proposte in letteratura e di uso più frequente.
Utilizzando la simbologia

	+	-	**Totale**
A	a	b	r_1
B	c	d	r_2
Totale	c_1	c_2	t

e disponendo diversamente le varie informazioni, è possibile ricavare:

1 - Test χ^2, test G^2, test T^2
2 - Odds ratio (OR) = $(a/b)/(c/d)$
3 - Relative Risk (RR) = $(a/r_1)/(c/r_2)$
4 - Overall Fraction Correct = $(a+d)/t$
5 - Mis-classification Rate = 1 - *Overall Fraction Correct*
6 - Sensitivity = a/c_1
7 - Specificity = d/c_2
8 - Positive Predictive Value (PPV) = a/r_1
9 - Negative Predictive Value (NPV) = d/r_2
10 - Difference in Proportions = $(a/r_1)-(c/r_2)$
11 - Absolute Risk Reduction (ARR) = $(c/r_2)-(a/r_1)$
12 - Relative Risk Reduction (RRR) = $ARR/(c/r_2)$

13 – Positive Likelihood Ratio (+LR) = $Sensitivity/(1-Specificity)$

14 – Negative Likelihood Ratio (-LR) = $(1-Sensitivity)/Specificity$

15 – Diagnostic Odds Ratio =

$$[Sensitivity/(1-Sensitivity)]/[(1-Specificity)/Specificity]$$

16 – Error Odds Ratio =

$$[Sensitivity/(1-Sensitivity)]/[Specificity/(1-Specificity)]$$

17 – Youden's J = $Sensitivity + Specificity - 1$

18 – Number Needed to Diagnose (NND) =

$$1/[Sensitivity-(1-Specificity)] = 1/Youden's\ \ J$$

19 – Kappa di Cohen

20 – Coefficienti di contingenza, quali il Phi di Cramer, il Q di Yule e altri riportati nel capitolo relativo

5.3. PERCHE' LA VARIANZA DI P E' PQ E SUE CONSEGUENZE; VARIANZA E ERRORE STANDARD DI UNA FREQUENZA RELATIVA O ASSOLUTA, IN UNA POPOLAZIONE INFINITA E FINITA

Dopo la presentazione di alcuni termini del linguaggio tecnico, per l'**analisi statistica di una proporzione** è necessario riprendere i concetti presentati nella d**istribuzione binomiale**.

Nei fenomeni binari, quindi con risposte Si-No, Vivo-Morto, che per l'analisi statistica sono tradotti in **numeri con 1 – 0**, in **una popolazione** di **N** individui nella quale **X** presentano la caratteristica **A (indicata con 1),** per essa

la proporzione π nella popolazione è

$$\pi = \frac{X}{N}$$

Quando da questa popolazione si estrae un campione di dimensione n,

la **<u>proporzione campionaria</u> p** è

$$p = \frac{X}{n}$$

e la proporzione **q** della caratteristica alternativa **B (indicata con 0)**

è

$$q = 1 - p \qquad \text{oppure} \qquad q = \frac{n - X}{n}$$

All'infuori dei due casi estremi, in cui $X = \mathbf{0}$ oppure $X = \mathbf{1}$

- se dalla popolazione si estraggono casualmente vari campioni di dimensione n,
- si hanno altrettante stime campionarie p**.**

La proporzione p, anche se può apparire una singola osservazione, in realtà **è una media di un fenomeno binario,** che come misure singole ha **0** oppure **1**.

La sua varianza è la varianza di una media e può essere ricavata rapidamente con le due formule abbreviate seguenti:

- per **la popolazione** di n individui, σ_p^2 è

$$\sigma_p^2 = \frac{p \cdot q}{n}$$

- per **un campione** di n individui, s_p^2 è

$$s_p^2 = \frac{p \cdot q}{n-1}$$

Quando n è grande, come quasi sempre richiesto per una stima sufficientemente accurata di una proporzione, le differenze tra le due formule sono minime. Per tale motivo, molti testi suggeriscono la prima anche per un campione.

Per comprendere in modo semplice che **queste formule sono equivalenti a quelle classiche per il calcolo della varianza**, è utile una dimostrazione elementare. Si supponga di avere somministrato un tossico ad un gruppo di 20 cavie e che tra esse 5 abbiano presentato sintomi di intossicazione.

La **proporzione** di individui intossicati

$$p = \frac{X}{n} = \frac{5}{20} = 0{,}25$$

è $p = \mathbf{0{,}25}$

e con la formula abbreviata

la sua **varianza** s_p^2

$$s_p^2 = \frac{p \cdot q}{n-1} = \frac{0{,}25 \cdot 0{,}75}{19} = \frac{0{,}1875}{19} = 0{,}009868$$

è $s_p^2 = 0{,}009868$

(sono utilizzati più decimali, solo per dimostrare empiricamente l'uguaglianza dei due risultati).

Se a ognuno dei **5 soggetti intossicati** si attribuisce convenzionalmente valore **1** e a ognuno degli altri **15 non intossicati** valore **0**,

- la **media** $\overline{X}$ del fenomeno

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{(1 \cdot 5) + (0 \cdot 15)}{5 + 15} = \frac{5}{20} = 0{,}25$$

è uguale a 0,25

- **la sua devianza SQ**

$$SQ = \sum_{i=1}^{n} \left(X_i^2\right) - \frac{\left(\sum_{i=1}^{n} Xi\right)^2}{n} = \left(1^2 + 1^2 + 1^2 + 1^2 + 1^2\right) - \frac{5^2}{20} = 5 - 1{,}25 = 3{,}75$$

è uguale a 3,75

- e, da questa, **la sua varianza** s^2

$$s^2 = \frac{SQ}{n-1} = \frac{3{,}75}{19} = 0{,}197368$$

è uguale a 0,197368.

Ma questa è la **varianza dei dati** (0 e 1), mentre nel caso della proporzione **p** la varianza prima calcolata con la formula abbreviata è riferita alla media.

Poiché la **deviazione standard della media** (quindi l'errore standard),

è

$$e.s. = \frac{s}{\sqrt{n}}$$

e la **varianza di una media** $s_{\overline{X}}^2$ è

$$s_{\overline{X}}^2 = \frac{s^2}{n}$$

con i dati del problema, si ottiene che

la varianza della proporzione media p

$$s_p^2 = \frac{0{,}197368}{20} = 0{,}009868$$

è $s_p^2 = 0{,}009868$.

Il risultato ($s_p^2 = 0{,}009868$) è identico a quello ottenuto con il calcolo abbreviato.

Evidenziando nuovamente **i concetti principali,** questo risultato mostra che

- la varianza s_p^2

$$s_p^2 = \frac{p \cdot q}{n-1}$$

è la varianza di una media p.

Nello stesso modo,

- la sua deviazione standard (s_p)

$$s_p = \sqrt{\frac{p \cdot q}{n-1}}$$

in realtà **è l'errore standard di p**

e misura la dispersione delle medie p **di** n **elementi**, intorno al valore centrale.

Questi stessi concetti sono dimostrati in **modo più scolastico**, con una serie di passaggi logici e matematici da George W. **Snedecor** e William G. **Cochran** nel loro testo del 1974 ***Statistical Methods*** (6th ed., The Iowa State University Press, Ames, Iowa, XIV + 593, vedi pag. 208).

(1)	(2)	(3)	(4)	(5)	(6)
X	f_X	Xf_X	$X-\mu$	$(X-\mu)^2$	$(X-\mu)^2 \cdot f_X$
0	q	0	$-p$	p^2	$p^2 q$
1	p	p	$1-p$	q^2	$q^2 p$
		$\mu = p$			$\sigma_X^2 = pq$

La loro successione può essere schematizzata in 6 punti, riportati nelle colonne.

1) In una scala a intervalli o di rapporti la misura è continua e viene indicata con X; in una classificazione binaria in cui si misura la assenza - presenza del fattore, i valori possono essere tradotti in numeri con 0 e 1.

2) Raggruppando i dati per gli stessi valori, la frequenza relativa per la variabile continua X è f_X ; in una classificazione binaria, le frequenze relativa di 0 è q e la frequenza della relativa di 1 è p

3) La media μ, in una variabile continua è Xf_X; in una classificazione binaria è $1 \cdot p = p$

Quindi la prima conclusione è che $\mu_X = p$

4) Per calcolare la varianza, si deve partire dagli scarti: per la variabile continua sono $X - \mu$, mentre per la variabile discreta sono $-p$ per i valori 0 e $1 - p$ per i valori 1.

5) Questi valori devono essere elevati al quadrato, ottenendo rispettivamente

- per la variabile continua $(X - \mu)^2$,
- per la variabile binaria p^2 per i valori 0 e q^2 per i valori 1.

6) Considerato che f_X, p e q sono frequenze relative, si ricava che la varianza $\sigma_X^2 = p \cdot q$

E' la seconda conclusione, che si voleva dimostrare.

La **varianza della popolazione** e quella **campionaria della proporzione q** (con $q = 1 - p$)

sono identiche a quelle di **p**

$$\sigma_q^2 = \sigma_p^2 \qquad \text{e} \qquad s_q^2 = s_p^2$$

Ne deriva che anche la **deviazione standard** di una proporzione **p** o **q,**

$$\sigma_p = \sqrt{\frac{p \cdot q}{n}} \qquad \text{e} \qquad s_p = \sqrt{\frac{p \cdot q}{n-1}}$$

è totalmente determinata dal suo valore medio.

I calcoli sono semplici:

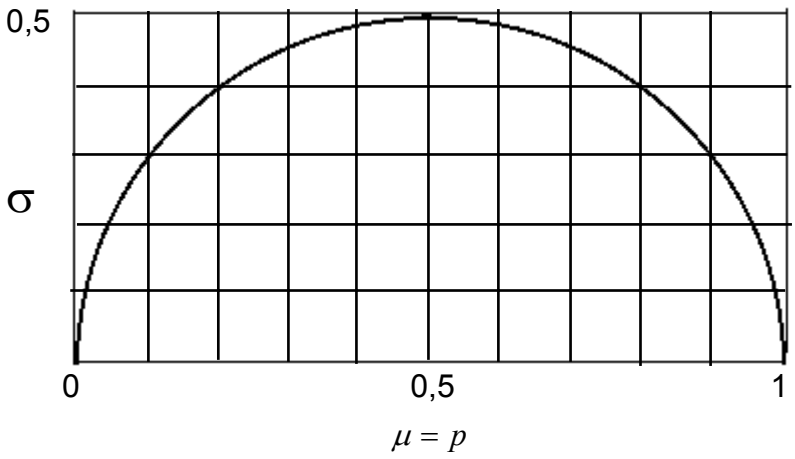
$\mu = p$	0,5	0,4 o 0,6	0,3 o 0,7	0,2 o 0,8	0,1 o 0,9	0,05 o 0,95	0,01 o 0,99
$\sigma = \sqrt{pq}$	0,500	0,490	0,458	0,400	0,300	0,218	0,099

In una rappresentazione grafica, la corrispondenza risulta visivamente molto chiara.

Se in un diagramma cartesiano si riportano

- la media **p** sull'asse delle ascisse
- la sua deviazione standard $\sigma = \sqrt{pq}$ sull'asse delle ordinate

è evidente il loro rapporto in **campioni costanti di dimensioni** n .



Questa relazione pone problemi rilevanti, quando alle proporzioni o percentuali sono applicati test parametrici per il confronto tra le medie quali p , quali il test **t** di **Student** e il test **F** di **Fisher-Snedecor**. Infatti non può essere rispettata una delle **condizioni fondamentali di validità** per confrontare due o più medie, rispetto alla quale questi test sono poco robusti:

- i vari gruppi a confronto devono **avere la stessa varianza**, anche quando le loro medie sono differenti.

Per confrontare la significatività della differenza tra medie diverse, si dovrà quindi

- nella statistica parametrica, ricorrere a **trasformazioni dei dati**, come quella in **arcoseno**;
- in altri casi, che saranno discussi in seguito, sarà più conveniente utilizzare **test non parametrici**;

Più recentemente, diffusi in particolare dai programmi informatici, ma spesso criticati nella letteratura statistica, sono proposti **test t tra due medie** e **test ANOVA tra due o più medie**, che possono essere utilizzati anche quando le varianze sono differenti, sulla base della proposta di **Beherens-Fisher**.

Un'altra conseguenza molto importante di **questa relazione nelle proporzioni** o **frequenze relative**

- tra la media $\mu = p$
- e la sua varianza $\sigma^2 = p \cdot q$

che nel caso di **conteggi o frequenze assolute** diventa la relazione

- tra la media $\mu = n \cdot p$

- e la sua varianza $\sigma^2 = n \cdot p \cdot q$

è che **non è necessario avere misure ripetute per calcolare la varianza** di un campione, ma è **sufficiente conoscere la sua media**.

CORREZIONE PER UNA POPOLAZIONE FINITA

Già diffusa da W. G. **Cochran** nel 1977 nel volume ***Sampling Techniques*** (3rd ed., John Wiley, New York, p. 428) e, fra i testi internazionali di statistica applicata a maggior diffusione, riportata da Jerrold H. **Zar** nel volume del 1999 ***Biostatistical Analysis*** (4th ed., Prentice Hall, Upper Saddle River, New Jersey, XII + 663 p. + App. 212)

- quando la **proporzione è stimata con un campione,** che è una **parte non trascurabile della popolazione** intera, la **varianza deve essere corretta** utilizzando

1 - per la **frequenza relativa p**

- **la varianza**

$$s_p^2 = \frac{p \cdot q}{n-1} \cdot \left(1 - \frac{n}{N}\right)$$

- **l'errore standard**

$$s_p = \sqrt{\frac{p \cdot q}{n-1} \cdot \left(1 - \frac{n}{N}\right)}$$

2 – **per la frequenza assoluta F** con

$$F = p \cdot N$$

- **la varianza**

$$s_F^2 = \frac{N \cdot (N-n) \cdot p \cdot q}{n-1}$$

- **l'errore standard**

$$s_F = \sqrt{\frac{N \cdot (N-n) \cdot p \cdot q}{n-1}}$$

dove

- n = numero di unità che formano il **campione raccolto**,
- N = numero di unità che formano la **popolazione finita**.

Il concetto è semplice:

- Se si calcola una **proporzione** p ,
- utilizzando **tutti gli elementi di una popolazione finita** (quindi **π**),
- **non esiste l'errore di campionamento**.

Ad esempio, se una popolazione è composta solo di $N = 200$ individui, come possono essere quelli sottoposti a una operazione all'anca presso la stessa clinica, e a un controllo successivo 120 pazienti si dimostrano totalmente riabilitati, dal **conteggio su tutti gli individui della popolazione** risulterà sempre che la **proporzione di guariti** della popolazione è $\pi = 120/200 = 0{,}60$ anche se viene ripetuta da persone differenti.

La varianza della proporzione π è uguale a 0.

Ma se, per **stimare la proporzione** π **vera o reale** di guariti, **si utilizza solamente un campione di** n **pazienti** (con $n < N$) e si ripete l'operazione **ricampionando,** tutte le volte le proporzioni p rilevate saranno differenti tra loro e dalla proporzione vera π .

Nella formule precedenti,

- $\frac{n}{N}$ è chiamata **quota di campionamento** (***sampling fraction***),
- $1-\frac{n}{N}$ può essere scritta anche come $N \cdot (N-n)$ ed è chiamata **correzione per la popolazione finita (*finite population correction*).**

Dalle formule precedenti è ovvio dedurre che, **quando** $n = N$,

- **la varianza e l'errore standard diventano uguali a 0**,
- sia nella **frequenza relativa** che nella **frequenza assoluta**.

ESEMPIO 1. In una popolazione di 350 pazienti sottoposti a una operazione all'anca presso la stessa clinica, si vuole valutare dopo 6 mesi dall'operazione quale è la proporzione di individui che presentano ancora difficoltà di deambulazione. Dato il costo dell'indagine, è stato utilizzato un campione più ridotto. Dall'elenco completo, sono stati estratti casualmente 160 individui e sottoposti a controllo; tra essi 28 presentavano ancora difficoltà.

Calcolare **la varianza** e **l'errore standard**

- (a) **della proporzione** p
- (b) della **frequenza assoluta** F

Risposta. Prima di tutto, con $N = 350$ e $n = 160$ occorre calcolare la proporzione o frequenza relativa p e la frequenza assoluta F

$$p = \frac{28}{160} = 0{,}175 \qquad F = 0{,}175 \cdot 350 = 61{,}25$$

ottenendo $p = 0{,}175$ e $F = 61{,}25$.

Con i dati dell'esperimento, **si stima** che **nella popolazione** di 350 pazienti

- la **frequenza relativa** di persone ancora non guarite è $p = 0{,}175$, anche se in realtà è stata calcolata solamente su 160 individui;
- la **frequenza assoluta** di persone ancora non guarite è $F = 61{,}25$, nell'ipotesi che la proporzione calcolata sul campione di 160 individui sia vera anche nella popolazione totale di 350.

Sono **stime**; quindi **hanno un errore** o meglio una variabilità, che dipende da chi erano i 160 individui controllati.

A) Per la **frequenza relativa** stimata $p = 0{,}175$

- la varianza è

$$s_p^2 = \frac{p \cdot q}{n-1} \cdot \left(1 - \frac{n}{N}\right) = \frac{0{,}175 \cdot 0{,}825}{160-1} \cdot \left(1 - \frac{160}{350}\right) = 0{,}00091 \cdot 0{,}54286 = 0{,}00049$$

$s_p^2 = 0{,}00049$

- l'errore standard è

$$s_p = \sqrt{\frac{p \cdot q}{n-1} \cdot \left(1 - \frac{n}{N}\right)} = \sqrt{0{,}00049} = 0{,}02223$$

(B) Per la **frequenza assoluta** stimata $F = 61{,}25$

- la varianza è

$$s_F^2 = \frac{N \cdot (N-n) \cdot p \cdot q}{n-1} = \frac{350 \cdot (350-160) \cdot 0{,}175 \cdot 0{,}825}{160-1} = \frac{9600{,}9}{159} = 60{,}38$$

$s_p^2 = 60{,}38$

- l'errore standard è

$$s_F = \sqrt{\frac{N \cdot (N-n) \cdot p \cdot q}{n-1}} = \sqrt{60{,}38} = 7{,}771$$

$s_F = 7{,}771$.

Nel caso di conteggi e proporzioni, la **distribuzione normale è un approssimazione asintotica** (per n che tende all'infinito)

- della **distribuzione binomiale,** quando p e q sono lontani dagli estremi 0 e 1.
- della **distribuzione poissoniana**, quando p tende a 0;
- della **distribuzione ipergeometrica** che, come in questo caso, si applica a una popolazione di dimensione N , finita e piccola.

5.4. INTERVALLO DI CONFIDENZA DI UNA FREQUENZA RELATIVA O ASSOLUTA CON LA NORMALE, IN UNA POPOLAZIONE INFINITA O FINITA; METODI GRAFICI PER L'INTERVALLO FIDUCIALE E LA STIMA DEL NUMERO DI DATI.

Per stimare i **limiti di confidenza di una proporzione o frequenza relativa** p , la procedura è analoga a quella per la media. La differenza fondamentale deriva dal fatto che dalla **media** p si **ricava direttamente l'errore standard**: non è necessario calcolarlo su una serie di proporzioni. Con un campione di dimensioni n

- l'**errore standard** di p è

$$\sqrt{n \cdot p \cdot q}$$

Nella ricerca statistica, qualche volta è nota la **proporzione vera** o **reale**, detta più tecnicamente anche **proporzione della popolazione** (π). Ad esempio, in un processo industriale di selezione della frutta per scartare quella troppo piccola o immatura, può essere nota quale sia la proporzione di scarti di quella annata almeno a grandi linee. Ma con una macchina o un gruppo di operai che selezionano n oggetti ogni ora, la proporzione p oraria di scarti non è sempre uguale.

Conoscendo la proporzione reale π di una popolazione, è possibile **stimare la distribuzione della proporzione campionaria** p **, in un gruppo di** n **oggetti**,

mediante la relazione

$$\mathrm{P}\left(\pi - Z_{\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} < p < \pi + Z_{\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \right) = 1\text{-}\alpha$$

Essa significa che,

- con una probabilità di affermare il vero uguale a $1-\alpha$,
- il valore della proporzione campionaria p
- si trova tra la proporzione vera π più e meno il valore di Z per l'errore standard di π.

Per la probabilità del 95% ($\alpha = 0.05$) può essere scritta come

$$\mathrm{P}\left(\pi - 1{,}96 \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} < p < \pi + 1{,}96 \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \right) = 0{,}95$$

Gli stessi concetti sull'**intervallo di confidenza della proporzione** p sono definiti più rapidamente con la formula seguente

$$p = \pi \pm Z_{\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

dove,

in una **distribuzione normale bilaterale** (quindi **α/2** in ogni coda), il valore di **Z**

- per la probabilità del 95% è Z = 1,96
- per la probabilità del 99% è Z = 2,576 (spesso arrotondato nei testi in 2,58).

ESEMPIO 1 (DALLA POPOLAZIONE AL CAMPIONE). Con numerose ricerche è stato dimostrato che un tossico diluito in acqua alla concentrazione standard determina mediamente la morte del 30% degli individui della specie A.
Alla probabilità del 95% entro quali limiti sarà compresa la frequenza relativa dei decessi in un esperimento con 80 individui?

Risposta. Con **Z** = 1,96 associata alla probabilità **α = 0.05**, con **π = 0,3** e **n = 80** come risulta dai dati dell'esempio

$$p = \pi \pm Z \cdot \sqrt{\frac{p \cdot q}{n}} = 0{,}3 \pm 1{,}96 \cdot \sqrt{\frac{0{,}3 \cdot 0{,}7}{80}} = 0{,}3 \pm 1{,}96 \cdot 0{,}051 = 0{,}3 \pm 0{,}10$$

si ottiene

- una proporzione p di decessi che, con una probabilità del 95%, sarà compreso nell'intervallo che
- come limite inferiore ha $L_1 = 0{,}2$
- come limite superiore ha $L_2 = 0{,}4$.

ESEMPIO 2 (DALLA POPOLAZIONE AL CAMPIONE). Il tossico X determina la morte del 4% delle cavie utilizzate. Entro quali limiti alla probabilità del 99% sarà compresa la percentuale di decessi in un esperimento con 500 individui?

Risposta. Con **Z = 2,58** associata alla probabilità bilaterale **α = 0.01** e con **π = 0,04** e **n = 500**

$$p = \pi \pm Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}} = 0{,}04 \pm 2{,}58 \cdot \sqrt{\frac{0{,}04 \cdot 0{,}96}{500}} = 0{,}04 \pm 2{,}58 \cdot 0{,}0088 = 0{,}04 \pm 0{,}023$$

per il valore campionario di frequenza relativa *p*

- si stima un intervallo che al 99% di probabilità è compreso tra
- il limite inferiore $\mathbf{L_1 = 0{,}017}$
- il limite superiore $\mathbf{L_2 = 0{,}063}$.

Per un uso più immediato, sovente i testi di statistica applicata riportano, in forma grafica oppure in tabelle, il campo di variazione (alla probabilità **1-α** prefissata) di una percentuale campionaria **p,** estratta da una popolazione con percentuale vera **π**. Un esempio dei valori, indicati come proporzioni, sono quelli della tabella successiva.

INTERVALLO DI VARIAZIONE DI *p* IN RAPPORTO ALLA PROPORZIONE **π**
DELLA POPOLAZIONE E ALLA DIMENSIONE *n* DEL CAMPIONE,
ALLA PROBABILITÀ DEL 95 % .

DIMENSIONI (n) DEL CAMPIONE

π	**20**	**50**	**100**	**200**	**500**	**1000**
0.050	--- ---	--- ---	.007 - .093	.020 - .080	.031 - .069	.036 - .064
0.100	--- ---	.017 - .183	.041 - .159	.058 - .142	.074 - .126	.081 - .119
0.200	.025 - .375	.089 -.311	.122 - .278	.145 - .255	.165 - .235	.175 - .225
0.300	.099 - .501	.173 - .427	.210 -.390	.236 - .364	.260 - .340	.272 - .328
0.400	.185 – 615	.264 - .536	.304 - .496	.332 - .468	.357 - .443	.370 - .430
0.500	.281 – 719	.361 – 639	.402 - .598	.431 – 569	.456 - .544	.469 - .531
0.600	.385 - .815	.464 – 736	.504 - .696	.532 - .668	.557 - .643	.570 - .630
0.700	.499 - .901	.573 – 827	.610 - .790	.636 - .764	.660 – 740	.672 - .728
0.800	.625 - .975	.689 - .911	.722 - .878	.745 - .855	.765 - .835	.775 - .825
0.900	--- ---	.817 – 983	.841 - .959	.858 - .942	.874 - .926	.881 - .919
0.950	--- ---	--- ---	.907 - .993	.920 - .980	.931 - .969	.936 - .964

La sua lettura è semplice.

Per esempio, **estraendo da una popolazione** che ha una proporzione π = 0.30 un campione di 20 individui, la percentuale campionaria **p** con probabilità del 95% è compresa nell'intervallo tra .099 e .501. E' un intervallo obiettivamente molto grande. Ma deriva dal fatto che una classificazione qualitativa fa perdere molta informazione, rispetto a una misura quantitativa, come utilizzata nel capitolo precedente.

Mantenendo **costante la probabilità α di un errore di I Tipo**, **all'aumentare del numero di osservazioni (n) il campo di variazione della stessa percentuale campionaria p si riduce**.

Continuando l'esempio sempre per **π = 0.30** e **α = 0.05,**

- con 50 osservazioni **p** è compresa tra 0,173 e 0,427;
- con 100 osservazioni tra 0,210 e 0,390;
- con 200 osservazioni tra 0,236 e 0,364;
- con 500 osservazioni tra 0,260 e 340;
- con 1000 osservazioni tra 0,272 e 0,328.

La tabella mostra anche che, alla stessa probabilità di affermare il vero del 95% e con lo stesso numero (**n**) di osservazioni,

- **il campo di variazione di p è massimo** quando **π= 0,50**
- **e minimo verso gli estremi 0 e 1, in modo simmetrico.**

Nella tabella, è utile osservare che **non sono stati riportati** i valori dell'**intervallo fiduciale o di confidenza** per le proporzioni π vicine a **0** né per quelle vicine a **1**, con dimensioni campionarie (**n**) ridotte.

Il motivo è che

- quando i **campioni sono piccoli** e π è vicino agli **estremi,**
- la distribuzione **non può essere approssimata alla normale standardizzata**.

Nella stima dell'intervallo di confidenza, essa potrebbe fornire estremi $\mathbf{L_1}$ e $\mathbf{L_2}$ negativi oppure superiori a 1, che sono **valori privi di significato per una proporzione**. Questa anomalia deriva dal fatto che con valori vicino agli estremi, la distribuzione delle probabilità p non è simmetrica. Di conseguenza, **si deve ricorrere alla distribuzione binomiale**, già illustrata nel Capitolo 2 sulle distribuzioni teoriche e riportata anche in una paragrafo successivo per questo uso specifico.

Molto spesso, negli esperimenti in laboratorio e nella raccolta dei dati in natura, **la situazione è opposta a quella appena illustrata**: con un esperimento,

- è frequente ottenere la **stima di una proporzione campionaria p (r/n),**
- dalla quale si vuole ricavare la **stima della frequenza relativa π,** chiamata proporzione vera oppure proporzione della popolazione.

Come suggerito da vari autori di testi di statistica, tra i quali W. G. **Cochran** (vedi del 1977 il testo ***Sampling Techniques***, 3rd ed. John Wiley, New York, 428 pp.), il modo più semplice

- per stimare l'**intervallo di confidenza di una proporzione campionaria p**,
- che sia stata calcolata su **n** dati,
- estratti casualmente da una popolazione teoricamente infinita e con proporzione reale **π**,

utilizza la **distribuzione normale** e la sua **deviazione standard**:

$$\pi = p \pm Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n-1}}$$

dove

- per la probabilità del 95% ($\alpha = 0.95$) il valore di **Z** è 1,96
- per una probabilità del 99% ($\alpha = 0.99$) il valore di **Z** è 2,58.

Scritto in modo più formale,

$$\mathrm{P}\left(p - Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n-1}} < \pi < p + Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n-1}} \right) = 1\text{-}\alpha$$

ESEMPIO 3 (DAL CAMPIONE ALLA POPOLAZIONE). In un campione di 80 fumatori, il 35% ha presentato sintomi di polmonite. Quali sono i limiti entro i quali alla probabilità del 95% e del 99% si troverà la media reale (**π**) di individui con sintomi di polmonite, nella popolazione dei fumatori?

Risposta. Dopo aver individuato i termini della domanda

$p = 0{,}35$ $\quad n = 80$ $\quad Z_{0.05}$ bilaterale = 1,96 $\quad Z_{0.01}$ bilaterale = 2,58

si calcola entro quale intervallo si troverà la proporzione vera π.

1 – Con probabilità di affermare il vero del **95%** ($1 - \alpha = 0.95$)

la **proporzione vera π** si troverà tra

$$\pi = 0{,}35 \pm 1{,}96 \cdot \sqrt{\frac{0{,}35 \cdot 0{,}65}{79}} = 0{,}35 \pm 1{,}96 \cdot 0{,}0537 = 0{,}35 \pm 0{,}1052 = \begin{matrix} 0{,}2448 \\ 0{,}4552 \end{matrix}$$

- il limite inferiore $L_1 = 0{,}2448$
- il limite superiore $L_2 = 0{,}4552$.

2 - Con probabilità di affermare il vero del 99% ($1 - \alpha = 0.99$)
la **proporzione vera π** si troverà tra

$$\pi = 0{,}35 \pm 2{,}58 \cdot \sqrt{\frac{0{,}35 \cdot 0{,}65}{79}} = 0{,}35 \pm 2{,}58 \cdot 0{,}0537 = 0{,}35 \pm 0{,}1385 = \begin{matrix} 0{,}2115 \\ 0{,}4885 \end{matrix}$$

- il limite inferiore $L_1 = 0{,}2115$
- il limite superiore $L_2 = 0{,}4885$.

La **probabilità di errore α o di I Tipo** che è associata **all'intervallo fiduciale** di π ha un significato identico a quello della probabilità α per l'intervallo di confidenza della media vera μ:
- se dalla popolazione si estraessero tutti i possibili campioni e si costruissero tutti i possibili intervalli di confidenza,
- una **frazione uguale a 1-α** comprenderebbe il valore reale di π,
- mentre la **rimanente frazione α** non lo comprenderebbe.

ESEMPIO 4 (DAL CAMPIONE ALLA POPOLAZIONE). Su un campione di 148 individui che vivono in un'area ad alto inquinamento atmosferico, 31 hanno presentato sintomi di malattie dell'apparato respiratorio.
Stimare l'intervallo di confidenza della **proporzione π**, detta proporzione vera o della popolazione, al 95% di probabilità.

Risposta. Per utilizzare la formula appena presentata, il calcolo dell'intervallo, entro il quale si troverà la proporzione reale **π** con una probabilità del 5% di errare, richiede di conoscere
- **p** = proporzione del campione, che è 31/148 = **0,209**
- **n** = numero di dati del campione, che è **148**
- **Z** per la probabilità **α = 0.05 bilaterale**, che è **1,96**

Da essi, si stima l'**intervallo fiduciale** o **intervallo di confidenza** (***confidence interval***) di π **:**

- per **limite inferiore**

$$L_1 = 0{,}209 - 1{,}96\sqrt{\frac{0{,}209 \cdot 0{,}791}{148-1}} = 0{,}209 - 1{,}96 \cdot 0{,}034 = 0{,}209 - 0{,}067 = 0{,}142$$

è $L_1 = 0{,}142$

- per **limite superiore**

$$L_2 = 0{,}209 + 1{,}96\sqrt{\frac{0{,}209 \cdot 0{,}791}{148-1}} = 0{,}209 + 1{,}96 \cdot 0{,}034 = 0{,}209 + 0{,}067 = 0{,}276$$

è $L_2 = 0{,}276$.

UN METODO GRAFICO

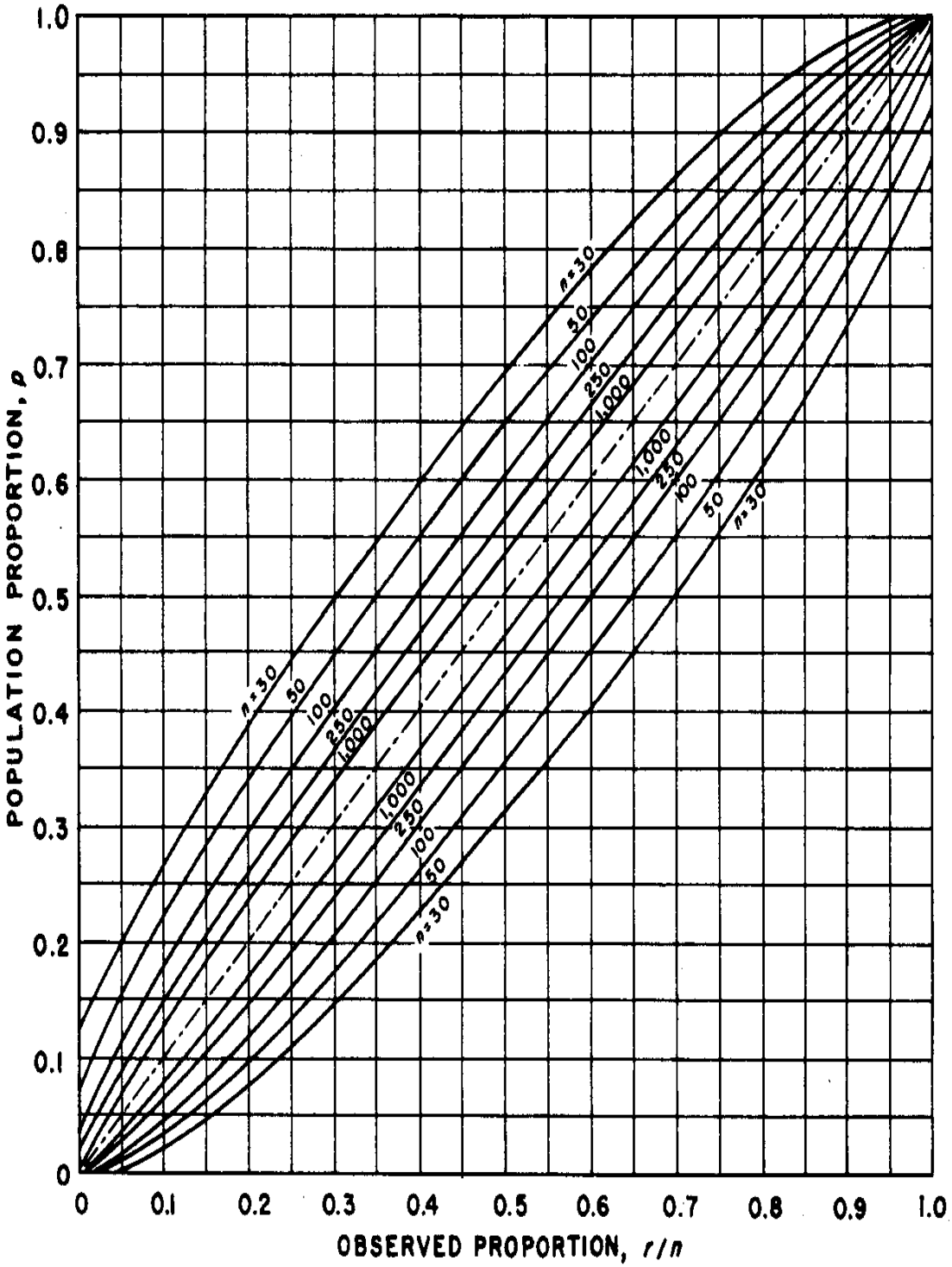
In modo molto più rapido, seppure più approssimato, è possibile ottenere gli stessi risultati sull'intervallo confidenza di π ricorrendo a tabelle, come quella illustrata in precedenza. In altro metodo simile alle tabelle, operativamente più lungo ma concettualmente altrettanto semplice, è l'uso di grafici, come i due riportati nelle pagine seguenti.

Tratti dall'articolo di C. J. **Clopper** e E. S. **Pearson** del 1934 ***The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial*** (pubblicate su **Biometrika** Vol. 26, pp.: 404-413) sono riportati anche nel manuale del Dipartimento di ricerca della Marina militare Americana, pubblicato nel 1960, dal titolo ***Statistical Manual*** (by Edwin L. **Crow**, Frances A. **Davis**, Margaret W. **Maxfield**, Research Department U. S: Naval Ordnance Test Station, Dover Pubblications, Inc., New York, XVII + 288 p.).

E' un metodo che ora è superato dall'uso dei computer, con i quali è possibile una stima sia rapida, sia precisa. Ma è sempre utile una conoscenza dei vari metodi storici, seppure a volte obsoleti, per giustificare in modo più completo la scelta del test.

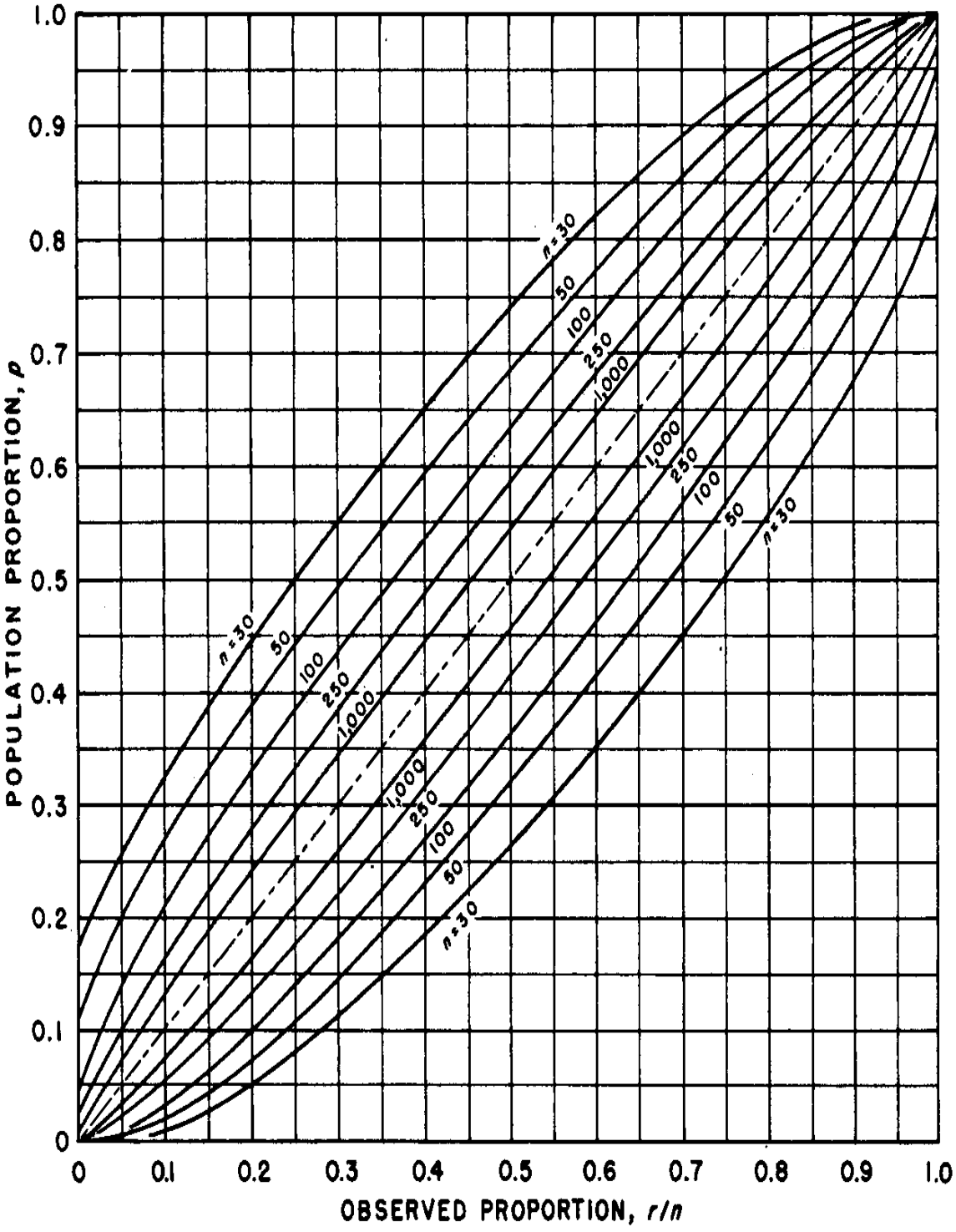
Queste **curve di confidenza delle proporzioni** (***confidence belts for proportions***), delle quali sono state riportate solamente i grafici di uso più frequente (α **= 0.05** e α **= 0.01**), sono valide per **campioni abbastanza grandi**. In questo caso, gli autori del testo definiscono tale limite quando n **> 30**.

L'uso delle **curve di confidenza** è semplice.



$\alpha = 0.05$

Strisce di confidenza per le proporzioni campionarie $p = r / n$



α = 0.01

Strisce di confidenza per le proporzioni campionarie $p = r / n$

Calcolata la **proporzione campionaria** $p = \frac{r}{n}$ (il cui valore può **variare da 0 a 1.0)**

e dopo aver scelto il grafico per la **probabilità α desiderata,** esse servono:

1 - per trovare **i limiti l'intervallo di confidenza** di π, quando sia nota la dimensione campionaria n,

2 – per valutare quale sia **la dimensione campionaria** n, dopo che sia stata scelta l'ampiezza massima possibile dell'intervallo di confidenza.

Ovviamente si ottengono **misure approssimate**,

- sia per la natura stessa del **metodo grafico**,
- sia per la **natura discreta** dei **dati originali**, i **conteggi** r e la **dimensione** n del campione.

L'uso del grafico è illustrato nei due esempi successivi, con la presentazione di due situazioni classiche della ricerca applicata:

1 – (esempio 5) dopo aver trovato una proporzione p con un campione di dimensioni n, calcolare l'intervallo di confidenza della proporzione π della popolazione, con probabilità α di errare;

2 – (esempio 6) stimare quale deve essere la dimensione n del campione da raccogliere, per ottenere una proporzione vera π con un intervallo fiduciale di ampiezza massima prestabilita, **senza avere alcuna idea di quello che sarà il valore della proporzione** p **del campione**;

ESEMPIO 5 (DAL CAMPIONE ALLA POPOLAZIONE) L'analisi di un campione di 250 sacche di plastica per la conservazione del sangue ha rilevato che, dopo un mese di custodia in frigo, quelle degradate erano esattamente 30, corrispondenti al 12% del campione analizzato. Con una probabilità del 95% di affermare il vero, indicare quale è la proporzione vera di scarti con quel metodo di conservazione.

Risposta. Dopo aver scelto la figura per $\alpha = 0{,}05$

- sull'asse delle ascisse si individua il punto che identifica $p = 0.12$;
- salendo verticalmente, si incontra la curva per $n = 250$ due volte:
- la prima in un punto che sull'**asse delle ordinate** corrisponde alla **proporzione** $\pi = 0.08$,
- la seconda in un punto che sull'**asse delle ordinate** corrisponde alla **proporzione** $\pi = 0.17$.

In conclusione, nella popolazione la percentuale di sacche degradate è compreso tra l'8% e il 17%. Questa affermazione ha una probabilità $\alpha = 0{,}05$ di essere errata (o del 95% di essere vera).

E' importante osservare che, a differenza di quanto succede con la distribuzione normale, i due limiti dell'intervallo di confidenza non sono simmetrici rispetto alla proporzione p del campione. E' quindi più rispettosa della forma reale di distribuzione delle probabilità: lontano dal valore centrale p = 0,5 l'intervallo di confidenza della proporzione π è sempre più asimmetrica.

ESEMPIO 6 (STIMARE n CON p IGNOTA E CONFRONTO CON NORMALE). Prima di effettuare il controllo, il responsabile della conservazione del sangue non aveva alcuna idea su quale sarebbe stata la proporzione p di sacche degradate, che avrebbe potuto trovare. Ma gli era stata chiesta una misura abbastanza precisa di π; più esattamente che l'intervallo tra il limite inferiore e il limite superiore non superasse il valore di 0,10 con una probabilità del 95% di affermare il vero.

Risposta. Non avendo alcuna idea sul valore che è possibile trovare, **per il principio di cautela** occorre mettersi nella **condizione meno favorevole**. **Con le proporzioni, è quando il campione è p = 0,5 poiché ha la varianza ($n \cdot p \cdot q$) massima**.

Scelto il grafico per α = 0.05 e salendo verticalmente da p = 0,5 si trova che,

- per le due curve n = 250, l'intervallo massimo è 0,14 (infatti incontra la prima curva per una proporzione della popolazione pari a 43 e la seconda per una proporzione pari a 57);
- per le due curve n = 1000, l'intervallo massimo è 0,06 (infatti incontra la prima curva per una proporzione della popolazione pari a 47 e la seconda per una proporzione pari a 53).

Il primo intervallo (0,14) è troppo grande, rispetto al valore massimo desiderato di 0,10; quindi un campione di dimensioni n = 250 è troppo piccolo.

Il secondo intervallo (0,06) è piccolo, rispetto al valore massimo desiderato di 0,10; quindi un campione di dimensioni n = 1000 è troppo grande.

Si deve ricavare una stima, utilizzando l'interpolazione lineare.

Poiché la dimensione massima individuata mediante le curve è n = 1000, l'interpolazione è fatta rispetto a $1/n$ con i seguenti calcoli:

Dopo aver valutato che le dimensioni del campione variano tra 250 e 1000,

- si calcola che, nell'unità di misura $1000/n$,
- la quantità minore n = 250 equivale a 4 e la quantità maggiore n = 250 equivale a 1;
- per cui la distanza tra i due rapporti è: 3 = 4-1.

Successivamente, per la lunghezza degli intervalli, si stima la distanza: 0,14 - 0,10 = 0,04

- e la distanza 0,14 – 0,06 = 0,08

n	$1000 / n$	Distanza 4-1	Intervallo Massimo	Distanza 0,14-0,10	Distanza 0,14-0,06
250	4		0,14		
400	***2,5***	3	0,10	0,04	0,08
1000	1		0,06		

Dalla relazione lineare

$$4 - \frac{0,04}{0,08} \cdot 3 = 2,5$$

si ricava che il valore $1000 / n$ è uguale a 2,5.

Da questo rapporto si perviene

alla **stima conclusiva**:

$$n = \frac{1000}{2,5} = 400$$

Il campione deve avere n = 400 osservazioni.

In un paragrafo successivo, questa stima è effettuata con **l'uso della distribuzione normale**. Per una sua presentazione dettagliata si rimanda ad esso. Tuttavia per un confronto dei risultati,

con essa

$$n = \frac{Z_{\alpha/2}^2 \cdot p \cdot q}{\delta^2} = \frac{1,96^2 \cdot 0,5 \cdot 0,5}{0,05^2} = \frac{0,9604}{0,0025} = 384,2$$

si **ottiene una stima abbastanza simile sulla dimensione del campione** richiesto con l'uso del grafico e l'interpolazione lineare: n = 384,2, arrotondato a 385 unità.

Nell'ultima formula,

- $Z_{\alpha/2}$ è il valore di Z per il rischio α bilaterale, che la differenza tra **p** campionario e π reale non sia superiore a δ.

Quando **a priori,** almeno in modo approssimato, **la proporzione** p che presumibilmente verrà rilevata nel campione **è nota**, la metodologia è del tutto identica a quella appena presentata per una proporzione p ignota. Ma (con la sola eccezione del caso in cui p = 0,5) si ha una varianza minore e quindi si determina un numero n di osservazioni che risulta minore.

Ad esempio (usando la distribuzione normale), se il tecnico avesse avuto una indicazione esterna, come aver letto su un rapporto oppure aver ricavato da un esperimento preliminare che la quantità di sacche degradate approssimativamente era del 12%, mantenendo costanti $Z = 1{,}96$ e $\delta = 0.05$

avrebbe ricavato

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2} = \frac{1{,}96^2 \cdot 0{,}12 \cdot 0{,}88}{0{,}05^2} = \frac{0{,}4057}{0{,}0025} = 162{,}3$$

una stima $n = 163$.

FORMULE PER UNA FREQUENZA ASSOLUTA O CONTEGGIO

L'**intervallo di confidenza** può essere calcolato anche per la **frequenza assoluta o conteggio**, con una formula più complessa di quella utilizzata per la frequenza relativa, ma sulla base di concetti del tutto uguali.

E' sufficiente illustrare la metodologia con un esempio.

Riprendendo i dati dell'esempio 5, nell'analisi di un lotto di 250 sacche di plastica per la conservazione del sangue si ipotizzi di voler stimare il numero o frequenza assoluta di quelle che saranno quelle da scartare, in lotti di 250 sacche, sempre alla probabilità $\alpha = 0.05$ che tale affermazione sia errata.

Nel Manuale della Marina Americana, già citato,

è proposta la formula

$$\frac{r + \frac{Z^2_{\alpha/2}}{2} \pm \sqrt{\left(r + \frac{Z^2_{\alpha/2}}{2}\right)^2 - \frac{\left(n + Z^2_{\alpha/2}\right) \cdot r^2}{n}}}{n + Z^2_{\alpha/2}}$$

dove

- r = conteggio o frequenza assoluta di sacche difettose
- n = numero di unità che formano il campione.

Nel caso di una popolazione finita, come può essere un lotto di $N = 600$ unità per il quale si disponga solo dell'analisi di un campione di $n = 250$ unità,

tale formula diventa

$$\frac{r + \frac{Z^2_{\alpha/2}}{2} \cdot \frac{N-n}{N-1} \pm \sqrt{\left(r + \frac{Z^2_{\alpha/2}}{2} \cdot \frac{N-n}{N-1}\right)^2 - \frac{\left(n + Z^2_{\alpha/2} \cdot \frac{N-n}{N-1}\right) \cdot r^2}{n}}}{n + Z^2_{\alpha/2} \cdot \frac{N-n}{N-1}}$$

dove,

- rispetto al prima formula, la quantità $Z^2_{\alpha/2}$ è sostituita da $Z^2_{\alpha/2} \cdot \frac{N-n}{N-1}$

5.5. INTERVALLO DI CONFIDENZA DI UNA PROPORZIONE, MEDIANTE LA DISTRIBUZIONE F

Ritornando al metodo di **calcolo dell'intervallo di confidenza di π** che utilizza la **distribuzione normale,** è necessario evidenziare che tale metodo **perde accuratezza**,

- sia quando **n** è piccolo,
- sia quando **p** è vicino a **0** oppure a **1**

Ovviamente la perdita di accuratezza è maggiore, quando si verificano entrambe queste condizioni.

Per **stime più accurate** di quelle che sono permesse dall'uso della distribuzione normale, illustrate nella prima parte del paragrafo precedente, si può ricorrere

- alla **distribuzione binomiale**, il cui intervallo di confidenza è presentato in un paragrafo successivo e tra i test per un campione di statistica non parametrica
- alla **distribuzione F,** che richiede la conoscenza dei **gradi di libertà** e quindi al **numero** n **di osservazioni** sulle quali **p** è calcolato.

Già nel 1963

- R. A. **Fisher** e F. **Yates** (nel volume ***Statistical Tables for Biological, Agricultural and Medical Research*****,** 6th ed. Hafner, New York, 146 pp.) avevano evidenziato le relazioni tra **distribuzione F** e **distribuzione binomiale**.

Per questa stima più accurata rispetto alla distribuzione normale, da utilizzare nei casi estremi per valori di p e/o n piccolo, è possibile fare uso del metodo riportato da

- **C. I. Bliss** nel 1967 nel testo ***Statistics in Biology*** (Vol. 1 McGraw-Hill, New York, 558 pp.) e riproposto più recentemente da
- Jerrold H. **Zar** nel 1999 nel testo ***Biostatistical Analysis*** (4th ed. Prentice Hall, New Jersey, 663 pp.).

In un **campione** di n individui, dei quali un numero X presentano la caratteristica in oggetto,

- i limiti di confidenza (**L_1** e **L_2**) della proporzione π

possono essere determinati con le due formule seguenti:

- per il **<u>limite inferiore</u>** **L_1**

$$L_1 = \frac{X}{X + (n - X + 1) \cdot F_{\alpha/2,\nu1,\nu2}}$$

dove i **df ν_1** e **ν_2** sono

$$\nu_1 = 2(n - X + 1) \quad \text{e} \quad \nu_2 = 2X$$

- per il **limite superiore** $\mathbf{L_2}$

$$L_2 = \frac{(X+1)\cdot F_{a/2,'\nu 1,'\nu 2}}{n - X + (X+1)\cdot F_{a/2,'\nu 1,'\nu 2}}$$

dove i **df** '$\boldsymbol{\nu_1}$ e '$\boldsymbol{\nu_2}$ sono

$$'\nu_1 = 2(X+1) = \nu_2 + 2 \quad \text{e} \quad '\nu_2 = 2(n - X) = \nu_1 - 2$$

ESEMPIO 1. Determinare l'intervallo di confidenza al 95% di probabilità della proporzione **p**, stimata su un campione casuale di 200 individui, dei quali 4 presentano la caratteristica in esame.
La domanda potrebbe essere posta anche con un linguaggio differente, più tecnico:
- calcolare i limiti dell'intervallo entro il quale si trova la proporzione reale $\boldsymbol{\pi}$ con probabilità del 95% (oppure con una probabilità $\alpha = 0.05$).

Risposta. Con **n = 200** e **X = 4**
la proporzione campionaria **p** di individui con la caratteristica in esame
risulta

$$p = \frac{X}{n} = \frac{4}{200} = 0.02$$

uguale a 0,02.
Per il limite inferiore $\mathbf{L_1}$ alla probabilità P = 0.95 dapprima
si stimano ν_1 e ν_2

$$\nu_1 = 2(n - X + 1) = 2(200 - 4 + 1) = 394$$

$$\nu_2 = 2X = 2 \cdot 4 = 8$$

che permettono di individuare il valore di F; ma poiché è raro disporre di tabelle della distribuzione **F** con gdl $\boldsymbol{\nu_1 = 394}$ si può utilizzare $\boldsymbol{\nu_1 = \infty}$ (cioè **F = 3,67)** che se ne discosta per una quantità minima;
successivamente con

$$F_{0.025,\infty,8} = 3{,}67$$

si ottiene

$$L_1 = \frac{X}{X + (n - X + 1)\cdot F_{0.025,394,8}} \approx \frac{4}{4 + (200 - 4 + 1)\cdot 3{,}67} = 0{,}0055$$

$\mathbf{L_1 = 0{,}0055}$.

Per il limite superiore $\mathbf{L_2}$, sempre alla probabilità P = 0.95,

dapprima si stimano ν_1 e ν_2

$$'\nu_1 = 2(X+1) = 2(4+1) = 10 \quad \text{oppure} \quad '\nu_1 = \nu_2 + 2 = 8 + 2 = 10$$

$$'\nu_2 = 2(n-X) = 2(200-4) = 392 \quad \text{oppure} \quad '\nu_2 = \nu_1 - 2 = 394 - 2 = 392$$

che permettono di individuare il valore di F; ma ugualmente è raro disporre di tabelle della distribuzione F con gdl $\nu_2 = 392$. Per ν_2, le tabelle riportate nei testi solitamente sono più dettagliate, seppure senza il valore esatto qui richiesto; si può quindi adoperare $\nu_2 = 300$ oppure **400**. Poiché per il principio di cautela è preferibile errare stimando un intervallo di confidenza maggiore, è conveniente scegliere **F** con $\nu_2 = 300$;

successivamente con

$$F_{0.025,10,300} = 2{,}09$$

si ottiene

$$L_2 = \frac{(X+1)\cdot F_{0.025,10,392}}{n - X + (X+1)\cdot F_{0.025,10,392}} \approx \frac{(4+1)\cdot 2{,}09}{200 - 4 + (4+1)\cdot 2{,}09} = 0{,}0506$$

L_2 = 0,0506.

Al 95 % di probabilità, la proporzione reale π si trova tra il limite inferiore **0,0055** e il limite superiore **0,0506**.

In modo convenzionale, si scrive

$$P(0.0055 \le \pi \le 0.0506) = 0.95$$

Per **la proporzione q**, l'intervallo di confidenza può essere stimato utilizzando la differenza a **1** ed invertendo **L_1** e **L_2**.

Pertanto, quando **p = 0,02** e l'intervallo fiduciale è compreso tra **L_1 = 0,0055** e **L_2 = 0,0506**

- si ha che la proporzione **q = 1 - 0,02 = 0,98** e il suo intervallo fiduciale è compreso tra
- **L_1**(di q) **= 1 - L_2** (di p)**= 1- 0,0506 = 0,9494**
- **L_2** (di q) **= 1 – L_1** (di p) **= 1- 0,0055 = 0,9945**

In questa stima dell'intervallo di confidenza, come già evidenziato la difficoltà maggiore consiste nel poter disporre di una tabella completa e molto dettagliata dei valori di F, comprendente anche i gradi di libertà grandi sia al numeratore che al denominatore. E' una tavola molto più analitica di quella che solitamente si usa per il suo impiego più frequente nella statistica applicata, il test ANOVA per il confronto tra più medie. Ma ora esistono programmi informatici che li possono generare.
Anche per calcolare l'intervallo di confidenza con il test F, è vantaggioso disporre di un programma informatico. Questa presentazione serve solamente per esporre i concetti sui quali è fondata.

Nella ricerca ambientale avviene, forse con frequenza più alta rispetto ad altre discipline, che la popolazione sia formata di un numero limitato di soggetti. I grandi laghi di una provincia, i pozzi che alimentano un acquedotto, le aziende che emettono certi fumi possono essere poche decine.
Nella produzione industriale, i prodotti sono inscatolati in lotti, formati a volte da poche centinaia o poche decine di oggetti. Ma, per questione di costi, l'analisi delle caratteristiche di un lotto avviene analizzandone solamente poche decine o poche unità.
Per stimare la proporzione di quelli che rispettano i limiti di legge, non è necessario avere una valutazione di tutti. E' possibile analizzare solo un campione. Ma, trattandosi di un campione estratto da una popolazione finita, la varianza effettiva è minore di quella stimata con le formule precedenti.

I limiti di confidenza della **proporzione p**

- in un **campione di dimensione n**,
- estratto **da una popolazione finita di N individui**,

sono minori di quelli per una proporzione di un campione estratto da una popolazione infinita. Come proposto da **H. Burstein** nel 1975 (vedi articolo ***Finite population correction for binomial confidence limits*** in Journal Amer.Statist. Assoc. vol 70, pp. 67-69) possono essere stimati apportando una correzione alla formula fondata sulla distribuzione binomiale e la distribuzione F.
Con l'uso della **distribuzione F**
1 - per il **<u>limite inferiore</u>** ,
- dopo aver calcolato $\mathbf{L_1}$ con

$$L_1 = \frac{X}{X + (n - X + 1) \cdot F_{\alpha/2,\nu1,\nu2}}$$

si ottiene il valore corretto $\mathbf{L_1(corretto)}$
con

$$L_1(corretto) = \frac{X - 0.5}{n} - \left(\frac{X - 0.5}{n} - L_1 \right) \cdot \sqrt{1 - \frac{n}{N}}$$

2 - per il **limite superiore**,

- dopo aver calcolato $\mathbf{L_2}$ con

$$L_2 = \frac{(X+1)\cdot F_{a/2,'\nu 1,'\nu 2}}{n - X + (X+1)\cdot F_{a/2,'\nu 1,'\nu 2}}$$

si ottiene il valore corretto $\mathbf{L_2(corretto)}$

con

$$L_2(corretto) = \frac{X'}{n} + \left(L_2 - \frac{X'}{n}\right)\cdot\sqrt{1-\frac{n}{N}}$$

dove

$$X' = X + \frac{X}{n}$$

ESEMPIO 2. Gli organismi di controllo della qualità dei prodotti, quasi sempre devono verificare un numero di soggetti che non è infinito. Le aziende o gli artigiani che producono rifiuti particolari (quali batterie o gomme) in una provincia, il numero di cassonetti per la raccolta differenziata di carta o erba dei giardini collocati da un'azienda, i negozi di alimentari in un distretto sanitario sono quantità limitate, spesso formato solamente da poche decine.

Una verifica raramente è estesa a tutti. Spesso, per limitare i costi o il tempo richiesto, viene campionata una frazione importante, ma compresa fra il 10 e il 20 percento dell'universo considerato.

Successivamente, nella presentazione dei risultati, per il confronto tra realtà territoriali differenti o per evidenziare più correttamente l'evoluzione temporale, è utile riportare anche **l'intervallo di confidenza della proporzione vera** π, oltre alla proporzione p di casi trovati.

Riprendendo l'esempio 1, in cui su un campione di 200 verifiche sono stati trovati 4 casi positivi, si stimi l'intervallo di confidenza alla stessa probabilità del 95%, se la popolazione totale è composta da 750 unità.

Risposta. Con 4 casi positivi su un campione di 200 (**X = 4** e **n = 200**)

la proporzione campionaria **p**

$$p = \frac{X}{n} = \frac{4}{200} = 0.02$$

è uguale a **0,02**.

Alla probabilità del 95% l'intervallo di confidenza per una popolazione infinita è risultato compreso tra

- il limite inferiore $\mathbf{L_1 = 0{,}0055}$

- il limite superiore **L_2 = 0,0506**.

Trattandosi di un universo composto da 750 (**N**) casi, in cui la frazione campionata (**n = 200**) rappresenta una quota non trascurabile, i valori corretti dell'intervallo di confidenza sono:

- per il **<u>limite inferiore</u>** (dove era uguale a **0,0055**) **il valore L_1(corretto)**

diventa

$$L_1(corretto) = \frac{4-0.5}{200} - \left(\frac{4-0.5}{200} - 0{,}0055\right)\cdot\sqrt{1-\frac{200}{750}}$$

$$L_1(corretto) = 0{,}0175 - (0{,}0175 - 0{,}0055)\cdot 0{,}8564 = 0{,}0175 - 0{,}0103 = 0{,}0072$$

uguale a 0,0072;

- per il **<u>limite superiore</u>** (dove era uguale a **0,0506**) **il valore L_2(corretto)**

dopo aver stimato

$$X' = X + \frac{X}{n} = 4 + \frac{4}{200} = 4{,}02$$

diventa

$$L_2(corretto) = \frac{4{,}02}{200} + \left(0{,}0506 - \frac{4{,}02}{200}\right)\cdot\sqrt{1-\frac{200}{750}}$$

$$L_2(corretto) = 0{,}0201 + (0{,}0506 - 0{,}0201)\cdot 0{,}8563 = 0{,}0201 + 0{,}0261 = 0{,}0462$$

uguale a 0,0462.

Come già evidenziava la formula, con il campionamento in una popolazione finita **l'intervallo fiduciale diventa minore**, fino ad annullarsi quando **n = N**.

Il confronto tra i risultati ottenuti con la distribuzione F e quelli con la distribuzione normale Z, effettuato con un campione estratto da una popolazione infinita, nel quale

$n = 200$ $\quad p = 0{,}02$ $\quad Z = 1{,}96$ per $\alpha = 0.05$

evidenzia

$$\pi = p \pm Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n-1}} = 0{,}02 \pm 1{,}96 \cdot \sqrt{\frac{0{,}02 \cdot 0{,}98}{200-1}} = 0{,}02 \pm 0{,}0195$$

un valore π che varia tra

- il minimo $L_1 = 0{,}0005$
- il massimo $L_2 = 0{,}0395$.

Presentano una **simmetria rispetto al valore centrale** campionario $p = 0{,}02$ **che è errata**, in quanto non esiste quando la proporzione p è vicino a un suo valore limite (0 e 1).

5.6. CALCOLO DEL CAMPIONE MINIMO NECESSARIO, PER LA STIMA DI UNA PROPORZIONE CAMPIONARIA CON UN ERRORE MASSIMO PREFISSATO

Quando **p** e **q** sono **non troppo vicini** a **0** oppure a **1**, è possibile utilizzare la **distribuzione normale** anche per calcolare il numero **n** di osservazioni, che è necessario per ottenere una stima **p** (quindi anche **q**) con un errore inferiore alla quantità δ prefissata.

A tale scopo, W. G. **Cochran** (nel testo già citato del 1977 ***Sampling Techniques***, 3rd ed. John Wiley, New York, 428 pp.) propone

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2}$$

dove

- $Z_{\alpha/2}$ è il valore di Z per il rischio α bilaterale; in altri termini, è la probabilità che la differenza tra **p** campionario e π reale sia superiore alla quantità δ **prefissata**.

Quando, ovviamente **in una popolazione finita di dimensione N**, il numero **n** del campione ne rappresenta una frazione non trascurabile, si può pervenire ad una stima corretta **n'**,

con

$$n' = \frac{n}{1 + \frac{n-1}{N}}$$

ESEMPIO 1. Nella popolazione, la proporzione di individui affetti da allergie è **p = 0,12**. La presenza di sostanze inquinanti aumenta tale proporzione. Quanti individui occorre analizzare, per ottenere una stima di **p** che abbia un errore massimo di 0,06 con un rischio di sbagliare α **= 0.05**?

Risposta. Con

p = 0,12 (quindi **q = 0,88**) δ **= 0,06** **Z = 1,96** per α **= 0.05** bilaterale

è richiesto

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2} = \frac{(1{,}96)^2 \cdot 0{,}12 \cdot 0{,}88}{(0{,}06)^2} = \frac{0{,}4057}{0{,}0036} = 112{,}7$$

un campione di almeno n = 113 persone (ovviamente arrotondato all'unità superiore per il principio di cautela).

Nell'utilizzo di queste formule per il calcolo di n, occorrono almeno quattro avvertenze.

1 - Il valore δ, cioè l'**errore massimo accettabile**, deve essere **espresso in proporzione** come il valore della media p. La quantità δ non è l'errore in percentuale della media, ma la differenza massima che può esistere in ognuna delle due code. Vale a dire che

- se p = 0,25 e δ = 0,06

si indica che il valore reale π alla probabilità α deve essere compreso tra

- il limite inferiore L_1= 0,19 (0,25 - 0,006)
- il limite superiore L_2 = 0,31 (0,25 + 0,06).

2 - Un altro aspetto importante del valore δ da tenere in considerazione è il suo effetto sulla dimensione n del campione. Poiché al denominatore si ha δ^2,

- un **dimezzamento dell'errore δ** massimo accettato
- comporta che **le dimensioni n del campione siano moltiplicate per 4**.

Ad esempio

con **δ = 0,03** al posto di δ = 0,06 precedente

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2} = \frac{(1{,}96)^2 \cdot 0{,}12 \cdot 0{,}88}{(0{,}03)^2} = \frac{0{,}4057}{0{,}0009} = 450{,}77$$

si ricava n = 451.

Quindi attenzione a **non diminuire eccessivamente il valore δ**, poiché la quantità n aumenta in modo eccessivo. La quantità δ, come già illustrato nel capitolo precedente per le scale a intervalli o di rapporti, dovrebbe essere scelta sulla base dei suoi effetti biologici o ambientali.

3 - Il valore δ tuttavia è legata al **valore** p. Infatti può avere senso un errore di più o meno 0,06 quando la proporzione media è p = 0,25, ma non quando la media è p = 0,04. In questo ultimo caso, ovviamente si deve diminuire l'errore massimo accettabile δ.

Diminuire δ quando p diventa piccolo ha un effetto minore sulla crescita il numero minimo di dati n. Infatti abbassando p diminuisce anche la varianza $p \cdot q$.

Ad esempio

con **δ = 0,02** e **p = 0,04**

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2} = \frac{(1{,}96)^2 \cdot 0{,}04 \cdot 0{,}96}{(0{,}02)^2} = \frac{0{,}1475}{0{,}0004} = 368{,}7$$

si ricava n = 369.

Gli effetti sono simmetrici, per cui il discorso è ugualmente valido per la proporzione **p = 0,96**.

4 – Ha un peso non trascurabile, seppure inferiore, anche la **scelta dell'errore** α. Quando è molto basso, ha ricadute non trascurabili sui costi dell'esperimento. Con gli stessi dati dell'esempio 1, ma con α = 0.01 al posto di α = 0.05 e quindi

con Z = 2,576

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2} = \frac{(2{,}576)^2 \cdot 0{,}12 \cdot 0{,}88}{(0{,}06)^2} = \frac{0{,}7007}{0{,}0036} = 194{,}6$$

si ricava n = 194,6 invece delle 112,7 unità precedenti: è un aumento del 73%.

ESEMPIO 2. In piccolo centro abitato nel quale vivono 450 persone adulte, è presente un'azienda che scarica fumi ritenuti tossici. Nella popolazione, la proporzione di individui affetti da allergie è **p = 0,12**. La presenza di sostanze inquinanti aumenta tale proporzione. Quanti individui occorre analizzare, per ottenere una stima di **p** che abbia un errore massimo di 0,06 con un rischio di sbagliare **α = 0.05**?

Nel caso di una popolazione infinita, si era stimato n = 113. Nel caso di una popolazione con **N = 450**, quanti individui occorrerà sottoporre a visita di controllo?

Risposta. Con **n = 113** e **N = 450**, il numero minimo di individui da sottoporre a visita per mantenere gli stessi parametri dell'esempio precedente

è

$$n' = \frac{n}{1 + \frac{n-1}{N}} = \frac{113}{1 + \frac{113-1}{450}} = \frac{113}{1 + 0{,}2489} = 90{,}5$$

n' = 91.

Il numero di individui da controllare è minore: scende da 113 a 91.

Nella prassi della ricerca, è frequente **il caso in cui a priori non è possibile indicare il valore di p,** neppure in modo approssimato. Quando si analizzano gli effetti di una nuova sostanza tossica, somministrata in dose elevata a cavie, non sempre è possibile stimare quale sarà la proporzione **p** di individui che risulteranno positivi, neppure in modo largamente approssimato: potrebbe essere solamente p = 0,03 ma anche p = 0,41.

Anche in questa condizione, è possibile stimare il numero **n** di dati minimo per misurare **p** con un errore che non superi la quantità δ prefissata alla probabilità α. Ma occorre porsi nella situazione sperimentale peggiore, cioè quella con la varianza massima.
Poiché la varianza è massima quando **p = 0,5**
la formula precedente

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2}$$

diventa

$$n = \frac{Z^2_{\alpha/2} \cdot 0{,}5 \cdot 0{,}5}{\delta^2}$$

Essa può essere semplificata in

$$n = \frac{Z^2_{\alpha/2}}{4 \cdot \delta^2}$$

Anche in questo caso, se la popolazione è finita e **n** è una quota non trascurabile di **N**, si può pervenire ad una stima corretta **n',**
utilizzando la formula già presentata:

$$n' = \frac{n}{1 + \frac{n-1}{N}}$$

ESEMPIO 3. Riprendendo i dati dell'esempio 1, quanti individui occorre analizzare per ottenere con un rischio di sbagliare α **= 0.05** una stima campionaria della proporzione **p** di individui affetti da allergie, che abbia una differenza massima δ **= 0,06** rispetto al valore ignoto della popolazione?

Risposta. Nell'ipotesi peggiore per le dimensioni del campione e quindi
- con **p = 0,5**
- per una differenza massima δ **= 0,06**
- e alla probabilità α **= 0.05** bilaterale quindi con **Z = 1,96**

è richiesto

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{\delta^2} = \frac{Z^2_{\alpha/2}}{4 \cdot \delta^2} = \frac{(1{,}96)^2 \cdot 0{,}5 \cdot 0{,}5}{(0{,}06)^2} = \frac{(1{,}96)^2}{4 \cdot (0{,}06)^2} = 266{,}7$$

un campione di almeno 267 persone.

Nell'esempio 1 (con **p = 0,12**), si era stimato n = 113.

ESEMPIO 4. Se la popolazione complessiva N è di 450 individui e il campione stimato n = 267, quante persone occorrerà sottoporre a visita di controllo?

Risposta. Con **n = 267** e **N = 450**, il numero minimo di individui da sottoporre a visita è

$$n' = \frac{n}{1 + \frac{n-1}{N}} = \frac{267}{1 + \frac{267-1}{450}} = \frac{267}{1 + 0{,}5911} = 167{,}8$$

n' = 168.

Nell'esempio 2 (con **p = 0,12**), si era ottenuto n = 91.

Le due ultime conclusioni sono due dimostrazioni empiriche dell'importanza e della convenienza di analizzare seriamente il problema, prima di effettuare un campionamento. Avere un'idea su quale sarà la proporzione p campionaria che presumibilmente verrà calcolata, permette un risparmio non trascurabile in tempo e denaro.

Nella ricerca applicata, è quanto si richiede quasi sempre a uno statistico professionale. Da alcuni anni, con una sempre maggiore richiesta di cultura statistica, per ogni progetto serio di ricerca si richiede sempre che venga espressa la dimensione del campione da raccogliere, per rispettare i parametri fissati. Un errore in queste ipotesi può determinare conseguenze gravi, tali da annullare buona parte della validità della ricerca. E' semplice dedurre che, per esempio, se si ipotizza una proporzione p = 0,12 e poi nell'esperimento a posteriori si trova p = 0,47 la stima della proporzione reale π avrà un errore α o un intervallo δ maggiori di quanto richiesto nella ricerca.

L'impostazione di una ricerca richiede anche una competenza elevata nel settore, non solamente buone conoscenze statistiche.

5.7. IL CONFRONTO TRA UNA PROPORZIONE CAMPIONARIA E UNA PROPORZIONE ATTESA CON IL TEST Z; DIMENSIONE MINIMA DEL CAMPIONE, PER L'USO DELLA DISTRIBUZIONE NORMALE.

Nel capitolo 3 sul chi quadrato, è stato presentato l'uso della **distribuzione normale Z**

- sia per il confronto di una **proporzione osservata con una proporzione attesa**,
- sia per il confronto tra **due proporzioni osservate**,

quando i **campioni sono di grandi dimensioni**.

Infatti, in tali condizioni, esiste corrispondenza (già illustrata) tra distribuzione χ^2 con un gdl e la distribuzione Z:

$$\chi^2_{(1)} = Z^2$$

Per una verifica empirica, è sufficiente controllare sulle tabelle dei valori critici che per α **= 0.05**

- nella **distribuzione normale bilaterale** si ha **Z = 1,96**
- nella distribuzione chi-quadrato con un gdl si ha χ^2 **= 3,84**

Nell'analisi statistica dei dati, è quindi possibile utilizzare indifferentemente metodi che si rifanno alle due distribuzioni.

Ad esempio, nello studio dell'effetto di 4 farmaci con i dati seguenti

Risposta	Farmaci			
	A	B	C	D
Guariti	37	45	48	15
Non guariti	72	87	105	58
Campione	109	132	153	73
p	0,339	0,341	0,314	0,205

mediante la **distribuzione normale o test Z** è possibile verificare

1 - se la proporzione p di guariti con il farmaco A (dove $p = 39/109$ = 0,339 o 33,9%) è statisticamente differente da una proporzione π prefissata;

2 – se la proporzione di guariti del farmaco C (con $p_C = 48/153 = 0{,}314$) è statisticamente differente da quella del farmaco D (con $p_D = 15/73 = 0{,}205$).

Per **confrontare tra loro l'effetto di tre o più farmaci** sulle frequenze di guarigione, è opportuno ricorrere al test χ^2 o al test G, già presentati. Quindi si utilizzano i dati della **tabella 2 x k** seguente:

Risposta	Farmaci			
	A	B	C	D
Guariti	37	45	48	15
Non guariti	72	87	105	58

Ma, come illustrato nel capitolo 3, è possibile servirsi di queste due distribuzioni (χ^2 o G) anche per risolvere i primi due problemi.

Con $k > 2$ i test sono sempre bilaterali, poiché le ipotesi alternative all'ipotesi nulla

$$H_0: \pi_A = \pi_B = \pi_C = \pi_D$$

sono sempre multiple, esprimibili con la frase

H_1: almeno una π è diversa dalle altre

oppure

H_1: almeno due π sono differenti tra loro χ^2 o il test G

Rifiutata l'ipotesi nulla, con metodi più sofisticati è possibile poi verificare tra quali proporzioni p_i campionarie esiste una differenza significativa. La prassi di applicare il test generale prima di passare a confronti a coppie, è **un principio di cautela** illustrato nei paragrafi dedicati ai confronti multipli. Non tutti gli autori concordano su questa cautela, ritenendo che i test siano equivalenti. In realtà la prassi di applicare un test generale, prima di utilizzare i confronti singoli, dipende dalla probabilità α complessiva, che deve essere calcolata tenendo presenti le singole probabilità α di tutti i confronti effettuati. Con il metodo generale, la stima è data molto semplicemente dal valore di **α prescelto**.

Se in un esperimento con n casi, un numero r di essi è risultato positivo,
- per confrontare se è **statisticamente significativa**

- la **differenza** tra la **proporzione** p (con $p = \frac{r}{n}$) e una **proporzione** π **di una popolazione,**

- è possibile **utilizzare** sia (1) la **frequenza assoluta,** sia (2) la **frequenza relativa:**

1 – **con la frequenza assoluta** r

$$Z = \frac{|r - n\pi_0| - 0{,}5}{\sqrt{npq}}$$

2 – con la **frequenza relativa** p

$$Z = \frac{|p - \pi_0| - \frac{1}{2n}}{\sqrt{\frac{pq}{n}}}$$

La parte **– 0,5** (spesso scritta $-\frac{1}{2}$)

- riportata al numeratore nella formula con la frequenza assoluta,
- è chiamata **correzione per la continuità** o correzione di **Yates**

(Frank Yates, inglese 1902–1994. Laureato in matematica, assistente di Ronald Fisher nel 1931 quando lavora nell'istituto di ricerche agrarie **Rothamsted Agricultural Research Institute**. Diventatone direttore, nel 1954 dirige l'installazione del primo computer inglese, l'Elliot 401, scrivendo programmi per l'analisi della varianza).

La **correzione di Yates** è motivata dal fatto che r è un conteggio, quindi un numero intero che viene collocato su una scala continua, come è il valore $n\pi$.

Ad esempio, se nella tabella precedente prendiamo

- che i guariti con il farmaco A sono r = 37 e il valore atteso era $n\pi$ = 40,3
- occorre considerare che il valore 37 non è da intendere come un valore continuo (37,00) ma un valore di una unità, collocato sul 37;
- quindi l'unità 37 su una scala continua occupa lo spazio unitario da 36,5 a 37,5.

Ne deriva che la distanza $r - np$

in realtà è

$$|37 - 40{,}3| - 0{,}5 = 2{,}8$$

I suoi effetti sono ovvii: il valore di Z risulta minore.

Con l'uso di una frequenza relativa, la **correzione di Yates** diventa $-\frac{1}{2n}$

Per utilizzare la **distribuzione Z** in **test sulle proporzioni**, si richiede che le dimensioni n del campione siano **abbastanza grandi**. Questo valore n non è costante, ma dipende da quello della proporzione p.

Quando p è compreso tra 0,3 e 0,7 è ritenuto adeguato un campione di dimensioni $n \geq 30$.

Negli altri casi, è ritenuto adeguato solamente

un campione di dimensioni n

$$n > \frac{10}{\min(p, 1-p)}$$

dove al denominatore

- è da intendere il valore minore tra p e $1-p$.

Devono essere esclusi i valori estremi, minori di 0,02 oppure maggiori di 0,98.

Ovviamente, come è possibile vedere in altre parti del testo, non esiste uniformità tra le diverse scuole di statistica anche su questo piccolo problema su quando un campione può essere definito **abbastanza grande**: alcune richiedono un campione n ancore maggiore, altre si accontentano di un valore minore.

ESEMPIO 1. Calcolare le dimensioni minime n del campione per poter utilizzare la distribuzione normale Z, quando (a) $p = 0{,}5$ e (b) $p = 0{,}92$.

Risposte.

A) Quando $p = 0{,}5$ si ha che anche $1 - p = 0{,}5$;

Con il calcolo

$$n = \frac{10}{0{,}5} = 20$$

si stima che il campione minimo sarebbe $n = 20$.

B) Quando $p = 0{,}92$ si ha che anche $1 - p = 0{,}08$;

Ne deriva

$$n = \frac{10}{0{,}08} = 125$$

che il campione minimo sarebbe $n = 125$.

Quando ci confronta il valore sperimentale p con un valore atteso o teorico π, questo test per un campione può essere sia **bilaterale o a due code**, sia **unilaterale o a una coda**.

Nel caso di un **test bilaterale**, le ipotesi sono formalmente scritte come

$$H_0: \pi = \pi_0 \qquad \text{contro} \qquad H_1: \pi \neq \pi_0$$

dove l'ipotesi nulla significa che

- la proporzione π della popolazione, dalla quale è stato estratto il campione di proporzione p, è uguale a quella π_0 della popolazione di confronto;

- oppure, espressa con altre parole, il campione con proporzione p è stato estratto dalla popolazione con proporzione π_0.

Nel caso di un **test unilaterale**, l'ipotesi alternativa è scritta come

$$H_1: \pi > \pi_0 \qquad \text{oppure} \qquad H_1: \pi < \pi_0$$

La loro ipotesi nulla,

- in alcuni testi, è scritta come nel caso bilaterale $H_0: \pi = \pi_0$

- in altri e formalmente più corretta, considera anche il segno opposto.

Pertanto le ipotesi unilaterali possono essere scritte come

$$H_0: \pi \leq \pi_0 \qquad \text{contro} \qquad H_1: \pi > \pi_0$$

oppure

$$H_0: \pi \geq \pi_0 \qquad \text{contro} \qquad H_1: \pi < \pi_0$$

ESEMPIO 2. Un farmaco contro l'asma determina una netta riduzione dei sintomi nel 28% dei pazienti, già dopo una settimana di somministrazione. Per valutare gli effetti di un nuovo principio attivo, su 150 pazienti ai quali è stato somministrato nelle stesse condizioni, il numero di individui con la stessa riduzione dei sintomi è stato 57. Il nuovo principio attivo è statisticamente migliore?

Risposta. La proporzione di individui che hanno risposto positivamente alla somministrazione del farmaco nel campione con n = 150 è stato

$$p = \frac{57}{150} = 0{,}380 .$$

Il test è unilaterale, poiché si tratta di decidere se la **proporzione reale** π del nuovo farmaco, della quale **p = 0,38** è solamente la risposta di un campione, è statisticamente maggiore di π_0 = 0,280.

In termini più formali si scrive

$$H_0: \pi \leq \pi_0 \qquad \text{contro} \qquad H_1: \pi > \pi_0$$

Usando la formula con

- la **frequenza relativa**

$$Z = \frac{|p - \pi_0| - \frac{1}{2n}}{\sqrt{\frac{pq}{n}}} = \frac{|0{,}380 - 0{,}280| - \frac{1}{2 \cdot 150}}{\sqrt{\frac{0{,}38 \cdot 0{,}62}{150}}} = \frac{0{,}100 - 0{,}003}{0{,}0396} = 2{,}44$$

si stima Z = 2,44.

Usando la formula con

- la **frequenza assoluta**

$$Z = \frac{|r - n\pi_0| - 0{,}5}{\sqrt{npq}} = \frac{|57 - 150 \cdot 0{,}28| - 0{,}5}{\sqrt{150 \cdot 0{,}38 \cdot 0{,}62}} = \frac{15 - 0{,}5}{5{,}94} = 2{,}44$$

si stima lo stesso valore Z = 2,44 (a meno delle approssimazioni nei calcoli)

In una distribuzione **normale unilaterale**, a **Z = 2,44** corrisponde una probabilità **P = 0,007**.

Questo risultato significa che,

- se **H_0 fosse vera,** cioè se la proporzione vera π di guarigione del farmaco nuovo fosse 0,28 come per il farmaco vecchio,
- esiste una probabilità P = 0,007 che il farmaco nuovo dia in risultato come quello ottenuto o ancora migliore.

Una probabilità P = 0,007 (o del 7 per mille) è oggettivamente bassa. Di conseguenza, si decide di rifiutare l'ipotesi nulla e quindi implicitamente di accettare l'ipotesi alternativa.

La dizione estesa di tale conclusione è: **con probabilità P = 0,007 di errare** (perché può essere che H_0 sia vera e che il risultato sia stato ottenuto effettivamente solo per caso) **rifiuto l'ipotesi nulla e accetto l'ipotesi alternativa**.

Oppure, più sinteticamente: **con probabilità P = 0,007 il farmaco nuovo è statisticamente migliore**.

5.8. LA POTENZA A POSTERIORI E A PRIORI DI UN TEST SULLA PROPORZIONE PER UN CAMPIONE, CON L'USO DELLA NORMALE.

Se

- **p** non è troppo vicino a **0** oppure a **1** e
- il numero **n** di osservazioni è abbastanza grande,
- anche la potenza (**1-β**) di un test sulla proporzione di un campione può essere ottenuta con la **distribuzione normale Z.**

Per calcolarla, non esiste una formula unica, ma tre formule che tra loro differiscono, in rapporto alla **direzionalità dell'ipotesi alternativa** H_1:

- se è **bilaterale** (1) o **unilaterale**,
- nel caso in cui sia **unilaterale**, se **destra** (2) oppure **sinistra** (3).

Indicando con

- **p** la proporzione trovata sperimentalmente e con **q = 1- p**
- π_0 la proporzione attesa o teorica di confronto,

la formula diventa:

1) nel caso di un **test bilaterale** la potenza è ottenuta con

$$\text{1-}\beta = P\left(Z < \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} - Z_{\alpha/2}\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right) + P\left(Z > \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} + Z_{\alpha/2}\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right)$$

2) nel caso di un **test unilaterale, con** ipotesi nulla

$$\mathbf{H_0: \pi \leq \pi_0} \qquad \text{contro} \qquad H_1: \mathbf{\pi > \pi_0}$$

è ottenuta con

$$\text{1-}\beta = P\left(Z > \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} + Z_{\alpha}\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right)$$

3) per l'ipotesi nulla

$$\mathbf{H_0: \pi \geq \pi_0} \qquad \text{contro} \qquad H_1: \mathbf{\pi < \pi_0}$$

è ottenuta con

$$\text{1-}\beta = P\left(Z < \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} - Z_{\alpha}\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right)$$

ESEMPIO 1 (CALCOLO DELLA POTENZA IN UN TEST BILATERALE). In una ricerca antecedente entro un'area ad alto inquinamento, il 50% dei campioni d'acqua superava i limiti di legge. A distanza di tempo, si intende effettuare una nuova verifica, programmando 50 prelievi.
Quale è la probabilità **1-β** di trovare che una differenza di 0,10 nella proporzione di laghi inquinati risulti significativa alla probabilità $\alpha = 0.05$?

Risposta. E' un test bilaterale, in cui l'ipotesi nulla è

$$\mathbf{H_0: \pi = \pi_0 = 0.5}$$

con ipotesi alternativa bilaterale

$$\mathbf{H_1: \pi \neq \pi_0}$$

Con la formula

$$\mathbf{1-\beta} = P\left(Z < \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} - Z_{\alpha/2}\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right) + P\left(Z > \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} + Z_{\alpha/2}\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right)$$

dove

- per $\alpha = 0.05$ si ha $Z_{\alpha/2} = 1{,}96$ $\pi_0 - p = 0{,}1$ π_0 0,5
- p e q sono uguale rispettivamente a 0,4 e 0,6 (o viceversa) mentre $n = 50$

si ottiene

$$\mathbf{1-\beta} = P\left(Z < \frac{0{,}1}{\sqrt{\frac{0{,}4 \cdot 0{,}6}{50}}} - 1{,}96\sqrt{\frac{0{,}5 \cdot 0{,}5}{0{,}4 \cdot 0{,}6}} \right) + P\left(Z > \frac{0{,}1}{\sqrt{\frac{0{,}4 \cdot 6}{50}}} + 1{,}96\sqrt{\frac{0{,}5 \cdot 0{,}5}{0{,}4 \cdot 0{,}6}} \right)$$

$$\mathbf{1-\beta} = P\left(Z < \frac{0{,}1}{0{,}0693} - 1{,}96 \cdot 1{,}0206 \right) + P\left(Z > \frac{0{,}1}{0{,}0693} + 1{,}96 \cdot 1{,}0206 \right)$$

$$\mathbf{1-\beta} = P(Z < 1{,}4430 - 2{,}0004) + P(Z > 1{,}4430 + 2{,}004)$$

$$\mathbf{1-\beta} = P(Z < -0{,}5574) + P(Z > 3{,}4434)$$

Questa somma deve essere effettuata attraverso le probabilità P corrispondenti.

In una distribuzione normale

- a un valore di Z = 0,56 in una coda della distribuzione corrisponde una probabilità P = 0.288

$$P(Z < -0{,}5574) = 0.2880$$

- a un valore di Z = 3,44 in una coda della distribuzione corrisponde una probabilità P = 0.0003

$$P(Z > 3{,}4434) = 0.0003$$

Di conseguenza, la **potenza del test**

$$\mathbf{1\text{-}\beta} = 0.2880 + 0.0003 = 0.2883$$

è **1 - β = 0.2883**.

Vi sarà solamente una probabilità del 29% che il campione raccolto risulti significativo con i parametri indicati. Simmetricamente, vi sarà una probabilità del 71% di commettere un errore β, vale a dire di non trovare una differenza che in realtà esiste.

ESEMPIO 2 (CALCOLO DELLA POTENZA IN UN TEST UNILATERALE). In un'area ad alto inquinamento, il 50% dei prelievi superava i limiti di legge. Dopo un'azione di risanamento, si intende effettuare una nuova verifica, programmando 50 prelievi.

Quale è la probabilità **1-β** che una proporzione p = 0,40 di laghi inquinati risulti significativa alla probabilità $\alpha = 0.05$?

Risposta. Nella domanda si ha $\pi_0 = \mathbf{0{,}50}$ e la proporzione campionaria p = 0,40

E' un test unilaterale, in cui l'ipotesi nulla è

$$\mathbf{H_0: \pi \geq \pi_0 = 0.5}$$

e l'ipotesi alternativa unilaterale è

$$\mathbf{H_1: \pi < \pi_0}$$

Con la formula

$$\mathbf{1\text{-}\beta} = P\left(Z < \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} - Z_\alpha \sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right)$$

dove

- per $\alpha = 0.05$ si ha $Z_\alpha = 1{,}645$ $\quad \pi_0 = 0{,}5$ $\quad p = 0{,}4$ $\quad n = 50$

si ottiene

$$\textbf{1-}\beta = P\left(Z < \frac{0{,}5 - 0{,}4}{\sqrt{\frac{0{,}4 \cdot 0{,}6}{50}}} - 1{,}645\sqrt{\frac{0{,}5 \cdot 0{,}5}{0{,}4 \cdot 0{,}6}} \right)$$

$$\textbf{1-}\beta = P\left(Z < \frac{0{,}1}{0{,}0693} - 1{,}645 \cdot 1{,}0206 \right) = P\left(Z < 1{,}4430 - 1{,}6789\right) = P\left(Z < -0{,}2359\right)$$

un valore di Z = -0,2359.

In una coda della distribuzione a **Z = 0,24** corrisponde una probabilità **P = 0.405**.

La potenza di questo **test unilaterale** è **1-β = 0.405**.

Con un test unilaterale, pure mantenendo costanti tutti gli altri parametri utilizzati nell'esempio precedente, vi sarà una probabilità del 40,5% che il campione raccolto risulti significativo. Simmetricamente, vi sarà una probabilità del 59,5% di commettere un errore β, vale a dire di non trovare una differenza che in realtà esiste.

Per un **test unilaterale nell'altra direzione**, cioè per rendere significativo un aumento di 0,10 si sarebbe utilizzata la formula

$$\textbf{1-}\beta = P\left(Z > \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} + Z_\alpha \sqrt{\frac{\pi_0 \cdot (1 - \pi_0)}{pq}} \right)$$

ottenendo

$$\textbf{1-}\beta = P\left(Z > \frac{0{,}5 - 0{,}6}{\sqrt{\frac{0{,}6 \cdot 0{,}4}{50}}} + 1{,}645\sqrt{\frac{0{,}5 \cdot 0{,}5}{0{,}6 \cdot 0{,}4}} \right)$$

$$\textbf{1-}\beta = P\left(Z > \frac{-0{,}1}{0{,}0693} + 1{,}645 \cdot 1{,}0206 \right) = P\left(Z > -1{,}4430 + 1{,}6789\right) = P\left(Z > 0{,}2359\right)$$

un valore di Z = 0,2359 identico al valore precedente, ma con segno opposto.

Con le formule presentate è possibile anche stimare **n** o **potenza a priori**, cioè

- le dimensioni minime **n** del campione
- affinché la differenza tra una **proporzione attesa** π_0 e una **proporzione osservata p** risulti significativa,
- alla **probabilità** α e con il **rischio** β **prefissati**.

Dovendo considerare, come riportato nell'ultima riga, contemporaneamente due parametri, quali

- la **probabilità** α o **errore di I Tipo**,
- la **probabilità** β**,** detto anche **rischio** β o **errore di II Tipo**,

un metodo per calcolare **n** consiste nell'uso delle formule prima presentate per la potenza, ma procedendo per tentativi, in modo iterativo.

Con una presentazione più dettagliata del metodo,

- dopo aver scelto i valori di **p**, π_0 e $\mathbf{Z_\alpha}$
- si fissa un valore di **n** e se ne calcola **la potenza (1-**β**),**
- utilizzando una delle tre ultime formule presentate, in rapporto all'**ipotesi da verificare**.

Se la potenza risulta inferiore a quella prefissata, si aumenta **n**; se la potenza risulta maggiore, si può abbassare **n.**

Il metodo risulta più facilmente comprensibile in tutti i suoi passaggi logici e operativi, con lo svolgimento completo e dettagliato di un esempio.

ESEMPIO 3 (STIMA DI n CON I DATI DELL'ESEMPIO 2). In un'area ad alto inquinamento, il 50% dei prelievi superava i limiti di legge. Dopo un'azione di risanamento, si intende effettuare una nuova verifica. E' stato dimostrato che, con **n = 50,** la probabilità **1-**β che un abbassamento di 0,10 nella proporzione di laghi inquinati risulti significativa alla probabilità $\alpha = 0.05$ è uguale a 0,405.
Quanti dati occorre raccogliere, affinché la potenza sia almeno uguale o superiore a 0,80?

Risposta. Si intende applicare un test unilaterale, in cui l'ipotesi nulla è

$$\mathbf{H_0: \pi = \pi_0 = 0.5}$$

e l'ipotesi alternativa unilaterale è

$$\mathbf{H_1: \pi < \pi_0}$$

La stima della potenza **1-**β del test, con

- per $\alpha = 0.05$ unilaterale $Z_\alpha = 1{,}645$
- $\pi_0 = 0{,}5$ e $p = 0{,}4$

- n scelto intuitivamente a priori uguale a 120 (serve solo l'esperienza per indicare come primo numero un valore vicino a quello che risulterà dai calcoli),

attraverso

$$\mathbf{1\text{-}\beta} = P\left(Z < \frac{\pi_0 - p}{\sqrt{\frac{pq}{n}}} - Z_\alpha \sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{pq}} \right)$$

permette di pervenire

$$\mathbf{1\text{-}\beta} = P\left(Z < \frac{0{,}5 - 0{,}4}{\sqrt{\frac{0{,}4 \cdot 0{,}6}{120}}} - 1{,}645 \sqrt{\frac{0{,}5 \cdot 0{,}5}{0{,}4 \cdot 0{,}6}} \right) = P\left(Z < \frac{0{,}1}{\sqrt{0{,}002}} - 1{,}645\sqrt{1{,}0417} \right)$$

$$\mathbf{1\text{-}\beta} = P\left(Z < \frac{0{,}1}{0{,}0447} - 1{,}645 \cdot 1{,}0206 \right) = P(Z < 2{,}2371 - 1{,}6789) = P(Z < 0{,}5582)$$

a un valore di Z = 0,5582.

E' in risultato con Z positivo. Arrotondato a 0,56 nella coda destra della distribuzione corrisponde a una probabilità P = 0,288.

Poiché 0,4 (frequenza campionaria) è minore di 0,5 (frequenza dell'ipotesi nulla) e quindi nella distribuzione normale si trova alla sua sinistra, la potenza del test è stimata dalla probabilità complessiva che si trova a sinistra del valore Z calcolato (+0,56).

Ne deriva che la potenza 1-β di questo test è dato dalla somma della probabilità 0,50 (la parte negativa) + 0,212 (la parte positiva della probabilità, inferiore a Z = 0,56) risultando uguale a 0,712.

Più rapidamente,

$$\mathbf{1 - \beta = 1 - 0{,}288 = 0{,}712}$$

Il valore alla potenza richiesta (0,80) era superiore. Di conseguenza, i 120 dati ipotizzati sono insufficienti e serve un numero minimo **n** superiore. Si deve indicare un numero maggiore, come 160, che deve essere verificato mediante una seconda stima della potenza.

Con n = 160

$$\mathbf{1\text{-}\beta} = P\left(Z < \frac{0{,}5 - 0{,}4}{\sqrt{\frac{0{,}4 \cdot 0{,}6}{160}}} - 1{,}645 \sqrt{\frac{0{,}5 \cdot 0{,}5}{0{,}4 \cdot 0{,}6}} \right) = P\left(Z < \frac{0{,}1}{0{,}0015} - 1{,}645\sqrt{1{,}0417} \right)$$

$$\mathbf{1\text{-}\beta} = P\left(Z < \frac{0{,}1}{0{,}03873} - 1{,}645 \cdot 1{,}0206 \right) = P\left(Z < 2{,}5819 - 1{,}6789\right) = P\left(Z < 0{,}903\right)$$

si ottiene un valore di Z = 0,903.

Arrotondato a 0,90 (in difetto), nella coda destra della distribuzione ad esso corrisponde una probabilità uguale a 0,184. Di conseguenza, la potenza **1 - β** di questo test è 1 – 0,184 = 0,816.

La potenza stimata è leggermente superiore a quella richiesta e quindi può essere accettata: si devono raccogliere **n = 160** dati. E' possibile un campione leggermente minore, forse di 5 dati; ma, per affermarlo con maggiore sicurezza, occorrerebbe una terza stima con n = 155.

5.9. TEST PER UNA PROPORZIONE: LA BINOMIALE PER CAMPIONI PICCOLI E L'INTERVALLO DI CONFIDENZA CON F PER CAMPIONI GRANDI.

Calcolata una **proporzione sperimentale** p , si pone il problema di

- verificare se essa **si discosta significativamente da una proporzione teorica** od attesa π_0,
- ricorrendo a un **test bilaterale** oppure **unilaterale**.

Ma quando il **campione è piccolo,**

cioè quando

$$n \leq \frac{10}{\min\left(p, \quad 1 - p \right)}$$

secondo alcuni autori di testi di statistica non si possono **utilizzare le metodologie precedenti** fondate su distribuzioni continue, neppure con la correzione, ma

- si deve si ricorrere alla **distribuzione binomiale,** che è discreta.

La **distribuzione binomiale,** che si deve utilizzare appunto

- **quando** p **è grande e** n **è piccolo,**

è già stata illustrata nel capitolo II dedicato alle distribuzioni teoriche più importanti per le applicazioni della statistica nella ricerca biologica e ambientale.

E' da ricorda tuttavia che, sotto l'aspetto teorico, i metodi per le proporzioni p_1 e p_2 di due campioni indipendenti quali

- il test **chi-quadrato**, il metodo esatto di **Fisher** e il **metodo G**, si rifanno alla **distribuzione poissoniana**, che è
- **valida quando** p **è piccolo e** n **è grande**.

In questo paragrafo, l'ulteriore esposizione è limitata a un esempio sull'utilizzo della distribuzione binomiale, per un **test sulla significatività di una proporzione campionaria** p .

ESEMPIO 1 (TEST UNILATERALE CON LA BINOMIALE, PER UN CAMPIONE PICCOLO). In un'area altamente inquinata, esattamente il 50% dei numerosi campioni prelevati nei corsi d'acqua superava i limiti di legge.
Dopo un'azione di risanamento, è stata condotta una prima verifica con un campione molto piccolo: su 12 prelievi, effettuati in zone scelte con estrazione casuale, solo 2 superano i limiti di legge.
Con questi pochi dati, si può affermare che la proporzione di aree inquinate si è abbassata in modo significativo?

Risposta. In termini più formali, indicando con
- $\pi_0 = 0{,}5$ la proporzione reale del primo periodo,
- π = la proporzione reale della nuova situazione, per la quale si ha solo il campione di 12 dati categoriali (classificati in due gruppi, in funzione del fatto che il valore rilevato è superiore o inferiore ai limiti di legge) con **p = 2 / 12 = 0,167**

il problema richiede di **verificare l'ipotesi nulla** $H_0: \pi \geq \pi_0$
contro l'ipotesi alternativa unilaterale $H_1: \pi < \pi_0$

A questo scopo, attraverso la distribuzione binomiale

$$P_{(i)} = C_{12}^{i} \cdot 0{,}5^{i} \cdot 0{,}5^{12-i}$$

dove i varia da 0 a 12, si deve

1 - stimare
- la probabilità complessiva di trovare solo 2 casi positivi su 12
- oppure una situazione ancora più estrema (un solo caso positivo e zero casi positivi),

nella condizione che l'ipotesi nulla sia vera (quindi, con i dati del problema, **p = 0.5** in quanto è il valore di π_0).

Nella formula della distribuzione binomiale appena citata, variando i da 0 a 12, si ottiene la seguente serie di probabilità:

Risposte positive (i)	P(i)
0	0,00024
1	0,00293
2	0,01611
3	0,05371
4	0,12085
5	0,19336
6	0,22559
7	0,19336
8	0,12085
9	0,05371
10	0,01611
11	0,00293
12	0,00024

2 – Successivamente, di devono sommare le tre probabilità $P_{(i)}$ relative ai tre valori minori di i, ottenendo

P_0	P_1	P_2	Totale
0,00024	0,00293	0,01611	0,01928

una probabilità complessiva P = 0,01928.

Il valore totale rappresenta la **probabilità complessiva** di

- trovare per caso due campioni positivi o un solo campione oppure nessun campione positivo,
- nella ipotesi che la proporzione reale di **p** (cioè π) sia uguale a 0,5.

3 - Poiché la probabilità P di questo evento è piccola (in percentuale, esattamente P = 1,928%), si può rifiutare l'ipotesi nulla, accettando implicitamente l'ipotesi alternativa.

In conclusione, la nuova proporzione π di zone con inquinamento superiore ai limiti di legge è significativamente minore del precedente $\pi_0 = 0,5$.

Se l'ipotesi alternativa fosse stata bilaterale (vale a dire chiedersi se esiste differenza tra la situazione attuale e quella precedente, senza sapere se è migliorata o peggiorata),

- alla probabilità calcolata in una coda della distribuzione

- si sarebbe dovuto sommare la probabilità nell'altra coda.

Trattandosi di una distribuzione simmetrica, come evidenzia la tabella precedente che riporta tutte le 13 probabilità esatte, la probabilità complessiva sarebbe stata esattamente il doppio (3,856%).

Anche in questo caso si sarebbe pervenuti al rifiuto dell'ipotesi nulla, con una significatività $\alpha < 0.05$.

Un **metodo alternativo (**più complesso ma che utilizza concetti già illustrati nei paragrafi precedenti**)** per rispondere a questa ultima domanda bilaterale è

- calcolare **l'intervallo di confidenza** della nuova proporzione **p = 2/12 = 0,167.**

In questo caso, poiché la proporzione campionaria è inferiore a quella dell'ipotesi nulla, è possibile limitare il calcolo al solo

- limite superiore L_2

$$L_2 = \frac{(X+1)\cdot F_{a/2,'v1,'v2}}{n - X + (X+1)\cdot F_{a/2,'v1,'v2}}$$

utilizzando le modalità già illustrate in un altro paragrafo di questo capitolo.

Se L_2 risulta inferiore a π_0 (in questo caso 0,5) si può concludere,

- con la probabilità di errare uguale ad α scelta per il valore di F,
- che la nuova proporzione π è significativamente minore dalla precedente proporzione π_0.

Teoricamente, **la distribuzione binomiale potrebbe essere utilizzata anche per grandi campioni**.

Ma è un metodo che diventa praticamente inapplicabile, se svolto manualmente. Per questo, con grandi campioni, nella prassi della statistica che risale ai primi decenni del '900, si ricorre alla **distribuzione normale ridotta**.

Il problema di **quando il campione** di dati raccolti **sia abbastanza grande** e come comportarsi di conseguenza, è affrontato dagli autori di testi di statistica non sempre nello stesso modo. Pertanto è utile, nella pratica della ricerca e nell'uso dei test statistici, conoscere anche queste idee, che differiscono da quanto riportato nel paragrafo precedente e che rappresentano la prassi più diffusa.

Nel caso di **campioni grandi** (**n > 12** in altri testi più cautelativi **n > 20**), per i **motivi pratici** derivanti dalla difficoltà dei calcoli e dal tempo richiesto dall'uso della distribuzione binomiale, è conveniente utilizzare l'**approssimazione alla distribuzione normale**.

Ma il test Z **è sempre meno potente di quello che ricorre alla distribuzione binomiale**, come dimostrano **P. H. Ramsey** e **P. P. Ramsey** nel 1988 (vedi articolo ***Evaluating the normal approximation to the binomial test***, pubblicato su **Journal Educ**. **Statist**. Vol. 13, pp.: 264 – 282).

Quindi oltre al problema della validità del test, si pone quello della sua potenza.
Per risolvere il problema della validità, si sceglie un **comportamento cautelativo.** In inglese è detto anche **comportamento conservatore** e è contrapposto al **comportamento liberale**, che cerca la maggior potenza del test.
Come già presentato, per un comportamento cautelativo vari testi consigliano
- la **correzione per la continuità**, riducendo lo scarto tra osservato ed atteso ($X - n\pi_0$) di 0,5.

Si ottiene
- un **risultato più prudenziale**;
- ma, come **Ramsey** e **Ramsey** dimostrano, **la potenza del test diminuisce** e quindi a questo scopo (poter rifiutare l'ipotesi nulla) sarebbe **preferibile il valore di Z non corretto**.

Nella **situazione più comune** di **p = 0,5**
- che ha **applicazione in molti test non parametrici**, tra i quali uno dei più diffusi è il test dei segni, (presentato nel capitolo dei test non parametrici per un campione) si ha **buona approssimazione della binomiale alla normale**
- per α **= 0.05** e **p = 0,5** se **n ≥ 27**,
- per α **= 0.01** e **p = 0,5** se **n ≥ 19**.

Un altro indicatore sintetico utilizzato per affermare la **bontà dell'approssimazione della normale alla binomiale**
- è fornito dal valore di $\sqrt{np_0q_0}$.

In modo più dettagliato, si ottiene una buona approssimazione quando
- per α **= 0.05** se né **p** né **q** sono inferiori a **0,01** e $\sqrt{np_0q_0} \geq 10$;
- per α **= 0.01** se né **p** né **q** sono inferiori a **0,10** e $\sqrt{np_0q_0} \geq 35$.

5.10. LA POTENZA DI UN TEST PER UNA PROPORZIONE, CON L'USO DELLA DISTRIBUZIONE BINOMIALE

La **potenza di un test sulla significatività della differenza di una proporzione** sperimentale **p** rispetto ad una proporzione attesa π_0 può essere misurata in
A) una **distribuzione binomiale**,

B) una **distribuzione normale**.

I concetti sono identici, ma la procedura è differente, a motivo delle caratteristiche delle due distribuzioni. In questo caso, le più importanti sono che la binomiale fornisce **probabilità esatte**, mentre la normale fornisce **probabilità cumulate**.

A - Con una **distribuzione binomiale**, è necessario calcolare

- le probabilità di trovare ogni risposta (i) con **p** uguale al **valore di π_0 espresso nell'ipotesi nulla**,
- le probabilità di trovare ogni risposta (i) con **p** uguale al **valore p sperimentale**.

Successivamente, si individuano

- nella prima distribuzione, la zona di rifiuto,
- nella seconda distribuzione, le probabilità per le stesse risposte (i).

La cumulata di questa seconda serie di probabilità fornisce la stima della potenza (**1-β**) del test.

La cumulata delle restanti probabilità, cioè per le risposte che nella prima distribuzione cadono nella zona di accettazione, forniscono la stima dell'errore β.

Questa metodologia può essere spiegata in modo più semplice e più facilmente comprensibile con una applicazione.

ESEMPIO 1. (POTENZA DEL TEST CON I DATI DELL'ESEMPIO DEL PARAGRAFO PRECEDENTE). In un'area altamente inquinata, esattamente il 50% dei campioni prelevati nei corsi d'acqua superava i limiti di legge.

Dopo un'azione di risanamento, è stata condotta una prima verifica: su 12 prelievi in zone scelte con estrazione casuale, solo 2 superano i limiti di legge.

Stimare la potenza del test binomiale, per

A) **$\alpha = 0.05$**

B) **$\alpha = 0.01$**

Risposta. E' un **test unilaterale**, nel quale

- la probabilità dell'ipotesi nulla è **p = 0,5**
- la probabilità sperimentale di confronto è **p = 0,167**.

La procedura richiede alcuni passaggi.

Per entrambe le probabilità (**$\alpha = 0.05$** e **$\alpha = 0.01$**) con la distribuzione binomiale

$$P_{(i)} = C_{12}^{i} \cdot p^{i} \cdot q^{12-i}$$

si calcolano tutte le probabilità esatte $P_{(i)}$ per i che varia da 0 a 12 (tabella seguente).

Risposte positive (i)	Prob(i) con P = 0,5	Prob(i) con P = 0,167
0	0,0002	0,1116
1	0,0029	0,2685
2	0,0161	0,2975
3	0,0537	0,1996
4	0,1209	0,0919
5	0,1934	0,0286
6	0,2256	0,0062
7	0,1934	0,0012
8	0,1209	0,0001
9	0,0537	0,0000
10	0,0161	0,0000
11	0,0029	0,0000
12	0,0002	0,0000

Successivamente,

A) per $\alpha = 0.05$,

si individua l'**area di rifiuto dell'ipotesi nulla** nella **colonna di p = 0,5**.

Essa risulta **i = 2,** in quanto **la somma di questa probabilità insieme con i = 1 e i = 0 è inferiore a 0.05**; infatti

$$(0,0002 + 0,0029 + 0,0161) = 0,0192$$

il totale delle prime tre probabilità risulta uguale a 0,0192.

Infine, nella **colonna di p = 0,167** si sommano le probabilità con i = 0 e i = 1 e 1= 2

$$(0,1116 + 0,2685 + 0,2975) = 0,6776$$

Questa probabilità p = 0,6776 è la potenza del test (1-β).

Infatti, se il numero di risposte positive nel campione è al massimo 2, il test risulta significativo con probabilità $\alpha \leq 0.05$.

B) per $\alpha = 0.01$,

come in precedenza si individua l'**area di rifiuto dell'ipotesi nulla** nella **colonna di p = 0,5**.

In questo caso, essa risulta **i = 1** poiché **la somma di questa probabilità insieme con i = 0 è inferiore a 0.01**.

Di conseguenza, nella **colonna di p = 0,167** si sommano le probabilità con i = 0 e i = 1

$$(0,1116 + 0,2685) = 0,3801$$

Questa probabilità **p = 0,3801** è la potenza del test (**1-β**).

Infatti, con i dati campionari se il numero di risposte positive è al massimo 1, si rifiuta l'ipotesi nulla alla probabilità $\alpha \leq 0.01$.

In **test bilaterali**, la probabilità α **considera i due estremi nella distribuzione teorica,** in ognuna delle quali si valutano i valori di *i* necessari alla stima della potenza. Inoltre occorre ricordare che la probabilità di β ha sempre una distribuzione unilaterale.

1) Per $\alpha = 0.05$ la probabilità di trovare per caso uno dei tre valori estremi nelle due code è 0,0384 (dato da 0,0192 x 2)
Tuttavia, con i dati di questo esempio, in pratica i valori di *i* non vengono modificati, rispetto ad un test unilaterale (la probabilità $\alpha = 0{,}0192$ anche se moltiplicata per due è sempre inferiore a **0.05**); quindi la potenza non subisce variazioni.

Anche con $\alpha = 0.01$ la potenza del test non varia tra ipotesi unilaterale e bilaterale, a causa della forte discontinuità nelle stime di probabilità con n piccolo.
Con **n = 12**, la probabilità α complessiva per **i = 0** e **i = 1** è uguale a **0.0031**.
Anche se moltiplicata per due, resta inferiore alla probabilità prefissata di $\alpha = 0.01$.
Di conseguenza, la potenza **1-β** del test è ancora **0,3801**.

Per la stima delle dimensioni minime del campione, con la binomiale si richiedono molte coppie di distribuzioni. Il tempo richiesto dai calcoli diventa molto lungo e quindi è necessario utilizzare programmi informatici.

5.11. TEST PER LA BONTA' DELL'ADATTAMENTO DI UNA DISTRIBUZIONE OSSERVATA E LA DISTRIBUZIONE BINOMIALE, COSTRUITA CON UNA PROPORZIONE NOTA E CON UNA PROPORZIONE IGNOTA

Nel precedente capitolo III è già stato affrontato il problema, frequente nella ricerca biologica e ambientale, del

- **confronto di una distribuzione campionaria** con una **distribuzione attesa o nota.** che può essere

La distribuzione attesa può essere di qualsiasi natura, vale a dire che può essere costruita su una legge matematica oppure una legge biologica. Ad esempio, nel capitolo III è stato presentato il confronto tra il risultato di un esperimento sulla segregazione di un diibrido e la legge di Mendel.

Disponendo delle 4 classi fenotipiche, ottenute dall'incrocio tra due diibridi $AaBb \;\; x \;\; AaBb$,

AB	Ab	aB	ab
315	101	108	32

- si è verificato se esse possono essere ritenute statisticamente in accordo con la legge di Mendel, secondo la quale le quattro classi dovrebbero seguire i rapporti di

AB	Ab	aB	ab
9	3	3	1

Con il test si è voluto rispondere alla domanda
- se le differenze riscontrate tra la distribuzione osservata e quella attesa sono imputabili al caso
- oppure se quella osservata è troppo distante dall'attesa e quindi probabilmente segue una legge differente.

La medesima procedura può essere applicata alle **frequenze di un fenomeno binario**, per il quale la distribuzione teorica delle frequenze attese è fornita dalla distribuzione binomiale.
Quando il numero di classi o categorie è superiore a due, non si utilizza il metodo presentato nei paragrafi precedenti che utilizzavano una sola frequenza, ma
- il confronto tra una distribuzione osservata e la corrispondente distribuzione binomiale attesa è esteso a tutta le classi.

Con la **distribuzione binomiale**

$$P_{(i)} = C_n^i p^i q^{n-i}$$

è possibile stimare una serie di frequenze attese,
- sia quando la **proporzione** p **è nota**, sulla base di una ipotesi, legge o teoria di qualsiasi natura,
- sia quando la **proporzione** p **è ignota a priori** e quindi è calcolata sui dati campionari.

La differenza fondamentale è che
- nel **primo caso**, la distribuzione osservata e quella attesa possono avere una p **differente**,
- nel **secondo caso**, le due distribuzioni hanno la **stessa** p, per costruzione di quella teorica.

Ad esempio, in molte specie animali con differenze tra i due cromosomi sessuali, il rapporto tra i sessi è di **1:1**

- quindi la proporzione di femmine è **p** = 0.5 come atteso dalla proporzione di spermi con il cromosoma X.

In nidiate di n figli, la frequenza attesa di quelle con un numero i di femmine (con i che varia da 0 a n) può essere stimata con

$$P_{(i)} = C_n^i p^i q^{n-i}$$

dove **p = 0,5** è la probabilità attesa o teorica che ogni figlio sia femmina (e dove ovviamente q = 0,5 che sia maschio).

Per nidiate di **6 figli,** sviluppando la formula precedente della binomiale, la probabilità di avere 0 femmine, 1 femmina, 2 femmine, ecc. … è riportata nella tabella successiva

Femmine (i)	0	1	2	3	4	5	6	Totale
$P(i)$ teoriche	0,0156	0,0937	0,2344	0,3126	0,2344	0,0937	0,0156	1,00

Ora si supponga che una ricerca condotta su 68 casi (68 nidiate di 6 figli) abbia dato il seguente risultato

Femmine (i)	0	1	2	3	4	5	6	Totale
Freq.(osservate)	0	6	14	19	17	8	4	68

E' possibile chiedersi: "Le frequenze osservate sono complessivamente in accordo con le probabilità stimate nella tabella precedente?"

E' un esempio di

- verifica della **bontà di adattamento di una distribuzione osservata a una distribuzione binomiale attesa o teorica** (***goodness of fit for the binomial distribution***) in cui la p è nota a priori.

Essa (vedi tabella sottostante) richiede

- dapprima, la stima delle probabilità attese (seconda riga) avendo prefissato p = 0,5,
- rapportate alla dimensione totale del campione sperimentale (terza riga);

Femmine (i)	0	1	2	3	4	5	6	Totale
$P(i)$ attese	0,0156	0,0937	0,2344	0,3126	0,2344	0,0937	0,0156	1,00
Freq.(attese)	1,06	6,37	15,94	21,26	15,94	6,37	1,06	68,00

- Successivamente, si deve effettuare il confronto tra le frequenze osservate (seconda riga) e le frequenze attese (terza riga)

Femmine (i)	0	1	2	3	4	5	6	Totale
Freq. (osservate)	0	6	14	19	17	8	4	68
Freq. (attese)	1,06	6,37	15,94	21,26	15,94	6,37	1,06	68,00

ricorrendo al test di **Kolmogorov-Smirnov** (presentato nei capitoli di statistica non parametrica per uno e per due campioni), oppure al test χ^2 o al test **G**.
Tra i tre, il test **più potente** in questo caso è **il test di Kolmogorov-Smirnov**. Inoltre, con questi dati caratterizzati da un totale inferiore a 100 e con valori attesi nelle classi estreme molto piccoli, il test G appare preferibile al χ^2.

Utilizzando il **test** χ^2 e/o il **test G occorre** ricordare che **hanno gdl = k-1,** poiché alle frequenze attese è stato posto il solo vincolo di avere lo stesso totale della distribuzione osservata.

In altre situazioni sperimentali, **la probabilità p dell'evento è ignota** e viene quindi stimata utilizzando i risultati dell'esperimento. Per esempio, nell'uomo la probabilità che alla nascita un bambino sia maschio o femmina **non è esattamente** p **= 0,5** nonostante quanto atteso sulla base della segregazione dei cromosomi sessuali. In realtà, in tutte le popolazioni nascono più maschi che femmine. I dati raccolti su grandi popolazioni dimostrano che la probabilità che un figlio sia maschio è p = 0,515 e pertanto che sia femmina è q = 0,485.
Riprendendo l'esempio precedente condotto su 68 casi di nidiate di 6 figli,

Femmine (i)	0	1	2	3	4	5	6	Totale
Freq. (osservate)	0	6	14	19	17	8	4	68
Femmine (totali)	0	6	28	57	68	40	24	223

è semplice osservare che

- la proporzione di maschi e di femmine non è esattamente p = 0,5 ma che
- su 408 figli (68 x 6), il numero di femmine è 223.

Quindi,

- la media del numero di femmine non è 3,0 ma in realtà è 3,28 (223/68) e, mantenendo la stessa legge,
- la probabilità che un neonato sia femmina **non è p = 0,5** ma è **p = 0,547** (3,28/6 oppure 223/408).

Con **p = 0,547** (noto sulla base del risultato dell'esperimento), è possibile chiedersi se tutta la distribuzione, cioè ogni classe della distribuzione osservata, è in accordo con le frequenze teoriche di una distribuzione binomiale fondata sulla probabilità **p = 0,547**. Potrebbe infatti avvenire che,

- pure rispettando questo vincolo della proporzione media,
- la distribuzione osservata non sia in accordo con la distribuzione binomiale.

Ad esempio potrebbe avvenire che

- con frequenza nettamente maggiore dell'atteso alcune coppie abbiano solo figlie femmine e altre tutti figli maschi,
- oppure che tutte le coppie abbiano un numero equilibrato di maschi e di femmine, e quindi che le coppie con 0 e 1 femmina e quelle con 5 e 6 femmine siano pochissime o addirittura assenti.

Per risolvere questo problema,

- dopo aver stimato **p = 0,547**

1 - si calcola la probabilità che ogni nidiate di 6 figli sia formata da un numero i di femmine che varia da 0 a 6, attraverso lo sviluppo della binomiale

$$P_{(i)} = C_6^i \cdot 0{,}547^i \cdot 0{,}453^{6-i}$$

Si ottiene la serie seguente di probabilità P(i):

Femmine (i)	$C_6^i \cdot 0{,}547^i \cdot 0{,}453^{n-i}$	P(i)
0	$C_6^0 \cdot 0{,}547^0 \cdot 0{,}453^{6-0}$	0,0085
1	$C_6^1 \cdot 0{,}547^1 \cdot 0{,}453^{5-1}$	0,0627
2	$C_6^2 \cdot 0{,}547^2 \cdot 0{,}453^{6-2}$	0,1889
3	$C_6^3 \cdot 0{,}547^3 \cdot 0{,}453^{6-3}$	0,3045
4	$C_6^4 \cdot 0{,}547^4 \cdot 0{,}453^{6-4}$	0,2755
5	$C_6^5 \cdot 0{,}547^5 \cdot 0{,}453^{6-5}$	0,1332
6	$C_6^6 \cdot 0{,}547^6 \cdot 0{,}453^{6-6}$	0,0267
Totale di tutte le probabilità		1,0000

2 - Successivamente, sulla base di queste probabilità stimate, si calcolano le frequenze attese, rapportate allo stesso ammontare totale (terza riga della tabella successiva)

Femmine (*i*)	0	1	2	3	4	5	6	Totale
$P(i)$	0,0085	0,0627	0,1889	0,3045	0,2755	0,1332	0,0267	1,00
Freq. (attese)	0,58	4,26	12,85	20,70	18,73	9,06	1,82	68,00

3 - Infine, il confronto avviene tra le frequenze osservate e le nuove frequenze attese (seconda e terza riga della tabella successiva)

Femmine (*i*)	0	1	2	3	4	5	6	Totale
Freq. (osservate)	0	6	14	19	17	8	4	68
Freq. (attese)	0,58	4,26	12,85	20,70	18,73	9,06	1,82	68,00

Anche ad occhio, è possibile evidenziare come le nuove frequenze attese siano più vicine a quelle osservate nel caso precedente, in cui le frequenze attese erano state calcolate utilizzando la probabilità teorica **p = 0,5**.

In questo caso, per valutare se esiste una differenza significativa **il test G appare preferibile** al χ^2 , che richiede valori attesi maggiori nelle classi estreme.

Un altro aspetto distintivo importante di questo test rispetto al caso della **p nota a priori,** dove i gdl sono k-1, è che

- con una **proporzione p osservata**, il numero di **gdl è k-2** (dove **k** è il numero di gruppi).

Infatti la distribuzione delle frequenza attese in questo caso è stata calcolata tenendo presente due informazioni:

- il numero totale di dati **n** (68),
- la proporzione reale π (0,547).

Il test di **Kolmogorov-Smirnov**, che **non considera i gdl ma solo le dimensioni del campione**, non permette di utilizzare un valore critico inferiore a quello del caso precedente, come qui viene richiesto; di conseguenza, può risultare meno potente del test **G**.

Per l'uso del test G,

1 - dopo aver aggregato la prima classe estrema perché formata da una frequenza troppo piccola, addirittura con 0 nella classe osservata per 0 figlie femmine che con il test G impedisce i calcoli,

Femmine (i)	0-1	2	3	4	5	6	Totale
Freq. (osservate)	6	14	19	17	8	4	68
Freq. (attese)	4,84	12,85	20,70	18,73	9,06	1,82	68,00

2 – si stima il valore di G con

$$G = 2 \cdot \left(6 \ln \frac{6}{4,84} + 14 \ln \frac{14}{12,85} + 19 \ln \frac{19}{20,70} + 17 \ln \frac{17}{18,73} + 8 \ln \frac{8}{9,06} + 4 \ln \frac{4}{1,82} \right)$$

$$G = 2 \cdot \left(6 \ln 1,2397 + 14 \ln 1,0895 + 19 \ln 0,9179 + 17 \ln 0,9076 + 8 \ln 0,8830 + 4 \ln 2,1978 \right)$$

$$G = 2 \cdot \left(1,2892 + 1,2001 - 1,6277 - 1,6482 - 0,9954 + 3,1498 \right)$$

e sommando tra loro positivi e negativi

$$G = 2 \cdot \left(5,6391 - 4,2713 \right) = 2 \cdot \left(1,3678 \right) = 2,7356$$

si ottiene G = 2,7356.

3 – Il valore è molto piccolo e quindi le frequenze osservate e le frequenze attese nelle singole classi hanno complessivamente differenze piccole. Tuttavia per una presentazione più completa del metodo in tutti i passaggi logici, trattandosi di un numero di osservazioni non elevato (in totale 68 casi), è possibile apportare la **correzione di Williams**, cioè il **valore q**

stimato con la formula

$$q = 1 + \frac{k+1}{6N}$$

dove

k = 5 e N = 68

$$q = 1 + \frac{5+1}{6x68} = 1 + \frac{6}{408} = 1 + 0{,}0147 = 1{,}0147$$

risulta q = 1,0147

4 - Quindi il valore corretto di G (G_{adj})

$$G_{adj} = \frac{2{,}7356}{1{,}0147} = 2{,}696$$

risulta G_{adj} = 2,696.

Poiché il valore critico del χ^2 con 5 gdl alla probabilità α **= 0,05** è uguale a 11,071 non è possibile rifiutare l'ipotesi nulla.

Anzi, con una lettura più particolareggiata della tabella dei valori critici del chi-quadrato, si può osservare che la probabilità di trovare per solo effetto del caso uno scarto complessivo tra valori osservati e valori attesi simile a quello calcolato è vicino a quello per la probabilità α = 0,75. Pertanto

- si può concludere l'analisi affermando non solo che non è possibile rifiutare l'ipotesi nulla, ma che probabilmente l'**ipotesi nulla è vera,** poiché

- **la distribuzione osservata è molto vicina alla distribuzione binomiale teorica,** centrata sulla stessa probabilità p = 0.

5.12. TEST SULLA DIFFERENZA TRA DUE PROPORZIONI, CON IL METODO DI FELDMAN E KLUGER, PER ABBREVIARE IL METODO ESATTO DI FISHER.

Nel caso del confronto tra due proporzioni sperimentali (p_1 e p_2), per un test bilaterale oppure unilaterale un metodo consiste nel confronto tra le frequenze assolute presentate in una tabella di contingenza 2 x 2, già illustrate nel capitolo III.

Nel caso di **campioni molto piccoli**, si può utilizzare il **metodo esatto di Fisher** (***Fisher exact test***), derivato dalla distribuzione ipergeometrica.

Nel caso di **campioni intermedi**, formati complessivamente da alcune decine di osservazioni (tra 30 e 100), si possono usare sia il test **G** sia il test χ^2, eventualmente con le relative correzioni per la continuità.
Nel caso di **campioni grandi**, sono ritenuti validi il test **G**, il test χ^2 e l'approssimazione alla distribuzione normale.

Nel caso di **campioni piccoli**, il **metodo esatto di Fisher** pone il problema pratico di effettuare calcoli con i fattoriali per valori superiori a 20-30 unità; non ha soluzioni semplici e rapide, neppure ricorrendo alla trasformazione logaritmica. A questo scopo, sono state proposte varie formule abbreviate, tra le quali la **formula abbreviata** proposta da **S. E. Feldman** e **E. Kluger** nel 1963 (nell'articolo ***Short cut calculation of the Fisher-Yates "exact test"*** pubblicato su **Psychometrika** vol. 28, pp.: 289 - 291).
Riprendendo la stessa simbologia utilizzata nel capitolo precedente e gli stessi dati per meglio evidenziare il confronto diretto dei risultati

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

con il metodo esatto di Fisher la probabilità di ogni singola risposta è data da

$$P_{(i)} = \frac{n_1! n_2! n_3! n_4!}{a! b! c! d! N!}$$

Applicata all'esempio della tabella sottostante

DATI OSSERVATI	**Animali Sopravvissuti**	**Animali Morti**	**Totale**
Pesticida A	7	1	**8**
Pesticida B	3	6	**9**
Totale	**10**	**7**	**17**

si ricava che

- la probabilità di avere per caso la risposta osservata nell'esperimento, nella quale il valore più piccolo osservato nelle quattro caselle a, b, c, d è 1,

$$P_{(1)} = \frac{8!\cdot 9!\cdot 10!\cdot 7!}{7!\cdot 1!\cdot 3!\cdot 6!\cdot 17!} = 0{,}03455$$

- e la risposta successiva più estrema nella stessa direzione è

RISPOSTA PIU' ESTREMA	**Animali Sopravvissuti**	**Animali Morti**	**Totale**
Pesticida A	8	0	**8**
Pesticida B	2	7	**9**
Totale	**10**	**7**	**17**

con probabilità $P_{(0)}$ uguale a

$$P_{(0)} = \frac{8!\cdot 9!\cdot 10!\cdot 7!}{8!\cdot 0!\cdot 2!\cdot 7!\cdot 17!} = 0{,}00185$$

Secondo **il metodo di Feldman e Kluger**, questa ultima probabilità può essere ottenuta dalla precedente, in modo più rapido di quanto sia possibile con i calcoli fondati sulla distribuzione ipergeometica, che sono effettivamente lunghi da effettuare manualmente

Indicando con

- a il valore minore della prima tabella (uguale a 1 nell'esempio)
- d il valore corrispondente nella diagonale (uguale a 3 nell'esempio)
- b e c i due valori nell'altra diagonale sempre della prima tabella (uguali a 7 e a 6)

questa seconda probabilità (P_{i-1}) è ottenuta dalla precedente (P_i),

attraverso la relazione

$$P_{i+1} = P_i \cdot \frac{a \cdot d}{b' \cdot c'}$$

dove

- $b' = b + 1$
- $c' = c + 1$

ESEMPIO. Con gli stessi dati dell'ultima tabella, la seconda probabilità ($P_{(0)}$ = 0,00185) è ricavata in modo più rapido dalla precedente ($P_{(1)}$ = 0,03455),

attraverso la relazione

$$P_{i-1} = 0{,}03455 \cdot \frac{1 \cdot 3}{(7+1)\cdot(6+1)} = 0{,}03455 \cdot \frac{3}{56} = 0{,}00185$$

Nel caso di un **test bilaterale**, la distribuzione delle probabilità quasi mai è simmetrica, soprattutto quando i campioni sono molto piccoli. Come calcolare la probabilità complessiva, con il **metodo esatto di Fisher** considerando ambedue le code della distribuzione, vede gli statistici divisi. Esistono due scuole di pensiero:

- alcuni ritengono corretto **moltiplicare per due la probabilità calcolata in precedenza,** cioè stimata per la coda alla quale appartiene il valore minore della tabella;
- altri ritengono che **questo non sia un procedimento corretto**, in quanto la distribuzione spesso non è simmetrica e la probabilità, quando calcolata da un estremo fino al valore centrale, potrebbe essere maggiore di 0,5 e quindi superare 1, se moltiplicata per due. Per una probabilità, è un risultato assurdo.

Per facilitare il calcolo delle probabilità anche in un test bilaterale, **Feldman e Kluger** hanno proposto una procedura che permette di calcolare la probabilità di ognuna delle possibili risposte, a partire da un estremo.

Prima della diffusione dei computer, le proposte per una stima semplificata delle probabilità esatte in tabelle 2 x 2 e in tabelle più ampie, di dimensioni M x N, sono state numerose. Tra quelle che hanno avuto maggiore successo è da ricordare **il metodo dei coefficienti binomiali**.

In letteratura è stato discusso da vari autori, dei quali un breve elenco comprende:

- **Leslie** P. H. per il suo articolo del 1955 (***A simple methods of calculating the exact probability in 2x2 contingency tables with small marginal totals*** pubblicato su **Biometrika** Vol. 42, pp.: 522 – 523);
- **Leyton** M. K. per il suo articolo del 1968 (con ***Rapid calculation of exact probabilities for 2 x 3 contingency tables***, pubblicato da **Biometrics** vol. 24, pp.: 714 – 717);
- **Ghent** A. W. per il suo articolo del 1972 (con ***A method for exact testing of 2 x 2, 2 x 3, 3 x 3, and other contingency tables, employing binomial coefficients pubblicato*** su **Amer. Midland Natur**. Vol. 88, pp.: 15 – 27);
- **Carr** W. E. per il suo articolo del 1980 **(**con ***Fisher's exact test extended to more than two samples of equal size***, pubblicato da **Technometrics** vol. 22, pp.. 269- 270).

Attualmente, questo problema è superato dalla possibilità di calcolo dei computer.

5.13. SIGNIFICATIVITA' E INTERVALLO DI CONFIDENZA DELLA DIFFERENZA TRA DUE PROPORZIONI, CON LA DISTRIBUZIONE NORMALE.

Quando i **campioni** sono **grandi**, oltre le 200 unità secondo le indicazioni di vari autori recenti, la **significatività della differenza tra due proporzioni campionarie** $p_1 - p_2$ può essere verificata

- sia con il test χ^2 e il test **G**,
- sia con la **distribuzione normale ridotta Z,** a motivo dell'approssimazione alla normale.

Anche in questo caso, viene riassunto quanto esposto già in modo dettagliato nel capitolo III.

Per verificare l'ipotesi di una diversa incidenza delle malattie polmonari in aree ad alto e a basso inquinamento, ai fini dell'inferenza sulla differenza tra le due proporzioni, quindi per la verifica di

$$H_0: \pi_1 = \pi_2 \qquad \text{oppure} \qquad H_0: \pi_1 - \pi_2 = 0$$

è possibile presentare gli stessi dati

- sia in tabelle di contingenza 2 x 2 come la seguente

	Persone **con** malattie	Persone **senza** malattie	**Totale**
Zona a **alto** inq.	145	291	**436**
Zona a **basso** inq.	81	344	**425**
Totale	**226**	**635**	**861**

- sia con le proporzioni, come nella tabella seguente

	Persone **con** malattie	Totale persone visitate	**Proporzione**
Zona a **alto** inq.	145	436	**0,333**
Zona a **basso** inq.	81	425	**0,191**
Totale	**226**	**861**	**0,262**

La prima è l'impostazione dei dati per la formula classica del χ^2;

la seconda, per applicare la formula che utilizza

la distribuzione normale,

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

dove

- p^* è la proporzione media ponderata dei 2 gruppi a confronto.

e il risultato è uguale, poiché

$$\chi^2_{(1)} = Z^2 \text{ oppure } \sqrt{\chi^2_{(1)}} = Z$$

A differenza dei metodi classici del χ^2 e del metodo delle probabilità esatte di Fisher, con la distribuzione Z è possibile

- **valutare se la differenza tra le due proporzioni campionarie ($p_1 - p_2$) è significativamente diversa da una proporzione attesa π_0**; è la formula generale per la verifica di una differenza

con

$$Z = \frac{|p_1 - p_2| - \pi_0}{\sqrt{p^* \cdot (1 - p^*) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}$$

Con la **correzione per la continuità di Yates,** la formula del χ^2

- per la **significatività della differenza $p_1 - p_2$ rispetto a una differenza nulla**

H_0: $\pi_1 = \pi_2$ equivalente a H_0: $\pi_1 - \pi_2 = 0$

è

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

Nel test Z essa diviene

$$z = \frac{|p_1 - p_2| - \frac{1}{2}(\frac{1}{n_1} + \frac{1}{n_2})}{\sqrt{p^* \cdot (1 - p^*) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}$$

Per il semplice **confronto tra due proporzioni con un test bilaterale**, i metodi tradizionali sono **il test esatto di Fisher** (***the Fisher's exact test***) e **il test chi-quadrato con la correzione per la continuità di Yates** (***the chi-square test with Yate's continuity correction***).
Tuttavia, il ricorso al**la distribuzione normale** è frequente, poiché presenta 5 vantaggi rispetto al χ^2. Infatti essa permette

1 – la verifica di **ipotesi unilaterali** oltre a quelle bilaterali,

2 – il confronto della differenza osservata tra due proporzioni ($p_1 - p_2$) con una **differenza attesa** (π),

3 – la stima dell'**intervallo fiduciale** della differenza tra le due proporzioni,

4 – di comprendere **i parametri per il calcolo della potenza (1-β) del test**, detta potenza a posteriori,

5 - di comprendere **i parametri per il calcolo del numero minimo (*n*)** di dati necessario affinché il test risulti significativo, detto potenza a priori.

I primi due punti sono già stati illustrati nel capitolo III e rapidamente richiamati in questo paragrafo. Il punto 3 è presentato in questo paragrafo; i punti 4 e 5 saranno illustrati nel paragrafo successivo.

L'intervallo di confidenza della differenza reale $\pi_1 - \pi_2$ tra due proporzioni a partire da quelle campionarie ($p_1 - p_2$)
è dato da

$$\pi_1 - \pi_2 = (p_1 - p_2) \pm \left[Z_{\alpha/2} \cdot \sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}} + \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \right]$$

dove
- p* è la frequenza media ponderata
- **$\alpha/2$** è la probabilità prescelta in una distribuzione a due code

Questa procedura può essere utilizzata anche per **verificare la significatività della differenza in un test bilaterale**, poiché
- se una differenza tra due proporzioni è esclusa da questo intervallo, essa è significativamente diversa dalla differenza (p_1-p_2) intorno al quale è stata costruito l'intervallo fiduciale, alla probabilità α prescelta.

ESEMPIO 1. Con un sondaggio presso medici di famiglia, è stata rilevata la proporzione di persone affette da malattie polmonari, tra coloro che vivono da almeno 10 anni in zone ad inquinamento atmosferico alto o basso della stessa città. La rilevazione ha fornito i seguenti risultati

	Persone visitate	Persone **con** malattie	**Proporzione**
Zona a **alto** inq.	436	145	**0,333**
Zona a **basso** inq.	425	81	**0,191**
Totale	**861**	**226**	**0,262**

Calcolare l'intervallo di confidenza della differenza vera tra le due proporzioni, con probabilità del 95% di affermare il vero.

Risposta. Con

- $p_1 = 0{,}333$ e $n_1 = 436$
- $p_2 = 0{,}191$ e $n_2 = 425$
- $p^* = 0{,}262$ e $Z = 1{,}96$ (per $\alpha = 0.05$ considerando ambedue le code della distribuzione)

l'intervallo fiduciale della differenza

è

$$\pi_1 - \pi_2 = (0{,}333 - 0{,}191) \pm \left[1{,}96 \cdot \sqrt{\frac{0{,}262 \cdot 0{,}738}{436} + \frac{0{,}262 \cdot 0{,}738}{425}} + \frac{1}{2}\left(\frac{1}{436} + \frac{1}{425} \right) \right]$$

$$\pi_1 - \pi_2 = 0{,}142 \pm \left[1{,}96 \cdot \sqrt{0{,}000443 + 0{,}000455} + \frac{1}{2}(0{,}002294 + 0{,}002353) \right]$$

$$\pi_1 - \pi_2 = 0{,}142 \pm [1{,}96 \cdot 0{,}02997 + 0{,}002324] = 0{,}142 \pm 0{,}061$$

uguale a $0{,}141 \pm 0{,}061$.

Quindi, con probabilità del 95% di affermare il vero, la differenza vera $\boldsymbol{\pi_1 - \pi_2}$ è compresa tra

- il **limite inferiore $L_1 = 0{,}080$** (0,141 – 0,061),
- il **limite superiore $L_2 = 0{,}202$** (0141 + 0,061).

Ai fini dell'inferenza con un test bilaterale, si afferma che

- qualunque differenza risulti esclusa da questo intervallo, è significativamente differente da questa, in un **test bilaterale** alla stessa probabilità $\boldsymbol{\alpha = 0.05}$.

5.14. POTENZA A POSTERIORI (1-β) E A PRIORI (n) DEI TEST SULLA DIFFERENZA TRA DUE PROPORZIONI; BILANCIAMENTO DI DUE CAMPIONI.

Anche nel confronto tra le proporzioni ($\mathbf{p_1}$ e $\mathbf{p_2}$) di due campioni indipendenti, per valutare la significatività della loro differenza ($\mathbf{p_1 - p_2}$), è possibile commettere errori di due tipi.

Il primo, chiamato **errore di I Tipo** (***Type I error***) o di prima specie, consiste nel

- dichiarare che la differenza tra le due proporzioni è significativa, quando in realtà è nulla.

E' l'errore che ha avuto l'attenzione maggiore nelle pubblicazioni di statistica e nello studio dell'inferenza.

Ma, secondo vari autori di testi di statistica applicata, è un punto di vista puramente teorico. Nella realtà della ricerca, **è una preoccupazione eccessiva,** in quanto

- tale errore non è mai commesso nella pratica sperimentale.

Come già sottolineava **Joseph L. Fleiss** nel 1973 (nel cap. 3 del volume ***Statistical Methods for Rates and Proportion***, John Wiley & Sons, New York), molti autori di testi di statistica applicata mettono in evidenza che

- quasi mai due popolazioni sono identiche, poiché inevitabilmente esiste sempre una differenza, per quanto piccola e insignificante essa possa essere**.**

Nella programmazione di un esperimento e nella analisi statistica dei dati, quando si cerca di dimostrare che una differenza è necessario non sbagliare neppure nell'altra direzione. Con questo non intendendo l'errore β o di II Tipo, ma

l'errore di ricercare la significatività di qualsiasi differenza, per quanto piccola possa essere.

E' perciò indispensabile saper distinguere e combinare i concetti di **significatività statistica** e **significatività biologica**: è utile

- ricercare la **significatività statistica** solamente per differenze che hanno **rilevanza biologica, ambientale od ecologica**.

Nella programmazione di un esperimento, un ricercatore dovrebbe **evidenziare solo le differenze δ che assumono una importanza reale nella sua disciplina**, non astrattamente una differenza di qualsiasi entità, anche trascurabile agli effetti pratici e inutile nella interpretazione dei risultati

Di conseguenza, nella programmazione dell'esperimento si **deve evitare di raccogliere un campione molto più grande di quanto sia necessario per non commettere l'errore di II tipo, ma solo in rapporto al valore δ** predeterminato. Con un aumento ingiustificato del campione, si avrebbe un incremento dei costi e dei tempi oltre quanto è utile per conseguire il risultato con la significatività desiderata.

Per **non commettere un errore di I tipo** alla probabilità α,

- in un **test bilaterale occorre che il valore critico di Z per la probabilità $\alpha/2$ sia superiore al valore calcolato.**

Il concetto è scritto come

$$|z| > C_{\alpha/2}$$

e in un **test unilaterale**

$$|z| > C_{\alpha}$$

dove **C**

è il **valore critico** riportato nella tabella della distribuzione **Z.**

Quando $\alpha = 0.05$ tale valore è

- Z = 1,96 per un test bilaterale,
- Z = 1,645 per un test unilaterale.

Nello stesso tempo, quando si raccolgono i dati del campione, per **non commettere un errore di II tipo** alla probabilità **β, che è sempre unilaterale,** il valore critico di **Z deve essere inferiore** a quello corrispondente alla probabilità β

$$|Z| < C_{\beta}$$

Da questi concetti deriva la stima della potenza (1- β).

Secondo quanto proposto da L. A. **Marascuilo** e M. **McSweeney** nel loro testo di statistica non parametrica del 1977 (vedi: ***Nonparametric and Distribution-free Methods for the Social Sciences***, edito da **Brooks/Cole, Monterey, California**, pp. 556),

- nel caso di un test **bilaterale** quindi con

$$\mathbf{H_0: \pi_1 = \pi_2} \text{ contro } \mathbf{H_1: \pi_1 \neq \pi_2}$$

la potenza **1-β** è ottenuta con

$$1-\beta = P\left[Z \leq \frac{-Z_{\alpha/2}\sqrt{\frac{p*q*}{n_1}+\frac{p*q*}{n_2}}-(p_1-p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1}+\frac{p_2 \cdot q_2}{n_2}}} \right] + P\left[Z \geq \frac{Z_{\alpha/2}\sqrt{\frac{p*q*}{n_1}+\frac{p*q*}{n_2}}-(p_1-p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1}+\frac{p_2 \cdot q_2}{n_2}}} \right]$$

- nel caso di un test **unilaterale**, dipende dalla **direzionalità**:

a) per l'ipotesi nulla $\mathbf{H_0: \pi_1 \leq \pi_2}$ contro l'ipotesi alternativa $\mathbf{H_1: \pi_1 > \pi_2}$

la potenza **1-β** è ottenuta con

$$1-\beta = P\left[Z \geq \frac{Z_\alpha \cdot \sqrt{\frac{p^*q^*}{n_1} + \frac{p^*q^*}{n_2}} - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}} \right]$$

b) per l'ipotesi nulla $\mathbf{H_0: \pi_1 \geq \pi_2}$ contro l'ipotesi alternativa H_1: $\mathbf{H_0: \pi_1 < \pi_2}$

è ottenuta con

$$1-\beta = P\left[Z \leq \frac{-Z_\alpha \cdot \sqrt{\frac{p^*q^*}{n_1} + \frac{p^*q^*}{n_2}} - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}} \right]$$

Nel suo testo del 1999 più volte citato (***Biostatistical Analysis***, fourth ed. edito da **Prentice Hall, Upper Saddler River, New Jersey**, pp. 663 + App. pp. 212), Jerrold H. **Zar** scrive: ***questo calcolo della potenza è basato sull'approssimazione al test esatto di Fisher e tende a produrre un risultato conservativo: la potenza reale è maggiore di quella calcolata con questa formula***.

ESEMPIO 1 (APPLICAZIONE A UN TEST BILATERALE). In un quartiere con forte intensità di traffico veicolare, su 50 rilevazioni il 50% hanno superato i livelli di attenzione per $\mathbf{NO_2}$.
In un periodo successivo, sono state effettuate 45 osservazioni. Quale è la potenza del test, perché una differenza di 0,15 risulti significativa alla probabilità $\alpha = 0.05$?

Risposta. Per verificare l'ipotesi nulla

$\mathbf{H_0: \pi_1 = \pi_2}$ contro l'ipotesi alternativa bilaterale $\mathbf{H_1: \pi_1 \neq \pi_2}$

è necessario dapprima calcolare p* con $p_1 = 0,5$ e p_2 supposto uguale a 0,35 (oppure uguale a 0,65).
Con $n_1 = 50$ e $n_2 = 45$

$$P^* = \frac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2} = \frac{50 \cdot 0,5 + 45 \cdot 0,35}{50 + 45} = \frac{25,00 + 15,75}{95} = 0,439$$

si ricava che il valore medio delle due frequenze è p* = 0,439.

Poiché $Z_{0.05/2} = 1{,}96$

si ottiene

$$1-\beta = P\left[Z \le \frac{-1{,}96\sqrt{\frac{0{,}439\cdot 0{,}561}{50}+\frac{0{,}439\cdot 0{,}561}{45}} - (0{,}50-0{,}35)}{\sqrt{\frac{0{,}5\cdot 0{,}5}{50}+\frac{0{,}35\cdot 0{,}65}{45}}} \right] +$$

$$+ P\left[Z \ge \frac{1{,}96\sqrt{\frac{0{,}439\cdot 0{,}561}{50}+\frac{0{,}439\cdot 0{,}561}{45}} - (0{,}50-0{,}35)}{\sqrt{\frac{0{,}5\cdot 0{,}5}{50}+\frac{0{,}35\cdot 0{,}65}{45}}} \right]$$

$$1-\beta = P\left[Z \le \frac{-1{,}96\sqrt{0{,}0049+0{,}0055} - (0{,}15)}{\sqrt{0{,}0050+0{,}0051}} \right] + P\left[Z \ge \frac{1{,}96\sqrt{0{,}0049+0{,}0055} - (0{,}15)}{\sqrt{0{,}0050+0{,}0051}} \right]$$

$$1-\beta = P\left[Z \le \frac{-1{,}96\cdot 0{,}1020 - (0{,}15)}{0{,}1005} \right] + P\left[Z \ge \frac{1{,}96\cdot 0{,}1020 - (0{,}15)}{0{,}1005} \right]$$

un valore della potenza (**1-β**)

$$1-\beta = P(Z \le -3{,}48) + P(Z \ge 0{,}497)$$

uguale a Z = -3,48 nella coda sinistra e Z = 0,497 nella coda destra.

Al valore $Z \le -3{,}48$ **nella coda sinistra** corrisponde una probabilità β uguale a P < 0.0001.

Al valore $Z \ge 0{,}497$ (arrotondato a +0,50) nella coda destra della distribuzione corrisponde una probabilità β uguale a P = 0.309.

Di conseguenza, la potenza 1-β del test è uguale a 0.3091 (0.309 + 0.0001).

ESEMPIO 2 (APPLICAZIONE A UN TEST UNILATERALE). In un quartiere con forte intensità di traffico automobilistico, su 50 rilevazioni esattamente il 50% hanno superato i livelli di attenzione per **NO_2**.

In un periodo successivo, sono state effettuate 45 osservazioni. Quale è la potenza del test, per dimostrare che una riduzione della proporzione di 0,15 risulta significativa alla probabilità α = 0.05?

Risposta. E' un test unilaterale, con ipotesi nulla

H_0: $\pi_1 \le \pi_2$ contro l'ipotesi alternativa **H_1: $\pi_1 > \pi_2$**

Con

- $p_1 = 0,5$ e $n_1 = 50$
- $p_2 = 0,35$ e $n_2 = 45$
- $p^* = 0,439$ e $Z_{0.05} = 1,645$

mediante

$$1-\beta = P\left[Z \geq \frac{Z_\alpha \cdot \sqrt{\frac{p^*q^*}{n_1} + \frac{p^*q^*}{n_2}} - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}} \right]$$

si ottiene

$$1-\beta = P\left[Z \geq \frac{1,645 \cdot \sqrt{\frac{0,439 \cdot 0,561}{50} + \frac{0,439 \cdot 0,561}{45}} - (0,50 - 0,35)}{\sqrt{\frac{0,5 \cdot 0,5}{50} + \frac{0,35 \cdot 0,65}{45}}} \right]$$

$$1-\beta = P\left[Z \geq \frac{1,645 \cdot \sqrt{0,0049 + 0,0055} - (0,15)}{\sqrt{0,0050 + 0,0051}} \right] = P\left[Z \geq \frac{0,168 - 0,15}{0,1005} \right] = Z \geq 0,18$$

un valore di $Z \geq 0,18$.

Al valore $Z \geq 0,18$ nella coda destra della distribuzione corrisponde una probabilità P uguale a 0.429.

E' la potenza 1-β del test.

Anche per stimare **la potenza a priori**, cioè per calcolare **n**, **il numero minimo di dati da raccogliere per ognuno dei due campioni** in un esperimento bilanciato (in quanto fornisce la potenza maggiore), le proposte sono numerose.

E' possibile, come dimostrato per una sola proporzione, utilizzare le formule appena presentate per la potenza, procedendo per tentativi, in modo iterativo.

Con una presentazione più dettagliata di questo metodo,

- dopo aver scelto i valore di $\mathbf{p_1}$, $\mathbf{p_2}$ e $\mathbf{Z_\alpha}$ per un test **unilaterale** oppure **bilaterale**
- si fissa un valore di **n** e si calcola **la potenza (1-β),**
- utilizzando una delle tre ultime formule presentate, in rapporto all'ipotesi da verificare.

Se la potenza risulta inferiore a quella prefissata, si aumenta **n**; se la potenza risulta maggiore, si può abbassare **n.**

Con la stessa logica, invece di modificare la potenza, si può cercare il numero di dati necessari in funzione di α oppure dei valori di $\mathbf{p_1}$ e $\mathbf{p_2}$ e della loro differenza, tenendo in considerazione anche la direzionalità dell'ipotesi. E' un metodo che può servire per stimare ognuno degli altri parametri che entrano nella funzione, dopo aver predeterminato gli altri.

Un problema che si pone con frequenza è la necessità di raccogliere un campione di dati per il confronto con un campione già raccolto in precedenza: avendo già a disposizione il campione con $\mathbf{n_1}$ e $\mathbf{p_1}$ fissati, si vuole determinare il numero di dati da raccogliere nel secondo campione($\mathbf{n_2}$), affinché una determinata proporzione $\mathbf{p_2}$ risulti significativa alla probabilità α e con il rischio β, in un test unilaterale o bilaterale. E' il caso di un controllo a distanza di tempo dopo una prima analisi, per verificare un abbassamento significativo dei livelli d'inquinamento, conseguente a un'azione di risanamento; oppure un aumento, dopo l'attivazione di una potenziale sorgente d'inquinamento, quale un forno inceneritore per l'aria, una discarica per le falde acquifere, un deposito di idrocarburi per il suolo.

Ritornando alla stima del numero minimo di dati (**n**) per il confronto di due proporzioni ($\mathbf{p_1}$ e $\mathbf{p_2}$) in campioni indipendenti, in modo che la loro differenza ($\mathbf{p_1}$ - $\mathbf{p_2}$) risulti significativa alla probabilità α con un certo rischio β, sono state proposte varie **formule abbreviate**. Tra quelle ricorrenti con frequenza maggiore nei testi di statistica applicata, è possibile ricordare le proposte di J. L. **Fleiss** del 1981 (vedi testo di edizione più recente, rispetto alle citazioni precedenti: ***Statistical Methods for Rates and Proportions***. John Wiley and Sons , New York. 677 pp.)

Con 2 campioni bilanciati, quindi

$$p^* = \frac{p_1 + p_2}{2}$$

dopo aver determinato

- $\mathbf{p_1}$ e $\mathbf{p_2}$ e calcolato la loro media $\mathbf{p^*}$
- la probabilità α (errore di I Tipo) e la probabilità β (errore di II Tipo) oppure la potenza $\mathbf{1-\beta}$,

- in **un test con bilaterale**

la formula è

$$\mathbf{n} = \frac{\left(Z_{\alpha/2} \cdot \sqrt{2p^* \cdot q^*} + Z_{\beta} \cdot \sqrt{p_1 \cdot q_1 + p_2 \cdot q_2}\right)^2}{(p_2 - p_1)^2}$$

- in **<u>un test unilaterale</u>**

è sufficiente nella formula sostituire $\mathbf{Z_{\alpha/2}}$ con $\mathbf{Z_{\alpha}}$.

$$\mathbf{n} = \frac{\left(Z_{\alpha} \cdot \sqrt{2p^{*} \cdot q^{*}} + Z_{\beta} \cdot \sqrt{p_1 \cdot q_1 + p_2 \cdot q_2}\right)^2}{\left(p_2 - p_1\right)^2}$$

ESEMPIO 3 (TEST BILATERALE, CON FORMULA ABBREVIATA). Due serie di analisi campionarie sulla qualità dell'aria hanno dimostrato che nella zona A il 50% delle rilevazioni supera i limiti di attenzione per SO_2, mentre nella zona B esse sono il 35%.
Quanti dati occorre raccogliere affinché **questa differenza risulti significativa** alla probabilità α = 0.05 e il test abbia una potenza 1-β = 0.90?

Risposta. Con

- $Z_{0.025} = 1{,}96$ per la probabilità α bilaterale,
- $Z_{0.10} = 1{,}28$ per la probabilità β unilaterale,
- $p_1 = 0{,}5$ $p_2 = 0{,}35$ $p^* = 0{,}425$

si ottiene

$$\mathbf{n} = \frac{\left(1{,}96 \cdot \sqrt{2 \cdot 0{,}425 \cdot 0{,}575} + 1{,}28 \cdot \sqrt{0{,}5 \cdot 0{,}5 + 0{,}35 \cdot 0{,}65}\right)^2}{\left(0{,}5 - 0{,}35\right)^2}$$

$$\mathbf{n} = \frac{\left(1{,}96 \cdot 0{,}699 + 1{,}28 \cdot 0{,}691\right)^2}{\left(0{,}15\right)^2} = \frac{\left(1{,}37 + 0{,}88\right)^2}{\left(0{,}15\right)^2} = \frac{5{,}06}{0{,}0225} = 224{,}8$$

un numero minimo pari a 225 misure per gruppo.

ESEMPIO 4 (TEST UNILATERALE, CON FORMULA ABBREVIATA). Due serie di analisi campionarie sulla qualità dell'aria hanno dimostrato che nella zona A il 50% delle rilevazioni supera i limiti di attenzione per SO_2, mentre nella zona B esse sono il 35%.
Quanti dati occorre raccogliere perché **la proporzione della zona B risulti inferiore a quella della zona A** alla probabilità α = 0.05 e il test abbia una potenza 1-β = 0.90?

Risposta. Con

- $Z_{0.05} = 1{,}645$ per la probabilità α unilaterale,

- $Z_{0.10} = 1{,}28$ per la probabilità β unilaterale,
- $p_1 = 0{,}5$ $p_2 = 0{,}35$ $p^* = 0{,}425$

si ottiene

$$\mathbf{n} = \frac{\left(1{,}645 \cdot \sqrt{2 \cdot 0{,}425 \cdot 0{,}575} + 1{,}28 \cdot \sqrt{0{,}5 \cdot 0{,}5 + 0{,}35 \cdot 0{,}65}\right)^2}{(0{,}5 - 0{,}35)^2}$$

$$\mathbf{n} = \frac{(1{,}645 \cdot 0{,}699 + 1{,}28 \cdot 0{,}691)^2}{(0{,}15)^2} = \frac{(1{,}15 + 0{,}88)^2}{(0{,}15)^2} = \frac{4{,}12}{0{,}0225} = 183{,}1$$

un numero minimo pari a 184 misure per gruppo.

Stimato il numero minimo (**n**) di osservazioni per ognuno dei due gruppi, affinché il test abbia la potenza desiderata, sorgono due problemi:

A) quando il campione richiesto non è di grandi dimensioni, ma rimane inferiore alle duecento osservazioni, la stima effettuata (**n**) dovrebbe essere corretta, in quanto il test successivo dovrebbe utilizzare formule con la correzione per la continuità, che alzano il valore della probabilità α;

B) il numero **n** calcolato è la quantità minima di osservazioni necessarie in ognuno dei due gruppi a confronto; ma non sempre è possibile o conveniente effettuare un esperimento bilanciato.

A) Quando, con queste formule, si stima che sono sufficienti alcune decine di osservazioni o un numero di poco superiore al centinaio, sorge un problema. Con un **campione di piccole dimensioni**, come possono essere definiti questi, nel test di verifica dell'ipotesi nulla con il test χ^2 o con il corrispondente test Z, si apporta la correzione per la continuità o correzione di Yates.

Ad esempio, una formula corretta utilizzando la distribuzione Z, già presentata in questo capitolo e nel precedente, è

$$Z = \frac{|\, p_1 - p_2 \,| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{p^*(1 - p^*) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Essa ha uno scopo cautelativo: abbassa la significatività del test, appunto perché con pochi dati le conclusioni sono meno attendibili e nella logica statistica non si vuole rifiutare l'ipotesi nulla quando la risposta è incerta.

Nella stima della dimensione **n** del campione da raccogliere, per rendere il test significativo alla stessa probabilità, è quindi necessario aumentare la quantità **n,** calcolata con la formula generale già descritta.

Nel 1959, **M. Kramer e S. W. Greenhouse** (nell'articolo ***Determination of sample size in treatment-control comparison for chronic disease studies in which drop-out or non-adherence in a problem***, pubblicato dalla rivista **J. Chronic. Dis**., n. 20, pp. 233-239) hanno proposto

- n' = stima corretta di **n**,

che tiene appunto presente la correzione per la continuità.

Dopo avere stimato **n**, **per considerare la correzione per la continuità di** Yates, si perviene ad una sua valutazione corretta **n'** mediante la relazione

$$\text{n'} = \frac{n}{4} \cdot \left[1 + \sqrt{1 + \frac{4}{n \cdot |p_2 - p_1|}} \right]^2$$

ESEMPIO 5. (TEST BILATERALE, CON CORREZIONE PER LA CONTINUITA'). Calcolare le dimensioni minime (**n**) del campione, affinché la differenza tra le proporzioni $\mathbf{p_2 = 0,28}$ e $\mathbf{p_1 = 0,12}$ risulti significativa alla probabilità $\boldsymbol{\alpha = 0.05}$ con un rischio $\boldsymbol{\beta = 0.10.}$

Successivamente, apportare la correzione se il campione non è grande (inferiore a 200).

Risposta

1 - Dapprima si calcola **n**, ricavando dal problema che è un test bilaterale,

in cui $\mathbf{Z_{\alpha/2} = 1,96}$ e $\mathbf{Z_{\beta} = 1,28}$ con $\bar{p}$ **= (0,28 + 0,12)/2 = 0,2**

Il numero minimo **n** di dati per ognuno dei due gruppi è

$$\mathbf{n} = \frac{\left(1{,}96 \cdot \sqrt{2 \cdot 0{,}2 \cdot 0{,}8} + 1{,}28 \cdot \sqrt{(0{,}28 \cdot 0{,}72) + (0{,}12 \cdot 0{,}88)}\right)^2}{(0{,}28 - 0{,}12)^2}$$

$$\mathbf{n} = \frac{(1{,}96 \cdot 0{,}5657 + 1{,}28 \cdot 0{,}5543)^2}{0{,}16^2} = \frac{(1{,}1088 + 0{,}7075)^2}{0{,}16^2} = \frac{3{,}299}{0{,}0256} = 128{,}9$$

uguale a 128,9. In ognuno dei due gruppi a confronto sono necessari 129 osservazioni.

2 - Trattandosi di campioni inferiori a 200 osservazioni, si deve apportare una correzione al numero **n** calcolato, stimando **n'**; in questo caso, esso risulta

$$n' = \frac{129}{4} \cdot \left[1 + \sqrt{1 + \frac{4}{129 \cdot |0{,}28 - 0{,}12|}} \right]^2$$

$$n' = 32{,}25 \cdot \left(1 + \sqrt{1 + 0{,}194}\right)^2 = 32{,}25 \cdot (1 + 1{,}093)^2 = 141{,}3$$

uguale a 142.

Vari programmi informatici, insieme con la stima di **n** (il valore della potenza a priori), nell'output forniscono anche la significatività del test a posteriori, qualora si realizzassero esattamente le condizioni supposte per la stima di **n**.

Se prima della raccolta dei dati il valore di α poteva essere uguale a 0.05, una volta raccolti i dati (quindi a posteriori) la significatività del test è di gran lunga superiore e perciò avrà un valore di α nettamente minore. Infatti, **il calcolo a priori di** n **inglobava il rischio che la differenza ipotizzata tra le due proporzioni fosse, per variazione casuale, minore dell'atteso.**

Nella stima a priori di **n** è compresa la probabilità β di commettere un errore di II Tipo; a posteriori questo rischio non esiste più.

B) Stimato **n**, è conveniente formare due campioni bilanciati: garantisce la potenza massima del test, con un numero totale di dati uguale a **2n**.

Ma non sempre è possibile o conveniente. Un gruppo di dati potrebbe essere già stato raccolto; le osservazioni dei due gruppi potrebbero non avere gli stessi costi morali, quali esperimenti con il placebo o il farmaco su due gruppi di ammalati; più semplicemente, potrebbero avere costi economici diversi, quale la somministrazione di un nuovo farmaco, di prezzo più elevato rispetto al precedente già sul mercato e da tempo in produzione (questo argomento è trattato in modo più ampio nel capitolo 5, illustrando il **test t di Student** per due campioni indipendenti).

Si pone quindi il problema di formare due campioni non bilanciati, ricercando la convenienza massima, senza che il test perda in potenza.

Sulla base dei principi succintamente enunciati,

affinché

$$n_2 = r \cdot n_1$$

dove n_1 è il campione minore, si deve

1 - dapprima calcolare **r**

$$r = \frac{n_2}{n_1}$$

e **p***

$$p^* = \frac{p_1 + r \cdot p_2}{r + 1}$$

2 - successivamente stimare **n**

$$n = \frac{\left[Z_\alpha \cdot \sqrt{(r+1) \cdot p^* \cdot q^*} + Z_\beta \cdot \sqrt{r \cdot p_1 \cdot q_1 + p_2 \cdot q_2} \right]^2}{r \cdot (|p_2 - p_1|)^2}$$

3 - infine calcolare n_1

$$n_1 = \frac{n}{4} \cdot \left[1 + \sqrt{1 + \frac{2(r+1)}{r \cdot n \cdot |p_2 - p_1|}} \right]^2$$

4 - da quale derivare anche n_2

$$n_2 = r \cdot n_1$$

ESEMPIO 6 (DIMENSIONI DEI CAMPIONI, SENZA BILANCIAMENTO – Prima parte). Un ricercatore deve verificare la differenza della qualità delle falde idriche di due aree. Analisi preliminari hanno dimostrato che nella zona 1 il 45% dei prelievi supera i limiti di attenzione per almeno un parametro; mentre nella zona 2 tali limiti sono superati solo dal 25% dei prelievi.
Quanti dati deve raccogliere in ognuna delle due zone, per dimostrare che tale differenza è significativa alla probabilità α **= 0.05** con un rischio β **= 0.10**?

Risposta. Con

- $Z_{0.025} = 1{,}96$ per la probabilità α bilaterale,
- $Z_{0.10} = 1{,}28$ per la probabilità β unilaterale,
- $p_1 = 0{,}45$ $p_2 = 0{,}25$ $p^* = 0{,}35$

si ottiene

$$\mathbf{n} = \frac{\left(1{,}96 \cdot \sqrt{2 \cdot 0{,}35 \cdot 0{,}65} + 1{,}28 \cdot \sqrt{0{,}45 \cdot 0{,}55 + 0{,}25 \cdot 0{,}75}\right)^2}{(0{,}45 - 0{,}25)^2}$$

$$\mathbf{n} = \frac{(1{,}96 \cdot 0{,}675 + 1{,}28 \cdot 0{,}660)^2}{(0{,}2)^2} = \frac{(1{,}323 + 0{,}845)^2}{(0{,}2)^2} = \frac{4{,}7}{0{,}04} = 117{,}5$$

un **numero minimo** pari a 118 rilevazioni per gruppo.

Trattandosi di due campioni non grandi, per cui nel test si dovrebbe utilizzare la correzione per la continuità, è conveniente raccogliere un numero di dati più alto.

La stima corretta **n'**

diventa

$$n' = \frac{118}{4} \cdot \left[1 + \sqrt{1 + \frac{4}{118 \cdot |0{,}45 - 0{,}25|}}\right]^2$$

$$n' = 29{,}5 \cdot \left(1 + \sqrt{1 + 0{,}169}\right)^2 = 29{,}5 \cdot (1 + 1{,}081)^2 = 127{,}75$$

uguale 127,75.

Servono almeno 128 osservazioni per gruppo.

ESEMPIO 6 (DIMENSIONI DEI CAMPIONI, SENZA BILANCIAMENTO - Seconda parte). Si supponga che la zona 2 sia vicina al laboratorio di analisi, mentre la zona 1 sia distante e quindi con costi maggiori per i prelievi. Una stima più precisa quantifica in circa 10mila lire il costo di ogni prelievo nella zona 2 e di 30mila nella zona 1. Con gli stessi dati della prima parte, quanti prelievi occorre effettuare nella zona 1 e quanti nella zona 2?

Risposta. Dopo aver supposto r = 3

- si stima dapprima **p***

$$p^* = \frac{p_1 + r \cdot p_2}{r + 1} = \frac{0{,}45 + 3 \cdot 0{,}25}{3 + 1} = \frac{1{,}2}{4} = 0{,}3$$

che risulta uguale a 0,3

e successivamente **n**

$$n = \frac{\left[1{,}96 \cdot \sqrt{(3+1) \cdot 0{,}3 \cdot 0{,}7} + 1{,}28 \cdot \sqrt{3 \cdot 0{,}45 \cdot 0{,}55 + 0{,}25 \cdot 0{,}75}\right]^2}{3 \cdot \left(|0{,}45 - 0{,}25|\right)^2}$$

$$n = \frac{\left[1{,}96 \cdot \sqrt{0{,}84} + 1{,}28 \cdot \sqrt{0{,}93}\right]^2}{3 \cdot (0{,}2)^2} = \frac{(1{,}796 + 1{,}234)^2}{0{,}12} = \frac{9{,}18}{0{,}12} = 76{,}5$$

che risulta 76,5.

Infine da esso si ricava n_1

$$n_1 = \frac{76{,}5}{4} \cdot \left[1 + \sqrt{1 + \frac{2(3+1)}{3 \cdot 76{,}5 \cdot |0{,}45 - 0{,}25|}}\right]^2$$

$$n_1 = 19{,}125 \cdot \left(1 + \sqrt{1 + \frac{8}{45{,}9}}\right)^2 = 19{,}125 \cdot (1 + 1{,}084)^2 = 83{,}06$$

che risulta uguale a 83,06.

Per il campione 1 servono almeno 84 prelievi e quindi per il campione 2 (3 x 83,06 = 249,18) almeno 250 prelievi.

E' importante osservare che, **con due campioni sbilanciati**, il **numero totale di osservazioni aumenta**:

- è diventato uguale a **334** (84 + 250),
- mentre con due campioni bilanciati era uguale a **256** (128 + 128).

Ma è **diminuito il costo totale delle analisi**, anche se in questo caso in modo molto limitato.

Con due campioni bilanciati era

(128 x 10.000) + (128 x 30.000) = 1.280.000 + 3.840.000 = 5.120.000

uguale a lire 5.120.000

mentre con la nuova programmazione il costo

(84 x 30.000) + (250 x 10.000) = 2.520.000 + 2.500.000 = 5.020.000

risulta uguale a 5.020.000.

Ma esistono soluzioni migliori?

La soluzione è cercata nell'esempio successivo.

ESEMPIO 6 (DIMENSIONI DEI CAMPIONI, SENZA BILANCIAMENTO - Terza parte). Con gli stessi dati della prima e della seconda parte dell'esercizio 6, verificare i costi con r = 2,2.

Risposta. Con r = 2,2

- si stima dapprima **p***

$$p^* = \frac{0{,}45 + 2{,}2 \cdot 0{,}25}{2{,}2 + 1} = \frac{0{,}45 + 0{,}55}{3{,}2} = 0{,}3125$$

che risulta uguale a 0,3125

- e successivamente **n**

$$n = \frac{\left[1{,}96 \cdot \sqrt{(2{,}2+1) \cdot 0{,}3125 \cdot 0{,}6875} + 1.28 \cdot \sqrt{2{,}2 \cdot 0{,}45 \cdot 0{,}55 + 0{,}25 \cdot 0{,}75}\right]^2}{2{,}2 \cdot \left(|0{,}45 - 0{,}25|\right)^2}$$

$$n = \frac{\left[1{,}96 \cdot \sqrt{0{,}6875} + 1{,}28 \cdot \sqrt{0{,}732}\right]^2}{2{,}2 \cdot (0{,}2)^2} = \frac{(1{,}625 + 1{,}095)^2}{0{,}088} = \frac{7{,}3984}{0{,}088} = 84{,}07$$

che risulta 84,07.

- Infine si stima n_1

$$n_1 = \frac{84{,}07}{4} \cdot \left[1 + \sqrt{1 + \frac{2(2{,}2+1)}{2{,}2 \cdot 84{,}07 \cdot |0{,}45 - 0{,}25|}}\right]^2$$

$$n_1 = 21{,}0175 \cdot \left(1 + \sqrt{1 + \frac{6{,}4}{36{,}99}}\right)^2 = 21{,}0175 \cdot (1 + 1{,}083)^2 = 91{,}19$$

che risulta uguale a 91,19.

Per il campione 1 servono almeno 92 osservazioni.

Quindi per il campione 2 ne sono necessarie almeno 201 (ricavato da 2,2 x 91,19 = 200,6).

E' importante osservare che,

- mentre con due campioni bilanciati il numero totale di osservazioni era uguale a 256 (128 + 128) e
- con due campioni sbilanciati con un rapporto di 3 a 1 tale numero era aumentato a 334 (84 + 250),
- ora con un rapporto di 2,2 a 1 il numero minimo richiesto è 293 (92 + 201).

Lo sbilanciamento tra i due campioni è minore e quindi il numero complessivo di dati richiesti per un test che abbia la stessa potenza è inferiore.

In questo caso, il costo totale delle analisi è

(92 x 30.000) + (201 x 10.000) = 2.760.000 + 2.010.000 = 4.770.000

risulta uguale a 4.770.000.

E' più conveniente dei due precedenti.

Un altro aspetto interessante nella **programmazione di un esperimento** è la **curva dei costi**.

In rapporto ai costi per la rilevazione dei dati in un esperimento con due campioni, la funzione quasi sempre è a U:

- diminuisce allontanandosi dal rapporto di 1 a 1 fino ad un livello minimo,
- per aumentare successivamente con lo "sbilanciamento" sempre più accentuato dei due campioni, che inevitabilmente porta a richiedere un numero totale di osservazioni sempre maggiore.

Per **ridurre al minimo i costi complessivi**, è necessario trovare

- il punto ottimale di rapporto numerico tra i due campioni,

eventualmente per tentativi e in modo iterativo, anche se esistono funzioni che abbreviano i calcoli.

<u>Un altro modo</u> per stimare le dimensioni del campione 2, fissate quelle del campione 1, è il ricorso alla formula già presentata:

$$1-\beta = P\left[Z \geq \frac{Z_\alpha \cdot \sqrt{\frac{p^*q^*}{n_1} + \frac{p^*q^*}{n_2}} - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}} \right]$$

A causa delle relazioni esistenti tra i parametri implicati nella formula,

- è possibile stimare un parametro, dopo aver fissato tutti gli altri.

E' un metodo che si rivela utile in molti casi, per risolvere altri problemi di bilanciamento. Un caso classico è quando si deve **confrontare la situazione del passato con quella attuale**.

Ma **i dati del passato sono già stati raccolti**. Si vuole quindi sapere **quanti nuovi dati è necessari raccogliere**.

Ad esempio, questa ultima formula permette di calcolare quanti nuovi dati è necessario raccogliere se, con gli stessi dati dell'esempio 6, il confronto dovesse avvenire tra la situazione attuale (tempo 2) e una situazione precedente (tempo 1), per la quale furono raccolte 80 osservazioni.

5.15. IL RAPPORTO TRA DUE PROPORZIONI (R): INTERVALLO DI CONFIDENZA E SIGNIFICATIVITA'; FORMULA TEST-BASED DI MIETTINEN PER R.

Vari concetti illustrati in questo capitolo sono presentati anche in altri. La differenza tra due proporzioni, trattata nei paragrafi precedenti, è già stata esposta nel capitolo sul chi-quadrato; il rapporto tra due proporzioni e tra due odds, discusso in questo paragrafo, è riproposto nel capitolo sulle misure di associazione. Non si tratta di una banale duplicazione.

Anche quando i concetti sono identici e i metodi sono sovrapponibili, l'approccio è differente. Il confronto tra essi serve per conseguire una visione più ampia del problema, che è didatticamente utile per evidenziare le differenze tra scuole e apprendere come giustificare, in modo più completo, la scelta di un test o di una variante nelle formule, tra i vari che sono stati proposti in 50 anni di sviluppo della metodologia. Anche i pacchetti informatici, presenti su un mercato sempre più ampio ed esigente, quando propongono gli stessi test spesso si rifanno a metodi o formule differenti. In conclusione, conoscere approcci diversi è utile per raggiungere quella cultura statistica che permette di giustificare le differenze tra metodi. Serve, nella presentazione di un rapporto scientifico o una di pubblicazione, anche per controbattere quelle chiusure ideologiche, non insolite nei referee di questa disciplina, che accettano come valida solamente una impostazione statistica. E spesso senza motivazioni, senza giudizi sulla potenza o sulla robustezza del test, sul tipo di scala oppure sulle caratteristiche della distribuzione dei dati, sul rischio α oppure sulle dimensioni del campione.

In questo settore della statistica, le differenze fondamentali tra i test derivano dall'essere fondati su probabilità esatte o asintotiche, dal fatto che le soluzioni siano più o meno approssimate, dal richiedere metodi lunghi e difficili oppure fondati su soluzioni rapide.

Un primo aspetto della ricerca è quasi sempre l'uso di un linguaggio scientifico. Nella ricerca epidemiologica e ambientale, sovente si usano **termini equivoci**.

Ad esempio, se la proporzione di persone che soffrono di allergia in un determinato periodo è del 30% ($p_1 = 0,30$) e si afferma che nei 10 anni successivi hanno avuto un aumento del 15%, si intende dire che:

1 - sono diventati il 45% ($p_2 = p_1 + d = 0,30 + 0,15 = 0,45$)?

Oppure che

2 - sono diventati il 34,5% ($p_2 = p_1 \times R = 0,30 \times 1,15 = 0,345$)?

Nel primo caso, per confrontare il valore finale con quello iniziale, è stata utilizzata la **differenza tra due proporzioni**:

$$d = p_2 - p_1 = 0,45 - 0,30 = 0,15$$

Nel secondo, il **rapporto tra due proporzioni:**

$$R = \frac{p_2}{p_1} = \frac{0{,}345}{0{,}300} = 1{,}15$$

Da questa osservazione, derivano due conseguenze.

- La prima è banale: per evitare fraintendimenti, è utile riportare tre informazioni, in particolare le prime due: (a) il valore iniziale, (b) il valore finale, (c) il valore dell'accrescimento, che può essere la differenza oppure il rapporto; ma, insieme con i primi due, è sempre comprensibile senza equivoci.
- La seconda è un problema tecnico: come si analizza **un rapporto tra due proporzioni** e come si **confrontano due rapporti**, dopo che nei paragrafi precedenti sono state presentate le tecniche per l'**analisi di una differenza tra proporzioni.**

Collegato al concetto di **rapporto tra due proporzioni** nei testi di statistica applicata spesso è presente anche il concetto del **rapporto tra due odds**.

Sono differenti, ma quando un fenomeno è raro, quindi le proporzioni sono basse, i risultati dei due metodi sono simili. Ne consegue che in letteratura è facile vedere l'utilizzo di uno al posto dell'altro, inducendo le persone con poca esperienza tecnica a credere che essi siano uguali, una semplice variante matematica come la formula abbreviata e la formula euristica che sono stati presentati per alcuni test.

Il **rapporto tra due odds** (***odds ratio***), che a prima vista appare meno semplice, in alcune analisi statistiche offre il vantaggio tecnico non trascurabile di permettere l'uso della regressione logistica. E' un metodo importante nella interpretazione statistica degli studi caso-controllo, frequenti in medicina, farmacologia ed ecotossicologia.

Utilizzando la simbologia riportata schematicamente nella tabella successiva

	Campione 1	Campione 2	Totale
Conteggio positivi	r_1	r_2	r
Conteggio negativi	$n_1 - r_1$	$n_2 - r_2$	$n - r$
Totale	n_1	n_2	n
Proporzione di successi	$p_1 = \frac{r_1}{n_1}$	$p_2 = \frac{r_2}{n_2}$	$p = \frac{r}{n}$
Odds di successo	$o_1 = \frac{r_1}{n_1 - r_1}$	$o_2 = \frac{r_2}{n_2 - r_2}$	$o = \frac{r}{n - r}$

è evidente

- sia la **differenza** tra una **proporzione** $p = \frac{r}{n}$ e un **odds** $o = \frac{r}{n-r}$,

- sia il significato delle due proporzioni p_1 e p_2

e quindi quello del rapporto tra esse

$$R = \frac{p_2}{p_1}$$

Quando due proporzioni sono uguali, il rapporto è $R = 1$

Ma se $p_2 < p_1$, il rapporto R tende a 0;

mentre se $p_2 > p_1$, il rapporto R tende all'infinito positivo.

Ne deriva che la **distribuzione di** R ha una **forte asimmetria destra**.

Approssimativamente, è una **distribuzione log-Normale,** come dimostrano i dati successivi.

R	1/32	1/16	1/8	1/4	1/2	1	2	4	8	16	32
Valore	0,031	0,062	0,125	0,250	0,500	1	2	4	8	16	32
$\ln R$	-3,47	-2,77	-2,08	-1,39	-0,69	0	+0,69	+1,39	+2,08	+2,77	+3,47

Con due proporzioni misurate in due campioni indipendenti,

1 - R può assumere valori come quelli riportati nella prima riga: **i rapporti** tra p_2 e p_1 **variano in modo bilanciato**;

2 - ma se si calcolano i rapporti, come nella seconda riga, e con essi si costruisce una distribuzione in classi di frequenza con passo 1, è semplice dedurre che tutti i rapporti minori di 1 saranno nella prima classe e gli altri formeranno 32 classi, con molte di esse vuote; risulta visivamente evidente che i valori R determinano una **distribuzione con forte asimmetria destra**.

3 – Infine, applicando a questa ultima distribuzione di dati la **trasformazione logaritmica**, in questo caso la **log normale** ($\ln R$) come nella terza riga, si ottiene una **distribuzione simmetrica**, **approssimativamente normale**.

Con R, si indica un **rapporto campionario** tra due proporzioni; il **rapporto reale**, quello della popolazione, è indicato con il simbolo greco **ρ (rho** minuscolo**,** anche se il precedente è maiuscolo**)**.

Dopo la **trasformazione di R** in $\ln$ **R**, è possibile utilizzare la **distribuzione normale ridotta Z**,
- sia per costruire l'intervallo di confidenza di ρ,
- sia per confrontare due R.

Nel primo caso, per stimare l'intervallo di confidenza di ρ a partire da un valore campionario $\ln R$, serve la varianza di $\ln R$.
Dato che

$$\ln R = \ln \frac{p_2}{p_1} = \ln p_2 - \ln p_1$$

e poiché le due proporzioni p_2 e p_1 sono indipendenti

$$\text{var}(\ln p_2 - \ln p_1) = \text{var}(\ln p_2) + \text{var}(\ln p_1)$$

si ricava che
- **la varianza della differenza tra due proporzioni è uguale alla somma delle loro varianze**.

Questo concetto è facilmente comprensibile con una dimostrazione elementare.
Se è vera l'**ipotesi nulla H_0**, le due **proporzioni reali** sono uguali ($\pi_2 = \pi_1$).
Quindi le **proporzioni campionarie** p_2 e p_1 possono avere variazioni casuali di entità simile, che
- a volte saranno nella stessa direzione $p_2 + \varepsilon$ e $p_1 + \varepsilon$ oppure $p_2 - \varepsilon$ e $p_1 - \varepsilon$, con il risultato che i loro effetti nella differenza si annullano $(p_2 + \varepsilon) - (p_1 + \varepsilon) = 0$ e $(p_2 - \varepsilon) - (p_1 - \varepsilon) = 0$
- altre volte saranno nella direzione opposta come $p_2 + \varepsilon$ e $p_1 - \varepsilon$, con il risultato che i loro effetti si sommano $(p_2 + \varepsilon) - (p_1 - \varepsilon) = +2\varepsilon$ in modo positivo o negativo $(p_2 - \varepsilon) - (p_1 + \varepsilon) = -2\varepsilon$.

Nello stesso modo della differenza tra due medie, questi ultimi due passaggi dimostrano che
- **la varianza di una differenza è uguale alla somma delle due varianze**.

In conclusione,
1 - per la **proporzione** $p = r/n$

la **varianza stimata** di $\ln p$ è uguale a $\frac{1-p}{r}$ scritto anche $\frac{q}{r}$ oppure $\frac{q}{np}$

2 – per il $\ln R$ la **varianza stimata** diventa

$$\text{var}(\ln R) = \frac{q_1}{r_1} + \frac{q_2}{r_2}$$

e con la radice quadrata

$$\sqrt{\text{var}(\ln R)} = \sqrt{\frac{q_1}{r_1} + \frac{q_2}{r_2}}$$

diventa **l'errore standard** (ES) di $\ln R$.

Da questa stima dell'errore standard, si ricava che per la **probabilità α**,

A) i limiti dell'**intervallo di confidenza** di $\ln R$ sono

- il **limite inferiore**

$$L_1 = \ln R - Z_{\alpha/2} \cdot \sqrt{\frac{q_1}{r_1} + \frac{q_2}{r_2}}$$

- il **limite superiore**

$$L_2 = \ln R + Z_{\alpha/2} \cdot \sqrt{\frac{q_1}{r_1} + \frac{q_2}{r_2}}$$

B) i limiti dell'**intervallo di confidenza di ρ** (quindi del valore $R = \frac{p_2}{p_1}$) sono

1 - il **limite inferiore**: e^{L1} scritto anche $\exp(L_1)$,

2 - il **limite superiore:** e^{L2} scritto anche $\exp(L_2)$;

C) la **significatività del rapporto R** è determinata mediante

$$Z = \frac{\ln R}{ES(\ln R)} = \frac{\ln R}{\sqrt{\frac{q}{r_1} + \frac{q}{r_2}}}$$

Questa ultima formula dell'errore standard, che

- richiede l'uso di q al posto di q_1 e q_2 presenti nella formula già indicata per l'intervallo di confidenza,
- deriva dal fatto che l'**ipotesi nulla** che si intende verificare è

$$\mathbf{H_0}: \pi_2 = \pi_1 = \pi$$

- nella quale la stima migliore di π è fornita da

$$p = \frac{r_1 + r_2}{n_1 + n_2}$$

- quando si utilizzano **i dati di due campioni indipendenti** e dove $q = 1 - p$.

Il test per la **significatività del rapporto R** spesso è scritto

come

$$Z = \frac{\ln R}{\sqrt{q \cdot \left(\frac{1}{r_1} + \frac{1}{r_2} \right)}}$$

evidenziando ancor meglio il suo **errore standard** dipende dal **valore medio ponderato di** p .

ESEMPIO 1. (RAPPORTO R E SUOI LIMITI DI CONFIDENZA) Dalle due proporzioni p_1 e p_2 ricavate da due campioni indipendenti, dove p_1 = 108/180 e p_2 = 60/120,

- calcolare il rapporto R e i limiti dell'intervallo di confidenza alla probabilità α = 0.05.

Risposta. Dopo aver calcolato p_1 = 60/120 = 0,5 e p_2 = 108/180 = 0,6

1 - si ottiene il rapporto $R = 0{,}6 / 0{,}5 = 1{,}2$.

Ma per avere, almeno approssimativamente, una **distribuzione normale** delle risposte campionarie possibili e quindi poter calcolare **l'intervallo di confidenza** mediante **la distribuzione Z**,

2 - tale rapporto deve essere trasformato in

$$\ln R = \ln 1{,}2 = 0{,}183$$

3 - il cui **errore standard** (ES di $\ln R$) con

$q_1 = 1 - p_1 = 1 - 0{,}5 = 0{,}5$

$q_2 = 1 - p_2 = 1 - 0{,}6 = 0{,}4$

$r_1 = 120$ e $r_2 = 180$

è

$$ES(\ln R) = \sqrt{\frac{q_1}{r_1} + \frac{q_2}{r_2}} = \sqrt{\frac{0{,}5}{60} + \frac{0{,}4}{108}} = \sqrt{0{,}00833 + 0{,}00370} = 0{,}1097$$

uguale a 0,1097.

Poiché per $\alpha = 0.05$ in una **distribuzione normale ridotta bilaterale** è riportato $Z = 1,96$

4 – per l'**intervallo di confidenza** di $\ln R = 0,183$

- **il limite inferiore**

$$L_1 = \ln R - Z_{\alpha/2} \cdot \sqrt{\frac{q_1}{r_1} + \frac{q_2}{r_2}} = 0,183 - 1,96x0,1097 = 0,183 - 0,215 = -0,032$$

è $L_1 = -0,032$

- **il limite superiore**

$$L_2 = \ln R + Z_{\alpha/2} \cdot \sqrt{\frac{q_1}{r_1} + \frac{q_2}{r_2}} = 0,183 + 1,96x0,1097 = 0,183 + 0,215 = 0,398$$

è $L_2 = 0,398$.

con probabilità del 95% che quanto affermato sia vero.

5 - Infine, dall'intervallo di confidenza di $\ln R$ si ritorna all'**intervallo di confidenza di** R.

Quindi, con i dati dell'esempio, intorno al **valore medio campionario** $R = 1,2$ si hanno

- il **limite inferiore** $e^{L1} = 2,718^{-0,032} = 0,969$

- il **limite superiore** $e^{L2} = 2,718^{0,398} = 1,489$.

In conclusione i **limiti dell'intervallo fiduciale** di ρ sono 0,969 e 1,489.

Ovviamente, con la trasformazione da $\ln R$ al rapporto R, l'intervallo non è più simmetrico.

ESEMPIO 2 (SIGNIFICATIVITA' DEL RAPPORTO R CON DATI ESEMPIO 1). Valutare la significatività del rapporto tra le due proporzioni p_1 e p_2 ricavate da due campioni indipendenti, dove $p_1 = 108/180$ e $p_2 = 60/120$.

Risposta. In un **test bilaterale** con

$$H_0: \pi_1 = \pi_2 \quad \text{contro} \quad H_1: \pi_1 \neq \pi_2$$

e dove

- $r_1 = 60$ e $r_2 = 108$

- $n_1 = 120$ e $n_2 = 180$

dopo aver calcolato

- $p_1 = 60/120 = 0,5$ e $p_2 = 108/180 = 0,6$

- $p = \frac{r_1 + r_2}{n_1 + n_2} = \frac{60 + 108}{120 + 180} = \frac{168}{300} = 0,56$

- $q = 1 - p = 1 - 0{,}56 = 0{,}44$

il **rapporto R** è

$$R = \frac{p_2}{p_1} = \frac{0{,}6}{0{,}5} = 1{,}2$$

e la sua **significatività** è verificata

con

$$Z = \frac{\ln R}{ES(\ln R)} = \frac{\ln R}{\sqrt{\frac{q}{r_1} + \frac{q}{r_2}}} = \frac{\ln 1{,}2}{\sqrt{\frac{0{,}44}{60} + \frac{0{,}44}{108}}} = \frac{0{,}183}{\sqrt{0{,}0114}} = \frac{0{,}183}{0{,}107} = 1{,}71$$

ottenendo Z = 1,71.

In una distribuzione normale ridotta bilaterale, corrisponde alla probabilità P = 0,087.

Quindi **non permette di rifiutare l'ipotesi nulla** se, come prassi, la soglia di significatività minima è stata indicata in $\alpha = 0.05$.

Come tutti gli intervalli di confidenza, pure quello precedente dovrebbe servire anche per valutare la **significatività del rapporto**

$$R = \frac{p_2}{p_1}$$

in un **test bilaterale** con ipotesi

$$H_0: \pi_1 = \pi_2 \quad \text{contro} \quad H_1: \pi_1 \neq \pi_2$$

In questi test, si rifiuta l'ipotesi nulla H_0,

- quando **nell'intervallo di confidenza di** R **non è compreso il valore 1** (che si dovrebbe ottenere quando l'ipotesi nulla è vera).

Di norma, l'intervallo di confidenza calcolato con la **distribuzione** normale ridotta Z e il test Z forniscono **risposte identiche**. Ma **non nel caso del rapporto R** e **del test per la significatività di R**, a motivo delle diverse formule utilizzate per calcolare l'errore standard di $\ln R$.

Esistono differenze; ma quasi sempre sono molto piccole, quando i campioni hanno dimensioni non troppo diverse. In pratica, anche per il rapporto R l'intervallo di confidenza è utilizzato per l'inferenza sulla sua significatività. La dimostrazione dell'esistenza di differenze trascurabili è data dalle due conclusioni precedenti, qui riportate:

A) Nell'esempio 1 del paragrafo precedente, con R che varia tra

- il limite inferiore $L_1 = 0{,}969$
- il limite superiore $L_2 = 1{,}489$
- il valore $R = 1{,}0$ è compreso nell'intervallo e quindi l'**ipotesi nulla non è stata rifiutata**, sempre con probabilità $\alpha = 0.05$ di un errore di Tipo I e in un test bilaterale.

B) Per verificare la stessa ipotesi

$$H_0: \pi_1 = \pi_2 \quad \text{contro} \quad H_1: \pi_1 \neq \pi_2$$

con il **test Z**

$$Z = \frac{\ln R}{ES(\ln R)} = \frac{\ln R}{\sqrt{\frac{q}{r_1} + \frac{q}{r_2}}} = \frac{\ln 1{,}2}{\sqrt{\frac{0{,}44}{60} + \frac{0{,}44}{108}}} = \frac{0{,}183}{\sqrt{0{,}0114}} = \frac{0{,}183}{0{,}107} = 1{,}71$$

nel quale si è ottenuto Z = 1,71

- **non è stato possibile rifiutare l'ipotesi nulla**, poiché corrisponde alla probabilità P = 0,087.
- sempre in una distribuzione bilaterale e con la soglia di significatività minima $\alpha = 0.05$.

Come già affermato, i due risultati non coincidono poiché l'errore standard è calcolato con due formule differenti. Con i dati dell'esempio

- per l'**intervallo di confidenza**

$$ES(\ln R) = \sqrt{\frac{q_1}{r_1} + \frac{q_2}{r_2}} = \sqrt{\frac{0{,}5}{60} + \frac{0{,}4}{108}} = \sqrt{0{,}00833 + 0{,}00370} = \sqrt{0{,}01203} = 0{,}1097$$

si è ottenuto ES($\ln R$) = 0,1097

- per il **test di significatività**

$$ES(\ln R) = \sqrt{\frac{q}{r_1} + \frac{q}{r_2}} = \sqrt{\frac{0{,}44}{60} + \frac{0{,}44}{108}} = \sqrt{0{,}00733 + 0{,}00407} = \sqrt{0{,}0114} = 0{,}1068$$

si è ottenuto ES($\ln R$) = 0,1068

Ma è una differenza trascurabile, minore del 3% rispetto al valore inferiore.

FORMULA TEST BASED DI MIETTINEN

Un **metodo rapido e approssimato** per calcolare l'**intervallo di confidenza di** ρ, cioè del valore vero del rapporto R tra due proporzioni, è stata proposta da Olli S. **Miettinen** nel 1976 (con l'articolo ***Estimability and estimation in case referent studies*** pubblicato su **American Journal of Epidemiology** Vol. 103, p.: 226-235). In letteratura è chiamato *formula **test-based** di Miettinen*, in quanto ricorre all'errore standard utilizzato nella formula per verificare la significatività della differenza tra due proporzioni.

Tralasciando la lunga dimostrazione matematica e i passaggi logici che permettono di derivarla dalle formule precedenti, alla **probabilità del 95%** i **limiti dell'intervallo di confidenza di ρ** possono essere determinati

con la formula

$$R^{(1\pm 1,96/Z_1)}$$

dove

$$Z_1 = \frac{p_2 - p_1}{\sqrt{(p \cdot q) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

e in parole

- Z_1 è la Deviata Normale Standardizzata della **differenza** tra due proporzioni.

Questa riportata è la formula più semplice. Al posto della differenza, altre varianti sempre proposte da Miettinen utilizzano il **rapporto R tra due proporzioni**, tra **due odds** oppure tra **due tassi**. Ma appunto perché sono rapporti, hanno una **distribuzione log-Normale**, con forte asimmetria destra, che può essere ricondotta alla normale solamente con una trasformazione logaritmica. Il calcolo diventa più complesso e lungo, rispetto a questa formula. Per ulteriori informazioni sulla metodologia, si rimanda a testi specifici.

La corrispondenza con l'intervallo di confidenza calcolato in precedenza è dimostrata con l'esempio seguente.

ESEMPIO 3 (USO DELLLA FORMULA DI MIETTINEN, CON I DATI DELL'ESEMPIO 1). Dalle due proporzioni p_1 e p_2 ottenute con due campioni indipendenti, dove p_1 = 108/180 e p_2 = 60/120,

- ricavare il rapporto R e i suoi limiti di confidenza alla probabilità α = 0.05.

Risposta. Dopo aver calcolato p_1 = 60/120 = 0,5 e p_2 = 108/180 = 0,6

si ottiene il rapporto $R = 0{,}6 / 0{,}5 = 1{,}2$.

Successivamente si deve stimare

$$p = \frac{60 + 108}{120 + 180} = \frac{168}{300} = 0{,}56$$

e il valore

$$Z_1 = \frac{p_2 - p_1}{\sqrt{(p \cdot q) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0{,}6 - 0{,}5}{\sqrt{(0{,}56x0{,}44) \cdot \left(\frac{1}{120} + \frac{1}{180}\right)}} = \frac{0{,}1}{\sqrt{0{,}2464x0{,}01389}} = \frac{0{,}1}{0{,}0585} = 1{,}71$$

Infine con

$$R^{(1 \pm 1{,}96 / Z_1)} = 1{,}2^{(1 \pm 1{,}96 / 1{,}71)} = 1{,}2^{(1 \pm 1{,}146)}$$

si trovano

- il **limite inferiore** $L_1 = 1{,}2^{(1-1{,}146)} = 1{,}2^{-0{,}146} = 0{,}974$

- il **limite superiore** $L_2 = 1{,}2^{(1+1{,}146)} = 1{,}2^{2{,}146} = 1{,}479$.

E' semplice osservare che, con i dati dell'esempio 1, intorno al **valore medio campionario** $R = 1{,}2$ per il valore reale ρ, con la distribuzione normale applicata a $\ln R$, si erano stimati

- il **limite inferiore** $e^{L1} = 2{,}718^{-0{,}032} = 0{,}969$

- il **limite superiore** $e^{L2} = 2{,}718^{0{,}398} = 1{,}489$

E' una dimostrazione empirica dell'equivalenza dei due metodi.

In questo caso, la formula di Miettinen determina un intervallo leggermente minore.

5.16. IL RAPPORTO TRA DUE ODDS (OR): INTERVALLO DI CONFIDENZA E SIGNIFICATIVITA'; FORMULA TEST-BASED DI MIETTINEN PER OR

Nel paragrafo precedente è stato presentato come, in un esperimento con n pazienti, se la cura ha successo per r individui, si può calcolare che

- la **proporzione del successo è** $p = \frac{r}{n}$

- l'**odds del successo è** $o = \frac{r}{n - r}$,

L'uso di un **odds in sostituzione di una proporzione**, ancor più nel caso di un **rapporto tra due odds in sostituzione di un rapporto tra due proporzioni**, comporta un vantaggio e uno svantaggio, entrambi rilevanti:

- il **vantaggio** deriva dalla **proprietà matematiche degli odds** che permettono elaborazioni più sofisticate di quanto è possibile con le proporzioni, come nel caso della regressione logistica;
- lo **svantaggio** è che **l'odds è un concetto privo di senso,** mentre **una proporzione è un concetto facilmente comprensibile**.

La **definizione di odds** è: **il rapporto del numero di eventi diviso il numero di non eventi**.

Quindi nella ricerca spesso **i concetti sono espressi in proporzioni** per stimare il **rischio relativo,** mentre **i calcoli** sono effettuati **con gli odds**, trasferendo le analisi e i risultati dall'uno all'altro.

L'odds è usato **in modo proprio** negli studi di **caso-controllo** quando **non è nota la prevalenza della malattia**. In tale caso, il **rapporto campionario tra due percentuali** fornisce una indicazione solo apparente della prevalenza. Per evitare equivoci è quindi appropriato l'uso dell'odds, pubblicato per la prima volta nel 1950 come metodo per gli studi caso-controllo.

Per confrontare il successo della stessa cura in due campioni, si può calcolare
il **rapporto dei due odds** (***odds ratio***) utilizzando

- sia le **frequenze assolute**

$$OR = \frac{o_1}{o_2} = \frac{r_1 \cdot (n_2 - r_2)}{r_2 \cdot (n_1 - r_1)}$$

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

$$OR = \frac{p_1 \cdot (1 - p_2)}{p_2 \cdot (1 - p_1)}$$

Questo ultimo rapporto è scritto anche

$$OR = \frac{p_1 \cdot q_2}{p_2 \cdot q_1}$$

Per tutte queste formule, la simbologia schematizzata nella tabella:

	Campione 1	Campione 2	Totale
Conteggio positivi	r_1	r_2	r
Conteggio negativi	$n_1 - r_1$	$n_2 - r_2$	$n - r$
Totale	n_1	n_2	n
Proporzione di successi	$p_1 = \frac{r_1}{n_1}$	$p_2 = \frac{r_2}{n_2}$	$p = \frac{r}{n}$
Proporzione di insuccessi	$q_1 = 1 - p_1$	$q_2 = 1 - p_2$	$q = 1 - p$
Odds di successo	$o_1 = \frac{r_1}{n_1 - r_1}$	$o_2 = \frac{r_2}{n_2 - r_2}$	$o = \frac{r}{n - r}$

Anche in questo caso, come nel paragrafo precedente, il valore di **OR ha una distribuzione campionaria che è log-Normale**, almeno in modo approssimato.

Ne consegue che

- sia per costruire l'**intervallo di confidenza** di ρ,

- sia per verificare la **significatività di un** OR

si deve utilizzare **non il valore di OR** direttamente, ma la sua **trasformazione** in $\ln OR$.

L'**errore standard del logaritmo dell'odds ratio**

è

$$ES(\ln OR) = \sqrt{\frac{1}{r_1} + \frac{1}{n_1 - r_1} + \frac{1}{r_2} + \frac{1}{n_2 - r_2}}$$

Con la solita simbologia delle tabelle di contingenza 2 x 2, applicato allo studio caso-controllo,

	Successi +	Insuccessi -	Totale
Caso	a	b	$\mathbf{n_1}$ = a + b
Controllo	c	d	$\mathbf{n_2}$ = c + d
Totale	$\mathbf{n_3}$ = a + c	$\mathbf{n_4}$ = a + d	N = a+b+c+d

corrisponde alla formula

$$ES(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

L'**intervallo di confidenza** del **logaritmo del valore reale dell'odds ratio** detto anche **intervallo di confidenza** di $\ln OR$ è delimitato da

- il **limite inferiore**

$$L_1 = \ln OR - Z_{\alpha/2} \cdot ES(\ln OR)$$

- il **limite superiore**

$$L_2 = \ln OR + Z_{\alpha/2} \cdot ES(\ln OR)$$

Da essi è possibile ricavare l'**intervallo di confidenza di ρ** (quindi del valore OR prima della trasformazione in $\ln OR$) dove

1 - il **limite inferiore** è: e^{L1} scritto anche $\exp(L_1)$,

2 - il **limite superiore** è: e^{L2} scritto anche $\exp(L_2)$;

Per verificare l'ipotesi nulla H_0: $\pi_1 = \pi_2$

che è equivalente sia a H_0: **ρ = 1** sia a H_0: $\ln \rho = 0$

in un test che può essere sia **unilaterale** sia **bilaterale**

si utilizza la **deviata normale standardizzata**

$$Z = \frac{\ln OR}{ES(\ln OR)}$$

Alternativamente, fondata sulla prima ipotesi nulla qui espressa sulle proporzioni, e come nel paragrafo precedente è possibile utilizzare anche

- la Deviata Normale Standardizzata della **differenza** tra due proporzioni

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

Ma come per le **tabelle di contingenza 2 x 2** è possibile utilizzare pure

- il **test chi-quadrato**, il **test G**, il **metodo esatto di Fisher**, come illustrati nel capitolo 3

FORMULA TEST BASED DI MIETTINEN

Anche in questo caso come nel paragrafo precedente, alla **probabilità del 95%** i **limiti dell'intervallo di confidenza di ρ** possono essere determinati

con la **formula di Miettinen**

$$OR^{(1\pm 1,96/Z)}$$

dove

- Z è la Deviata Normale Standardizzata della **differenza** tra due proporzioni, calcolata con l'ultima formula riportata.

Come nel paragrafo precedente, i due approcci dovrebbero fornire risultati approssimativamente simili.

ESEMPIO 1 (IL RAPPORTO TRA DUE ODDS E SUOI LIMITI DI CONFIDENZA). Con gli stessi dati utilizzati per il rapporto R tra due proporzioni del paragrafo precedente, dove su 180 persone a rischio 108 presentavano patologie e su un campione di controllo dove su 120 persone 60 presentavano patologie,

- calcolare il rapporto dell'**odds ratio** (OR) e i suoi **limiti dell'intervallo di confidenza** alla probabilità $\alpha = 0.05$.

Risposta. Dopo aver calcolato aver impostato correttamente i dati in una tabella di contingenza 2 x 2 al fine di meglio comprendere termini del problema

	Successi +	Insuccessi -	Totale
Esposti	a **108** r_1	b **72** $n_1 - r_1$	180 n_1
Controllo	c **60** r_2	d **60** $n_2 - r_2$	120 n_2
Totale	168	132	300

e aver utilizzato entrambe le simbologie per evidenziarne le corrispondenze

1 - si calcola l'**odds ratio stimato** con l'esperimento

$$OR = \frac{o_1}{o_2} = \frac{r_1 \cdot (n_2 - r_2)}{r_2 \cdot (n_1 - r_1)} = \frac{a \cdot d}{c \cdot b} = \frac{108x60}{60x72} = \frac{6480}{4320} = 1{,}5$$

2 – Ma per avere, almeno approssimativamente, una **distribuzione normale** delle risposte campionarie possibili e quindi poter calcolare **l'intervallo di confidenza** mediante **la distribuzione Z**,

- tale odds ratio **OR** deve essere trasformato nel

logaritmo dell'**odds ratio** ($\ln OR$)

$$\ln OR = \ln 1{,}5 = 0{,}4055$$

3 – il cui **errore standard** (ES di $\ln OR$)

è

$$ES(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \sqrt{\frac{1}{108} + \frac{1}{72} + \frac{1}{60} + \frac{1}{60}} = \sqrt{0{,}056482} = 0{,}2377$$

oppure

$$ES(\ln OR) = \sqrt{\frac{1}{r_1} + \frac{1}{n_1 - r_1} + \frac{1}{r_2} + \frac{1}{n_2 - r_2}} = \sqrt{\frac{1}{108} + \frac{1}{72} + \frac{1}{60} + \frac{1}{60}} = \sqrt{0{,}056482} = 0{,}2377$$

Poiché per $\alpha = 0.05$ in una **distribuzione normale ridotta bilaterale** è riportato $Z = 1{,}96$

per l'intervallo di confidenza di $\ln OR = 0{,}4055$

4 - il **limite inferiore**

è

$$L_1 = \ln OR - Z_{\alpha/2} \cdot ES(\ln OR) = 0{,}4055 - 1{,}96x0{,}2377 = 0{,}4055 - 0{,}4659 = -0{,}0604$$

- il **limite superiore**

$$L_2 = \ln OR + Z_{\alpha/2} \cdot ES(\ln OR) = 0{,}4055 + 1{,}96x0{,}2377 = 0{,}4055 + 0{,}4659 = 0{,}8714$$

con probabilità del 95% che quanto affermato sia vero.

5 - Infine, dall'intervallo di confidenza di $\ln OR$ si stima l'**intervallo di confidenza di** OR.

Quindi, con i dati dell'esempio, intorno al **valore medio campionario** $OR = 1{,}5$ come limiti del **rapporto vero** ρ si hanno

- il **limite inferiore** $e^{L1} = 2{,}718^{-0{,}0604} = 0{,}941$
- il **limite superiore** $e^{L2} = 2{,}718^{0{,}8714} = 2{,}390$.

Con gli stessi dati,

- nel paragrafo precedente

il **rapporto tra due proporzioni** è stato

$$R = \frac{p_2}{p_1} = \frac{0,6}{0,5} = 1,2$$

- in questo paragrafo

il **rapporto tra due odds** è

$$OR = \frac{o_1}{o_2} = \frac{r_1 \cdot (n_2 - r_2)}{r_2 \cdot (n_1 - r_1)} = \frac{a \cdot d}{c \cdot b} = \frac{108x60}{60x72} = \frac{6480}{4320} = 1,5$$

Ma quando **le frequenze dei successi diventano piccole**, come nel caso seguente

	Successi +	Insuccessi -	Totale
Esposti	a **108** r_1	b **1692** $n_1 - r_1$	1800 n_1
Controllo	c **60** r_2	d **1140** $n_2 - r_2$	1200 n_2
Totale	168	2832	3000

dove $p_2 = \frac{108}{1800} = 0,06$ e $p_1 = \frac{60}{1200} = 0,05$

- sebbene il **rapporto tra le due proporzioni** sia stato mantenuto uguale

$$R = \frac{p_2}{p_1} = \frac{0,06}{0,05} = 1,200$$

- il **rapporto tra due odds** diventa

$$OR = \frac{o_1}{o_2} = \frac{r_1 \cdot (n_2 - r_2)}{r_2 \cdot (n_1 - r_1)} = \frac{a \cdot d}{c \cdot b} = \frac{108x1140}{60x1692} = \frac{123120}{101520} = 1,213$$

molto simile a quello tra due proporzioni.

E' una dimostrazione empirica di quanto affermato nella prima parte del paragrafo precedente:

- quando le proporzioni diventano piccole (inferiori a 0,04 - 0,03)**, il rapporto R tra le due proporzioni e il rapporto OR tra i due odds convergono:**

- quindi è possibile **usare il rapporto tra due odds (OR) che gode di proprietà matematiche migliori**, seppure i concetti restino diversi.

ESEMPIO 2 (SIGNIFICATIVITA' DEL RAPPORTO OR, CON I DATI DELL'ESEMPIO 1). L'**odds ratio stimato** con l'esperimento riportato nell'esempio precedente è stato $OR = 1{,}5$.

E' significativo?

Risposta. Il test è unilaterale e per valutare l'ipotesi

$$H_0: \rho \leq 1 \quad \text{contro} \quad H_1: \rho > 1$$

oppure l'equivalente

$$H_0: \ln\rho \leq 0 \quad \text{contro} \quad H_1: \ln\rho > 0$$

1 - servendosi della distribuzione normale ridotta occorre utilizzare la trasformazione di OR in

$$\ln OR = \ln 1{,}5 = 0{,}4055$$

2 - il cui **errore standard** (ES di $\ln OR$) è

$$ES(\ln OR) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{108}+\frac{1}{72}+\frac{1}{60}+\frac{1}{60}} = \sqrt{0{,}056482} = 0{,}2377$$

3 - Il test

$$Z = \frac{\ln OR}{ES(\ln OR)}\frac{0{,}4055}{0{,}2377} = 1{,}71$$

permette di stimare Z = 1,71 che in una coda della distribuzione normale ridotta corrisponde alla probabilità P = 0,044.

Si rifiuta l'ipotesi nulla: il valore di odds ratio è statisticamente significativo.

ESEMPIO 3 (USO DELLLA FORMULA DI MIETTINEN, CON DATI DI ESEMPIO 1). Dalla tabella di contingenza

	Successi +	Insuccessi -	Totale
Esposti	a **108** r_1	b **72** $n_1 - r_1$	180 n_1
Controllo	c **60** r_2	d **60** $n_2 - r_2$	120 n_2
Totale	168	132	300

- calcolare i limiti di confidenza dell'odds ratio OR = 1,50 alla probabilità $\alpha = 0.05$.

Risposta. Dopo aver calcolato

- $p_1 = 60/120 = 0,5$ e $p_2 = 108/180 = 0,6$

si stima la **proporzione media p**

$$p = \frac{60+108}{120+180} = \frac{168}{300} = 0,56$$

e il valore

$$Z = \frac{p_2 - p_1}{\sqrt{(p \cdot q) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,6 - 0,5}{\sqrt{(0,56x0,44) \cdot \left(\frac{1}{120} + \frac{1}{180}\right)}} = \frac{0,1}{\sqrt{0,2464x0,01389}} = \frac{0,1}{0,0585} = 1,71$$

Infine con OR = 1,5 e

$$OR^{(1\pm1,96/Z)} = 1,5^{(1\pm1,96/1,71)} = 1,5^{(1\pm1,146)}$$

si trovano

- il **limite inferiore** $L_1 = 1,5^{(1-1,146)} = 1,5^{-0,146} = 0,943$

- il **limite superiore** $L_2 = 1,5^{(1+1,146)} = 1,5^{2,146} = 2,387$.

E' semplice osservare che, con i dati dell'esempio 1, intorno al **valore medio campionario** $OR = 1,5$ per il valore reale ρ, con la distribuzione normale applicata a $\ln OR$ si erano stimati

- il **limite inferiore** $e^{L1} = 2,718^{-0,0604} = 0,941$

- il **limite superiore** $e^{L2} = 2,718^{0,8714} = 2,390$.

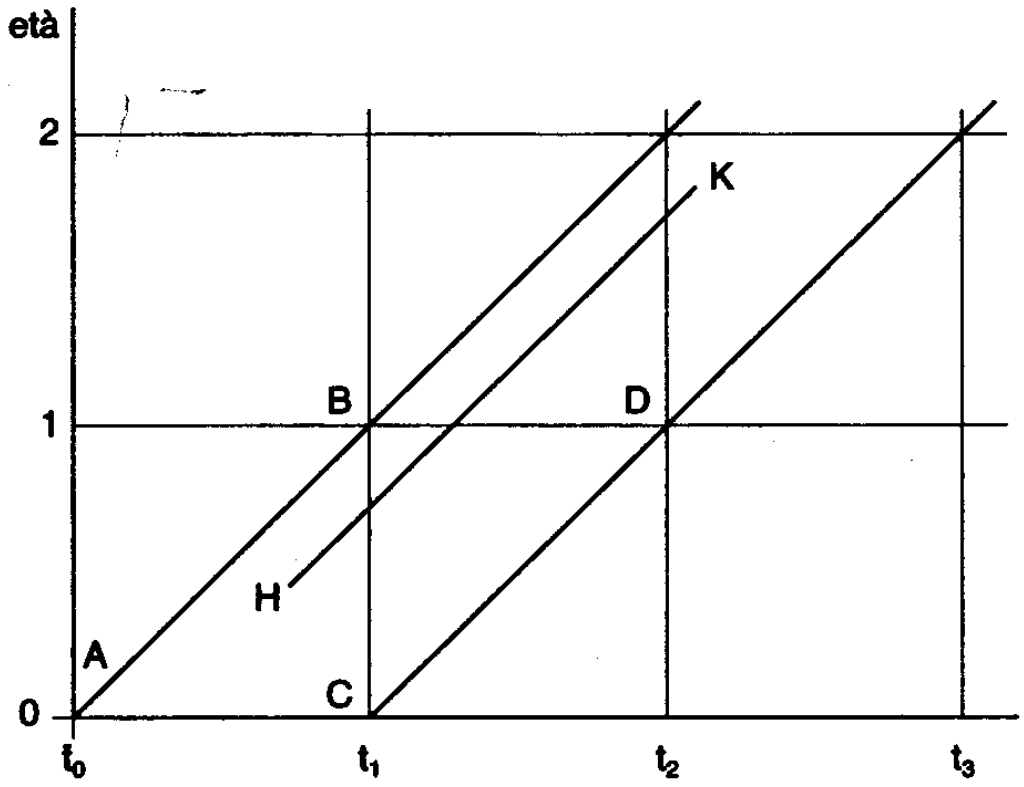
E' una dimostrazione empirica dell'equivalenza dei due metodi, data la differenza minima nei risultati. Anche in questo caso, la formula di Miettinen determina un intervallo leggermente minore. Ma per entrambe le formule sono calcoli effettuati con sole 3-4 cifre decimali.

5.17. IL RAPPORTO TRA DUE TASSI (RR): INTERVALLO DI CONFIDENZA E SIGNIFICATIVITA'; FORMULA TEST-BASED DI MIETTINEN

In demografia e in epidemiologia, **proporzione** e **tasso** (***rate***) hanno due significati chiaramente distinti, anche se spesso

- vengono utilizzati per analizzare gli stessi fenomeni e
- hanno valori simili quando le frequenza sono basse.

Per meglio comprenderne le differenze, è didatticamente vantaggioso presentare gli eventi demografici e epidemiologici nel **diagramma di Lexis**. (Wilhem Lexis 1837-1914 si laureò all'Università di Bonn nel 1859 in scienze e matematica e si trasferì a Parigi nel 1861 dove condusse studi sulle discipline sociali, pubblicando soprattutto nel quinquennio 1976-1880).



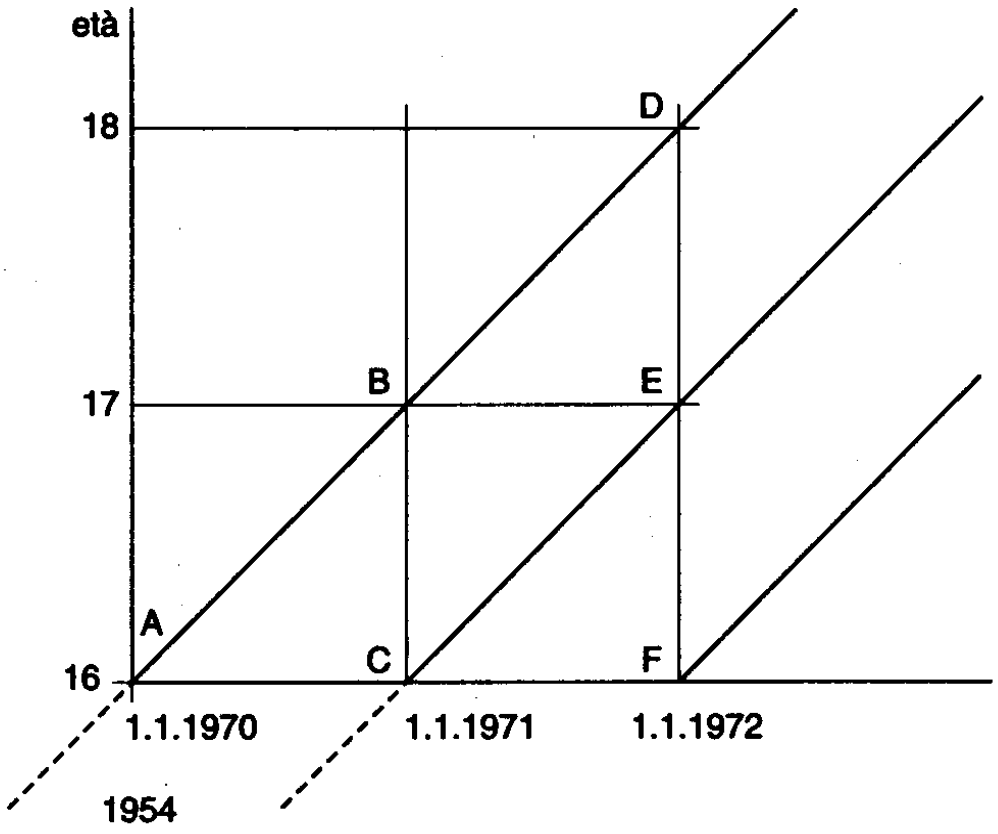
In letteratura, il diagramma di Lexis è presente in quattro versioni (quella originale proposta da Lexis, la versione americana, quella italiana utilizzata dall'ISTAT e quella la francese) delle quali quella più semplice, qui proposta, è quella francese. Ormai anche quasi tutti i testi italiani di demografia riportano la versione francese.

E' un diagramma cartesiano, come quello riportato nella pagina precedente. Per un individuo, permette di **rappresentare graficamente un evento nel tempo**, considerando contemporaneamente due parametri fondamentali che lo misurano,

- il tempo del calendario o le date, sull'asse delle ascisse;
- la durata dell'evento oppure l'età dell'individuo, sull'asse delle ordinate.

Ad esempio, nella figura precedente si può osservare la rappresentazione grafica della storia clinica di un bambino che è entrato in osservazione nel momento H e ne è uscito nel momento K (per guarigione, morte oppure trasferimento):

- la perpendicolare sull'asse delle ascisse dei due punti individua esattamente le due date,
- la perpendicolare sull'asse delle ascisse individua la sua età nei due momenti.



Quando da un individuo di passa a un gruppo di individui, le loro storie individuali si muovono nello spazio del diagramma cartesiano nello stesso modo e possono essere rappresentate con la figura della pagina precedente.

Ad esempio, si assuma che durante l'anno 1970, un gruppo di 180 giovani donne nel giorno del compimento del 16° anniversario (quindi appartenenti alla generazione nata nel 1954) si siano presentate a una visita di controllo per anemia, che siano state seguite tutte esattamente per un anno, fino al giorno del compimento del 17° anno. Se in quel periodo di osservazione 12 di esse hanno manifestato problemi gravi di anemia, nel grafico

- le 180 giovani sono riportate sul segmento AC,
- le 12 trovate positive sono riportate nel quadrilatero ABEC.

Il **rapporto ABEC/AC = 12/180 è la proporzione** di ragazze con problemi gravi di anemia.

Se riferita a una singola ragazza nelle stesse condizioni, **è la probabilità che essa si ammali**.

Se le 12 fossero giovani decedute, in questo modo si sarebbe misurata la loro probabilità di decesso.

Ma condurre un'analisi su un gruppo non sempre è così semplice.

Come prima, si supponga che durante l'anno 1970 un gruppo di 180 giovani donne nel giorno del compimento del 16° compleanno si siano presentate a una visita di controllo di anemia per essere seguite fino al compimento del 17° anno. Ma che non tutte siano state seguite esattamente per un anno, perché 30 di esse si sono trasferite in altra provincia e 10 nuove sono entrate. Se in quel periodo di osservazione 12 di esse hanno avuto problemi gravi di anemia, nel grafico

- il numero medio di ragazze seguite è identificato dal segmento CB.
- le 12 trovate positive come prima sono riportate nel quadrilatero ABEC.

Il **rapporto ABEC/CB è il tasso** di ragazze con problemi gravi di anemia.

Come calcolare il numero medio di ragazze seguite?

Esistono due modi ed entrambi richiedono una più esatta comprensione della trasformazione numerica del gruppo in osservazione:

1 - delle 180 ragazze iniziali, 30 sono emigrate dalla provincia; quindi le altre 150 sono state seguite per tutto l'anno;

2 – le altre 10 sono immigrate durante l'anno, per cui alla fine erano in 160.

Il **primo metodo** di calcolo è **la media tra il numero iniziale e quello finale**: (180 + 160) / 2 = 170

Nel grafico, questo 170 identifica il numero di persone che avremmo approssimativamente trovato facendo un loro censimento nel **momento BC**, cioè alla fine dell'anno 1970.

Il **secondo metodo** implica il **concetto di anni-vissuti**.

1) Poiché 150 ragazze sono stata eseguite per tutto l'anno, si può dedurre che insieme esse hanno vissuto 150 anni (150 ragazze x 1 anno) di osservazione;
2) ognuna delle 30 ragazze che si sono allontanate durante l'anno approssimativamente sono state in osservazione per metà anno; ne deriva che complessivamente gli anni vissuti in osservazione da questa ragazze sono stati 15 (30 ragazze x 0,5 anni);
3) le 10 ragazze entrate in osservazione durante l'anno approssimativamente sono state in osservazione per metà anno; il loro contributo agli anni vissuti in osservazione è 5 (10 ragazze x 0,5 anni);
4) in conclusione, **la somma di anni vissuti in osservazione** è 170 (150 + 15 + 5).

Questo secondo metodo è di grande utilità in quanto permette facilmente, sulla base dei tempi della malattia, di calcolare giorni, settimane o anni di osservazione. Quindi, conoscendo il numero di eventi di quel periodo, di calcolare il tasso giornaliero, settimanale oppure della malattia. Ma è utile soprattutto quando si segue un gruppo non molto grande di individui, (ad esempio 200 persone), per un periodo lungo (ad esempio 20 anni) di ***follow-up***. Il numero di anni persona (200 x 20) diventa 4000 e i valori dei tassi che si calcolano sono del tutto paragonabili al seguire 4000 persone per un anno. E' uno sforzo di analisi e controllo che pochi gruppi di medici sarebbero in grado di fare. Inoltre, se alcuni pazienti si allontanano e altri entrano in osservazione, in questo modo il calcolo del valore medio è fatto con correttezza e semplicità.

Nello schema della tabella

	Campione 1	Campione 2	Totale
Numero di eventi	d_1	d_2	d
Tempo vissuto (anni, mesi)	T_1	T_2	T
Tasso osservato	$m_1 = \frac{d_1}{T_1}$	$m_2 = \frac{d_2}{T_2}$	$m = \frac{d}{T}$

è evidenziato il **calcolo di due tassi**.
Il **rapporto tra tassi** (***Rate Ratio***) è

$$RR = \frac{m_1}{m_2}$$

E' utile ricordare due nozioni:

- questo RR non deve essere confuso con il Rischio Relativo; il ***Rate Ratio*** o **rapporto tra tassi** ne è solamente una delle misure;
- i tassi m non sono proporzioni o odds, perché al denominatore è riportata la popolazione media o gli anni vissuti, non la popolazione iniziale.

Come gli altri rapporti di rischio, quali i due presentati nei paragrafi precedenti, anche i rapporti tra tassi RR **variano da 1 a 0** in una coda e da **1 a +∞** nell'altra. Quindi hanno una **distribuzione log-Normale** che, per variare simmetricamente intorno a 0, richiede la **trasformazione** degli RR **in** $\ln RR$,

- sia per il **calcolo dei limiti di confidenza**,
- sia per il **test di significatività**.

Per entrambi, poiché **il logaritmo di un rapporto** tra due tassi è **uguale alla differenza tra i logaritmi dei due tassi**

$$\ln RR = \ln \frac{m_1}{m_2} = \ln m_1 - \ln m_2$$

si ricava che

$$\text{var}(\ln RR) = \text{var}(\ln m_1) + \text{var}(\ln m_2)$$

- **la varianza di una differenza è uguale alla somma delle due varianze**.

Quando un tasso m è calcolato come media ponderata di due campioni

Tasso osservato	$m = \frac{d}{T} = \frac{d_1 + d_2}{T_1 + T_2}$	$m_1 = \frac{d_1}{T_1}$	$m_2 = \frac{d_2}{T_2}$

la varianza stimata del logaritmo del tasso ($\ln m$) è uguale a $1/d$.

Pertanto, poiché $d = d_1 + d_2$,

la varianza è di $\ln RR$ è

$$\text{var}(\ln RR) = \frac{1}{d_1} + \frac{1}{d_2}$$

e l'errore standard di $\ln RR$ è

$$ES(\ln RR) = \sqrt{\frac{1}{d_1} + \frac{1}{d_2}}$$

L'intervallo di confidenza alla probabilità **α del logaritmo vero del rapporto di due tassi** m_1 e m_2 **(** $\ln \rho$) è

$$\ln \rho = \ln RR \pm Z_{\alpha/2} \cdot \sqrt{\frac{1}{d_1} + \frac{1}{d_2}}$$

La **significatività del logaritmo del rapporto di due tassi**, in un test unilaterale oppure bilaterale con ipotesi nulla

H_0: ρ = 1 oppure H_0: $\ln \rho = 0$

può essere verificata con

$$Z_2 = \frac{\ln RR}{ES(RR)} = \frac{\ln RR}{\sqrt{\frac{1}{d_1} + \frac{1}{d_2}}}$$

La stesse ipotesi nulle, che in questo caso sono meglio espresse con

H_0: $\mu_1 = \mu_2$ oppure H_0: $\mu_1 - \mu_2 = 0$

dove

- μ non indica la media ma è il simbolo greco del tasso m ,

possono essere verificate anche mediante la

- **significatività della differenza tra due tassi** m_1 e m_2

con

$$Z_1 = \frac{m_1 - m_2}{m \cdot \sqrt{\frac{1}{d_1} + \frac{1}{d_2}}}$$

dove

$$m = \frac{d_1 + d_2}{T_1 + T_2}$$

I **valori Z_1 e Z_2** dei due test di significatività **non sono identici**. Dovrebbero essere **approssimativamente uguali**, in particolare quando i tassi m_1 m_2 sono stati calcolati con **denominatori T_1 e T_2 abbastanza grandi.**

FORMULA TEST BASED DI MIETTINEN

La **deviata normale standardizzata Z_1** calcolata per la **differenza tra due tassi**

quindi ottenuta con

$$Z_1 = \frac{m_1 - m_2}{m \cdot \sqrt{\frac{1}{d_1} + \frac{1}{d_2}}}$$

serve nella formula **test-based** di Miettinen

$$\rho = RR^{(1 \pm 1{,}96 / Z_1)}$$

per calcolare **i limiti di confidenza del rapporto reale (ρ) tra tassi**, alla probabilità del 95%.

ESEMPIO 1 (RAPPORTO RR E SUOI LIMITI DI CONFIDENZA). Riprendendo i dati del paragrafo precedente con due campioni abbastanza grandi

	Campione 1	Campione 2	Totale
Numero di eventi	**108**	**60**	**168**
Tempo vissuto	**1800**	**1200**	**3000**

calcolare

1 - i tassi dei campioni 1 e 2 e il tasso totale;

2 - il rapporto dei tassi (RR);

3 - i limiti di confidenza al 95% del logaritmo del rapporto dei tassi;

4 - i limiti di confidenza al 95% del rapporto vero dei due tassi.

(Nella realtà, i tassi sono nettamente minori di quelli qui utilizzati. Spesso sono espressi non in percentuali ma in per-mille o per-diecimila, come attualmente i livelli di mortalità nelle classi giovanili o per cause specifiche che hanno livelli particolarmente bassi).

Risposte.

1 - I tassi dei campioni 1 e 2 e il tasso totale sono

	Campione 1	Campione 2	Totale
Tasso osservato	$m_1 = \frac{108}{1800} = 0{,}06$	$m_2 = \frac{60}{1200} = 0{,}05$	$m = \frac{168}{3000} = 0{,}056$

2 - Il **rapporto** RR tra i due **tassi** (***Rate Ratio***)

è

$$RR = \frac{m_1}{m_2} = \frac{0{,}06}{0{,}05} = 1{,}20$$

3 - Il **logaritmo del rapporto tra i tassi** ($\ln RR$)

è

$$\ln RR = \ln 1{,}2 = 0{,}1823$$

e i suoi **limiti di confidenza al 95%** sono

- il **limite inferiore L_1**

$$\ln RR - Z_{\alpha/2} \cdot \sqrt{\frac{1}{d_1} + \frac{1}{d_2}} = 0{,}1823 - 1{,}96 \cdot \sqrt{\frac{1}{108} + \frac{1}{60}} = 0{,}1823 - 0{,}3156 = -0{,}1333$$

- il **limite superiore L_2**

$$\ln RR + Z_{\alpha/2} \cdot \sqrt{\frac{1}{d_1} + \frac{1}{d_2}} = 0{,}1823 + 1{,}96 \cdot \sqrt{\frac{1}{108} + \frac{1}{60}} = 0{,}1823 + 0{,}3156 = 0{,}4979$$

4 - I **limiti di confidenza al 95%** del **rapporto vero dei due tassi** (con RR = 1,20) sono

- il **limite inferiore** $e^{L1} = 2{,}718^{-0{,}1333} = 0{,}875$
- il **limite superiore** $e^{L2} = 2{,}718^{0{,}4979} = 1{,}645$.

ESEMPIO 2 (SIGNIFICATIVITA' DEL RAPPORTO R, CON DATI ESEMPIO 1). Con gli stessi dati dell'esempio 1, verificare

1 - la **significatività del logaritmo del rapporto tra tassi**

2 - la **significatività della differenza tra tassi**

Risposta.

1 - La **significatività del logaritmo del rapporto tra tassi** ($\ln RR$) per verificare l'ipotesi nulla

$$H_0: \ln \rho = 0$$

in un test che può essere bilaterale oppure unilaterale

è data da

$$Z_2 = \frac{\ln RR}{ES(RR)} = \frac{\ln RR}{\sqrt{\frac{1}{d_1} + \frac{1}{d_2}}} = \frac{0{,}1823}{\sqrt{\frac{1}{108} + \frac{1}{60}}} = \frac{0{,}1823}{\sqrt{0{,}0259}} = \frac{0{,}1823}{0{,}1609} = 1{,}13$$

E' un valore di Z molto piccolo, lontano dai valori critici.

La stessa ipotesi,

- formulata in modo differente ma **concettualmente identica**,
- può essere verificata mediante la **differenza tra due tassi**,

sulla base del ragionamento che se due tassi sono tra loro statisticamente differenti dovrebbero essere differenti sia il loro rapporto, sia la loro differenza.

I due metodi devono dare lo stesso risultato.

2 – In modo più formale, la significatività della differenza tra tassi $m_1 - m_2$ con ipotesi nulla

$$H_0: \mu_1 = \mu_2 \quad \text{oppure} \quad H_0: \mu_1 - \mu_2 = 0$$

dove

- μ non indica la media, ma è il **simbolo greco del tasso** m , (con $m = 0{,}056$)

è

$$Z_1 = \frac{m_1 - m_2}{m \cdot \sqrt{\frac{1}{d_1} + \frac{1}{d_2}}} = \frac{0{,}06 - 0{,}05}{0{,}056 \cdot \sqrt{\frac{1}{108} + \frac{1}{60}}} = \frac{0{,}01}{0{,}056 \cdot \sqrt{0{,}02593}} = \frac{0{,}01}{0{,}009} = 1{,}11$$

Rispetto al metodo precedente, il risultato ha una differenza trascurabile, dovuta agli arrotondamenti.

ESEMPIO 3 (USO DELLLA FORMULA DI MIETTINEN, CON DATI DI ESEMPIO 1). Con gli stessi dato dell'esempio 1, con la formula test-based di Miettinen calcolare

- i **limiti di confidenza** alla probabilità del 95% **del rapporto reale tra tassi**.

Risposta. Con RR = 1,2 e $Z_1 = 1{,}11$

mediante

$$RR^{(1\pm1,96/Z_1)} = 1,2^{(1\pm1,96/1,11)} = 1,2^{(1\pm1,766)}$$

si trovano

- il **limite inferiore** $L_1 = 1,2^{(1-1,766)} = 1,2^{-0,766} = 0,870$

- il **limite superiore** $L_2 = 1,2^{(1+1,766)} = 1,2^{2,766} = 1,656.$

E' semplice osservare che, con i dati dell'esempio 1, intorno al **valore medio campionario** $RR = 1,2$ come stima del valore reale ρ con la distribuzione normale applicata a $\ln RR$ si erano stimati

- il **limite inferiore** $e^{L1} = 2,718^{-0,0604} = 0,875$

- il **limite superiore** $e^{L2} = 2,718^{0,8714} = 1,645$

confermando la corrispondenza tra i due metodi, anche nel rapporto tra due tassi.

5.18. DIMENSIONI DEI CAMPIONI E POTENZA, PER TEST SULLA DIFFERENZA E SULL'ODDS RATIO DELLE PROPORZIONI DI DUE CAMPIONI INDIPENDENTI.

Recentemente Hardeo **Sahai** e Anwer **Khurshid,** nell'articolo del 1996 ***Formulae and tables for the determination of sample sizes and power in clinical trials for the testing differences in proportions for the two-sample design: a review*** (pubblicato su **Statistics in Medicine** Vol. 15, pp.: 1-21), hanno presentato

- un compendio di **formule esatte** e di **formule asintotiche**,
- per test sulla **significatività della differenza tra due proporzioni (** $d = p_1 - p_2$ **)**,
- finalizzate a stimare sia le **dimensioni (** n **) del campione** minimo sia **la potenza** $(1-\beta)$ del test,
- con **campioni bilanciati** $(n_1 = n_2)$ e con **campioni** che hanno **un numero differente di osservazioni (** $n_1 \neq n_2$ **)**.

Tali formule servono per **rendere significativa la differenza minima** δ **prescelta tra le due proporzioni**. Presentano pure formule per verificare la significatività di **misure di associazione** oppure di **odds ratio**, che hanno una impostazione analoga ai test sulla differenza in quanto utilizzano sempre una tabella 2 x 2, ma con finalità differenti.

Per ottenere la **potenza massima** in test di **confronto tra i parametri di due o più campioni,** è sempre richiesto che l'**esperimento sia bilanciato**.

Ma **non sempre è possibile**. Può avvenire che uno dei due campioni (p_1) sia già stato raccolto e che le sue dimensioni (n_1) siano obiettivamente troppo piccole oppure eccessivamente grandi, per le

finalità della ricerca. Ne deriva che il secondo campione (p_2) dovrà avere dimensioni (n_2) tali da compensare questo difetto, in particolare se il primo è troppo piccolo.

Spesso **non è neppure desiderabile avere campioni con la stesso numero di dati**. Nella scelta delle dimensioni possono assumere importanza anche altri fattori, spesso ritenuti addirittura prioritari. Ad esempio, come verrà sviluppato nel capitolo sul test t di Student per due campioni indipendenti, non sempre l'attribuzione di un individuo a un gruppo è una scelta neutra, sotto l'aspetto etico od economico. Si pensi al confronto tra farmaco e placebo, quando ad alcuni ammalati viene somministrato il placebo; si consideri la sperimentazione di un farmaco nuovo, ritenuto più efficace, quando per valutare il miglioramento effettivo ad alcuni pazienti deve essere somministrato quello vecchio. Per l'aspetto economico, una raccolta di dati condotta nella propria azienda o nella zona di residenza spesso è meno costosa di una condotta fuori sede; una analisi chimica o biologica effettuata con la vecchia metodologia può avere costi differenti da quella nuova.

Calcolato il numero $\hat{n}$, cioè **il numero minimo necessario** per due campioni bilanciati, se già si dispone di **un primo campione di dimensioni** n_1

l'altro campione deve avere dimensioni n_2 determinate dalla

relazione

$$n_2 = \frac{\hat{n} \cdot n_1}{2n_1 - \hat{n}}$$

affinché il test mantenga la **stessa potenza** $(1-\beta)$ di quello **bilanciato** $(n_1 = n_2 = \hat{n})$.

In molti ricerche di epidemiologia, le differenze tra due proporzioni sono fornite su una **scala moltiplicativa**. Come ampiamente illustrato nei paragrafi precedenti,

- negli studi di coorti il **rischio relativo** è fornito come $\phi = p_2 / p_1$
- e negli studi per **confrontare caso - controllo**

si ricorre all'**odds ratio** ψ

$$\psi = \frac{p_2 q_1}{p_1 q_2}$$

In questo ultimo caso, quando i **tassi sono piccoli**, il valore ψ dell'**odds ratio approssima molto bene il valore del rischio relativo** $\phi = p_2 / p_1$.

Quindi, ma solo in queste circostanze, **le formule per stimare le dimensioni del campione con l'odds ratio e con il rischio relativo sono uguali**.

Nelle formule successive,

- le dimensioni dei due campioni con proporzioni p_1 e p_2 sono indicate rispettivamente con n_1 e n_2
- dove $n_2 = kn_1$ e $0 < k < \infty$
- e se $k = 1$, i due campioni sono bilanciati: $n_1 = n_2 = n$.

L'elenco di Hardeo **Sahai** e Anwer **Khurshid** riporta 13 formule per calcolare la dimensione minima n, anche se quelle effettivamente operative sono 11.
La formula per calcolare la potenza $1 - \beta$ è riportata solo in 7 casi.

Per primo è riportato il **metodo esatto di Fisher** per tabelle 2 x 2, in letteratura indicato anche come **metodo di Fisher-Irwin**, nella forma

- sia **condizionale (*exact conditional method*)**,
- che **non condizionale (*exact unconditional method*)**.

Il **metodo condizionale** è quello **classico proposto da Fisher**, illustrato nel capitolo sul chi- quadrato. E' fondato sulla costanza dei totali marginali e per i calcoli utilizza la **distribuzione ipergeometrica**.
In **metodo non condizionale** è fondato sulla osservazione che mantenere costanti tutti i totali marginali, in particolare **il numero di successi, non è un'ipotesi sempre credibile** come sostengono

- S. **Suissa** e J. J. **Shuster** nel 1985 con l'articolo ***Exact unconditional sample sizes for the 2 x 2 binomial trial*** (pubblicato su **Journal of the Royal Statistical Society**, Series A., Vol. 148, pp.: 317-327),
- J. T. **Casagrande**, M. C **Pike** e P. G. **Smith** con l'articolo del 1978 ***An improved approximate formula for comparing two binomial distributions*** (pubblicato su **Biometrics** Vol. 34, pp.: 483-486).

Con l'ipotesi condizionale, la soluzione per stimare la probabilità è basata sulla distribuzione binomiale.

Sia la formula fondata sulla distribuzione ipergeometrica sia quella che utilizza la binomiale sono concettualmente semplici. Ma per il calcolo delle dimensioni del campione esse richiedono procedimenti estremamente lunghi, poiché la stima di due probabilità congiunte è estesa dalla risposta campionaria a tutte quelle più estreme nella stessa direzione. Inoltre, nel caso di test bilaterale, quando non si accetta che la distribuzione delle probabilità sia simmetrica e quindi non è ritenuto valido il procedimento di raddoppiare la probabilità già stimata, il calcolo deve essere esteso anche a tutte le possibili risposte ugualmente estreme che sono collocate nella direzione opposta.
Ne consegue che non esiste una formula semplice e rapida,

- né per il calcolo di n in casi di esperimenti bilanciati,
- né di n_1 quando sia stato prestabilito il rapporto k con le dimensioni n_2 dell'altro campione.

La dimensione minima n è riportata

a_1 - sia per i casi di due campioni bilanciati,

a_2 – sia per quelli di due campioni con un numero differente di dati.

Le formule per stimare la potenza $1-\beta$ sono 7 e anch'esse riguardano i casi

b_1 – di due campioni bilanciati e

b_2 – di due campioni con dimensioni differenti.

Per facilitare eventuali calcoli, nella tabella successiva sono riportati i valori di Z per le probabilità α e β che più frequentemente vengono utilizzati nella ricerca applicata:

ALCUNI VALORI DI Z

PER α **BILATERALE** (***two-tailed*** = $\alpha/2$) OPPURE **UNILATERALE** (***one-tailed*** = α)

E PER β **UNILATERALE (sempre)**

$\alpha/2$	0.10	0.05	0.025	0.01	0.005	0.001
α e β	0.20	0.10	0.05	0.02	0.01	0.002
Z	1,6448	1,9600	2,2414	2,5758	2,8070	3,2905

ELENCO DEI METODI

1 – Il **metodo dell'arcoseno** (***the arcsine methods***).

E' fondato sull'**approssimazione alla distribuzione normale** (Z) di due proporzioni p_1 e p_2, dopo la loro trasformazione in **arcoseno.** Come più ampiamente illustrato nei paragrafi sulle trasformazioni, ai quali si rimanda per approfondimenti, essa serve per **omogeneizzare la varianza** delle due proporzioni.

Il problema della **trasformazione angolare** di una proporzione p compare per la prima volta nel dibattito scientifico alla fine degli anni '30. Secondo alcune pubblicazioni sulle metodologie statistiche, il primo articolo che lo discute in modo abbastanza esauriente è quello di

- W. O. **Kermack** e A. G. **Mckendrick** del 1940 ***The design and interpretion of experiment based on a four–fold table: the statistical assessment of the effects of treatment*** (pubblicato su **Proceeding of the Royal Society of Edinburgh,** Vol. 60, pp.: 362-375).

Un'altra pubblicazione che tratta l'argomento in modo relativamente completo è quella di

- E. **Paulson** e W. A. **Wallis** del 1947 ***Planning and analyzing experiments for comparing two percentages*** (un capitolo del volume (ed.) di C. **Eisenhart**, M. W. **Hastay**, W. A. **Wallis** intitolato **Selected Techniques of Statistical Analysis**, McGraw-Hill, New York, Chapter 7, pp.: 247-265).

Il primo a dimostrare che la trasformazione di una proporzione p **nel suo arcoseno ha l'**effetto di ampliare i valori agli estremi e comprimere quelli centrarli**, determinando una statistica con varianza approssimativamente unitaria, è**

- C. **Eisenhart** nel 1947 con il capitolo ***Inverse sine transformation of proportion*** (sul medesimo volume appena citato di (ed.) C. **Eisenhart**, M. W. **Hastay**, W. A. **Wallis** dal titolo **Selected Techniques of Statistical Analysis** , McGraw-Hill, New York, Chapter 16, pp.: 395-416).

Tuttavia, da alcuni autori e in vari testi recenti, la **trasformazione angolare** è ritenuta **non necessaria,**

- quando le **due proporzioni sono comprese tra 0,30 e 0,70.**

La motivazione è che oggettivamente le **loro varianze differiscono ugualmente di poco**.

Per stimare le dimensioni dei campioni, le due formule sono:

a_1 - **Numero (** n **) di dati in due campioni bilanciati**:

- se il **test è bilaterale**

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{2 \cdot \left(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}\right)^2}$$

- se il **test è unilaterale**

$$n = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{2 \cdot \left(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}\right)^2}$$

dove

- con $\alpha/2$ e α in questa e in tutte le formule successive si intende rispettivamente la probabilità di un errore di Tipo I per un test bilaterale ($\alpha/2$) e un test unilaterale (α),
- mentre con β si intende la probabilità di un errore di Tipo II, che è sempre unilaterale.

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\left(\frac{4k}{k+1}\right) \cdot \left(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}\right)^2}$$

Invertendo la formula per il calcolo di n si ricava quella di $Z_{1-\beta}$.

Per passare dal valore di Z alla potenza $1-\beta$, è sufficiente ricorrere alla tabella della normale (sempre e solo unilaterale), detta **funzione di distribuzione cumulativa della distribuzione normale standard** (***the cumulative distribution function of the standard normal distribution***).

b_1 – **Potenza ($1-\beta$) del test in due campioni bilanciati**:

$$Z_{1-\beta} = \sqrt{2n} \cdot \left(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}\right) - Z_{1-\alpha/2}$$

b_2 – **Potenza ($1-\beta$) del test in due campioni con dimensioni differenti**:

$$Z_{1-\beta} = \frac{4kn_1}{k+1} \cdot \left(arcsin\sqrt{p_1} - arcsin\sqrt{p_2}\right) - Z_{1-\alpha/2}$$

2 – Il **metodo dell'arcoseno con la correzione per la continuità** (***the arcsine with continuity correction***).

Quando il campione è piccolo, secondo vari autori il test di significatività richiede la **correzione per la continuità**, che ne abbassa la significatività. Se da una parte diminuisce la potenza del test, dall'altra lo rende più prudenziale. Di conseguenza, se si vuole mantenere la stessa potenza del test, occorre aumentare le dimensioni del campione.

I metodi che utilizzano la correzione per la continuità richiedono un procedimento iterativo, poiché la quantità n che deve essere stimata compare anche al denominatore. Insieme con quello della trasformazione angolare, questo problema è discusso da D. E. **Walter** in un articolo del 1979 ***In defense of the arcsine approximation*** (pubblicato su **The Statistician** Vol. 28, pp.: 219-222).

La sua formula fornisce una stima di n molto vicina a quella ottenuta con il metodo esatto di Fisher.

a_1 - **Numero di dati (n) in due campioni bilanciati**:

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{2 \cdot \left(arcsin\sqrt{p_1 - \frac{1}{2n}} - arcsin\sqrt{p_2 - \frac{1}{2n}} \right)^2}$$

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\left(\frac{4k}{k+1}\right) \cdot \left(arcsin\sqrt{p_1 - \frac{1}{2n}} - arcsin\sqrt{p_2 - \frac{1}{2n}} \right)^2}$$

Invertendo la formula per calcolare n, si ricava quella per $Z_{1-\beta}$.

b_1 – **Potenza ($1-\beta$) del test in due campioni bilanciati**:

$$Z_{1-\beta} = \sqrt{2n} \cdot \left(arcsin\sqrt{p_1 - \frac{1}{2n}} - arcsin\sqrt{p_2 - \frac{1}{2n}} \right) - Z_{1-\alpha/2}$$

b_2 – **Potenza ($1-\beta$) del test in due campioni con dimensioni differenti**:

$$Z_{1-\beta} = \frac{4kn_1}{k+1} \cdot \left(arcsin\sqrt{p_1 - \frac{1}{2n_1}} - arcsin\sqrt{p_2 - \frac{1}{2kn_1}} \right) - Z_{1-\alpha/2}$$

3 – Il **metodo non iterativo dell'arcoseno con la correzione per la continuità** (***a non-iterative version of the continuity corrected arcsine***).

La formula precedente con la correzione per la continuità prevede che nella stima di n le proporzioni p_1 e p_2 siano diminuite di una quantità $1/2n$. Per giungere al risultato, l'inserimento di n al denominatore richiede una procedura iterativa:

- dopo aver introdotto un primo valore di n opzionale al denominatore, con la formula si calcola un secondo valore n';
- questo nuova stima n' sostituisce il precedente valore n al denominatore, ricavando una seconda stima n'';
- tale nuovo risultato n'' sostituisce il valore n' al denominatore.

Solitamente, al secondo o al terzo tentativo la stima ottenuta è molto vicina al valore introdotto al denominatore: è il risultato n del test.

Allo scopo di evitare tale iterazione, A. E. **Dobson** e V. J. **Gebski** nel 1986 con l'articolo ***Sample sizes for comparing two independent proportions using the continuity corrected arcsine transformations***, (pubblicato su **The Statistician** Vol. 35, pp.: 51-53) forniscono una eccellente approssimazione della formula precedente, sia nella versione per due campioni bilanciati sia per quella di due campioni con un numero differente di osservazioni:

a$_1$ - **Numero (n) di dati in due campioni bilanciati**:

$$n = \frac{\left(Z + \sqrt{Z^2 + 2C\Delta}\right)^2}{8\Delta^2}$$

dove

- $Z = Z_{1-\alpha/2} + Z_{1-\beta}$
- $\Delta = arcsin\sqrt{p_1} - arcsin\sqrt{p_2}$
- $C = \frac{1}{p_1 q_1} + \frac{1}{p_2 q_2}$ con $q_1 = 1 - p_1$ e $q_2 = 1 - p_2$

a$_2$ - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{\left(Z^* + \sqrt{Z^{*2} + 4C^*\Delta}\right)^2}{16\Delta^2}$$

dove

- $Z^* = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right) \cdot \left(1 + \frac{1}{k}\right)$
- $\Delta = arcsin\sqrt{p_1} - arcsin\sqrt{p_2}$
- $C^* = \frac{1}{p_1 q_1} + \frac{k}{p_2 q_2}$ con $q_1 = 1 - p_1$ e $q_2 = 1 - p_2$

4 – Il **metodo di Poisson** (***the Poisson method***).

La distribuzione poissoniana, che può essere derivata dalla distribuzione binomiale (vedi capitolo II sulle distribuzioni teoriche) assumendo che

- $n \Rightarrow \infty$ e $p \Rightarrow 0$

è concettualmente la base chi-quadrato.

Secondo quanto affermato da M. **Gail** nel 1974, con l'articolo ***Power computations for designing comparative Poisson trials*** (su **Biometrics** Vol. 30, pp.: 231-237), questo metodo è raccomandato

- quando non è applicabile l'approssimazione alla distribuzione normale, perché le proporzioni sono vicine ai valori limite (0 oppure 1).

Il metodo fondato sulla distribuzione di Poisson dovrebbe essere utilizzato quando

- le **due proporzioni** p_1 **e** p_2 **molto piccole** (inferiori a 0.05),

- ma con n **abbastanza grande**, in modo che np_1 e np_2 siano entrambe ≥ 10;

oppure, simmetricamente, quando

- le **due proporzioni** p_1 **e** p_2 **molto grandi** (maggiori di 0.95)

- ma sempre con n **abbastanza grande** in modo che, con formula complementare alla precedente, i valori di nq_1 e nq_2 siano entrambi ≥ 10.

a$_1$ – **Numero (** n **) di dati in due campioni bilanciati**

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \cdot \left(p_1 + p_2\right)}{\delta^2}$$

dove

- δ = differenza minima tra due proporzioni che si vuole dimostrare significativa.

a$_2$ - **Numero (** n_1 **con** $n_2 = k \cdot n_1$ **) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \cdot \left(p_1 + \frac{p_2}{k}\right)}{\delta^2}$$

Invertendo la formula per calcolare n, si ricava per stimare $Z_{1-\beta}$.

b_1 – **Potenza ($1-\beta$) del test in due campioni bilanciati**:

$$Z_{1-\beta} = \frac{n\delta^2}{\sqrt{p_1 + p_2}} - Z_{1-\alpha/2}$$

b_2 – **Potenza ($1-\beta$) del test in due campioni con dimensioni differenti**:

$$Z_{1-\beta} = \frac{n_1\delta^2}{\sqrt{p_1 + \frac{p_2}{k}}} - Z_{1-\alpha/2}$$

5 – Il **metodo normale asintotico** (***Asymptotic normal method***).

E' il metodo più appropriato quando si utilizzano due proporzioni senza la trasformazione in arcoseno, poiché si assume che siano distribuite in modo asintoticamente normale. La formula tiene in considerazione che

- le due proporzioni p_1 e p_2 hanno varianze differenti ($\sigma_1^2 = p_1q_1$ e $\sigma_2^2 = p_2q_2$),

come discusso

- nell'articolo di M. **Halperin**, E. **Rogot**, J. **Gurian** e F. **Ederer** nel 1968 ***Sample size for medical trials with special reference to long term therapy*** (pubblicato su **Journal of Chronic Diseases** Vol. 21, pp.: 13-24)
- nel testo di P. **Armitage** e G. **Berry** del 1987 ***Statistical Methods in Medical Research*** (2nd ed. Blackwell Scientific Publications, Oxford),
- nel volume di J. L. **Fleiss** del 1981 ***Statistical Methods for Rates and Proportions*** (2nd ed. Wiley, New York).

Da questo approccio, deriva la formula più diffusa nei testi divulgativi, tra cui il volume di B. **Rosner** del 1994 ***Fundamentals of Biostatistics*** (4th ed. Duxbury Press, Belmont, California).

a_1 - **Numero (n) di dati in due campioni bilanciati**:

$$n = \frac{\left(Z_{1-\alpha/2} \cdot \sqrt{2\overline{pq}} + Z_{1-\beta} \cdot \sqrt{p_1q_1 + p_2q_2}\right)^2}{\delta^2}$$

dove

- $\overline{p} = \frac{p_1 + p_2}{2}$ e $\overline{q} = 1 - \overline{p}$

- $q_1 = 1 - p_1$ e $q_2 = 1 - p_2$

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{\left(Z_{1-\alpha/2} \cdot \sqrt{2\overline{pq} \cdot \left(1 + \frac{1}{k}\right)} + Z_{1-\beta} \cdot \sqrt{p_1 q_1 + \frac{p_2 q_2}{k}} \right)^2}{\delta^2}$$

dove

- $\overline{p} = \frac{p_1 + k p_2}{k + 1}$ e $\overline{q} = 1 - \overline{p}$

Invertendo la formula per il calcolo di n, si ricava quella di $Z_{1-\beta}$.

b_1 – **Potenza ($1 - \beta$) del test in due campioni bilanciati**:

$$Z_{1-\beta} = \frac{\sqrt{n\delta^2} - Z_{1-\alpha/2} \cdot \sqrt{2\overline{pq}}}{\sqrt{p_1 q_1 + p_2 q_2}}$$

b_2 – **Potenza ($1 - \beta$) del test in due campioni con dimensioni differenti**:

$$Z_{1-\beta} = \frac{\sqrt{n_1\delta^2} - Z_{1-\alpha/2} \cdot \sqrt{\left(1 + \frac{1}{k}\right)\overline{pq}}}{\sqrt{p_1 q_1 + \frac{p_2 q_2}{k}}}$$

6 – Il **metodo della normale con la correzione per la continuità** (***Normal with continuity correction***).

La formula precedente è equivalente a quella basata sul χ^2 di **Pearson**, senza la correzione per la continuità. Quando è **inserita tale correzione**, detta anche correzione di **Yates**, nel 1959 M. **Kramer** e S. W. **Greenhouse** con la pubblicazione ***Determination of sample sizes and selection of cases*** (nel volume di J. O. **Cole** e R. W. **Gerard** (eds.) **Psychopharmacology: Problems in Evaluations**,

National Academy of Science, National Research Council, Washington, D. C. pp.: 356-371) hanno proposto di stimare il numero di dati con:

a_1 - **Numero (n) di dati in due campioni bilanciati**:

$$n = \frac{n'}{4} \cdot \left(1 + \sqrt{1 + \frac{8}{n'\delta}} \right)^2$$

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**:

Nel 1980 J. L. **Fleiss**, A. **Tytun** e S. H. K. **Ury** con l'articolo ***A simple approximantion for calculing sample sizes for comparing two independent proportions*** (su **Biometrics** Vol. 36, pp.: 343-346) propongono anche la sua estensione a due campioni non bilanciati:

$$n_1 = \frac{n'_1}{4} \cdot \left(1 + \sqrt{1 + \frac{4 \cdot (k+1)}{k n'_1 \delta}} \right)^2$$

dove

- n'_1 è ottenuta con la formula per due campioni non bilanciati del metodo fondato sulla distribuzione asintoticamente normale (precedente formula $5a_2$).

7 – Il **metodo della normale modificato con la correzione per la continuità** (***Modified normal with continuity correction***).

Un miglioramento della formula precedente, per ottenere una stima ancora **più vicina a quella fornita dal metodo esatto di Fisher**, è stata proposta successivamente da J. T. **Casagrande**, M. C. **Pike** e P. G. **Smith** nel 1978 con l'articolo ***An improved approximate formula for comparing two binomial distributions*** (su **Biometrics** Vol. 34, pp.: 483-486) ed è stata ripresa dal testo a grande diffusione ***Biostatistical Analysis*** di J. H. **Zar** già nell'edizione del 1984 (2nd edn. Prentice-Hall, Inc., Englewood Cliffs, New Jersey):

a_1 - **Numero (n) di dati in due campioni bilanciati**:

$$n = \frac{n'}{4} \cdot \left(1 + \sqrt{1 + \frac{4}{n'\delta}} \right)^2$$

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**.

La formula per due campioni bilanciati è stata estesa al caso di due campioni non bilanciati

- nel 1981 da C. **Diegert** e K. V. **Diegert** con l'articolo ***Note on inversion of Casagrande-Pike-Smith approximate sample size formula for Fisher-Irwin test on 2 x 2 tables*** (su **Biometrics** Vol. 37, p.:595) e
- nel 1982 da J. L. **Fleiss**, A. **Tytun** e S. H. K. **Ury** con l'articolo ***Response to "The choice of relative group sizes for comparisons of independent proportions"*** (su **Biometrics** Vol. 38, pp.: 1093-1094):

$$n_1 = \frac{n'_1}{4} \cdot \left(1 + \sqrt{1 + \frac{2 \cdot (k+1)}{k n'_1 \delta}} \right)^2$$

dove

- n'_1 è ottenuta con la formula per due campioni non bilanciati del metodo fondato sulla distribuzione asintoticamente normale (precedente formula $5a_2$).

8 – **La formula abbreviata del metodo normale con la correzione per la continuità** (***Shortcut formula of modified normal with continuity correction***).

Sempre nell'articolo di J. L. **Fleiss**, A. **Tytun** e S. H. K. **Ury** del 1980 ***A simple approximation for calculating sample sizes for comparing two independent proportions*** (su **Biometrics** Vol. 36, pp.: 343-346) è dimostrato che un notevole grado di accuratezza può essere ottenuta anche con la formula molto semplice e rapida:

a_1 - **Numero (n) di dati in due campioni bilanciati**:

$$n = n' + \frac{2}{\delta}$$

dove

- n' corrisponde al valore n ottenuta con la formula del **metodo asintotico normale** per due campioni bilanciati (precedente formula $5a_1$).

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**.

Nello stesso articolo appena citato di J. L. **Fleiss**, A. **Tytun** e S. H. K. **Ury** del 1980 è presentata anche la formula per due campioni non bilanciati:

$$n_1 = n'_1 + \frac{k+1}{k\delta}$$

dove

- n'_1 corrisponde al valore n_1 ottenuta con la formula del **metodo asintotico normale** per due campioni non bilanciati (precedente formula $5a_2$).

Invertendo la formula per il calcolo di n, si ricava quella di $Z_{1-\beta}$.

b_1 – **Potenza ($1-\beta$) del test in due campioni bilanciati**:

$$Z_{1-\beta} = \frac{\sqrt{n\delta^2 - 2\delta} - Z_{1-\alpha/2} \cdot \sqrt{2\overline{pq}}}{\sqrt{p_1 q_1 + p_2 q_2}}$$

b_2 – **Potenza ($1-\beta$) del test in due campioni con dimensioni differenti**:

$$Z_{1-\beta} = \frac{\sqrt{n_1\delta^2 - \left(1+\frac{1}{k}\right)\delta} - Z_{1-\alpha/2} \cdot \sqrt{\left(1+\frac{1}{k}\right) \cdot \overline{pq}}}{\sqrt{p_1 q_1 + \frac{p_2 q_2}{k}}}$$

9 – Il **metodo del chi–quadrato con la correzione per la continuità** (***Chi-square with continuity correction***).

Un'altra formula approssimata e che permette un calcolo rapido è quella riportata da H. K. **Ury** e J. L. **Fleiss** nel 1980 nell'articolo ***On approximate sample sizes for comparing two independent proportions with use of Yates' corrections*** (su **Biometrics** Vol. 36, pp.: 347-251) per l'uso del χ^2 con la **correzione di Yates**:

a_1 - **Numero (n) di dati in due campioni bilanciati**:

$$n = \frac{n'}{4} \cdot \left(1 + \sqrt{\frac{2\delta}{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \cdot \overline{pq}}}\right)^2$$

dove

- $\overline{p} = \frac{p_1 + p_2}{2}$ e $\overline{q} = 1 - \overline{p}$

- $n_1^{'}$ è ottenuto con il metodo asintotico normale della formula $5a_1$.

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{n'_1}{4} \cdot \left(1 + \sqrt{\frac{(k+1) \cdot \delta}{k \cdot \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \cdot \overline{pq}}}\right)^2$$

dove

- $n_1^{'}$ è ottenuto con il metodo asintotico normale della formula $5a_2$.

10 – **Il metodo normale con l'ipotesi di omogeneità** (***Simple normal assuming homogeneity***).

Questa formula per stimare la dimensione n è un adattamento alle proporzioni della formula classica riportata da W. G. **Cochran** e G. M. **Cox** nel loro testo del 1957 ***Experimental Design*** (2nd eds. Wiley, New York), per la stima della dimensione di due campioni in una ANOVA a un criterio, quando si assume che le varianze siano uguali.

a_1 - **Numero (n) di dati in due campioni bilanciati**:

$$n = \frac{2\overline{pq} \cdot \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\delta^2}$$

a_2 - **Numero (n_1 con $n_2 = k \cdot n_1$) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{\left(1 + \frac{1}{k}\right) \cdot \overline{pq} \cdot \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\delta^2}$$

dove

- $\overline{p} = \frac{p_1 + kp_2}{k+1}$ e $\overline{q} = 1 - \overline{p}$

Invertendo la formula per calcolare n, si ricava quella per stimare $Z_{1-\beta}$.

b_1 – **Potenza ($1-\beta$) del test in due campioni bilanciati**:

$$Z_{1-\beta} = \sqrt{\frac{n\delta^2}{2\overline{pq}}} - Z_{1-\alpha/2}$$

b_2 – **Potenza ($1-\beta$) del test in due campioni con dimensioni differenti**:

$$Z_{1-\beta} = \sqrt{\frac{n_1\delta^2}{\left(1+\frac{1}{k}\right)\cdot\overline{pq}}} - Z_{1-\alpha/2}$$

11 – **Il metodo normale con ipotesi di eterogeneità** (***Simple normal assuming heterogeneity***):
Quando, con il modello parametrico fondato sulla normale, si assume che le varianze siano differenti, la formula precedente è trasformata in quella successiva. E' riportata anche nel testo classico di G. W. **Snedecor** e W. G. **Cochran** del 1989 ***Statistical Methods*** (8th edn. Iowa State University Press, Ames, Iowa) e nel volume di D. **Machin** e M. J. **Campbell** del 1987 ***Statistical Tables for the Design of Clinical Trials*** (Blackwell Scientific Publications, Oxford).

a_1 - **Numero (n) di dati in due campioni bilanciati**:

$$n = \frac{(p_1q_1 + p_2q_2)\cdot\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\delta^2}$$

a_2 - **Numero (n_1 con $n_2 = k\cdot n_1$) di dati in due campioni con dimensioni differenti**:

$$n_1 = \frac{\left(p_1q_1 + \frac{p_2q_2}{k}\right)\cdot\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\delta^2}$$

Invertendo la formula per calcolare n, si ricava quella per stimare $Z_{1-\beta}$.

b_1 – **Potenza ($1-\beta$) del test in due campioni bilanciati**:

$$Z_{1-\beta} = \sqrt{\frac{n\delta^2}{p_1q_1 + p_2q_2}} - Z_{1-\alpha/2}$$

b_2 – **Potenza ($1-\beta$) del test in due campioni con dimensioni differenti**:

$$Z_{1-\beta} = \sqrt{\frac{n_1\delta^2}{p_1q_1 + \frac{p_2q_2}{k}}} - Z_{1-\alpha/2}$$

A conclusione di questo elenco di formule, è conveniente ricordare che Hardeo **Sahai** e Anwer **Khurshid,** sempre nell'articolo del 1996 citato all'inizio, scrivono che i **test tradizionali** e **ricorrenti** nelle riviste di statistica applicata per **confrontare due proporzioni** sono il **metodo esatto di Fisher** e il **chi quadrato con la correzione per la continuità di Yates**. Ma, contrapposta a questa utilizzazione massiva, nella letteratura specialistica esiste una rilevante controversia sulla loro correttezza.

Molti ricercatori hanno dimostrato che

- il chi quadrato tradizionale di Pearson, **quello senza la correzione per la continuità,** fornisce una **difesa più che adeguata contro l'errore di Tipo I (errore α)**,
- mentre il **test esatto di Fisher** e **il chi quadrato con la correzione di Yates sono sistematicamente troppo conservativi.**

Ne deriva che questi due metodi sono troppo **poco potent**i (in inglese scritto anche poco ***liberal***), presentando un **errore sistematico che li rende troppo conservativi**: Essi… ***have an extremely conservative bias. This implies that the Fisher's exact test and the Pearson's chi-square with continuity correction are less powerful, and so have less chance of detecting a given difference in proportions than the chi-square test without the continuity correction*** (pag. 17).

La letteratura a favore di questa affermazione è numerosa. Tra gli articoli degli autori più importanti e che già nel titolo evidenziano l'approccio critico, è possibile ricordare

- di W. G. **Conover** del 1974 ***Some reasons for not using the Yates' continuity correction on 2 x 2 contigency tables (with comments and rejoinder)*** (pubblicato su **Journal of the American Statistical Association** Vol. 69, pp.: 374 – 384);

- di J. **Berkson** del 1978 ***In dispraise of the exact test: do the marginal totals of the 2 x 2 table contain relevant information respecting the table proportions*** (su **Journal od Statistical Planning and Inference** Vol. 2 pp.: 27 – 42);
- di R. B. **D'Agostino**, W. **Chase** e A. **Belanger** del 1988 ***The appropriateness of some common procedures for testing the equality of binomial parameters*** (in **The American Statistician** Vol. 42, pp.: 198 – 202).

Un altro aspetto del test esatto di Fisher che ha suscitato varie obiezioni è assumere l'ipotesi che tutti **i totali marginali si mantengano sempre fissi**. E' un concetto che è rifiuatato da chi ritiene che
- in un esperimento **il numero totale di successi non può essere prefissato**,
- ma che esso sia **una variabile random**.

Quindi il test chi-quadrato senza la correzione per la continuità sarebbe più appropriato del test esatto di Fisher, anche dal punto di vista logico. Tuttavia **mantenere costanti i totali** è un'ipotesi operativamente **utile per derivare un test esatto non parametrico**, che non sia fondato su distribuzioni asintotiche.

Inoltre è dimostrato che la **correzione di Yates porta a ricavare gli stessi valori del test esatto**.

E' quanto afferma lo stesso Frank **Yates** (1902- 1994, già assistente di Fisher nel 1931 nell'Istituto di ricerche agrarie Rothamsted di Londra)
- esattamente cinquant'anni dopo la sua proposta originaria del 1934 ***Contingency tables involving small numbers and the*** χ^2 ***test*** (su **Journal of the Royal Statistical Society** (Suppl. 1, pp.: 217 – 235),
- con l'articolo del 1984 ***Tests of significance for 2 x 2 contingency tables (with discussion)***, pubblicato sulla stessa rivista **Journal of the Royal Statistical Society** (Series A, Vol. 147, pp.: 426 – 463).

Il **dibattito** a favore o contro queste due formule **continua**.

La sua conclusione porta anche alla risposta su **quale sia la formula migliore**

La scelta del test più appropriato non è un esercizio meramente accademico. Ha una grande importanza pratica, poiché **le dimensioni del campione sono differenti** se è impiegato
- **il** χ^2 **con** oppure **senza la correzione per la continuità,**
- la **trasformazione in arcoseno** oppure la **distribuzione normale asintotica**,

ricordando sempre che **deve essere utilizzato il test per il quale è stata pianificata la raccolta dei dati**.