UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN
STATISTICA E GESTIONE DELLE IMPRESE

TESI DI LAUREA

Intervalli di confidenza per la stima di una probabilità

RELATORE: PROF. Alessandra Salvan

LAUREANDO: Gianluca Barbierato

ANNO ACCADEMICO 2009/2010

Indice

Introduzione

Nonostante si tratti di una questione elementare di inferenza statistica, le soluzioni più utilizzate per il problema del calcolo di un intervallo di confidenza per la probabilità di successo in una distribuzione binomiale hanno spesso gravi carenze, come evidenziato dalla letteratura recente.
Questa tesi presenterà principalmente i risultati dell'articolo di Lawrence D. Brown, T. Tony Cai, Anirban DasGupta pubblicato su *The Annals of Statistics* del febbraio 2002. Ciò vuol dire che, ove non espressamente scritto altrimenti, i risultati presentati sono da attirbuire all'articolo citato.
Si mostrerà che l'intervallo standard, comunemente detto *alla Wald*, usato solitamente, è molto poco efficiente, in quanto presenta probabilità di copertura costantemente distorte verso il basso rispetto a quelle nominali, anche per numerosità campionarie alte. In particolare, nel lavoro di tesi sono stati sviluppati i calcoli dei momenti di primo e secondo ordine della statistica test da cui deriva l'intervallo. Per questo motivo verranno presentati altri quattro intervalli di cui si studieranno le rispettive probabilità di copertura. Inoltre, gli intervalli verranno anche comparati in termini di lunghezza attesa.
Nel capitolo 1 si mostrerà con due esempi che l'intervallo standard non può essere ritenuto soddisfacente. Inoltre verranno fatte alcune considerazioni sui perchè dell'inefficienza a cui vengono associati i calcoli dei momenti della statistica di Wald. Verranno poi presentate quattro alternative: l'intervallo di Wilson, uno degli intervalli proposti da Agresti e Coull nel 1998, l'intervallo basato sul rapporto di verosimiglianza e l'intervallo bayesiano di Jeffreys.
Nel capitolo 2 si useranno gli sviluppi di Edgeworth per approssimare le

probabilità di copertura degli intervalli presi in considerazione e si confronteranno i risultati.

Nel terzo capitolo infine, sono riportati i risultati relativi al calcolo delle lunghezze attese di questi intervalli. Ciò permetterà di fare altre interessanti considerazioni che contribuiranno alla scelta di un intervallo appropriato.

Capitolo 1

Inefficienza dell'intervallo standard e alternative

1.1 Introduzione

Questo capitolo presenterà le debolezze che caratterizzano l'intervallo standard nel caso in cui si voglia fare inferenza sulla probabilità di successo p di una distribuzione binomiale $Bi(n, p)$, con n numero di prove effettuate. Definito X come il numero di successi ottenuti e $\hat{p} = X/n$, stimatore di massima verosimiglianza per p, l'intervallo in questione si esplicita come $\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, con $z_{1-\frac{\alpha}{2}}$ quantile di livello $1 - \frac{\alpha}{2}$ della distribuzione normale standard (d'ora in poi esso verrà indicato semplicemente con z). Innanzi tutto verranno mostrati alcuni esempi pratici che riveleranno che le debolezze di cui si sta trattando non sono trascurabili nemmeno per valori di p lontani da 0 e 1, né per campioni relativamente grandi. Si noterà che le probabilità di copertura di questo intervallo sono soggette a distorsioni sistematicamente negative e a oscillazioni significative. Nel prosieguo del capitolo si cercherà di spiegare, almeno intuitivamente, il perché di tutto ciò anche con il calcolo di media e varianza della statistica di Wald. Infine, verranno presentate le alternative all'intervallo standard che si prenderanno in considerazione.

1.2 Due esempi significativi

Di seguito verrà mostrato graficamente come l'intervallo standard non possa essere considerato soddisfacente per stimare probabilità di successo. Si consideri $p = 0.5$. La Figura 1 mostra la vera probabilità di copertura (calcolata cioè in modo esatto) dell'intervallo standard con livello di confidenza 0.95 per $n = 10, \ldots, 100$. La prima cosa che si nota è che la probabilità di copertu-

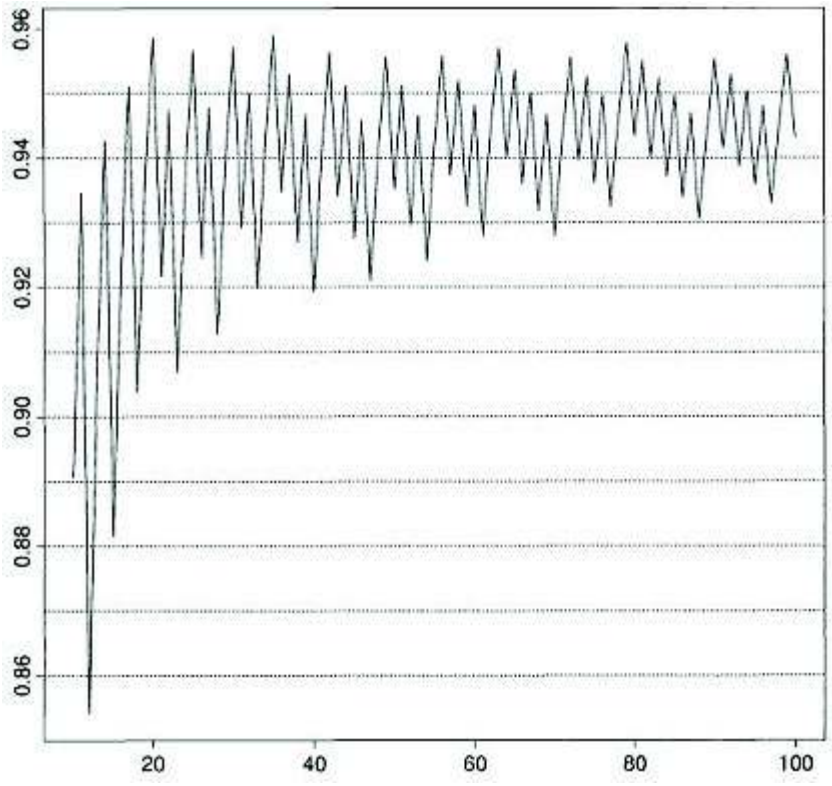

Figura 1: probabilità di copertura dell'intervallo standard al variare di n ($\alpha = 0.05$, $p = 0.5$).

ra non aumenta direttamente con l'aumentare di n ma è soggetta a numerose oscillazioni. Quella associata a $n = 17$, ad esempio, è 0.951, mentre, se si considera $n = 40$, si ottiene una probabilità molto più bassa (0.919). Inoltre, si può facilmente notare anche che, in media, la probabilità è decisamente minore di 0.95, anche considerando i soli valori di $n > 50$.
Nel secondo esempio, è stato fissato $n = 30$ ed è stato tracciato il grafico della

probabilità di copertura dell'intervallo con livello di confidenza 0.99. Come

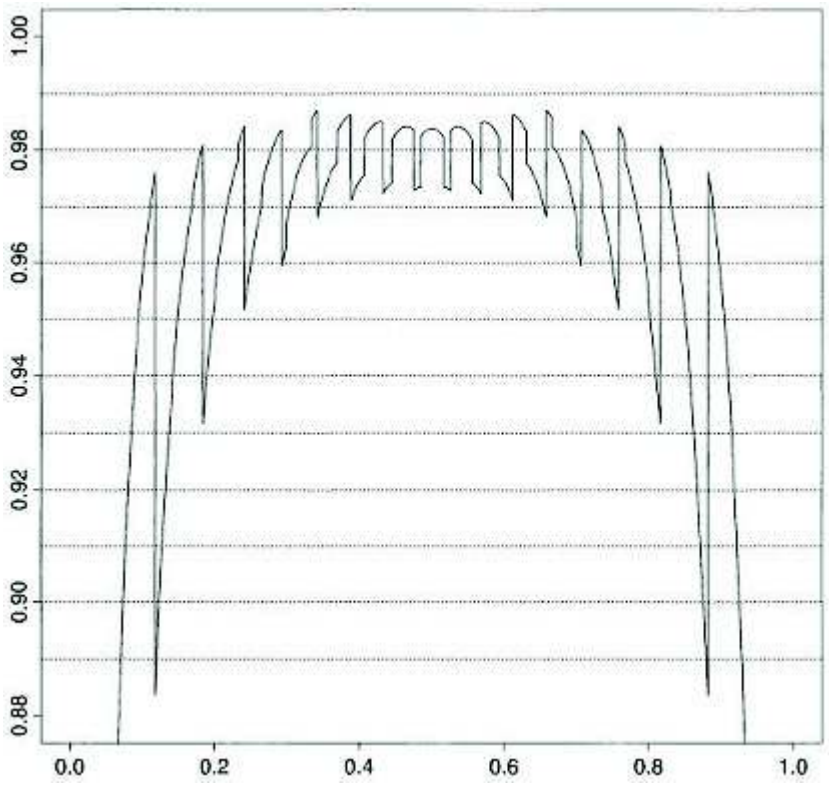

Figura 2: probabilità di copertura dell'intervallo standard al variare di p ($\alpha = 0.01$, $n = 30$).

si nota facilmente, il grafico è costantemente sotto la linea che rappresenta $p = 0.99$ (il valore medio è di 0.914). Inoltre si notano anche qui oscillazioni che rendono alcuni valori di p "più fortunati" di altri.

1.3 Ragioni delle distorsioni

Una delle ragioni principali delle sistematiche distorsioni negative, è l'errata scelta del centro dell'intervallo di confidenza. Nonostante $\hat{p}$ sia lo stimatore di massima verosimiglianza e sia non distorto, esso causa le distorsioni di cui si sta trattando se scelto come centro dell'intervallo. Come si potrà intuire dai risultati successivi, se al posto di $\hat{p}$ viene scelto $\tilde{p} = (X + z^2/2)/(n + z^2)$,

dove z rappresenta il quantile di livello $1 - \alpha/2$ della variabile normale standardizzata, si ottiene una probabilità di copertura sensibilmente più vicina al valore nominale, soprattutto per valori di p lontani da 0 e 1.

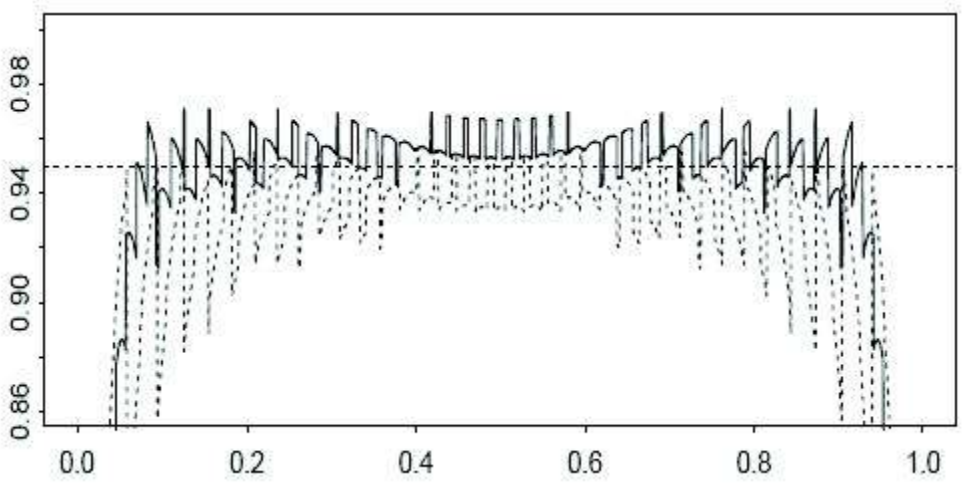

Figura 3: probabilità di copertura dell'intervallo standard confrontata con l'intervallo ricentrato (tratto continuo) al variare di p con $n = 50$ e $\alpha = 0.05$.

La debolezza di un altro importante assunto dell'intervallo di Wald contribuisce a spiegare le distorsioni. Infatti, si ha che

$$W_n = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

è significativamente non normale standard per valori di n anche abbastanza grandi. Si può, ad esempio, valutare analiticamente il valore atteso di W_n. Definita $Z_n = n^{1/2}(\hat{p} - p)/\sqrt{pq}$, si ha che

$$W_n = \frac{Z_n}{\sqrt{1 + (1-2p)Z_n/\sqrt{npq} - Z_n^2/n}}.$$

Per ottenere un'approssimazione per i momenti di W_n, si usa l'espansione di Taylor fino al terzo ordine del denominatore, ottenendo

$$W_n = \lambda(Z_n) = Z_n(1 - \frac{(1-2p)Z_n}{2\sqrt{npq}} + \frac{Z_n^2}{2n} + \frac{3(1-2p)^2Z_n^2}{8npq} - \frac{3(1-2p)Z_n^3}{4n\sqrt{npq}} - \frac{15(1-2p)^3Z_n^3}{48npq\sqrt{npq}}) + o(n^{-3/2}).$$

Da cui, usando i momenti di Z_n

$$\begin{aligned} E(Z_n) &= 0, \\ E(Z_n^2) &= 1, \\ E(Z_n^3) &= \frac{1-2p}{\sqrt{npq}}, \\ E(Z_n^4) &= \frac{1-6pq}{npq} + 3, \end{aligned}$$

si può facilmente ricavare un approssimazione del valore atteso di W_n:

$$\begin{aligned} E(W_n) = E(\lambda(Z_n)) =& E(Z_n - \frac{(1-2p)Z_n^2}{2\sqrt{npq}} + \frac{Z_n^3}{2n} + \frac{3(1-2p)^2Z_n^3}{8npq} - \frac{3(1-2p)Z_n^4}{4n\sqrt{npq}} \\ & - \frac{15(1-2p)^3Z_n^4}{48npq\sqrt{npq}}) + o(n^{-3/2})) \\ =& \frac{p-1/2}{\sqrt{npq}} - \frac{p-1/2}{n\sqrt{npq}} - \frac{6(p-1/2)^3}{2npq\sqrt{npq}} + \frac{9(p-1/2)}{2n\sqrt{npq}} \\ & + \frac{15(p-1/2)^3}{2npq\sqrt{npq}} + o(n^{-3/2}) \\ =& \frac{p-1/2}{\sqrt{npq}}\left(1 + \frac{7}{2n} + \frac{9(p-1/2)^2}{2npq}\right) + o(n^{-3/2}). \end{aligned}$$

E' chiaro quindi che si hanno distorsioni negative se $p < 0.5$ e distorsioni positive $p > 0.5$. Inoltre ci si può attendere che, spostando il centro dell'intervallo verso $1/2$, la probabilità di copertura aumenti come effettivamente mostrato in Figura 3.

Dai grafici (a) e (b) in Figura 4 si può notare come le distorsioni di $E(W_n)$ siano significative sia per valori di n relativamente grandi, sia per valori di p non lontani da $\frac{1}{2}$.

A questo punto, il calcolo dei momenti di ordine superiore è una questione puramente tecnica. Mantenendo lo stesso livello di approssimazione, ad esempio, la varianza della statistica di Wald è

$$\begin{aligned}
Var(W_n) =& E([\lambda(Z_n) - E(\lambda(Z_n))]^2) \\
=& E(Z_n^2 + \frac{(p-1/2)^2}{npq}(Z_n^2+1)^2 + \frac{2(p-1/2)}{\sqrt{npq}}(Z_n^3 - Z_n) + \frac{Z_n^4}{4n} \\
&+ \frac{3(p-1/2)^2}{npq}Z_n^4 + o(n^{-3/2})) \\
=& 1 + \frac{2(p-1/2)^2}{npq} - \frac{4(p-1/2)^2}{npq} + \frac{3}{4n} + \frac{9(p-1/2)^2}{4npq} + o(n^{-3/2}) \\
=& 1 + \frac{3}{4n} + \frac{(p-1/2)^2}{4npq} + o(n^{-3/2})
\end{aligned}$$

Anche in questo caso, si nota che l'errore di distorsione è maggiore quando il vero valore del parametro si allontana da 0.5.
In Figura 4, il grafico (c) mostra che la varianza di W_n si mantiene significativamente sopra il valore 1 anche per n elevato. Mentre in (d) si nota che anche con p vicino a 0.5 si ha una variabilità superiore a quella desiderata, il che dovrebbe sconsigliare di usare l'intervallo standard anche nei casi in cui il parametro è distante dai valori di frontiera.

1.4 Ragioni delle oscillazioni

La ragione principale delle oscillazioni è la struttura della distribuzione binomiale. La funzione di ripartizione presenta punti di discontinuità in corrispondenza degli interi appartenenti allo spazio campionario e questo causa delle oscillazioni dovute alla lunghezza dell'intervallo.
Infatti, se si risolvono in X le disequazioni $-z \leq \frac{\left(\frac{X}{n}-p\right)\sqrt{n}}{\sqrt{\hat{p}\hat{q}}} \leq z$, si individuano i valori di X che fanno accettare l'ipotesi nulla secondo l'intervallo standard. Evidenziando la dipendenza da p e z si ottiene

$$np - z\sqrt{\frac{X}{n}\frac{(n-X)}{n}} \leq X \leq np + z\sqrt{\frac{X}{n}\frac{(n-X)}{n}},$$

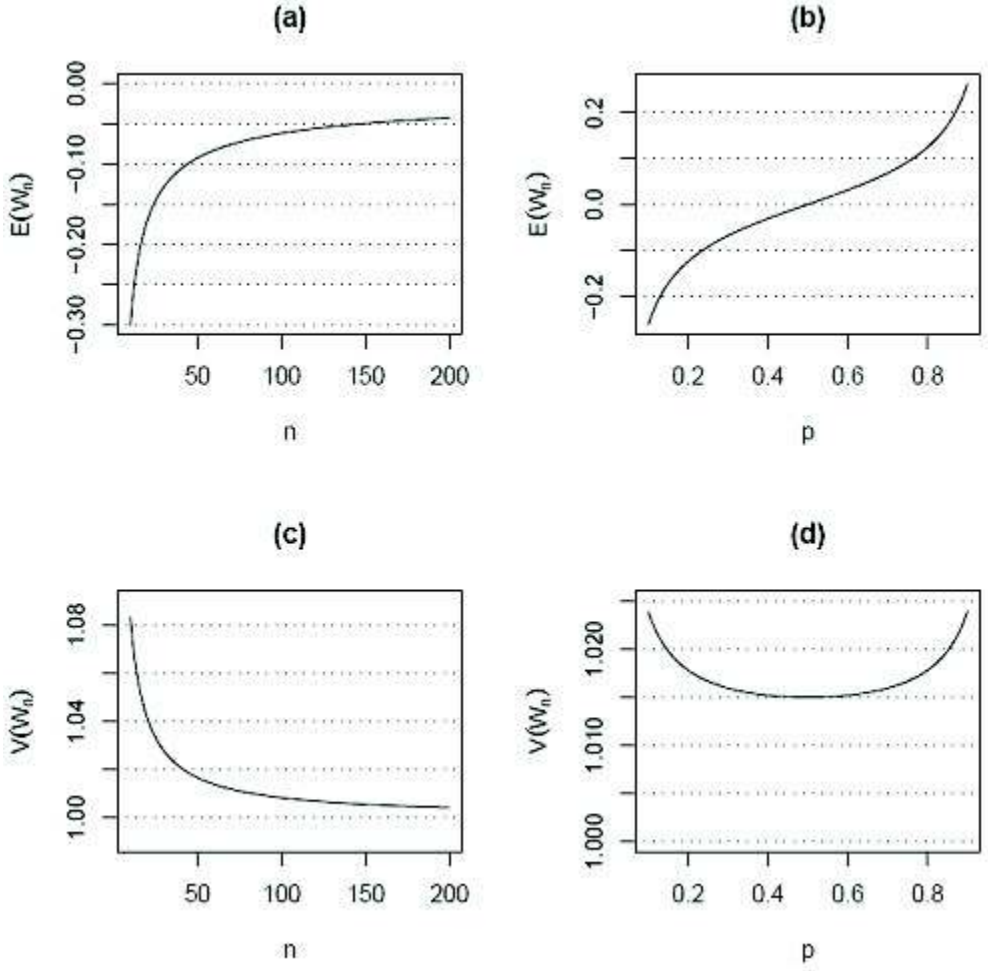

Figura 4: Media e varianza di W al variare di n, fissato $p = 0.25$ e di p con $n = 50$.

da cui elevando al quadrato e risolvendo separatamente si ottengono gli estremi

$$l_{n,p} = \frac{n(z^2 + 2np) - zn\sqrt{z^2 + 4np(1-p)}}{2(z^2+n)}$$

e

$$u_{n,p} = \frac{n(z^2 + 2np) + zn\sqrt{z^2 + 4np(1-p)}}{2(z^2+n)},$$

Quindi, la probabilità di copertura dell'intervallo standard, definita come $P_{n,p}(p \in CI_s)$, è uguale a $P_{n,p}(\lceil l_{n,p} \rceil \leq X \leq \lfloor u_{n.p} \rfloor)$ dove le funzioni $\lfloor k \rfloor$ e $\lceil k \rceil$ restituiscono rispettivamente la parte intera di k e la parte intera di k più uno. Si nota abbastanza facilmente che anche dei piccoli cambiamenti dei

valori di n o di p possono far cambiare valore a uno degli estremi, e quindi far cambiare sensibilmente la probabilità di copertura. Ad esempio, se si fissa $p = 0.5$, si ha che l'intervallo per $n = 39$ è $(14, 25)$, mentre invece, se solo si sceglie $n = 40$, si ottiene l'intervallo $(15, 25)$ e quindi un decremento della probabilità di copertura.

1.5 Alternative all'intervallo standard

Prima di presentare i quattro intervalli altenativi che verranno studiati nei capitoli successivi, è bene chiarire le notazioni utilizzate. Come già in precendenza, si indica con z il quantile della normale di livello $1 - \alpha/2$ (tutti gli intervalli considerati sono bilaterali). Inoltre, $\tilde{X} = X + z^2/2$, $\tilde{n} = n + z^2$ e $\tilde{p} = \tilde{X}/\tilde{n}$. Per praticità infine, si definisce $\hat{q} = 1 - \hat{p}$ e, analogamente, $\tilde{q}$.

- **Intervallo di Wilson**
  L'intervallo di Wilson usa come centro il valore $\tilde{p}$ ma differisce dall'intervallo standard ricentrato in quanto usa come errore standard $\sqrt{\frac{pq}{n}}$ al posto del valore stimato $\sqrt{\frac{\hat{p}\hat{q}}{n}}$. Questo intervallo si ricava invertendo l'approssimazione ottenuta dal Teorema Centrale del Limite applicato alla famiglia dei test bilaterali di $H_0 : p = \hat{p}$. L'intervallo ha pertanto la forma
  $$CI_S = \tilde{p} \pm \frac{z\sqrt{n}}{n + z^2} \sqrt{\left(\hat{p}\hat{q} + \frac{z^2}{4n} \right)}.$$

- **Intervallo di Agresti-Coull**
  Questo intervallo ha la stessa forma dell'intervallo standard, ma usa diversi valori per n, p, q.
  $$CI_{AC} = \tilde{p} \pm z\sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}.$$
  Se, al livello di confidenza $\alpha = 0.05$, si approssima il quantile della normale a 2, questo intervallo riconduce all'intervallo standard ottenuto aggiungendo due successi e due insuccessi.

- **Intervallo basato sul rapporto di verosimiglianza**
  Per ottenere questo intervallo, bisogna invertire il test basato sul rapporto di verosimiglianza che accetta l'ipotesi nulla $H_0 : p = p_0$ se

  $$-2(l(\hat{p}) - l(p)) \leq z^2$$

  con $l(p) = x\, log(p) + (n - x)log(1 - p)$ che rappresenta la funzione di log-verosimiglianza.

- **Intervallo di Jeffreys**
  E' un intervallo Bayesiano, dove come distribuzione a priori per p si sceglie $Beta(1/2, 1/2)$, pertanto l'intervallo è dato da

  $$CI_J = [B_{\alpha/2,X+1/2,n-X+1/2}, B_{1-\alpha/2,X+1/2,n-X+1/2}]$$

  con B_{α,m_1,m_2} quantile di livello α della distribuzione $Beta(m_1, m_2)$.

1.6 Perchè non usare un metodo esatto?

Gli intervalli proposti finora sono tutti di livello approssimato. Dovrebbe venire naturale chiedersi perchè non utilizzare un metodo che fornisca un intervallo esatto (almeno per quanto possibile).
In questo paragrafo vengono riprese in breve le considerazioni sulle probabilità di copertura dell'intervallo esatto ottenuto col metodo Clopper-Pearson esposte in Brown, Cai, Dasgupta, 2001.
In breve, se x è il valore osservato della variabile aleatoria $X \sim Bi(n, p)$ allora, l'intervallo di Clopper-Pearson è definito come $CI_{CP} = [l_{CP}, u_{CP}]$ dove gli estremi sono le soluzioni in p delle equazioni

$$P_p(X \geq x) = \alpha/2$$

$$P_p(X \leq x) = \alpha/2.$$

Con simbologia del tutto analoga a quella usata per l'intervallo di Jeffreys si ha che

$$CI_{CP} = [B_{\alpha/2,X,n-X+1}, B_{1-\alpha/2,X+1,n-X}].$$

Questo intervallo garantisce che la probabilità di copertura sia almeno del livello desiderato $1 - \alpha$. Il problema che rende poco utilizzabile questo intervallo è la sua eccessiva conservatività.
Il grafico seguente mostra la probabilità di copertura al variare di p con $n = 50$ e $\alpha = 0.05$. Il valore atteso della vera probabilità di copertura è circa di 0.97.
Per una trattazione più esauriente della questione si rimanda a Agresti e Coull, 1998.

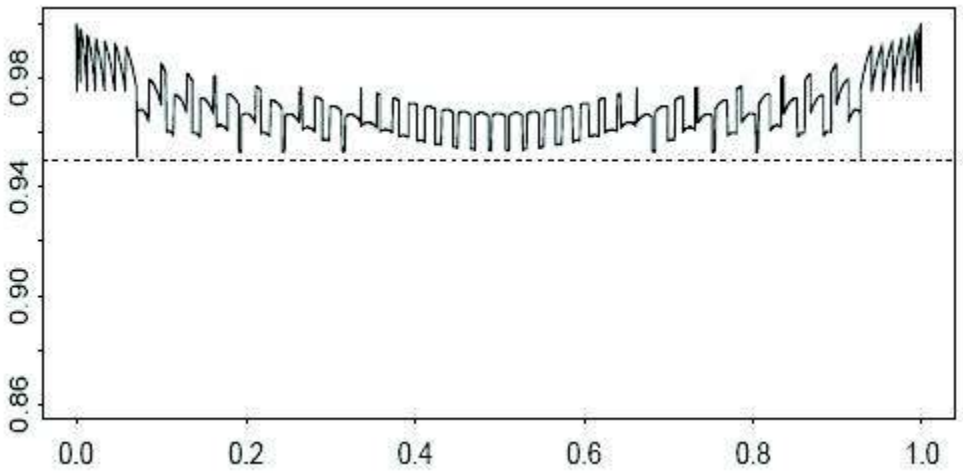

Figura 5: probabilità di copertura dell'intervallo ottenuto col metodo Clopper-Pearson al variare di p, con $n = 50$ e con $\alpha = 0.05$

Capitolo 2

Probabilità di copertura

2.1 Introduzione

In questo secondo capitolo si spiegherà come possono essere utilizzate le serie di Edgeworth per approssimare i valori delle probabilità di copertura dei vari intervalli presi in considerazione. Come presentato nell'articolo cui questa tesi fa riferimento, verrà mostrato come sia necessario ricorrere all'uso degli sviluppi a due termini. Infatti, se ci si ferma a considerare solo il primo termine della serie, si trovano valori approssimati sistematicamente maggiori di quelli reali.
Dopo aver calcolato in questo modo i valori delle probabilità di copertura di tutti gli intervalli considerati, si procederà a confrontare i valori ottenuti per iniziare a trarre le prime conclusioni.
In questo capitolo si presenteranno alcuni teoremi di cui non verranno date le dimostrazioni, per le quali si rimanda a Brown, Cai, Dasgupta, 2003. Una trattazione più approfondita sugli sviluppi di Edgeworth si trova invece in Pace, Salvan, 1996.

2.2 Sviluppi di Edgeworth a un termine

Viene ora introdotto un teorema da cui si può trarre un'approssimazione di Edgeworth a un termine per la probabilità di copertura dell'intervallo standard (d'ora in poi indicato con CI_S), di Wilson e di Agresti-Coull.
Sia CI un generico intervallo di confidenza per p. La sua probabilità di copertura è così definita:

$$C(p,n) = P_p(p \in CI) = \sum_{x=0}^{n} I(p,x) \binom{n}{x} p^x (1-p)^{1-x}$$

con $I(p,x)$ funzione indicatrice che ha valore 1 se l'intervallo contiene p quando $X = x$ e 0 altrimenti. Si definisce ora la funzione $h(x) = x - \lfloor x \rfloor$ che fornisce quindi la parte decimale di x e la funzione

$$g(p,k,n) = h(np - k\sqrt{npq}).$$

Il teorema in questione, tratto da Bhattacarya e Ranga Rao (1976), stabilisce che

$$P_p\left(\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \leq k\right) =$$

$$\Phi(z) + \left[\left(\frac{1}{2} - g(p,k,n)\right) + \frac{1}{6}(1-2p)(1-k^2)\right]\phi(z)(npq)^{-1/2} + O(n^{-1})$$

dove $\frac{1}{2} - g(p,z,n)$ rappresenta l'errore causato dalla discretezza mentre l'errore dovuto all'asimmetria è dato da $\frac{1}{6}(1-2p)(1-z^2)$.
Quindi, per esempio, tenuto conto dell'equivalenza tra eventi

$$\begin{aligned}
\{p \in CI_S\} &= \left\{-z \leq \frac{\left(\frac{X}{n} - p\right)\sqrt{n}}{\sqrt{\hat{p}\hat{q}}} \leq z\right\} \\
&= \left\{l_{n,p} \leq X \leq u_{n,p}\right\} \\
&= \left\{\frac{l_{n,p}/n - p}{\sqrt{\frac{pq}{n}}} \leq \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \leq \frac{u_{n,p}/n - p}{\sqrt{\frac{pq}{n}}}\right\} \\
&= \left\{l_S \leq \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \leq u_S\right\},
\end{aligned}$$

dove $l_{n,p}$ e $u_{n,p}$ sono definiti nel paragrafo 1.4 e

$$(l_S, u_S) = \left(\frac{\sqrt{n}(1/2 - p) \mp nz\sqrt{z^2/4n + pq}}{\sqrt{pq}(n + z^2)} \right),$$

sono funzioni di (p, n, z), si ottiene l'approssimazione

$$P_p(p \in CI_S) = \Phi(z) + [g(p, l_S, n) - g(p, u_S, n)]\phi(z)(npq)^{-1/2} + O(n^{-1}).$$

In modo analogo si procede con l'intervallo di Agresti-Coull e di Wilson.

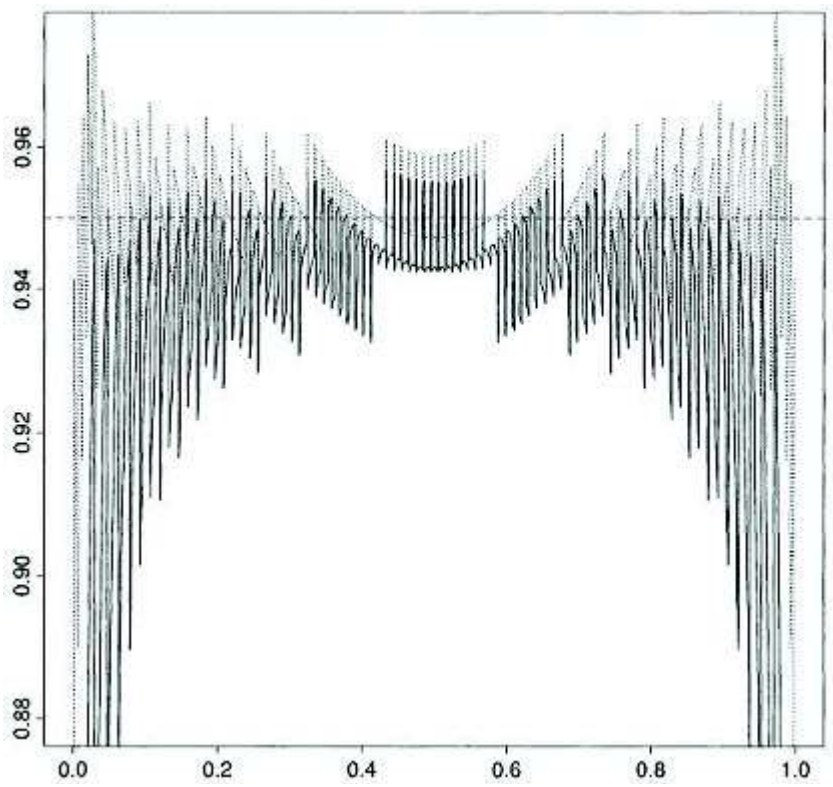

Figura 6: probabilità di copertura dell'intervallo standard al variare di p, con $\alpha = 0.05$ *e* $n = 100$. *A tratto uniforme la probabilità di copertura esatta, i puntini indicano invece l'approssimazione di Edgeworth a un termine.*

2.2.1 L'approssimazione a un termine sovrastima la probabilità di copertura

In Figura 6 l'approssimazione a un termine per l'intervallo standard appena ottenuta viene confrontata (al variare di p, fissato $\alpha = 0.05$ e $n = 100$) con la probabilità di copertura esatta. Si nota immediatamente che l'approssimazione ottenuta domina costantemente il vero valore della probabilità.
Il motivo si può capire studiando il termine di ordine $O(n^{-1})$. Esso è principalmente di segno negativo e non oscillatorio (si può infatti notare dal grafico come le oscillazioni della probabilità di copertura siano già catturate dall'approssimazione a un termine). Poichè questo termine di errore non è trascurabile, soprattutto per valori di n relativamente piccoli, è necessario considerare l'approssimazione a due termini, la quale è sensibilmente più precisa. Un altro motivo che porterebbe a questa decisione è che l'approssimazione a un termine per l'intervallo standard è sostanzialmente la stessa anche per alcuni degli altri intervalli presentati, mentre il secondo termine varia notevolmente a seconda dell'intervallo.

2.3 Sviluppi di Edgeworth a due termini

Le probabilità di copertura dei cinque intervalli saranno presentate sotto forma di teoremi a questo scopo. Si farà uso della notazione di seguito riportata. Con $g(p,k)$ si denoterà la funzione già presentata nel paragrafo precedente $g(p,k,n)$ dove si sottointende la dipendenza da n. Inoltre, siano:

$$
\begin{aligned}
w(k) &= \left(\frac{1}{9} - \frac{1}{36pq}\right)k^5 + \left(\frac{7}{36pq} - \frac{11}{18}\right)k^3 + \left(\frac{1}{6} - \frac{1}{6pq}k\right), \\
Q_{21}(l,u) &= 1 - g(p,l) - g(p,u), \\
Q_{22}(l,u) &= \frac{1}{2}\left[-g^2(p,l) - g^2(p,u) + g(p,l) + g(p,u) - \frac{1}{3}\right],
\end{aligned}
$$

dove con l e u andranno sostiuiti di volta in volta gli estremi dell'intervallo o le loro approssimazioni nel caso questi non si possano ottenere in forma esplicita.
Le approssimazioni a due termini sono estremamente accurate. Se si sceglie ad esempio $n = 40$ e $0.2 < p < 0.8$, l'errore massimo che si può commettere approssimando la probabilità di copertura effettiva dell'intervallo di Wilson è 0.0002. Nello stesso caso, per l'intervallo standard si ha un errore massimo di 0.0075 e per l'intervallo di Agresti-Coull esso non supera 0.0006.

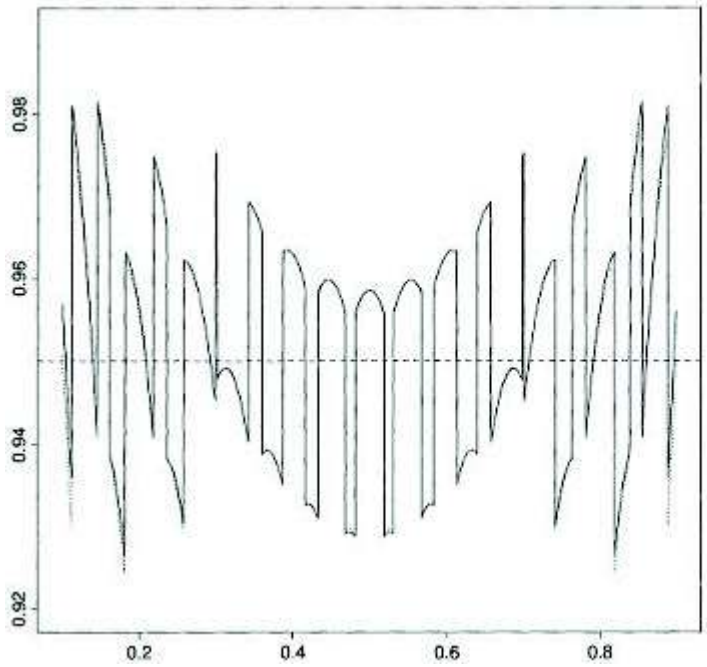

Figura 7: probabilità di copertura esatta dell'intervallo di Wilson (tratto uniforme) e sua approssimazione di Edgeworth a due termini (puntini). Fissati $n = 20$ e $\alpha = 0.05$, i due grafici appaiono quasi indistinguibili.

2.3.1 L'intervallo standard

Per studiare l'approssimazione di Edgeworth a due termini della probabilità di copertura dell'intervallo standard, è utile introdurre il seguente intervallo

generico:

$$CI_*(\beta) = \frac{X+\beta}{n+2\beta} \pm z\sqrt{\frac{pq}{n}}.$$

Esso riconduce all'intervallo standard CI_S se $\beta = 0$ ma può essere ricentrato cambiando il valore di β. Ad esempio, per $\beta = z^2/2$ si ottiene l'intervallo ricentrato introdotto nel paragrafo 1.3.

Teorema 1. Sia $0 < p < 1$ e $0 < \alpha < 1$. Definiti l_* e u_* in modo del tutto analogo a l_S e u_S, si supponga inoltre non intero $np + l_*\sqrt{npq}$. Allora la probabilità di copertura dell'intervallo generico $CI_*(\beta)$ soddisfa

$$\begin{aligned}
P_* = P_p(p \in CI_*(\beta)) &= (1-\alpha) \\
&+ [g(p,l_*) - g(p,u_*)]\phi(z)(npq)^{-1/2} \\
&+ \left\{ 2t_2 - zt_1^2 - (1-2p)\left(z - \frac{z^3}{3}\right) t_1 (pq)^{-1/2} + w(z) \right\} \phi(z) n^{-1} \\
&+ \left\{ \left[(1-2p)\left(\frac{z^2}{6} - \frac{1}{2}\right) - (pq)^{1/2} t_1 \right] Q_{21}(l_*,u_*) + Q_{22}(l_*,u_*) \right\} \frac{z\phi(z)}{pq} n^{-1} \\
&+ O(n^{-3/2}),
\end{aligned}$$

dove

$$\begin{aligned}
t_1 &= (z^2 - 2\beta)\left(\frac{1}{2} - p\right)(pq)^{-1/2}, \\
t_2 &= \left(\frac{1}{8pq} - 1\right) z^3 + \left(4 - \frac{1}{2pq} z\beta\right).
\end{aligned}$$

In particolare, per l'intervallo standard basta sostituire $\beta = 0$ e si ottiene

$$\begin{aligned}
P_S = P_p(p \in CI_S) &= 1-\alpha \\
&+ [g(p,l_S) - g(p,u_S)]\phi(z)(npq)^{-1/2} \\
&+ \left\{ -\frac{(1-2p)^2)}{12pq} z^5 - \frac{1}{4pq} z^3 + w(z) \right\} \phi(z) n^{-1} \\
&+ \left\{ -(1-2p)\left(\frac{z^2}{3} + \frac{1}{2}\right) Q_{21}(l_S,u_S) + Q_{22}(l_S,u_S) \right\} z\phi(z)(npq)^{-1} \\
&+ O(n^{-3/2}).
\end{aligned}$$

Il primo dei due termini di ordine $O(n^{-1})$ non ha carattere oscillatorio, per cui causerebbe un errore sistematico se venisse omesso. Al contrario, il secondo esprime le oscillazioni che derivano da due cause. Q_{22} rappresenta l'errore dovuto esclusivamente all'arrotondamento mentre il termine che contiene Q_{21} rappresenta l'errore dovuto all'asimmetria e alla non continuità della distribuzione binomiale.

2.3.2 L'intervallo di Wilson e quello di Agresti-Coull

A questo punto, la derivazione delle approssimazioni delle probabilità di copertura di questi intervalli è puramente di natura tecnica.
Per l'intervallo di Wilson non c'è bisogno di calcolare quantili specifici.
Teorema 2. Sia $0 < p < 1$ e sia $0 < \alpha < 1$. Si assuma $np - z(npq)^{1/2}$ non intero. Allora

$$\begin{aligned} P_W =& P_p(p \in CI_W) \\ &= (1-\alpha) + [g(p,-z) - g(p,z)]\phi(z)(npq)^{-1/2} + w(z)\phi(z)(n)^{-1} \\ &+ \left\{(1-2p)\left(\frac{z^2}{6} - \frac{1}{2}\right) Q_{21}(-z,z) + Q_{22}(-z,z)\right\} z\phi(z)(npq)^{-1} \\ &+ O(n^{-3/2}). \end{aligned}$$

Per l'intervallo di Agresti-Coull è necessario calcolare l_{AC} e u_{AC} in modo analogo a quanto fatto per l'intervallo standard.
Teorema 3. Sia $0 < p < 1$ e $0 < \alpha < 1$. Inoltre si supponga $np + l_{AC}(npq)^{1/2}$ non intero. Allora

$$\begin{aligned} P_{AC} &= P_p(p \in CI_{AC}) \\ &= (1-\alpha) + [g(p,l_{AC}) - g(p,u_{AC})]\phi(z)(npq)^{-1/2} \\ &+ \left[\left(\frac{1}{4pq} - 1\right) z^3 + w(z)\right] \phi(z) n^{-1} \\ &+ \left\{(1-2p)\left(\frac{z^2}{6} - \frac{1}{2}\right) Q_{21}(l_{AC},u_{AC}) + Q_{22}(l_{AC},u_{AC})\right\} z\phi(z)(npq)^{-1} \\ &+ O(n^{-3/2}). \end{aligned}$$

2.3.3 L'intervallo basato sul rapporto di verosimiglianza e quello di Jeffreys

Un discorso diverso va fatto per gli ultimi due intervalli considerati. Per essi, infatti, non esiste una forma esplicita degli estremi dell'intervallo. Per questo motivo bisogna prima trovare un'approssimazione degli estremi e poi procedere all'approssimazione della probabilità di copertura. Definite l_{LR}, u_{LR}, u_J, l_J le approssimazioni necessarie (con ovvio significato), possono essere presentate le approssimazioni delle probabilità di copertura

Teorema 4. Sia $0 < p < 1$ e $0 < \alpha < 1$. Inoltre si supponga $np + l_{LR}(npq)^{1/2}$ non intero. Allora vale l'approssimazione

$$\begin{aligned} P_{LR} &= P_p(p \in CI_{LR}) \\ &= (1-\alpha) + [g(p, l_{LR}) - g(p, u_{LR})]\phi(z)(npq)^{-1/2} \\ &+ \left(\frac{1}{6} - \frac{1}{6pq}\right) z\phi(z)n^{-1} \\ &+ \left\{\left(p - \frac{1}{2}\right) Q_{21}(l_{LR}, u_{LR}) + Q_{22}(l_{,LR}, u_{LR})\right\} z\phi(z)(npq)^{-1} \\ &+ O(n^{-3/2}). \end{aligned}$$

Teorema 5. Sia $0 < p < 1$ e $0 < \alpha < 1$. Inoltre si supponga $np + l_J(npq)^{1/2}$ non intero. Allora vale l'approssimazione

$$\begin{aligned} P_J &= P_p(p \in CI_J) \\ &= (1-\alpha) + [g(p, l_J) - g(p, u_J)]\phi(z)(npq)^{-1/2} \\ &- 12z\phi(z)(npq)^{-1} \\ &+ \left[\frac{2p-1}{3} Q_{21}(l_J, u_J) + Q_{22}(l_J, u_J)\right] z\phi(z)(npq)^{-1} \\ &+ O(n^{-3/2}). \end{aligned}$$

2.4 Confronti

Per confrontare le probabilità di copertura è utile notare la struttura comune a tutte le approssimazioni trovate.
Ognuna di esse è infatti composta di cinque parti:

1. $(1-\alpha)$, cioè il valore nominale dell'intervallo
2. un termine a carattere oscillatorio di ordine $O(n^{-1/2})$
3. un termine di ordine $O(n^{-1})$ che non oscilla e che, in valore assoluto, aumenta all'aumentare di z e per p che si allontana da 0.5. Questo termine è identificabile come la distorsione della probabilità di copertura dell'intervallo
4. un altro termine oscillatorio di ordine $O(n^{-1})$
5. l'errore di approssimazione di ordine $O(n^{-3/2})$

E' possibile dimostrare che, in media, i termini oscillatori sono di un ordine inferiore rispetto al termine di distorsione. Questo permette di confrontare le probabilità di copertura in termini del solo errore di distorsione presente nelle approssimazioni. Il seguente teorema formalizza quanto detto.
Teorema 6. Sia f una funzione di densità su un intervallo proprio $[a,b]$ tale che

$$|f(p_1) - f(p_2)| \leq M\,|p_1 - p_2| \;\; \forall\, p_1, p_2 \in [a,b].$$

Allora, per gli intervalli considerati (standard, standard ricentrato, di Wilson, di Agresti-Coull, basato sul rapporto di verosimiglianza e di Jeffreys), l'oscillazione integrata rispetto a f è asintoticamente trascurabile. In particolare

$$\int \text{''}oscillazione\; O(n^{-1/2})\text{''} f(p)dp = O(n^{-3/2}),$$
$$\int \text{''}oscillazione\; O(n^{-1})\text{''} f(p)dp = O(n^{-3/2}).$$

Per cui, detta $C(p,n)$ la probabilità di copertura reale,

$$\int \{C(p,n) - (1-\alpha) - \text{''}distorsione\; O(n^{-1})\text{''}\} f(p)dp = O(n^{-3/2}).$$

A questo punto, si possono usare le equazioni dei teoremi 1-5 per ottenere tutti i confronti necessari. Tralasciando gli errori di approssimazione e i termini oscillatori, si ottengono, tra gli altri:

$$P_{AC} - P_S = \left\{ \frac{(1-2p)^2}{12pq} z^5 + \left(\frac{1}{2pq} - 1 \right) z^3 \right\} \phi(z) n^{-1} \tag{2.1}$$

$$P_{AC} - P_W = \left(\frac{1}{4pq} - 1 \right) z^3 \phi(z) n^{-1} \tag{2.2}$$

$$P_{AC} - P_{LR} = \left\{ -\frac{(1-2p)^2}{36pq} z^5 + \left(\frac{4}{9pq} - \frac{29}{18} \right) z^3 \right\} \phi(z) n^{-1} \tag{2.3}$$

$$P_{AC} - P_J = \left\{ -\frac{(1-2p)^2}{36pq} z^5 + \left(\frac{4}{9pq} - \frac{29}{18} \right) z^3 + \left(\frac{1}{6} - \frac{1}{12pq} \right) z \right\} \phi(z) n^{-1} \tag{2.4}$$

Gli altri confronti possono essere ottenuti direttamente dai precedenti.

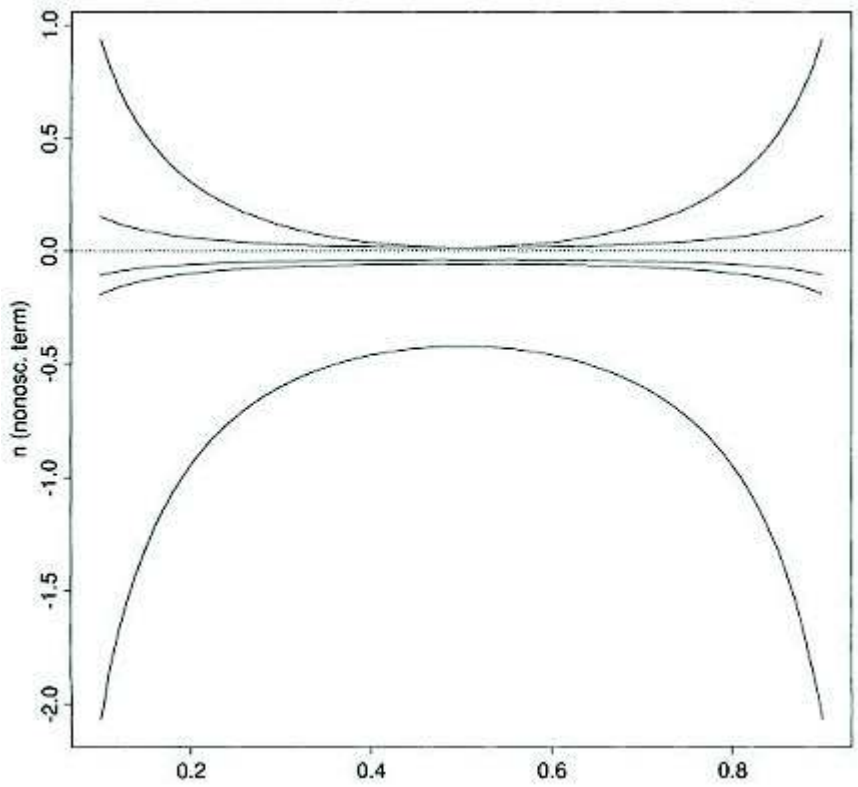

Figura 8: dall'alto al basso i termini di distorsione di P_{AC}, P_W, P_J, P_{LR}, P_S. α fissato a 0.05.

La cosa più importante da notare è che la (2.1) e la (2.2) sono positive per ogni p e a qualsiasi livello di confidenza. La (2.3) e la (2.4), invece, sono positive per ogni p se $z \leq 3.95$. Questo significa che l'intervallo di Agresti-Coull ha la probabilità di copertura maggiore rispetto agli altri intervalli. Il grafico in Figura 8 mostra l'andamento del termine non oscillatorio di ordine $O(n^{-1})$ di tutti gli intervalli (più precisamente nell'asse delle ordinate viene rappresentato il termine moltiplicato per n, in modo da eliminare la dipendenza dalla dimensione del campione), per tutti i valori di p, con $\alpha = 0.05$ e permette di fare altre osservazioni. Si nota infatti, immediatamente, come l'intervallo standard sia soggetto a distorsioni negative molto evidenti che aumentano in prossimità dei valori estremi di p. L'intervallo di Agresti-Coull, al contrario, è troppo conservativo quando p (o q) si avvicina a 0. Gli altri tre intervalli hanno una probabilità di copertura più stabile al variare di p, e sono sostanzialmente simili sotto questo aspetto.

Capitolo 3

Lunghezze attese

3.1 Introduzione

Nella comparazione degli intervalli di confidenza, oltre alla probabilità di copertura, può anche essere importante considerare la parsimoniosità della lunghezza dell'intervallo. Questo fattore deve essere naturalmente preso in considerazione solo in seconda battuta, in quanto, come abbiamo visto, alcuni intervalli hanno prestazioni molto simili in termini di copertura.
Perciò, in questo capitolo, le serie di Edgeworth verranno usate per studiare le lunghezze attese degli intervalli proposti. Ciò permetterà di fare interessanti considerazioni per la scelta degli intervalli a seconda del valore stimato del parametro. Come indice di prestazione verrà anche considerata la lunghezza attesa integrata definita come $\int_0^1 E(\overline{CI})dp$. Dove $\overline{CI}$ indica l'approssimazione di Edgeworth per la lunghezza attesa dell'intervallo di confidenza generico CI.

3.2 Considerazioni sulle lunghezze attese

Come si vedrà nel teorema seguente le approssimazioni per le lunghezze attese differiscono qualitativamente da quelle per le probabilità di copertura. I primi due termini della serie, infatti, sono di ordine $O(n^{-1/2})$ e $O(n^{-3/2})$.

Inoltre il primo termine è uguale per tutti gli intervalli presentati, per cui il confronto si riduce al secondo termine.

Teorema 7 Sia CI una notazione generica per uno degli intervalli CI_S, CI_W, CI_{AC}, CI_{LR}, CI_J. E sia $\overline{CI}$ il valore della lunghezza attesa dell'intervallo. Allora

$$\overline{CI} = 2z(pq)^{1/2}n^{-1/2}\left(1 - \frac{\delta(z,p)}{8npq}\right) + O(n^{-2})$$

dove

$$\delta(z,p) = \begin{cases} 1 & \text{per } CI_S, \\ 1 + z^2(8pq - 1) & \text{per } CI_W, \\ 1 + z^2(12pq - 2) & \text{per } CI_{AC}, \\ 1 + z^2\left(\frac{26}{9}pq - \frac{2}{9}\right) & \text{per } CI_{LR}, \\ 1 + z^2\left(\frac{26}{9}pq - \frac{2}{9}\right) + \frac{2}{9}(17pq - 2) & \text{per } CI_J. \end{cases}$$

L'approssimazione data dal Teorema 7 è molto accurata. L'errore che si commette per gli intervalli al livello di confidenza 0.95 con $n = 40$ non supera mai 0.0035 per i valori di $p \in [0.1, 0.9]$.

La prima considerazione da fare è che per ogni valore di z e per $p \in [0,1]$ si ha che $1 + z^2(8pq - 1) \geq 1 + z^2(12pq - 2)$, per cui l'intervallo di Agresti-Coull è più lungo in media di quello di Wilson (si poteva dimostrare anche direttamente dalle loro definizioni che CI_W è un sottoinsieme di CI_{AC}). Gli altri confornti dipendono da z e p.

In particolare si ha che, fissato $z = 1.96$ e $p \leq 0.5$, per $0.084 \leq p \leq 0.137$ l'intervallo più corto è quello basato sulla verosimiglianza, mentre il più breve è quello di Jeffreys se si considera $0.137 \leq p \leq 0.201$. L'intervallo di Wilson è invece il più breve per $0.201 \leq p \leq 0.5$. Ovviamente, vale la simmetria per $p > 0.5$. Inoltre, se si sceglie un altro livello di confidenza, i confronti fatti cambiano solo a livello qualitativo. Si può quindi dire che l'intervallo di Wilson è il meno conservativo per p lontano dai valori estremi, mentre per valori del parametro vicino a 0 o 1, sono più parsimoniosi gli intervalli CI_J e CI_{LR}. Naturalmente da queste considerazioni è stato escluso l'intervallo standard a causa della sua non accettabile probabilità di copertura.

3.2.1 Lunghezze attese integrate

Le considerazioni fatte finora sulla lunghezza degli intervalli portano a diverse scelte, in base al valore del parametro da stimare.
Può essere invece utile avere un indice di performance generale per l'intervallo. Brown, Cai e Dasgupta (2001) propongono la lunghezza attesa integrata, cioè l'integrale da 0 a 1, rispetto a p, delle lunghezze attese trovate nel Teorema 7. Si ottiene:

$$\int_0^1 \overline{CI_S}\, dp = \frac{z\pi}{4} n^{-1/2} - \frac{k\pi}{4} n^{-3/2} + O(n^{-2}),$$
$$\int_0^1 \overline{CI_W}\, dp = \frac{z\pi}{4} n^{-1/2} - \frac{k\pi}{4} n^{-3/2} + O(n^{-2}),$$
$$\int_0^1 \overline{CI_{AC}}\, dp = \frac{z\pi}{4} n^{-1/2} + \left(\frac{z^2}{2} - 1\right) \frac{k\pi}{4} n^{-3/2} + O(n^{-2}),$$
$$\int_0^1 \overline{CI_{LR}}\, dp = \frac{z\pi}{4} n^{-1/2} - \left(1 + \frac{5z^2}{36}\right) \frac{k\pi}{4} n^{-3/2} + O(n^{-2}),$$
$$\int_0^1 \overline{CI_J} dp = \frac{z\pi}{4} n^{-1/2} - \left(\frac{37}{36} + \frac{5z^2}{36}\right) \frac{k\pi}{4} n^{-3/2} + O(n^{-2}).$$

Come si può notare secondo questo indice non c'è differenza tra le prestazioni dell'intervallo standard e quello di Wilson. Ciò è dovuto al fatto che il primo è molto breve per valori di p vicino alla frontiera dello spazio parametrico, fino a ridursi a un punto se $p = 0$ o $p = 1$. Ovviamente le precedenti considerazioni sulla probabilità di copertura hanno la priorità riguardo alla scelta. Il pregio di questo indice è che fornisce una graduatoria che non dipende da z. CI_J è in ogni caso l'intervallo più breve e CI_{AC} quello più conservativo.

Conclusioni

In questa tesi si è mostrato per quali ragioni l'intervallo standard basato sulla statistica di Wald non può essere utilizzato per stimare probabilità di successo da una distribuzione binomiale. Ciò porta a cercare altri metodi per ottenere stime intervallari del parametro p. Gli sviluppi di Edgeworth permettono di studiare le alternative trovate in termini di probabilità di copertura e di lunghezza attesa in modo molto preciso, anche quando ci si trova in presenza di campioni relativamente piccoli.
Tralasciando momentaneamente le considerazioni fatte finora, l'intervallo di Agresti-Coull presenta sicuramente un paio di pregi da non sottovalutare. Esso è semplice da spiegare, anche per studenti che si trovano ai primi corsi di Statistica, e ha una forma analoga a quella dell'intervallo standard. Questa sua semplicità si traduce anche in una maggiore facilità computazionale rispetto agli altri intervalli alternativi.
Come visto, però, questo intervallo è un po' troppo conservativo (e lungo) soprattutto per valori del parametro vicino a 0 e 1, il che, a volte, può essere un problema che porta a scegliere l'intervallo di Wilson, di Jeffreys o quello basato sul rapporto di verosimigianza.
Questi tre intervalli sono molto simili sia a livello di copertura, sia confrontandone le lunghezze, per cui la scelta diventa quasi completamente soggettiva, soprattutto se le difficoltà computazionali non sono un problema.

Riferimenti bibliografici

Agresti, A. e Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119-126.

Bhattacharya R. N. e Ranga Rao R. (1976). *Normal approximation and asymptotic expansions*, Wiley, New York.

Brown D., Cai T. T. e DasGupta A. (2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *The Annals of Statistic* **30**, 159-201.

Brown D., Cai T. T. e DasGupta A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, **16**, 101-133.

Pace L. e Salvan A. (2001). *Introduzione alla Statistica. Inferenza, Verosimiglianza, Modelli*, CEDAM, Padova.

Pace L. e Salvan A. (1996). *Teoria della Statistica*, CEDAM, Padova.