Inferenze per le proporzioni

alcuni esempi

Singola proporzione. In un sondaggio svoltosi nel 1995, il 53.5% di 959 votanti del Quebec ha affermato che avrebbe votato 'sì' al referendum indipendentista svoltosi 5 giorni dopo.

- Lasciando da parte la possibilità di errori estranei al processo di campionamento, possiamo concludere che 5 giorni prima del referendum più della metà degli elettori aveva intenzione di votare 'sì' al referendum?

Due proporzioni. Due mesi prima, il 45.1% di un campione di 822 votanti aveva manifestato l'intenzione di votare 'sì'. La differenza tra le proporzioni dei due campioni è $0.535 - 0.451 = 0.084$ e dunque il sostegno per il 'sì' è più alto dell'8.4% nel secondo sondaggio.

- Possiamo concludere che, nel periodo di due mesi intercorso tra i due sondaggi, il supporto per il 'sì' è aumentato?

La proporzione di casi nella popolazione che esibiscono una data caratteristica è rappresentata dalla lettera π.

La corrispondente proporzione nel campione è

$$\hat{\pi} = \frac{\text{numero di "successi" nel campione}}{\text{numero di osservazioni nel campione}}$$

laddove un "successo" è uno dei due possibili esiti di un evento, come per esempio avere intenzione di votare 'sì' al referendum.

Distribuzione campionaria di $\hat{\pi}$

- Quali sono le proprietà di $\hat{\pi}$ quale stimatore della proporzione di casi π nella popolazione?

  Il numero di successi X in n prove segue la distribuzione binomiale.

  Dal momento che $\hat{\pi} = \frac{X}{n}$, la sua distribuzione campionaria sarà

$$P\left(\hat{\pi} = \frac{k}{n}\right) = P\left(X = k\right) = \binom{n}{k} \pi^k \left(1 - \pi\right)^{n-k}$$

Per un campione casuale semplice di grandi dimensioni (oltre le 200 unità) estratto da una popolazione di dimensioni molto maggiori è conveniente usare l'approssimazione normale alla distribuzione binomiale.

- La distribuzione campionaria di $\hat{\pi}$ è approssimativamente normale.
- L'approssimazione migliora al crescere delle dimensioni n del campione ed è tanto migliore quanto più la proporzione di casi π nella popolazione è simile a $\pi = 0.5$.

Valore atteso di $\hat{\pi}$

- consideriamo k "successi", in n prove, il cui modello probabilistico corrisponde ad una distribuzione discreta binomiale.
- Il valore atteso della distribuzione campionaria di $\hat{\pi} = \frac{k}{n}$ è uguale a π.

$$E(\hat{\pi}) = E\left(k\frac{1}{n}\right) = \frac{1}{n}E(k) = \frac{1}{n}np = \pi$$

- Ovvero, $\hat{\pi}$ è uno stimatore non distorto di π.

Deviazione standard di π

- La deviazione standard della distribuzione campionaria di $\hat{\pi}$ dipende da π ed è uguale a:

$$\sqrt{Var\left(\pi\right)} = \sqrt{Var\left(\frac{1}{n}k\right)} = \sqrt{\frac{1}{n^2}Var(k)} =$$

$$= \sqrt{\frac{1}{n^2}np(1-p)} =$$

$$= \sqrt{\frac{\pi\left(1-\pi\right)}{n}}$$

Si notino i seguenti fatti.

- Dato che n è al denominatore, la variabilità di $\hat{\pi}$ diminuisce al crescere della numerosità del campione.
- Essendo n all'interno della radice quadrata, per diminuire la deviazione standard di $\hat{\pi}$ di un fattore di due dobbiamo aumentare di quattro volte la numerosità del campione.
- Tenendo fisso n, la deviazione standard di $\hat{\pi}$ assume il suo valore massimo quando $\pi = 0.5$ (e quindi $\pi\left(1-\pi\right) = 0.5 \times 0.5 = 0.25$).

Due utili regole forniscono le basi per l'inferenza statistica per le proporzioni.

- La distribuzione normale fornisce una approssimazione appropriata alla distribuzione campionaria di $\hat{\pi}$ quando due condizioni sono soddisfatte.
  1. $n\pi \geq 10$ e $n\left(1-\pi\right) \geq 10$
  2. La deviazione standard di $\hat{\pi}$ è ben approssimata da $\sqrt{\pi\left(1-\pi\right)/n}$ quando la popolazione è almeno 10 volte più grande del campione. Questa condizione è quasi sempre soddisfatta.

- I risultati precedenti non possono essere direttamente usati in quanto, non conoscendo la proporzione della popolazione π, non possiamo calcolare la deviazione standard di $\hat{\pi}$.
- È quindi necessario sostituire alla deviazione standard di $\hat{\pi}$ una sua stima, l'errore standard stimato di $\hat{\pi}$.
- Per fare questo, utilizziamo $\hat{\pi}$ in luogo di π nella formula della deviazione standard di $\hat{\pi}$:

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}\,(1-\hat{\pi})}{n}}$$

Intervallo di confidenza

L'intervallo di confidenza per una proporzione π avrà dunque la forma

$$\hat{\pi} \pm z^* \hat{\sigma}_{\hat{\pi}}$$

dove z^* è il valore critico della distribuzione normale standardizzata al livello $(1-\alpha)$.

- Per un livello di confidenza al 95%, $z^* = 1.96$.

Illustrazione

Consideriamo nuovamente il sondaggio in cui, 5 giorni prima del referendum indipendentista del 1995, il 53.5% di 959 votanti del Quebec ha affermato che avrebbe votato 'sì'.

- Costruiamo l'intervallo di confidenza al 95% della proporzione π di individui nella popolazione che intendono votare 'sì':

$$\hat{\pi} = \frac{513}{959} = 0.535$$

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{0.535\,(1-0.535)}{959}} = 0.0161$$

- al livello di confidenza $1 - \alpha = 0.95$, per π si ha

$$\begin{aligned} \hat{\pi} \pm z^* \hat{\sigma}_{\hat{\pi}} &= 0.535 \pm 1.96 \times 0.0161 \\ &= 0.535 \pm 0.032 \\ &= (0.503; \quad 0.567) \end{aligned}$$

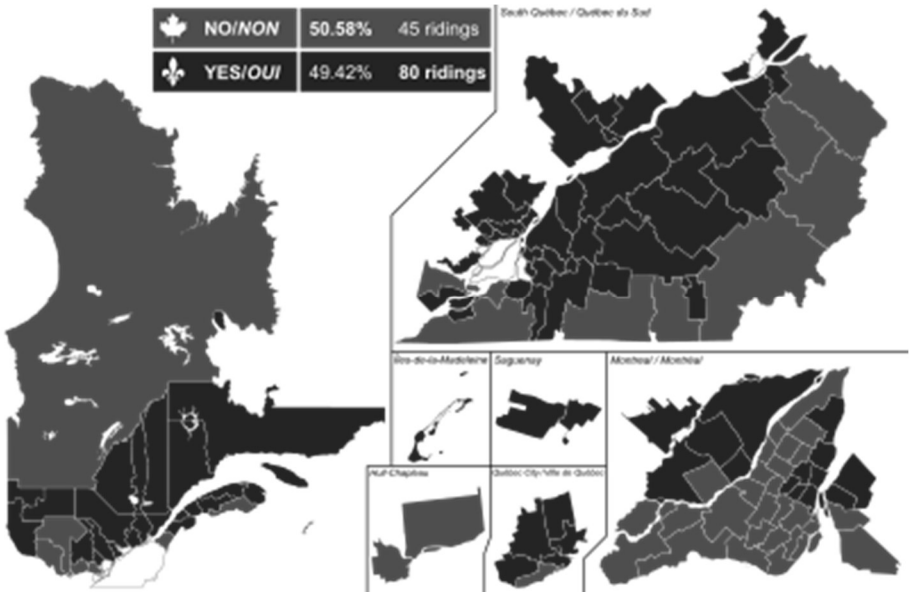
- possiamo quindi scrivere

$$P(0.503 \leq \pi \leq 0.567) = 0.95$$

Interpretazione dell'intervallo di confidenza

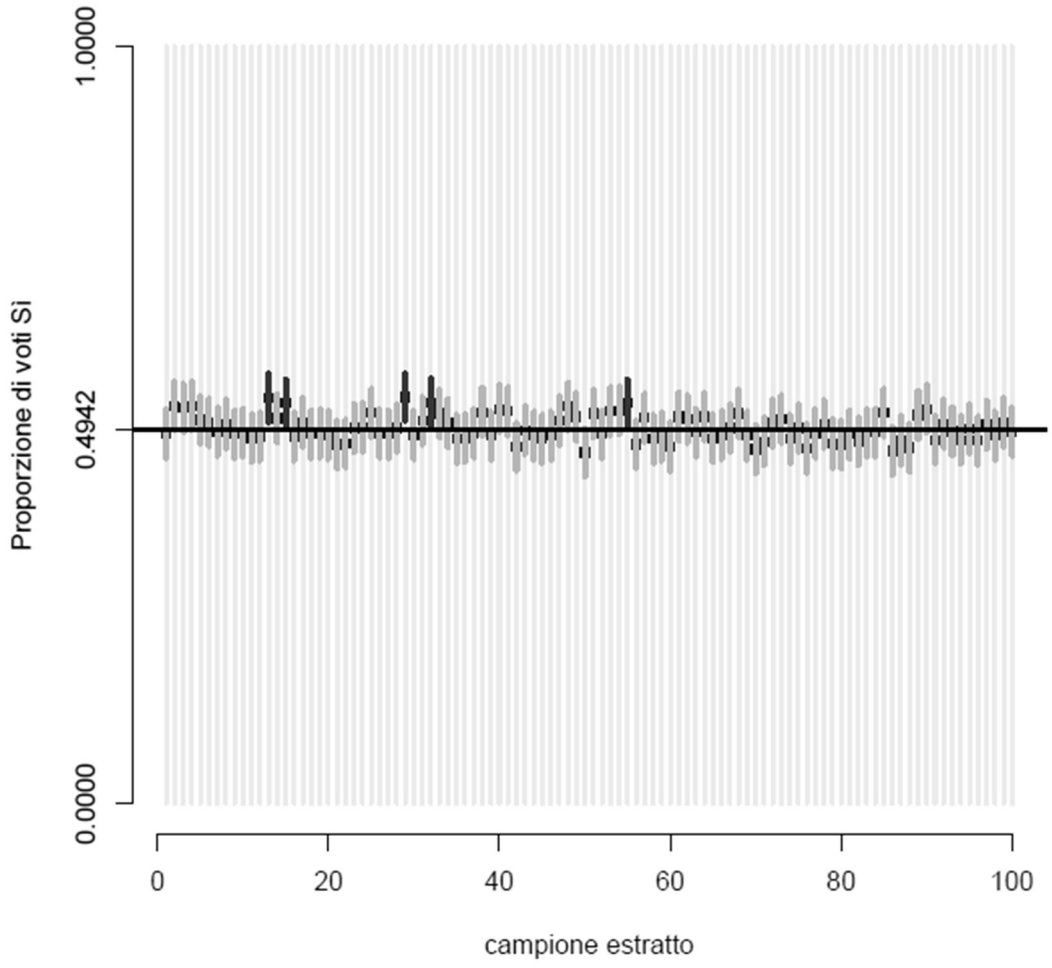
La probabilità $1 - \alpha$ dell'intervallo di confidenza per π ha un significato identico a quello della probabilità dell'intervallo di confidenza per μ:

- se dalla popolazione di votanti del Quebec venissero estratti tutti i possibili campioni di numerosità $n = 959$ e venissero costruiti tutti i possibili intervalli di confidenza per π usando la procedura sopra definita:
- una frazione uguale a $1 - \alpha$ comprenderebbe il valore reale di π e
- la rimanente frazione α non lo comprenderebbe.



Simulazione.

- 100 campioni casuali, di ampiezza $n = 959$, vengono estratti da una popolazione con $P(''SI'') = \pi = 0.4942$. [Ovviamente $P(''NO'') = 1 - \pi = 0.5058$]
- In ciascun campione si stima un intervallo di fiducia al 95% per la proporzione π
- Ci attendiamo una proporzione di (circa) 5 intervalli su 100 che non contengono la vera $\pi = 0.4942$ nota.



Test per una proporzione

- L'ipotesi nulla

$$H_0 : \pi = \pi_0$$

  specifica un valore π_0 per il parametro sconosciuto π.
- Dato che il test viene calcolato supponendo vera l'ipotesi nulla H_0, π_0 può essere usato in luogo di π per stimare la deviazione standard di $\hat{\pi}$:

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\pi_0 (1 - \pi_0)}{n}}$$

Sotto H_0, dunque, la statistica test diventa:

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0 (1 - \pi_0)}{n}}}$$

Se H_0 è vera, z seguirà la distribuzione normale standardizzata.

- Per trovare il $p-$valore della statistica test dobbiamo specificare un'ipotesi alternativa H_a.
- L'ipotesi alternativa può essere bilaterale

$$H_a : \pi \neq \pi_0$$

  o unilaterale

$$H_a : \pi < \pi_0 \quad \text{oppure} \quad H_a : \pi > \pi_0$$

Illustrazione

Consideriamo nuovamente i dati del sondaggio esaminato in precedenza e verifichiamo l'ipotesi nulla $H_0 : \pi = 0.5$ contro l'alternativa (unilaterale, destra) che il 'sì' sia maggioritario, $H_a : \pi > 0.5$.

La statistica test assume il valore:

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0 (1 - \pi_0)}{n}}} = \frac{0.535 - 0.5}{\sqrt{\frac{0.5 (1 - 0.5)}{959}}} = 2.17$$

Con un livello di significatività $\alpha = 0.05$ (assumendoci quindi questa probabilità per un errore di I tipo), il valore critico è $z_{\alpha=0.05} = 1.645 < 2.17$.

Poiché il p-valore della statistica test è basso (< 0.05), si rifiuta l'ipotesi nulla di uguaglianza degli orientamenti di voto:

> *possiamo essere ragionevolmente certi che a cinque giorni dal referendum la maggioranza dei votanti era orientata per il 'sì'.*

- Se avessimo specificato un'ipotesi alternativa bilaterale $H_a : \pi \neq 0.5$ (potrebbe essere in vantaggio sia il 'sì' che il 'no'), il $p-$valore sarebbe stato

$$P = 2 \times 0.015 = 0.030$$

Condizioni di validità

Affinché siano valide le inferenze statistiche su π basate sulla distribuzione normale è necessario che siano soddisfatte le seguenti condizioni

- I dati devono essere un campione casuale semplice estratto dalla popolazione di interesse.
- La popolazione deve essere almeno 10 volte più grande del campione.
- Per l'intervallo di confidenza, è necessario che $n\hat{\pi} \geq 10$ e $n (1 - \hat{\pi}) \geq 10$
- Per il test dell'ipotesi $H_0 : \pi = \pi_0$, è necessario che $n\pi_0 \geq 10$ e $n (1 - \pi_0) \geq 10$

- Nel caso dell'esempio sul referendum del Quebec, le dimensioni del campione $n = 959$ sono sicuramente molto minori di $1/10$ della popolazione, ma probabilmente il sondaggio non era basato su un campionamento casuale semplice.

- Per l'intervallo di confidenza, $n\hat{\pi} = 959(0.535) = 513$ e $n(1 - \hat{\pi}) = 959(1 - 0.535) = 446$ sono molto maggiori del valore richiesto di 10.
- Per il test dell'ipotesi $H_0 : \pi = \pi_0$, $n\pi_0 = n(1 - \pi_0) = 959(0.5) = 479.5 \gg 10$.

Numerosità del campione

- Consideriamo ora la numerosità del campione necessaria per la stima di una proporzione con errore prefissato.
- L'intervallo di confidenza per una proporzione al livello $(1 - \alpha)$ è

$$\hat{\pi} \pm z^* \hat{\sigma}_{\hat{\pi}}$$

- Quanti individui dobbiamo considerare per ottenere una stima campionaria della proporzione π con un errore che non superi la quantità m prefissata, alla probabilità α?
- La differenza massima rispetto a $\hat{\pi}$ è

$$m = z^* \sqrt{\frac{\hat{\pi}\,(1 - \hat{\pi})}{n}}$$

dove z^* è il valore critico della distribuzione normale standardizzata al livello di confidenza $1 - \alpha$ e $\hat{\pi}$ è la proporzione di "successi" nel campione.

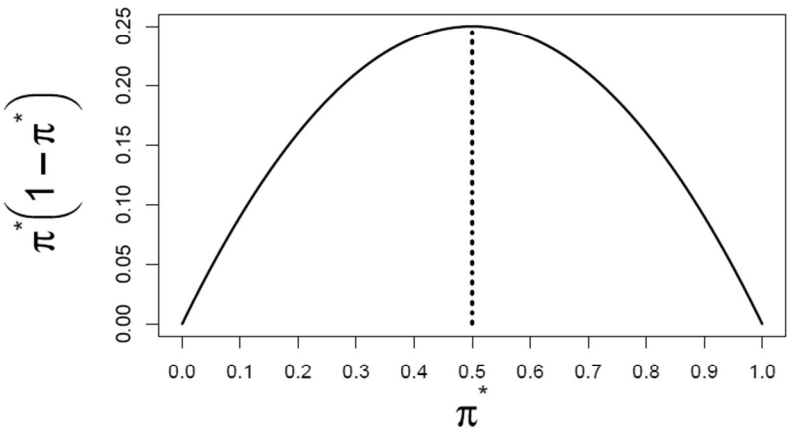
- Prima di raccogliere i dati, il valore di $\hat{\pi}$ è ignoto e, quindi, non è possibile risolvere l'equazione precedente per n al fine di determinare le dimensioni del campione.
- Una soluzione è quella di utilizzare un valore approssimativo π^* al posto di $\hat{\pi}$ per giungere alla seguente formula:

$$n = \left(\frac{z^*}{m}\right)^2 \pi^* (1 - \pi^*)$$

dove il valore π^* è desunto da ricerche precedenti (per es., un sondaggio precedente) o dalla conoscenza del fenomeno studiato.

- In alternativa, è possibile porsi nella situazione sperimentale peggiore, cioè con la varianza massima. Poiché la varianza è massima quando $\pi = 0.5$

$$\begin{aligned} \pi^* (1 - \pi^*) = &\quad 0.3\,(1 - 0.3) = 0.21 \\ = &\quad 0.5\,(1 - 0.5) = 0.25 \\ = &\quad 0.7\,(1 - 0.7) = 0.21 \end{aligned}$$



- la formula precedente diventa

$$n = \left(\frac{z^*}{m}\right)^2 0.5 \times 0.5$$

Illustrazione

Quanti votanti del Quebec è necessario intervistare per ottenere una stima di π che abbia un errore massimo di 0.01 (attorno alla $\hat{\pi}$ campionaria) con un rischio di sbagliare $\alpha = 0.05$?

Ponendo $\pi^* = 0.5$, il numero minimo di votanti da intervistare è

$$n = \left(\frac{z^*}{m}\right)^2 \pi^* (1 - \pi^*)$$

$$= \left(\frac{1.96}{0.01}\right)^2 0.25 = 9604 \approx 10000$$

Differenza tra due proporzioni

La notazione usata per confrontare due proporzioni provenienti da campioni indipendenti è simile a quella usata per confrontare due medie.

Popolazione	Dimensione del campione	Proporzione della popolazione	Proporzione del campione
1	n_1	π_1	$\hat{\pi}_1$
2	n_2	π_2	$\hat{\pi}_2$

- La differenza tra le proporzioni dei due campioni $\hat{\pi}_1 - \hat{\pi}_2$ viene usata quale stimatore della differenza tra le proporzioni $\pi_1 - \pi_2$ nella popolazione.
- Come in precendenza, la distribuzione campionaria di $\hat{\pi}_1 - \hat{\pi}_2$ fornisce le basi per l'inferenza statistica.

Si notino i seguenti tre fatti:

1. La differenza tra le proporzioni dei campioni $\hat{\pi}_1 - \hat{\pi}_2$ è uno stimatore non distorto della differenza tra le proporzioni $\pi_1 - \pi_2$ nella popolazione.

   In altri termini, $\pi_1 - \pi_2$ è la media della distribuzione campionaria di $\hat{\pi}_1 - \hat{\pi}_2$.

2. La varianza di $\hat{\pi}_1 - \hat{\pi}_2$ è uguale alla somma delle varianze di $\hat{\pi}_1$ e $\hat{\pi}_2$:

   $$\sigma^2_{\hat{\pi}_1 - \hat{\pi}_2} = \frac{\pi_1 (1 - \pi_1)}{n_1} + \frac{\pi_2 (1 - \pi_2)}{n_2}$$

   Quindi, la deviazione standard di $\hat{\pi}_1 - \hat{\pi}_2$ è

   $$\sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1 (1 - \pi_1)}{n_1} + \frac{\pi_2 (1 - \pi_2)}{n_2}}$$

3. Quando le dimensioni dei due campioni n_1 e n_2 sono sufficientemente grandi, la distribuzione campionaria di $\hat{\pi}_1 - \hat{\pi}_2$ è approssimativamente normale:

   $$\hat{\pi}_1 - \hat{\pi}_2 \sim N \left(\pi_1 - \pi_2, \quad \sqrt{\frac{\pi_1 (1 - \pi_1)}{n_1} + \frac{\pi_2 (1 - \pi_2)}{n_2}} \right)$$

Intervallo di confidenza per $\pi_1 - \pi_2$

- Come in precedenza, non possiamo utilizzare direttamente la distribuzione campionaria di $\hat{\pi}_1 - \hat{\pi}_2$ dato che la deviazione standard di $\hat{\pi}_1 - \hat{\pi}_2$ dipende dalle proporzioni sconosciute π_1 e π_2 nella popolazione.
- La soluzione è nuovamente quella di calcolare l'errore standard stimato di $\hat{\pi}_1 - \hat{\pi}_2$ utilizzando $\hat{\pi}_1$ e $\hat{\pi}_2$ quali stime di π_1 e π_2.

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1 (1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2 (1 - \hat{\pi}_2)}{n_2}}$$

L'intervallo di confidenza per $\pi_1 - \pi_2$ al livello $1 - \alpha$ è dunque

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z^* \hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}$$

dove z^* è il percentile di una variabile normale standardizzata z tale che $P(-z^* \leq z \leq z^*) = 1 - \alpha$

Illustrazione

Si costruisca l'intervallo di confidenza al 95% per la differenza tra due proporzioni utilizzando i risultati di due dei sondaggi presentati in precedenza.

Popolazione	Descrizione della popolazione	Proporzione del campione	Dimensione del campione
1	Votanti 25 ott.	$\hat{\pi}_1 = 0.535$	$n_1 = 959$
2	Votanti 12 sett.	$\hat{\pi}_2 = 0.451$	$n_2 = 822$

L'errore standard stimato di $\hat{\pi}_1 - \hat{\pi}_2$ è:

$$\begin{aligned} \hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} &= \sqrt{\frac{0.535\,(1 - 0.535)}{959} + \frac{0.451\,(1 - 0.451)}{822}} \\ &= 0.0237 \end{aligned}$$

L'intervallo di confidenza al 95% per $\pi_1 - \pi_2$ diventa

$$\begin{aligned} (\hat{\pi}_1 - \hat{\pi}_2) &\pm z^* \hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} \\ ((0.535 - 0.451)) &\pm 1.96 \times 0.0237 \\ 0.084 &\pm 0.046 = (0.038,\ 0.130) \end{aligned}$$

Test di ipotesi per $\pi_1 - \pi_2$

- Come nel caso di una singola proporzione, la stima dell'errore standard di $\hat{\pi}_1 - \hat{\pi}_2$ nel test di ipotesi è leggermente diversa da quella usata per la costruzione degli intervalli di confidenza.
- L'ipotesi nulla per la differenza tra due proporzioni è $H_0 : \pi_1 - \pi_2 = 0$ o, in maniera equivalente, $H_0 : \pi_1 = \pi_2$.
- L'ipotesi alternativa può essere unilaterale o bilaterale, a seconda di ciò che ci aspettiamo a proposito della differenza tra $\pi_1 - \pi_2$.

Dato che l'ipotesi nulla afferma che π_1 e π_2 sono uguali, per calcolare l'errore standard di $\hat{\pi}_1 - \hat{\pi}_2$ possiamo combinare i dati dei due campioni ottenendo così una stima migliore della proporzione comune π nella popolazione.

La stima $\hat{\pi}$ è data dal rapporto tra la somma dei successi in entrambi i campioni e la somma del numero di osservazioni in ciascun campione.

Test di ipotesi per $\pi_1 - \pi_2$

L'Errore standard stimato di $\hat{\pi}_1 - \hat{\pi}_2$ è

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\hat{\pi}\,(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

La statistica di test per l'ipotesi nulla $H_0 : \pi_1 - \pi_2 = 0$ è

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}}$$

e segue approssimativamente la distribuzione normale standardizza se l'ipotesi nulla è vera.

Illustrazione

Nell'esempio dei due sondaggi sull'indipendenza del Quebec, 513 su 959 votanti ($\hat{\pi}_1 = 0.535$) hanno manifestato l'intenzione di votare 'sì' nel sondaggio del 25 ottobre, mentre solo 371 su 822 votanti ($\hat{\pi}_2 = 0.451$) hanno manifestato l'intenzione di votare 'sì' nel sondaggio del 12 settembre.

- Verifichiamo l'ipotesi nulla secondo cui non c'è stato un cambiamento nelle intenzioni di voto, $H_0 : \pi_1 - \pi_2 = 0$, verso l'ipotesi alternativa secondo cui l'orientamento degli elettori è mutato, $H_a : \pi_1 - \pi_2 \neq 0$.

- Una stima della proporzione comune π è data da

$$\hat{\pi} = \frac{513 + 371}{959 + 822} = 0.496$$

- L'errore standard stimato di $\hat{\pi}_1 - \hat{\pi}_2$ è

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{0.496\,(1 - 0.496)\left(\frac{1}{959} + \frac{1}{822}\right)} = 0.0238$$

- La statistica di test per l'ipotesi nulla $H_0 : \pi_1 - \pi_2 = 0$ è

$$z = \frac{0.535 - 0.451}{0.0238} = 3.53$$

  a cui è associato il $p-$valore $P = 2 \times 0.0002 = 0.0004$
- Le evidenze empiriche sono a sostegno dell'ipotesi alternativa (bilaterale), secondo cui l'orientamento di voto degli elettori è mutato.

- `p-value` è il livello di significatività osservato.
- Poiché il $p-$valore è molto basso (< 0.001), si rifiuta l'ipotesi di uguaglianza delle proporzioni fra i due gruppi.
- La probabilità di errore di I tipo è quindi $\alpha < 0.001$

Condizioni di validità

La distribuzione campionaria di $\hat{\pi}_1 - \hat{\pi}_2$ può essere approssimata alla normale standardizzata se le seguenti condizioni risultano soddisfatte.

- Due campioni indipendenti casuali semplici vengono estratti dalle rispettive popolazioni.
- Ciascuna popolazione è almeno 10 volte più grande del campione.
- Per gli intervalli di confidenza, ciascuno dei termini $n_1\hat{\pi}_1$, $n_1\left(1-\hat{\pi}_1\right)$ e $n_2\hat{\pi}_2$, $n_2\left(1-\hat{\pi}_2\right)$ deve essere maggiore di 5.
- Per il test di ipotesi, ciascuno dei termini $n_1\hat{\pi}$, $n_1\left(1-\hat{\pi}\right)$ e $n_2\hat{\pi}$, $n_2\left(1-\hat{\pi}\right)$ deve essere maggiore di 5.

Campioni piccoli

- Le procedure che abbiamo considerato in precedenza sono adatte al caso di campioni di grandi dimensioni.
- Prenderemo ora in esame le procedure inferenziali per le proporzioni nel caso di campioni piccoli.
- La verifica di ipotesi per una proporzione fa uso del test binomiale.
- Il confronto tra le proporzioni di due campioni indipendenti fa uso del test esatto di Fisher.
- In generale, entrambe le procedure sono dette esatte dal momento che non si basano sull'approssimazione a distribuzioni note, ma piuttosto usano il vero modello probabilistico che descrive la generazione dei dati

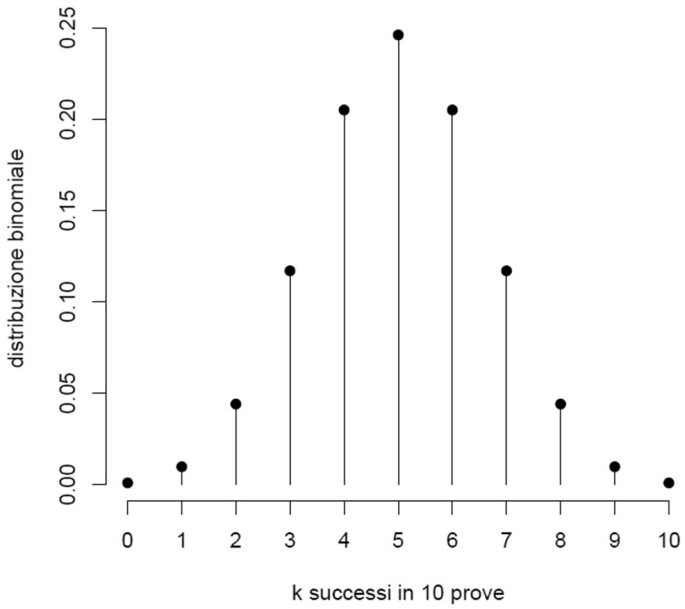
Il test binomiale

La distribuzione binomiale può essere usata per calcolare la probabilità esatta relativa ad una proporzione di successi nel caso di campioni di piccole dimensioni

- **Illustrazione.** Consideriamo l'esempio discusso nel testo e relativo ai possibili pregiudizi di genere nella selezione del personale con ruolo manageriale.
- 10 candidati devono essere selezionati da una popolazione suddivisa equamente tra maschi e femmine.
- Definiamo "successo" l'evento "selezione di una candidata".

- Se non ci fossero pregiudizi di genere, la probabilità che venga scelto un maschio sarebbe uguale alla probabilità che venga scelta una femmina.
- La scelta di un candidato (maschio o femmina) può essere considerata come una prova bernoulliana.
- In base all'ipotesi nulla (assenza di pregiudizi di genere), la probabilità di un "successo" è $p = 0.5$.
- La selezione di 10 candidati può essere considerata come una sequenza di 10 prove bernoulliane.
- La distribuzione binomiale fornisce la probabilità di osservare $k = 0; 1; \ldots; 9; 10$ successi in 10 prove bernulliane.

$X = k$	$\binom{10}{k}$	$0.5^{(k)}0.5^{(10-k)} = 0.5^{10}$	$P(X = k)$
0	1	0,000977	0,001
1	10	0,000977	0,010
2	45	0,000977	0,044
3	120	0,000977	0,117
4	210	0,000977	0,205
5	252	0,000977	0,246
6	210	0,000977	0,205
7	120	0,000977	0,117
8	45	0,000977	0,044
9	10	0,000977	0,010
10	1	0,000977	0,001



- La distribuzione binomiale con parametri $n = 10$ e $p = 0.5$ ci fornisce quindi la distribuzione campionaria della statistica corrispondente al numero X di "successi" nel campione, in assenza di pregiudizi di genere.
- Le ipotesi nulla e alternativa sono:

$$H_0 : p = 0.5 \qquad H_a : p \neq 0.5$$

- Supponiamo che nel campione sia presente una sola donna, ovvero $X = 1$. Sia $\alpha = 0.05$.

- Qual è la probabilità dell'evento $X = 1$ osservato nel campione, o di un evento ancora più estremo?
- In un test bilaterale, per una distribuzione binomiale con parametri $p = 0.5$ e $n = 10$, tale probabilità è:

$$\underbrace{P(X = 0) + P(X = 1)}_{\text{coda di sinistra} \approx 0.011} + \underbrace{P(X = 9) + P(X = 10)}_{\text{coda di destra} \approx 0.011} = 0.022$$

- Nella condizione che l'ipotesi nulla H0 sia vera, la probabilità dell'evento osservato, o ancora più estremo, è minore di α.
- Possiamo quindi rigettare l'ipotesi nulla $H_0 : p = 0.5$ e concludere che ci sono dei pregiudizi di genere nella selezione dei candidati.

Test esatto di Fisher

- Il test esatto di Fisher viene usato per confrontare due proporzioni relative a campioni indipendenti di piccole dimensioni.
- I calcoli necessari sono complessi ma possono essere eseguiti usando R.
- Il $p-$valore della statistica test, anche in questo caso, rappresenta la probabilità che venga osservata una statistica come quella calcolata sulla base dei due campioni, o ancora più estrema, se l'ipotesi nulla è vera.
- Tanto minore è il $p-$valore della statistica test, tanto maggiori sono le evidenze empiriche contrarie all'ipotesi nulla $H_0 : p_1 = p_2$.

Illustrazione. Agresti e Finlay discutono uno studio di Colombok e Tasker (1996) sull'orientamento sessuale di adulti in funzione dell'orientamento sessuale della madre. 25 figli di madri lesbiche e 20 figli di madri eterosessuali sono stati intervistati all'età 24 anni.

Una delle domande dell'intervista riguardava il loro orientamento sessuale, con risposte possibili "bisessuale/lesbica/gay" o "eterosessuale".

Illustrazione

Consideriamo innanzitutto il problema di leggere i dati all'interno di R nella forma opportuna.

```
  sexid <- matrix(c(2,23 ,0 , 20), byrow=T, ncol=2)
  sexid
     [,1] [,2]
[1,]    2   23
[2,]    0   20
  dimnames(sexid) <-
   list( mother=c("lesbian","heterosx"),
   identity=c("blg","hetero"))
  sexid
          identity
mother     blg hetero
  lesbian    2     23
  heterosx   0     20
```

Il test esatto di Fisher viene eseguito utilizzando la funzione `fisher.test`:

```
 fisher.test(sexid, alternative = "greater")

Fisher's Exact Test for Count Data

data:  sexid
p-value = 0.303
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.2316268       Inf
sample estimates:
odds ratio
       Inf
```

L'argomento `alternative = ''greater''` specifica un test unilaterale

- Il $p-$valore del test esatto di Fisher per un test unilaterale è $P = 0.303$.
- In questo caso non si può rifiutare l'ipotesi H_0: i dati forniscono evidenze insufficienti per potere concludere che un orientamento bisessuale/lesbico/gay sia più probabile per i figli di madri lesbiche.

Per un test bilaterale, $P = 0.495$.

```
 fisher.test(sexid, alternative = "two.sided")

Fisher's Exact Test for Count Data

data:  sexid
p-value = 0.4949
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1505797       Inf
sample estimates:
odds ratio
       Inf
```

Diversamente da quanto visto fino ad ora, il $p-$ valore del test bilaterale non è semplicemente il doppio di quello unilaterale: ciò dipende dall'asimmetria delle distribuzioni ipergeometriche soggiacenti al test.