alcuni esercizi (unità a-d)

guido masarotto

16 maggio 2004

Per la soluzione di alcuni degli esercizi è necessario essere in grado di calcolare i quantili e/o la funzione di ripartizione di una normale standard. Questo può essere fatto in R utilizzando le funzioni `qnorm` e `pnorm` oppure utilizzando la tabella riportata nell'ultima pagina.

---

1. In Inghilterra, dal 1200 fino a quando le monete sono state coniate in metallo prezioso, l'onestà del *Master of Mint* (il responsabile/appaltatore della zecca reale) è stata controllata mediante un controllo campionario, la cosidetta *Trial of Pyx* (*pyx* è un recipiente sacro). Il *Master of Mint*, nel caso la prova indicasse la sua disonestà, era soggetto a conseguenze non del tutto "gradevoli" (in alcuni secoli anche la condanna a morte).
I dettagli per le ghinee d'oro nel 1799 erano:
- 128 grani era il peso nominale di una ghinea (360 grani fanno un oncia);
- 100 ghinee venivano state estratte casualmente tra tutte quelle prodotte durante l'intero anno (due funzionari reali si recavano in giorni scelti casualmente disseminati durante tutto l'anno presso la zecca e sceglievano a caso una o più monete della produzione di quel giorno);
- le ghinee "estratte" venivano man mano conservate nel *pyx*;
- alla fine dell'anno il recipiente con le 100 ghinee veniva pesato;
- il *Master of Mint* "passava" la *Trial of Pyx* se il peso delle ghinee "estratte" era uguale al peso atteso più o meno 1/400 del peso atteso stesso.

Si ritiene che con la tecnologia dell'epoca il peso di ogni singola moneta si distribuisse come una normale di media controllabile dal *Master of Mint* e varianza unitaria.

(a) Calcolare la probabilità che un *Master of Mint* onesto sopravvivesse alla *Trial of Pyx*.
(b) Calcolare la probabilità che un *Master of Mint* disonesto e che avesse deciso di rubare mediamente 0,3 grani d'oro per ogni ghinea prodotta venisse scoperto.
(c) Implicitamente quale test statistico utilizzava il re per verificare l'onesta del suo *Master of Mint*?

(d) Nel contesto del test delineato al punto precedente che cosa sono e come si chiamano le due probabilità calcolate ai primi due punti dell'esercizio?

Schema di soluzione. Poniamo

$$\begin{aligned}
n &= 100 = \text{numero ghinee pesate},\\
y_i &= \text{peso della i-sima ghinea, } i = 1,\dots,n,\\
s &= \sum_{i=1}^{n} y_i = \text{peso totale delle delle ghinee estratte},\\
\overline{y} &= \frac{s}{n} = \text{peso medio delle ghinee estratte},\\
\mu &= \text{peso medio delle ghinee fissato dal } \textit{Master of Mint}.\\
\mu_0 &= 128 = \text{peso "nominale" di una ghinea}\\
\mu_1 &= 127{,}7 = \text{peso medio di una ghinea se il } \textit{Master}\\
&\qquad \text{cerca di "rubare" 0,3 grani d'oro per moneta}
\end{aligned}$$

Il *Master* passa la prova se

$$n\mu_0 - \frac{n\mu_0}{400} \leqslant s \leqslant n\mu_0 + \frac{n\mu_0}{400}.$$

Dividendo per n[1] possiamo riscrivere la "regola" utilizzata dal re come

"se $|\overline{y} - \mu_0| \leqslant \dfrac{\mu_0}{400}$ allora il *Master of Mint* viene dichiarato onesto".

Supponendo che i pesi delle singole monete siano tra di loro indipendenti[2] sappiamo che

$$\overline{y} \sim N\left(\mu, \frac{1}{n}\right) \quad \Leftrightarrow \quad \sqrt{n}(\overline{y} - \mu) \sim N(0,1). \tag{1}$$

Per rispondere alle prime due domande dobbiamo calcolare

P(il *master* venga dichiarato onesto quando fissa la media a μ)

ponendo nel caso della prima domanda $\mu = \mu_0$ e nel caso della seconda $\mu = \mu_1$. Ci conviene quindi calcolare la probabilità di sopra una volta per tutte per un valore di μ qualsiasi:

$$\begin{aligned}
&\text{P(il } \textit{Master} \text{ viene dichiarato onesto se } \mu \text{ è la media)} =\\
&= \text{P}\left(\mu_0 - \frac{\mu_0}{400} \leqslant \overline{y} \leqslant \mu_0 + \frac{\mu_0}{400}\right) =\\
&= \text{P}\left(\mu_0 - \mu - \frac{\mu_0}{400} \leqslant \overline{y} - \mu \leqslant \mu_0 - \mu + \frac{\mu_0}{400}\right) =\\
&= \text{P}\left(\sqrt{n}\left(\mu_0 - \mu - \frac{\mu_0}{400}\right) \leqslant \sqrt{n}(\overline{y} - \mu) \leqslant \sqrt{n}\left(\mu_0 - \mu + \frac{\mu_0}{400}\right)\right) =
\end{aligned}$$

[1] visto che a lezione abbiamo lavorato con le "medie" e non con i totali continuo a lavorare così

[2] questo non è scritto esplicitamente nel testo ma neanche negato e simile a quanto fatto nell'unità B

$$= \mathrm{P}\left(\sqrt{n}\left(\mu_0 - \mu - \frac{\mu_0}{400}\right) \leqslant \mathrm{N}(0,1) \leqslant \sqrt{n}\left(\mu_0 - \mu + \frac{\mu_0}{400}\right)\right) =$$
$$= \Phi\left(\sqrt{n}\left(\mu_0 - \mu + \frac{\mu_0}{400}\right)\right) - \Phi\left(\sqrt{n}\left(\mu_0 - \mu - \frac{\mu_0}{400}\right)\right) =$$

Alla terza riga ho diviso per lo scarto quadratico medio di $\overline{y}$[3] in maniera tale da poter utilizzare la (1) nei passaggi successivi. Nell'ultima riga $\Phi(\cdot)$ indica al solito la funzione di ripartizione di una $\mathrm{N}(0,1)$.

(a) Un *Master* onesto cerca di fissare μ uguale μ_0. Sperando (per la sua testa!) che non sbagli, troviamo

$$\mathrm{P}(\text{un } M. \text{ onesto viene dichiarato onesto}) =$$
$$= \Phi\left(\frac{128\sqrt{10}}{400}\right) - \Phi\left(-\frac{128\sqrt{10}}{400}\right) = \Phi(3{,}2) - \Phi(-3{,}2) = 2\Phi(3{,}2) - 1.$$

L'ultimo passaggio è una conseguenza del fatto che la simmetria della $\mathrm{N}(0,1)$ ci permette di scrivere

$$\Phi(-x) = 1 - \Phi(x).$$

Da una tabella dei quantili di una normale standard possiamo osservare che

$$3{,}2 > 3{,}09 = \text{quantile } 0{,}999 \text{ di una } \mathrm{N}(0,1).$$

Ma allora

$$\Phi(3{,}2) = \mathrm{P}(\mathrm{N}(0,1) \leqslant 3{,}2) > \mathrm{P}(\mathrm{N}(0,1) \leqslant 3{,}09) = 0{,}999$$

e quindi la probabilità cercata è maggiore di $2 \times 0.999 - 1 = 0{,}998$, ovvero un *Master of Mint* onesto rischia, al più, "di perdere la sua testa" una volta ogni 500 anni.
Utilizzando R per calcolare la probabilità troviamo

```
> 2*pnorm(3.2)-1
[1] 0.9986257
```

(b) In questo caso μ è posto dal *Master of Mint* uguale a $\mu_1 = \mu_0 - 0,3 = 127{,}7$. Ovvero $\mu_0 - \mu = 0{,}3$. Quindi

$$\mathrm{P}(\text{il } \textit{Master} \text{ viene dichiarato onesto}) =$$
$$= \Phi\left(\sqrt{100}\left(0{,}3 + \frac{128}{400}\right)\right) - \Phi\left(\sqrt{100}\left(0{,}3 - \frac{128}{400}\right)\right) =$$
$$= \Phi(6{,}2) - \Phi(-0{,}2)$$

Sappiamo che

$$\Phi(6{,}2) \approx 1.$$

[3] ovvero per $1/\sqrt{n}$

Inoltre, utilizzando la tabella dei quantili della normale e ricordandoci della simmetria della distribuzione, troviamo

$$\Phi(-0{,}2) = 1 - \Phi(0{,}2) \approx 1 - 0{,}58 = 0{,}42$$

e quindi

$$\mathrm{P}(\textit{Master} \text{ disonesto la faccia franca}) = \Phi(6{,}2) - \Phi(-0{,}2) \approx 1 - 0{,}42 = 0{,}58.$$

In R

```
> pnorm(6.2)-pnorm(-0.2)
[1] 0.5792597
```

La probabilità che "venga scoperto" è perciò, approssimativamente, $1 - 0{,}58 = 0{,}42$.

(c) Il re vuole verificare il sistema di ipotesi

$$\begin{cases} H_0 : \textit{Master} \text{ onesto} \\ H_1 : \textit{Master} \text{ disonesto} \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Il contesto e il test utilizzato sono quelli sulla media di una normale di varianza nota considerato nella seconda parte dell'unità B. Quello che sui lucidi è indicato con h[4] è in questo caso posto uguale a $\sqrt{n}\mu_0/400 = 3{,}2$. Questo, per quando visto al punto (a), garantisce che

$$0{,}998 < \mathrm{P}(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) < 1.$$

(d) Come appena notato la probabilità calcolata al punto (a) è

$$\mathrm{P}(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera})$$

che possiamo anche indicare come la probabilità di non commettere un errore di I tipo.
Al punto (b) dell'esercizio abbiamo viceversa calcolato la probabilità che il test non commetta un errore di II tipo quando la vera media è uguale a 127,7. Si osservi inoltre che se $\gamma(\mu)$ indica la funzione di potenza allora

$$\text{probabilità punto (a)} = 1 - \gamma(128)$$
$$\text{probabilità punto (a)} = \gamma(127{,}7)$$

---

2. Per misurare la concentrazione di Pb (in $\mu g g^{-1}$) si procede nella seguente maniera:

(i) il materiale originario viene diviso in n *pezzettini*;

[4] la soglia con cui confrontare il valore della statistica test

(ii) su ciascun *pezzettino* viene misurata la concentrazione utilizzando uno strumento appropriato;
(iii) la stima della concentrazione di Pb viene infine calcolata facendo la media aritmetica delle n misure ottenute.

Sapendo che gli errori commessi dallo strumento utilizzato si distribuiscono (almeno approssimativamente) come delle normali di media zero e scarto quadratico medio 0,2, dire quanto deve essere grande n affinchè almeno 9 volte su 10

$$|(\text{stima della concentrazione}) - (\text{vera concentrazione})| < 0{,}05.$$

Schema di soluzione. Poniamo

$$\begin{aligned}
\mu &= \text{vera concentrazione}\\
y_i &= \text{misura sull'i-simo pezzettino}\\
\overline{y} &= \sum_{i=1}^{n} y_i/n = \text{stima concentrazione}\\
\sigma &= 0{,}2 = \text{scarto quadratico medio dell'errore di una singola misura}\\
\delta &= 0{,}05 = \text{massimo errore da commettere 9 volte su 10}\\
1-\alpha &= 0{,}9 = \text{probabilità desiderata di un errore inferiore a } \delta.
\end{aligned}$$

Quello che viene richiesto è di determinare n in maniera tale che

$$P(|\overline{y} - \mu| < \delta) = 1 - \alpha.$$

Ma allora, per definizione di intervallo di confidenza,

$$\overline{y} \pm \delta$$

è un intervallo di confidenza di livello $1-\alpha$ per μ. Sappiamo che questi, nel presente contesto, sono del tipo

$$\overline{y} \pm \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}.$$

Quindi, affinchè sia soddisfatta la condizione desiderata dobbiamo scegliere n in maniera tale che

$$\frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} = \delta.$$

Risolvendo in n questa equazione troviamo

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{\delta}\right)^2.$$

Con i dati del problema, osservando nella tabella dei quantili di una $N(0,1)$ che $z_{0,95} = 1{,}645$, troviamo

$$n = \left(\frac{1{,}645 \times 0{,}2}{0{,}05}\right)^2 \approx 43{,}3.$$

Ovviamente un numero frazionale di misure non possiamo farle. Osservando che più piccolo è n più è piccolo l'intervallo possiamo scegliere $n = 44$. Infatti, così facendo ci garantiamo che

$$P(|\overline{y} - \mu| < \delta) \geqslant 1 - \alpha.$$

---

3. Nella contea di *Chiare Acque* da anni viene condotta un'indagine per stimare la proporzione di castelli abitati da fate. L'indagine è campionaria: ogni anno, sono considerati 100 castelli estratti a caso e per ogni castello viene rilevato se ci abitano o meno delle fate.

1. Sapendo che quest'anno il numero di castelli con fate nel campione è risultato pari a 38, costruire un intervallo che includa la vera proporzione con una probabilità (almeno approssimativamente) uguale a 0,90.
2. In un vecchio libro, il duca di *Dolci Acque* ha scritto che circa il 50% dei castelli della contea di *Chiare Acque* è abitato da fate. Utilizzare un appropriato test statistico per verificare se i dati sono compatibili con questa affermazione.
3. La vicina contea di *Fresche Acque* vuole da quest'anno condure una indagine analoga. Durante una riunione il ciambellano addetto ai castelli fatati dice: "*La nostra contea ha il doppio di castelli della contea di* Chiare Acque*. Quindi, affinchè la stima della nostra indagine abbia la stessa precisione dobbiamo ogni anno estrarre* 200 *castelli*". Siete d'accordo con il ciambellano?

Schema di soluzione. Certamente la densità dei castelli, fatati o meno, è altissima nelle contee di *Chiare e Fresche Acque*. Quindi, o esattamente se l'estrazione è fatta con reintroduzione o approssimativamente se l'estrazione avviene senza reintroduzione, possiamo supporre che per quanto ci riguarda siano nel *Reame della Binomiale*. Ovvero, posto

$$\begin{aligned}
y &= \text{numero castelli fatati estratti}\\
n &= \text{numero castelli estratti e ispezionati}\\
\vartheta &= \text{percentuale di castelli fatati nella contea}\\
\hat{\vartheta} &= \frac{y}{n} = \text{stima della percentuale di castelli fatati nella contea}
\end{aligned}$$

assumeremmo che

$$y \sim \mathrm{Bi}(n, \vartheta).$$

(a) La prima domanda richiede di calcolare, quando $y = 38$ e $n = 100$ un intervallo di confidenza che includa con probabilità uguale (almeno approssimativamente) a 90% il vero valore di ϑ. Sappiamo dall'unità C che una soluzione è l'intervallo

$$\hat{\vartheta} \pm z_{0,95}\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}$$

che nel nostro caso diventa

$$0{,}38 \pm 1{,}645\sqrt{\frac{0{,}38(1-0{,}38)}{100}} \approx [0{,}30 - 0{,}46].$$

(b) La seconda domanda richiede di verificare, utilizzando gli stessi dati di prima, il sistema di ipotesi

$$\begin{cases} H_0 : \vartheta = 0{,}5 (= \vartheta_0) \\ H_1 : \vartheta \neq 0{,}5 (\neq \vartheta_0) \end{cases}$$

Calcoliamo la statistica test usuale

$$(\hat{\vartheta} - \vartheta_0)/\sqrt{\vartheta_0(1-\vartheta_0)/n} = (0{,}38 - 0{,}5)/\sqrt{0{,}5 \times 0{,}5/100} = -2{,}4.$$

Il valore ottenuto deve essere confrontato con i valori previsti da una $N(0,1)$, distribuzione che descrive i valori che ci attendiamo per la statistica test quando è vera H_0.
[versione accetto/rifiuto] le probabilità di errore di I tipo (α) usualmente considerate in un approccio di questo tipo e i quantili corrispondenti sono

$$\begin{aligned} \alpha = 0{,}1 &\Rightarrow z_{1-\alpha/2} = 1{,}645 \\ \alpha = 0{,}05 &\Rightarrow z_{1-\alpha/2} = 1{,}960, \\ \alpha = 0{,}01 &\Rightarrow z_{1-\alpha/2} = 2{,}576. \end{aligned}$$

Il valore osservato per la statistica ci porterebbe quindi a rifiutare H_0 per $\alpha = 0{,}1$ e $\alpha = 0{,}05$ e ad accettare H_0 se fissiamo $\alpha = 0{,}01$. Risultati di questo tipo sono normalmente descritti come significativi (ma non altamente significativi) contro H_0.
[p-value] può essere calcolato come

$$2P(N(0,1) > 2{,}4).$$

Utilizzando R lo calcoliamo come

```
> 2*(1-pnorm(2.4))
[1] 0.01639507
```

Se abbiamo a disposizione solo una tabella dei quantili di una normale possiamo osservare, ad esempio, che

$$\begin{aligned} P(N(0,1) \geqslant 2{,}326) &= 0{,}01 \\ P(N(0,1) \geqslant 2{,}576) &= 0{,}005 \end{aligned}$$

e, visto che ovviamente

$$P(N(0,1) \geqslant 2{,}576) < P(N(0,1) \geqslant 2{,}4) < P(N(0,1) \geqslant 2{,}326)$$

concludere che

$$0{,}01 < (\text{livello significatività osservato}) < 0{,}02.$$

La interpretazione è la stessa di prima: i valori ottenuti fanno "sospettare" di H_0 ma non proprio "escluderla".
In definitiva i risultati suggeriscono che la situazione dei castelli fatati nella contea di *Chiare Acque* dovrebbe essere cambiata dagli anni del viaggio e del relativo libro del Duca di *Dolci Acque*. L'evidenza a questo proposito non è però fortissima.

(c) Il ciambellano è fuori strada. Quando lo incontreremo ricordiamoci di fargli osservare cose del tipo:
  (i) la distribuzione dell'errore di stima della indagine dipende dalla percentuale di castelli fatati nella sua contea (ϑ) e dal numero di castelli "ispezionati" (n);
  (ii) come conseguenza anche l'ampiezza degli intervalli di confidenza dipende solamente da $\hat{\vartheta}$ (e quindi indirettamente da ϑ), da n e ovviamente anche dalla copertura desiderata (α).

Nessuna traccia di dipendenza dal numero complessivo di castelli che, quindi, sembra irrilevante per scegliere un valore appropriato per n.
[nota importante] Le affermazioni precedenti dipendono in maniera cruciale dal tipo di campionamento adottato (e in questo caso ipotizzato). Sono vere nel caso di un campionamento con reintroduzione. Ma non nel caso di un campionamento senza reintroduzione[5]

---

4. Prima di un referendum sono state intervistate 2605 persone estratte a caso tra gli aventi diritto al voto. Di questi, 1207 hanno dichiarato di non avere intenzione di andare a votare. Sulla base di questi dati quale delle seguenti affermazioni fareste:
(A) il *quorum* sarà certamente raggiunto;
(B) è molto plausibile che il *quorum* venga raggiunto;
(C) non posso concludere se il *quorum* verrà o non verrà raggiunto.
(D) è poco plausibile che il *quorum* venga raggiunto.
Schema di soluzione. Possiamo pensare di essere nel contesto di un campionamento di tipo binomiale. Il parametro di interesse (la probabilità di successo della binomiale) è in questo caso la percentuale di elettori che si recheranno a votare. Una possibilità per cercare di rispondere alla domanda consiste nel calcolare e commentare opportunamente un intervallo di confidenza per questa percentuale. Scegliamo di farlo utilizzando una probabilità di copertura pari al 99%.

$$\hat{\vartheta} = \frac{y}{n} = \frac{1207}{2605} \approx 0{,}463$$

[5] si pensi alla situazione in cui ci siano 100 castelli nella contea; se "campioniamo con reintroduzione 100 castelli" li prendiamo tutti e quindi non abbiamo nessun errore di stima; viceversa se ne "campioniamo con reintroduzione 100 castelli" ma ce ne sono 1000 nella contea qualche errore possiamo commetterlo; quindi la distribuzione dell'errore di stima nei due casi non può essere la stessa.

$$\begin{aligned}
\alpha &= 0{,}01 \Rightarrow z_{1-\alpha/2} = 2{,}576 \\
z_{1-\alpha/2}\sqrt{\frac{\hat\vartheta(1-\hat\vartheta)}{n}} &\approx 2{,}576\sqrt{\frac{0{,}463(1-0{,}463)}{2605}} \approx 0{,}025 \\
\hat\vartheta \pm z_{1-\alpha/2}\sqrt{\frac{\hat\vartheta(1-\hat\vartheta)}{n}} &\approx 0{,}463 \pm 0{,}025 = [0{,}438 - 0{,}488]
\end{aligned}$$

L'intervallo di confidenza indica che i valori per la percentuale di votanti "plausibili" sulla base dei dati sono tutti più piccoli del 50%. Difficile dire con certezza che cosa accadrà. Però sembra poco plausibile sulla base di questi dati che il *quorum* venga raggiunto (affermazione D).

---

5. Dall'agenzia che conduce le indagini di mercato per l'azienda in cui lavorate ricevete un rapporto contenente la seguente frase:

> "Sulla base di <numero illegibile> interviste telefoniche, possiamo dire che la percentuale di donne tra i 18 e i 25 anni interessate al nuovo prodotto che state per immettere sul mercato è compresa tra il 24% e il 32% con una probabilità pari al 90%."

Secondo voi, quante interviste telefoniche sono state fatte?
Schema di soluzione. E' verosimile che sia stato utilizzato un intervallo di confidenza basato sulla binomiale del tipo

$$\hat\vartheta \pm z_{1-\alpha/2}\sqrt{\frac{\hat\vartheta(1-\hat\vartheta)}{n}}.$$

Il punto centrale di questo intervallo è $\hat\vartheta$. Quindi nel caso in esame deve essere

$$\hat\vartheta = \frac{0{,}24 + 0{,}32}{2} = 0{,}28.$$

Inoltre il livello di copertura dell'intervallo è 90%. Quindi,

$$\alpha = 0{,}1 \Rightarrow z_{1-\alpha/2} = 1{,}645.$$

Sfruttando la semiampiezza dell'intervallo dato possiamo scrivere

$$z_{1-\alpha/2}\sqrt{\frac{\hat\vartheta(1-\hat\vartheta)}{n}} = \frac{0{,}32 - 0{,}24}{2} = 0{,}04$$

dove l'unica incognita rimasta dopo le considerazioni precedenti è proprio n. Risolvendo quindi per n troviamo

$$n = \left(\frac{z_{1-\alpha/2}}{0{,}04}\right)^2 \hat\vartheta(1-\hat\vartheta) = \left(\frac{1{,}645}{0{,}04}\right)^2 0{,}28(1-0{,}28) \approx 350.$$

---

6. Un gruppo di medici vuole stimare l'efficacia di un protocollo di terapia recentemente proposto per curare una certa patologia. Ha quindi deciso di utilizzare il nuovo protocollo per i prossimi pazienti e di rilevare su di essi, dopo un tempo appropriato, il carattere dicotomico "guarito" o "non guarito".
Si indichi con n il numero di pazienti che "entreranno" nello studio e con y il numero di pazienti che guarirà. Si assuma inoltre che $y \sim \mathrm{Bi}(n, \vartheta)$ dove ϑ denota la probabilità che un paziente trattato guarisca e si ponga

$$\hat\vartheta = \frac{y}{n}$$

(ovvero $\hat\vartheta$ è l'usuale stima di ϑ calcolata dai dati).
Si determini n in maniera tale che

$$\mathrm{P}\left(\left|\hat\vartheta - \vartheta\right| \leqslant 0{,}02\right) \geqslant 0{,}99.$$

Schema di soluzione. Definiamo

$$\begin{aligned}
\delta &= 0{,}02, \\
\alpha &= 0{,}01. \\
z_{1-\alpha/2} &= 2{,}576 = \left(\begin{array}{c}\text{percentile } 1-\alpha/2 \\ \text{di una } \mathrm{N}(0,1)\end{array}\right)
\end{aligned}$$

Il problema chiede di determinare la numerosità campionaria (n) in maniera tale che l'errore di stima ($\hat\vartheta - \vartheta$) sia più piccolo, in valore assoluto, di δ con probabilità maggiore di $1-\alpha$.
Ricordando quello che sappiamo sugli intervalli di confidenza per la probabilità di successo di una binomiale quanto richiesto accade se

$$z_{1-\alpha/2}\sqrt{\frac{\vartheta(1-\vartheta)}{n}} \leqslant \delta.$$

Isolando n nella disequazione precedente troviamo

$$n \geqslant \left(\frac{z_{1-\alpha/2}}{\delta}\right)^2 \vartheta(1-\vartheta).$$

L'ultima disequazione deve essere soddisfatta per ogni θ quindi deve risultare

$$n \geqslant \left(\frac{z_{1-\alpha/2}}{\delta}\right)^2 \sup_{\vartheta\in[0,1]} \vartheta(1-\vartheta).$$

E' facile verificare che

$$\sup_{\vartheta\in[0,1]} \vartheta(1-\vartheta) = \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4}.$$

Infatti

$$\frac{d\vartheta(1-\vartheta)}{d\vartheta} = 1-2\vartheta = \begin{cases} > 0 \text{ se } 0 \leqslant \vartheta < \dfrac{1}{2} \\ = 0 \text{ se } \vartheta = \dfrac{1}{2} \\ < 0 \text{ se } \dfrac{1}{2} < \vartheta \leqslant 1 \end{cases}$$

e quindi $\vartheta(1-\vartheta)$ è crescente tra 0 e 1/2 e decrescente tra 1/2 e 1 ovvero ha un massimo quando $\vartheta = 1/2$.
In definitiva troviamo

$$n \geqslant \left(\frac{z_{1-\alpha/2}}{\delta}\right)^2 \frac{1}{4} = \left(\frac{2{,}576}{0{,}02}\right)^2 \frac{1}{4} \approx 4147{,}4.$$

e poichè, per considerazioni sia etiche che di tempo/costo, è meglio limitare il più possibile il numero di pazienti coinvolti sembra naturale scegliere la numerosità campionaria più bassa tra quelle che garantiscono la precisione richiesta ovvero porre $n = 4148$.
[nota] Il problema non precisava possibili valori per ϑ. Lo abbiamo risolto quindi "difendendoci" rispetto alla situazione "meno favorevole". Spesso nelle applicazioni esistono delle informazioni a priori su ϑ che possono essere utilizzate per determinare la numerosità campionaria. Ad esempio se ci aspetta che $\vartheta \approx 0{,}85$ potremmo porre

$$n \approx \left(\frac{z_{1-\alpha/2}}{\delta}\right)^2 0{,}85(1-0{,}85) = \left(\frac{2{,}576}{0{,}02}\right)^2 0{,}85(1-0{,}85) \approx 2115.$$

Ovviamente procedendo in questa maniera non siamo sicuri di riuscire a rispettare con certezza la condizione richiesta però, come si può vedere anche dall'esempio numerico, si può arrivare ad un valore di n inferiore.

---

7. Siano $(y_1, x_1), \ldots, (y_n, x_n)$ n determinazioni indipendenti tratte da una variabile casuale bivariata (X, Y) e si indichi con r_n in coefficiente di correlazione calcolato con queste osservazioni. Si indichi viceversa con ρ il coefficiente di correlazione esistente tra (Y, X) (ovvero il coefficiente di correlazione "nella popolazione").

(a) Sotto ipotesi deboli, che però non precisiamo, è possibile far vedere che quando ρ è uguale a zero la distribuzione di

$$z_n = \frac{r_n\sqrt{n-2}}{\sqrt{1-r_n^2}}$$

può essere approssimata, se n è sufficientemente grande, da una distribuzione normale standard[6]. Utilizzare questo risultato per costruire un test per il sistema d'ipotesi

$$H_0 \ : \ \rho = 0$$
$$H_1 \ : \ \rho \neq 0$$

ovvero per verificare l'ipotesi che nella popolazione non esista correlazione.

(b) Si supponga che $n = 100$ e che per certi dati sia risultato $r_n = 0{,}17$. A quali conclusione arrivate utilizzando il test delineato al punto precedente?

Schema di soluzione. L'esercizio da un lato vuole fornire uno strumento utile (come verificare la presenza di correlazione tra due variabili) e dall'altro illustrare come conoscendo una statistica test appropriata e conoscendone la distribuzione campionaria "sotto" l'ipotesi nulla si è in grado di costruire autonomamente un test statistico.

(a) Per il sistema d'ipotesi dato è intuitivamente plausibile pensare di utilizzare r_n come statistica test. Valori di r_n troppo lontani da zero (positivi o negativi) saranno ovviamente da interpretare come evidenza contro H_0.
Si ponga ora $f(x) = x/\sqrt{1-x^2}$ e si osservi che
(i) $f(0) = 0$;
(ii) $f(x) = -f(-x)$ ovvero che la funzione, come si usa dire, è dispari;
(iii) $f(x)$ è monotona crescente se $0 < x < 1$; infatti, derivando troviamo

$$\frac{df(x)}{dx} = \frac{1}{\sqrt{1-x^2}} + \frac{x^2}{(1-x^2)^{\frac{3}{2}}} > 0 \text{ se } 0 < x < 1.$$

Queste proprietà fanno si che un test che rifiuta per $|r_n|$ troppo grande sia equivalente ad un test che rifiuta per $|z_n|$ troppo grande. Possiamo quindi "ridefinire" la statistica test decidendo di utilizzare z_n al posto di r_n. Fatta questa scelta la meccanica del test diventa quella descritta nelle unità B e C con riferimento alla media della normale (con varianza nota) e alla probabilità di successo di una binomiale. Infatti la distribuzione sotto H_0 della statistica prescelta è anche in questo caso (solo asintoticamente) una $N(0,1)$.

(b) Calcoliamo

$$z_n = \frac{r_n\sqrt{n-2}}{\sqrt{1-r_n^2}} = \frac{0{,}17\sqrt{100-2}}{\sqrt{1-0{,}17^2}} \approx 1{,}71.$$

Questo valore va confrontato con i valori "attesi" per una $N(0,1)$. Poichè "lontano da H_0" vuol dire in questo caso $|z_n|$ grande il livello di significativita osservato in questo caso deve essere calcolato come

$$2(1-\Phi(1{,}71)) \approx 0{,}09.$$

Siamo nella coda destra dei valori attesi sotto l'ipotesi nulla: non sufficientemente a destra per concludere contro l'ipotesi nulla ma neanche sufficientemente vicino a zero per concludere che i dati non contengono "nessun suggerimento" contro H_0. Possiamo quindi concludere a favore di una "dubbiosa" accettazione che tra i due fenomeni considerati non ci sia correlazione.

[6] l'approssimazione è considerata ragionevole se $n > 50$.

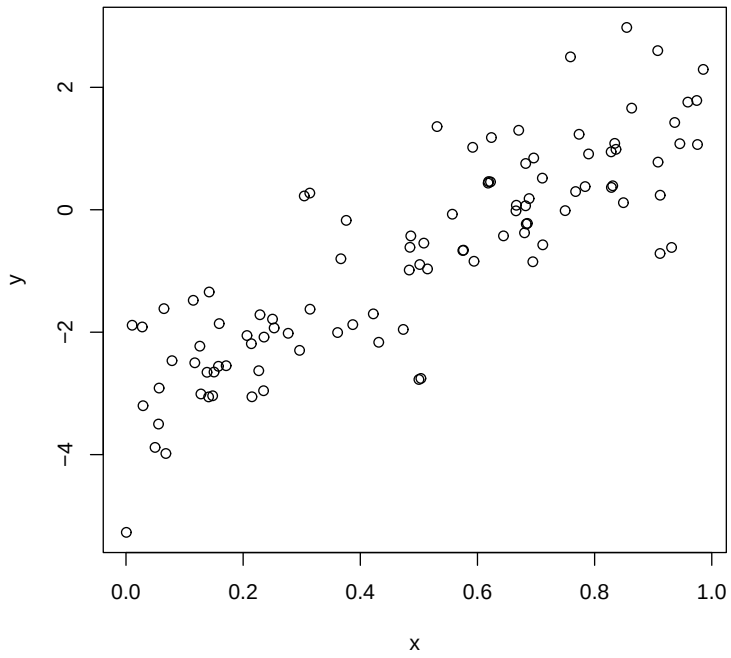

Figura 1: Diagramma di dispersione

8. Si considerino i dati bivariati della figura 1. Calcolando il livello di significatività osservato del test descritto all'esercizio precedente in quale dei seguenti intervalli

$$[-1;-0{,}5) \quad [-0{,}5;-0{,}05) \quad [-0{,}05;0) \quad [0;0{,}05) \quad [0{,}05;0{,}5) \quad [0{,}5;1]$$

vi aspettate che cada?

Schema di soluzione. Il livello di significatività osservato non può certamente essere negativo (è una probabilità). Questo eclude i primi tre intervalli.
Inoltre, il diagramma mostra chiaramente la presenza di una correlazione lineare. Ci aspettiamo quindi che il test la segnali. Il valore della statistica test che ci attendiamo sarà quindi più grande dei valori "prevedibili" da una normale standard. Per questo motivo non ci aspettiamo che il livello di significatività osservato cada in uno degli ultimi due intervalli.
In definitiva l'intervallo $[0;0{,}05]$ dovrebbe contenere il livello di significatività osservato.
[nota per i curiosi] il livello di significatività calcolato con i dati nella figura è dell'ordine di 10^{-60}.

quantili di una distribuzione normale standard

La tabella riporta i quantili di dimensione $p_0 + p_1$ di una $N(0,1)$. Ad esempio, 0,292 è il quantile 0,615 ovvero $\Pr(N(0,1) \leqslant 0{,}292) = 0{,}615$.

	p_0				
p_1	0.5	0.6	0.7	0.8	0.9
0	0.000	0.253	0.524	0.842	1.282
0.005	0.013	0.266	0.539	0.860	1.311
0.01	0.025	0.279	0.553	0.878	1.341
0.015	0.038	0.292	0.568	0.896	1.372
0.02	0.050	0.305	0.583	0.915	1.405
0.025	0.063	0.319	0.598	0.935	1.440
0.03	0.075	0.332	0.613	0.954	1.476
0.035	0.088	0.345	0.628	0.974	1.514
0.04	0.100	0.358	0.643	0.994	1.555
0.045	0.113	0.372	0.659	1.015	1.598
0.05	0.126	0.385	0.674	1.036	1.645
0.055	0.138	0.399	0.690	1.058	1.695
0.06	0.151	0.412	0.706	1.080	1.751
0.065	0.164	0.426	0.722	1.103	1.812
0.07	0.176	0.440	0.739	1.126	1.881
0.075	0.189	0.454	0.755	1.150	1.960
0.08	0.202	0.468	0.772	1.175	2.054
0.085	0.215	0.482	0.789	1.200	2.170
0.09	0.228	0.496	0.806	1.227	2.326
0.095	0.240	0.510	0.824	1.254	2.576
0.099	0.251	0.522	0.838	1.276	3.090