Un percorso: materiali per gli studenti

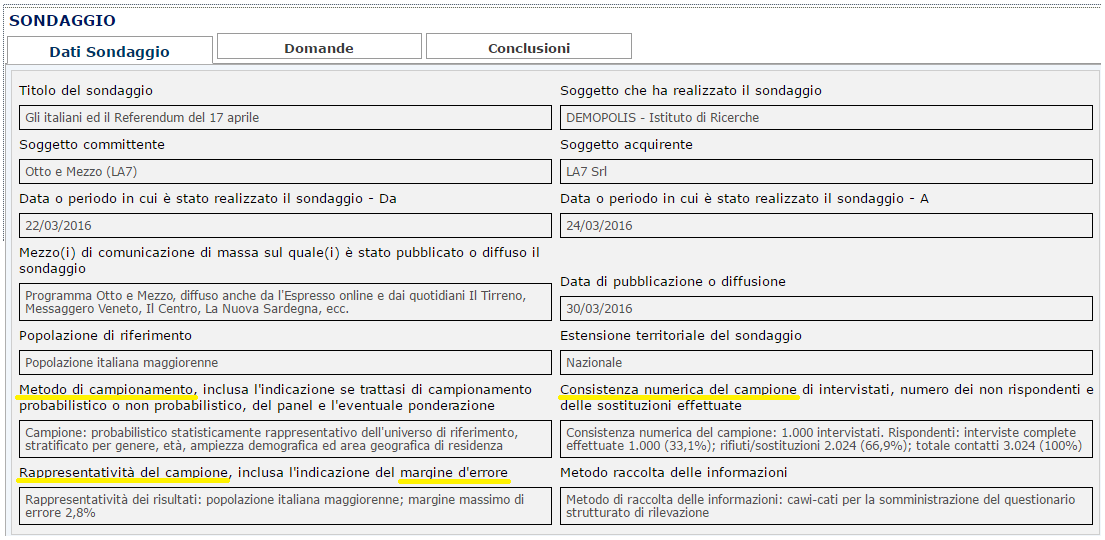
1 Motivazioni alla normale

1.1 I sondaggi

Per cercare di stabilire la posizione di una popolazione numerosa riguardo ad una determinata questione, sono molte le istituzioni che effettuano sondaggi[1] nei più svariati ambiti.

In particolare, i sondaggi che hanno valenza politica ed elettorale e che sono diffusi al pubblico, devono[2] essere pubblicati sul sito
www.sondaggipoliticoelettorali.it a cura del Dipartimento per l'Informatica e l'Editoria.

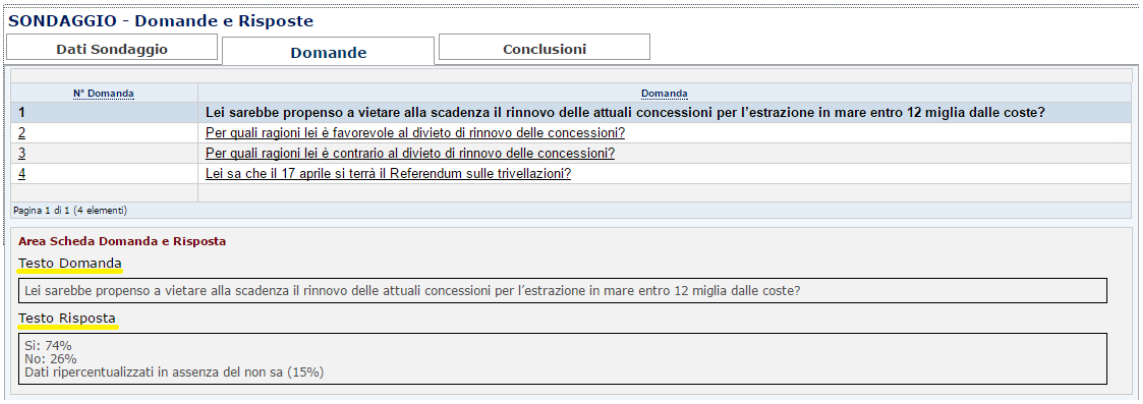
Vediamone uno: *Gli italiani ed il Referendum del 17 Aprile*



SONDAGGIO

Dati Sondaggio | Domande | Conclusioni

Titolo del sondaggio: Gli italiani ed il Referendum del 17 aprile	Soggetto che ha realizzato il sondaggio: DEMOPOLIS - Istituto di Ricerche
Soggetto committente: Otto e Mezzo (LA7)	Soggetto acquirente: LA7 Srl
Data o periodo in cui è stato realizzato il sondaggio - Da: 22/03/2016	Data o periodo in cui è stato realizzato il sondaggio - A: 24/03/2016
Mezzo(i) di comunicazione di massa sul quale(i) è stato pubblicato o diffuso il sondaggio: Programma Otto e Mezzo, diffuso anche da l'Espresso online e dai quotidiani Il Tirreno, Messaggero Veneto, Il Centro, La Nuova Sardegna, ecc.	Data di pubblicazione o diffusione: 30/03/2016
Popolazione di riferimento: Popolazione italiana maggiorenne	Estensione territoriale del sondaggio: Nazionale
Metodo di campionamento, inclusa l'indicazione se trattasi di campionamento probabilistico o non probabilistico, del panel e l'eventuale ponderazione: Campione: probabilistico statisticamente rappresentativo dell'universo di riferimento, stratificato per genere, età, ampiezza demografica ed area geografica di residenza	Consistenza numerica del campione di intervistati, numero dei non rispondenti e delle sostituzioni effettuate: Consistenza numerica del campione: 1.000 intervistati. Rispondenti: interviste complete effettuate 1.000 (33,1%); rifiuti/sostituzioni 2.024 (66,9%); totale contatti 3.024 (100%)
Rappresentatività del campione, inclusa l'indicazione del margine d'errore: Rappresentatività dei risultati: popolazione italiana maggiorenne; margine massimo di errore 2,8%	Metodo raccolta delle informazioni: Metodo di raccolta delle informazioni: cawi-cati per la somministrazione del questionario strutturato di rilevazione



SONDAGGIO - Domande e Risposte

Dati Sondaggio | Domande | Conclusioni

N° Domanda	Domanda
1	Lei sarebbe propenso a vietare alla scadenza il rinnovo delle attuali concessioni per l'estrazione in mare entro 12 miglia dalle coste?
2	Per quali ragioni lei è favorevole al divieto di rinnovo delle concessioni?
3	Per quali ragioni lei è contrario al divieto di rinnovo delle concessioni?
4	Lei sa che il 17 aprile si terrà il Referendum sulle trivellazioni?

Pagina 1 di 1 (4 elementi)

Area Scheda Domanda e Risposta

Testo Domanda: Lei sarebbe propenso a vietare alla scadenza il rinnovo delle attuali concessioni per l'estrazione in mare entro 12 miglia dalle coste?

Testo Risposta:
Si: 74%
No: 26%
Dati ripercentualizzati in assenza del non sa (15%)

Esaminiamo i termini per noi più significativi:

[1] I sondaggi vengono spesso fatti online mediante software quali cawi-cati, come nell'esempio riportato in figura (è indicato nella sezione "Metodo raccolta delle informazioni").

[2] Ciò è previsto dalla legge n. 28 del 22 Febbraio 2000, sulla par condicio.

- *Campione.* È importante avere informazioni sull'orientamento di voto di una popolazione o sull'opinione in merito ad una data questione. Un modo per ottenerlo è mediante le elezioni o i referendum; ma non si può coinvolgere l'intera popolazione tutte le volte. Pertanto, si considera un opportuno sottoinsieme della popolazione, detto appunto campione, e si registrano le preferenze espresse da esso. Le frequenze relative ottenute dal campione sono una stima delle preferenze dei singoli individui dell'intera popolazione.

- *Margine d'errore.* Se il sondaggio ha un margine massimo d'errore del $2,8\%$, significa che la probabilità dell'evento in questione (per esempio "l'individuo vota SI") si discosterà al massimo della quantità $0,028$ dalla stima ottenuta sul campione...con probabilità "grande"[3].

- *Consistenza numerica del campione.* Il numero di cittadini che costituiscono il campione è strettamente legato al margine di errore del sondaggio. Una questione significativa al riguardo è determinare la dimensione del campione affinché la stima fatta su esso sia "buona". Questa però è una questione articolata di cui non ci occupiamo.

La questione

Una popolazione costituita da **10.000** individui è chiamata a votare tra due candidati, diciamo A e B. Mediante un sondaggio effettuato su un campione della popolazione, si è stimato che la probabilità che l'**individuo** sia favorevole ad A è del **40**%.

Sulla base di ciò, si vuole stimare il numero **F** di individui che voteranno A. Precisamente, quanto vale

$$\mathbf{P}(\mathbf{3900} \leq \mathbf{F} \leq \mathbf{4100}) \ ?$$

Osservazioni:

- In sostanza, a partire dalla percentuale di individui del campione che sono a favore di A, si vuole stimare la probabilità che il generico individuo della popolazione sia a favore di A.

- Gli estremi di variabilità stabiliti per F nel problema rappresentano rispettivamente il 39% e il 41% della popolazione. Pertanto la richiesta si può così riformulare: qual è la probabilità che il numero F sia compreso tra $(40-1)\%$ e $(40+1)\%$ della popolazione?

[3] Precisamente con probabilità che di solito è del 95% (in fisica, invece, si usano margini molto più bassi!). Secondo un approccio frequentista, su molti campionamenti ci si aspetta che circa il 95% delle volte, la probabilità p sulla popolazione cada nell'intervallo $[\bar{p}-0,028; \bar{p}+0,028]$, dove $\bar{p}$ è la stima sul campione.

1.2 Un modello binomiale

Proviamo a modellizzare la situazione proposta mediante la *distribuzione binomiale*. Precisamente interpretiamo la questione nel modo seguente:

- una sequenza[4] di $\mathbf{n = 10.000}$ prove, ciascuna delle quali corrisponde ad una persona;
- ciascuna prova ha due soli esiti possibili: *"la persona vota A"* oppure *"la persona vota B"*; dall'esito del sondaggio stimiamo che la probabilità che l'individuo sia favorevole ad A è $\mathbf{p = 0.4}$. Essa è costante (cioè è la stessa) per ogni prova (individuo).

Affinché abbia senso modellizzare mediante una distribuzione binomiale la v.a. F, assumiamo inoltre che le prove siano indipendenti[5], ossia che il voto della persona non sia influenzato dal voto delle altre.

Secondo il nostro modello la probabilità che il numero di individui F sia, ad esempio, 3900 è[6]

$$P(F = 3900) = \binom{10000}{3900} \cdot 0,4^{3900} \cdot 0,6^{6100}$$

e la probabilità richiesta si può esprimere nella forma

$$P(\mathbf{3900 \leq F \leq 4100}) = P(F = \mathbf{3900}) + P(F = \mathbf{3901}) + ... + P(F = \mathbf{4100})$$

Osservazione. Senza ricorrere a formule ritenute a memoria, il valore di probabilità $P(F = 3900)$ si determina volta per volta in due passi, determinando:

1. la probabilità di una sequenza di 3900 voti a favore di A e 6100 voti a favore di B, quale

   $$\underbrace{AAA...A}_{3900 \text{ volte}}\underbrace{BBB...B}_{6100 \text{ volte}}$$

   per la legge della moltiplicazione (eventi indipendenti) la probabilità di tale sequenza è

   $$0,4^{3900} \cdot (1 - 0,4)^{6100}$$

2. il numero di tutte le sequenze di 3900 A e 6100 B
   esso è il numero di sottoinsiemi di 3900 *elementi (posizioni per A), contenuti in un insieme di* 10000 *elementi (posizioni possibili nella sequenza); dunque è*

   $$\binom{10000}{3900}$$

[4] Tale sequenza si indica spesso come "schema di Bernoulli" o "schema successo-insuccesso".

[5] Non possiamo dire che ciò avvenga realmente. Si tratta solamente di una nostra ipotesi che assumiamo, almeno in prima approssimazione, allo scopo di costruire un modello ragionevolmente semplice.

[6] Ricordiamo che la distribuzione di una variabile aleatoria binomiale S è data da

$$P(S = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

C'è un problema

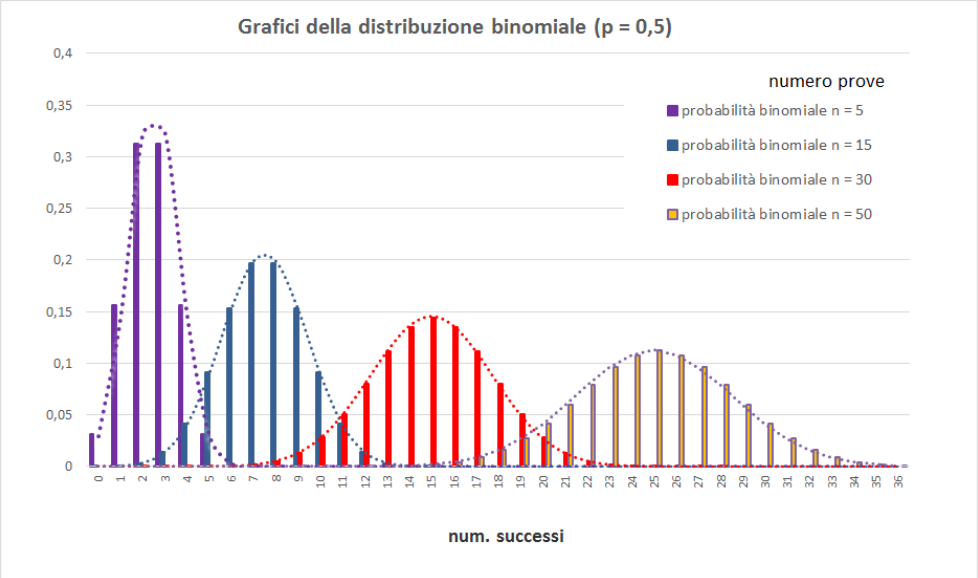
Il calcolo richiesto è assai articolato per la presenza dei coefficienti binomiali. Basti pensare che, ad esempio, $70! = 1,1979 \cdot 10^{100}$.
Come disse De Moivre nel Settecento, effettuare calcoli analoghi *"non è possibile senza un lavoro immenso, per non dire che è impossibile"*. E' vero che erano altri tempi e all'epoca non si disponeva dei calcolatori di oggi, ma anche noi siamo interessati a trovare un modo più efficiente per risolvere il problema.

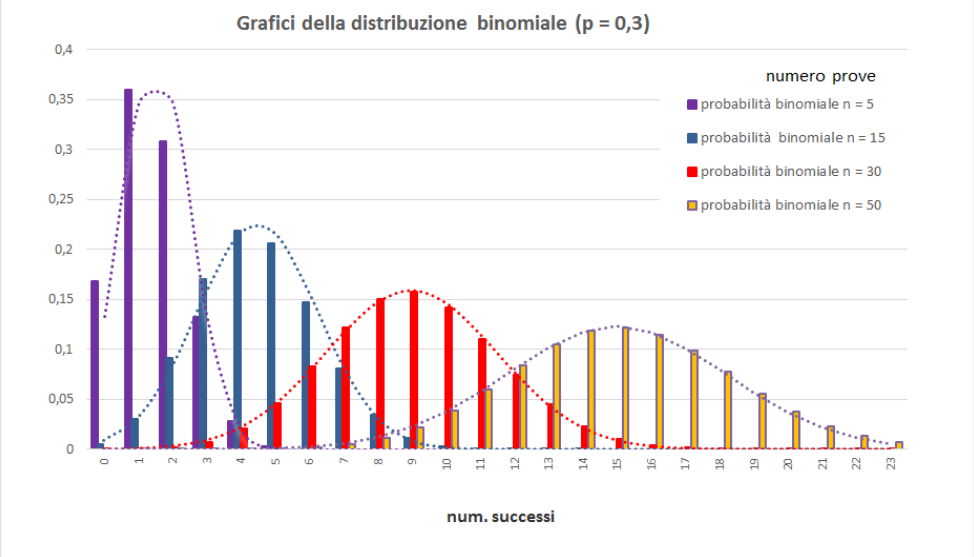
C'è un approccio meno dispendioso computazionalmente?

1.3 Un nuovo modello: verso il TLC

L'idea

Per comprendere come risolvere il problema, seguiamo un *approccio grafico*, cioè esaminiamo i grafici della distribuzione binomiale per p fissato (dove p è la probabilità di successo della prova) al crescere del numero n di prove.





- le figure suggeriscono che il grafico della distribuzione binomiale si può *approssimare* mediante una curva che ha forma di "*campana*"
- tale curva è[7] il grafico della *densità* di una certa variabile aleatoria

[7] Ciò si può dimostrare formalmente, come vedremo nel paragrafo 3.1.

- e i calcoli con la "nuova" distribuzione sono *più semplici*

Pertanto:

cercheremo di approssimare la binomiale mediante la distribuzione che resta definita dalla curva "a campana".

Un pò più precisamente

In effetti tale approssimazione è garantita da un teorema che precisa anche il senso in cui essa debba essere intesa. La sostanza di questo teorema si può formulare in prima approssimazione nel modo seguente e verrà precisata formalmente nella sezione 7.

TEOREMA LIMITE CENTRALE - TLC (una prima formulazione)

Sia S_n la v.a. binomiale relativa ad un numero n di prove e X una opportuna v.a. la cui densità ha come grafico la curva "a campana". Per n "grande", vale l'approssimazione:

$$\mathbf{P(a \leq S_n \leq b) \simeq P(a \leq X \leq b)}$$

Una nuova formulazione della questione

Possiamo allora usare il teorema appena visto per riformulare la nostra questione:

$$\textit{calcolare} \quad P(3900 \leq F \leq 4100)$$

dove F è il numero di individui della popolazione a favore del candidato A.
Siamo partiti modellizzando F mediante una variabile aleatoria binomiale relativa a $n = 10.000$ prove.
Consideriamo ora un grafico "a campana" che approssimi il grafico di tale distribuzione binomiale e indichiamo con $\mathbf{X}$ la v.a. la cui densità ha quella "campana" come grafico. Per il TLC la nostra questione diviene

$$\textit{calcolare} \quad \mathbf{P(3900 \leq X \leq 4100)}$$

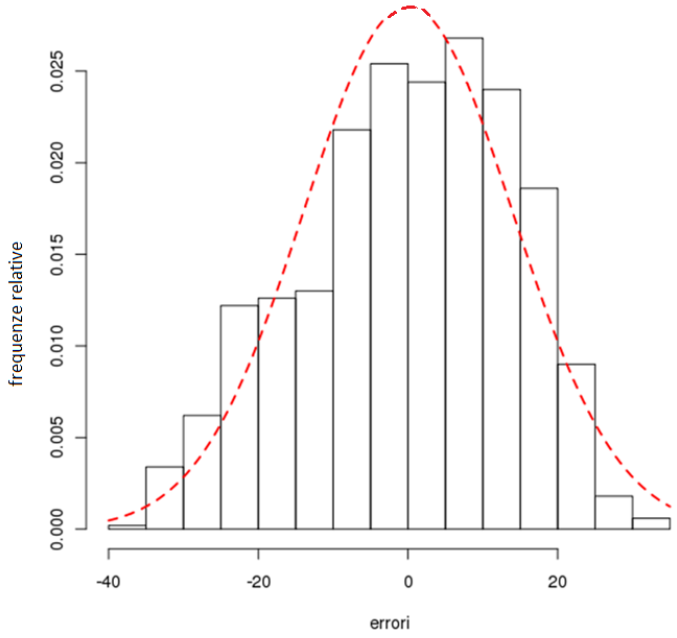
Nota. Questo è il nostro obiettivo; esso costituirà il filo conduttore di tutto il percorso. Vedremo, però, che sono molte le situazioni che si possono modellizzare mediante questo stesso schema. Ciò costituirà un motivo ulteriore per studiarlo a fondo.

1.4 Altre motivazioni alla normale

La curva "a campana" compare in molti altri contesti. Ne esaminiamo alcuni:

Gli errori accidentali nella misura

Quando si effettua la misura di una grandezza fisica mediante uno strumento, si compiono vari errori accidentali la cui distribuzione delle frequenze relative è del tipo in figura. In opportune ipotesi, la loro distribuzione può[8] essere approssimata con una curva "a campana".



Significativo e coinvolgente è il seguente scritto di Galton[9] a tal proposito: *"Difficilmente saprei indicare cosa alcuna altrettanto adatta a colpire l'immaginazione quanto la meravigliosa forma dell'ordine cosmico espressa dalla Legge di frequenza degli errori. Questa legge sarebbe stata personificata dai Greci, e deificata, se ne avessero avuto conoscenza. Essa regna con serenità ed in completa indifferenza tra la confusione più selvaggia. Più è immensa la folla più è grande l'apparente anarchia, e più è perfetto il suo governo. È la suprema* ***legge dell'Assenza di Ragione****.*
Ogni qualvolta un grande campione di elementi caotici viene preso in mano e disposto in ordine di grandezza, una insospettata e più bella forma di regolarità mostra di essere stata ivi latente. Le teste della riga ordinata formano una curva che scorre con proporzioni invariabili, ed ogni elemento quando vien messo a posto, trova, come se così fosse, una nicchia predisposta, accuratamente preparata a contenerlo."

[8] Da un punto di vista storico, Gauss introdusse la "nuova" distribuzione proprio per descrivere la legge degli errori casuali (1809).

[9] Galton F., "Order in Apparent Chaos", Natural Inheritance, n.39, 1889, p. 66; tratto da "President's Address" in Journal of Antropological Institute of Great Britain and Ireland, 15, 1886, pp. 489-499.

Le altezze di una popolazione

Esaminiamo la statura degli italiani. Tra il 1860 e il 1905, il generale Federico Torre raccolse, per ciascuna provincia italiana, i dati relativi ad oltre ventun milioni di giovani italiani chiamati alle armi e da questi creò tavole delle frequenze relative per altezze tra i 125 e 199 cm, suddivise in intervalli di 1 cm. Come si nota in figura 1, anche questa distribuzione è molto regolare e può essere approssimata con una curva "a campana".

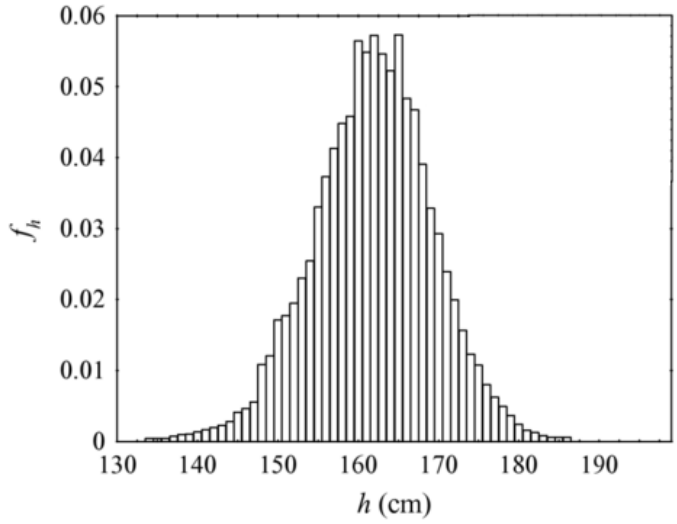


Figura 1: Distribuzione delle altezze dei circoscritti per la classe di leva 1900. La figura e i dati sono presi da [35].

1.5 Facciamo il punto

Abbiamo così visto come la curva "a campana" intervenga in tanti contesti. Pertanto:

- **studieremo** tale curva;

- ne esamineremo il **significato probabilistico**.

E per farlo, ci serve la sua espressione analitica.

La curva "a campana" è il grafico di una funzione della forma

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Tali funzioni costituiscono una famiglia al variare dei parametri reali μ, σ dove $\sigma > 0$.