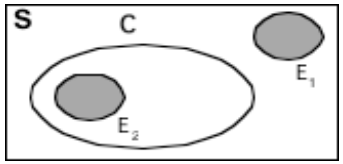
Probabità condizionata

Sia $C \subset W$ un evento di interesse dell' esperimento (pertanto: $P(C)>0$)

Come modificare lo spazio di probabilità nell'ipotesi che C si verifichi?



Se si è interessati ad E_1 si ha: $X \cap E_1 = \varnothing \rightarrow P(X \cap E_1)=0$; se invece l'evento di interesse è E_2 allora il fatto che $E_2 \subset X \rightarrow P(E_2)=1$.

Per comodità manteniamo lo stesso universo S anche se dal rettangolo S si è passati all'ellisse C e l'evento E_1 non può più verificarsi).

Riscalamento delle probabilità

Le probabilità vanno riscritte con la formula:

$$P(E|C) = \frac{P(E \cap C)}{P(C)}; \quad con \quad P(C) > 0$$

La probabilità di E sotto C è data dalla probabilità che i due eventi si presentino insieme (nello spazio di probabilità originario) rapportato alla probabilità assegnata (sempre nello spazio originario) all'evento condizionante.

Per comodità abbiamo mantenuto lo stesso simbolo "P" per indicare la funzione di probabilità condizionata, ma è chiaro che la funzione di non è più la stessa di quella originaria.

Esempio

Supponiamo che le facce di un dado siano equiprobabili. Abbiamo perciò le probabilità:

E	1	2	3	4	5	6	
$P(E)$	1/6	1/6	1/6	1/6	1/6	1/6	1

E	1	2	3	4	5	6	
$P(E)$	1/6	0	1/6	0	1/6	0	1/2

Se sappiamo che "è uscito un dispari" questo modifica la prova : alcuni eventi sono ancora possibili, altri no.

Le probabilità ridefinite alla luce di ciò che ciascuno aveva in comune con "A" e scalate in modo da sommare ad uno (probabilità dell'evento certo

E	1	2	3	4	5	6	
$P(E)$	(1/6)/(1/2)	0	(1/6)/(1/2)	0	(1/6)/(1/2)	0	(1/2)/(1/2)

$E\|A$	1	3	5	
$P(E\|A)$	1/3	1/3	1/3	1

Esempio

Si lanciano tre monete. Qual'è la probabilità che presentino la stessa faccia?

Prima soluzione

I casi possibili sono 8: (CCC, CCT, CTC, TCC, TTC, CTT, TCT, TTT); i casi favorevoli sono 2 e quindi la probabilità cercata è 2/8=1/4.

2ª soluzione

Due monete sono sicuramente uguali; quindi il risultato è determinato dalla 3ª; questa può essere testa o croce quindi la probabilità richiesta è 1/2.

La conoscenza dell'evento "almeno due monete uguali" non è rilevante dato che non modifica l'universo degli eventi originario.

Se E={tre facce uguali} e F={almeno due facce uguali} allora

$$P(E|F)=P(E \cap F)/P(F)=P(E \cap F)/1=P(E)$$

dato che E è già incluso in F.

Probabilità Composta

Un modo equivalente di scrivere la probabilità condizionata è

$$P(A \cap B) = P(A)P(B|A)$$

Vantaggi:

1) **Rimane valida anche se "A" è un evento impossibile**

2) **La definizione sul lato destro è più semplice rispetto a quella sul lato sinistro**

Esempio:
Un'urna contiene 4 biglie bianche e 2 nere. Si estraggono senza reimmissione due biglie. Qual è la probabililtà che entrambe siano bianche?

$$B1 = 1^a \ \ bianca; \ \ B2 = 2^a \ \ bianca$$

$$P(B1 \cap B2) = P(B1) = P(B2|B1) = \left(\frac{4}{6}\right)\left(\frac{3}{5}\right) = \frac{2}{5}$$

Probabilità nelle partizioni

Consideriamo una partizione (eventi necessari e incompatibili) dell'universo degli eventi

$$H_1, H_2, \ldots, H_k$$

Partizione significa che di questi eventi, in ogni prova, se ne verifica uno e solo uno.

Poiché $E = E \cap I = E \cap \{H_1 \cup H_2 \cup \ldots \cup H_k\}$

$$= (E \cap H_1) \cup (E \cap H_2) \cup \ldots \cup (E \cap H_k)$$

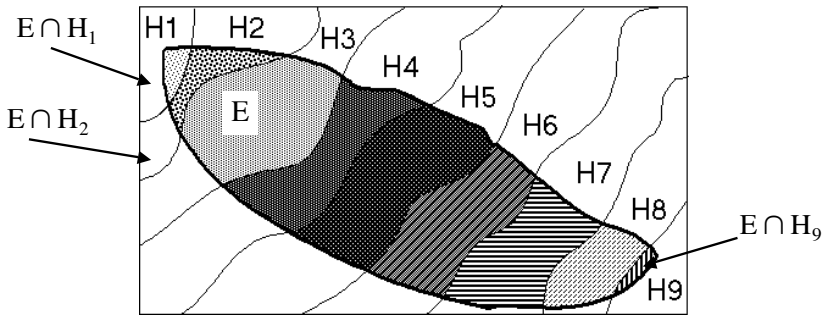
Ne consegue che

$$P(E) = P(E \cap H_1) + P(E \cap H_2) + \ldots + P(E \cap H_k)$$

$$= P(H_1) * P(E|H_1) + P(H_2) * P(E|H_2) + \ldots + P(H_k) * P(E|H_k)$$

La probabilità di "E" è pari alla somma PONDERATA delle probabilità condizionate di "E" . I pesi sono le probabilità incondizionate degli eventi elementari

Probabilità nelle partizioni/2

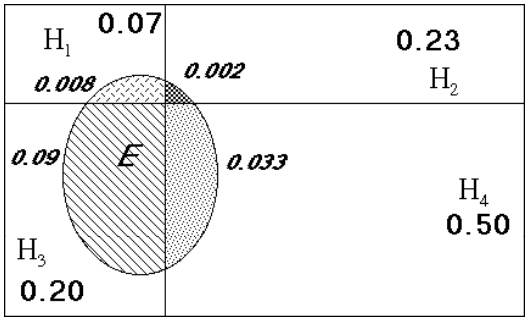


Ogni parte tratteggiata rappresenta l'intersezione dell'evento "E" con uno degli eventi della partizione

Poiché gli eventi della partizione sono incompatibili così saranno le loro parti toccate da "E"

Esempio con i diagrammi di Venn

Universo degli eventi



Scegliamo le seguenti probabilità

$P(H_1) =$	0.07	$P(E \cap H_1) =$	0.008
$P(H_2) =$	0.23	$P(E \cap H_2) =$	0.002
$P(H_3) =$	0.20	$P(E \cap H_3) =$	0.090
$P(H_4) =$	0.50	$P(E \cap H_4) =$	0.033

$$P(E) = 0.133$$

$$E = (H_1 \cap E) \cup (H_2 \cap E) \cup (H_3 \cap E) \cup (H_4 \cap E)$$

Ne consegue:

$$P(E) = P(E \cap H_1) + P(E \cap H_2) + P(E \cap H_3) + P(E \cap H_4)$$

$$= P(H_1) * P(E|H_1) + P(H_2) * P(E|H_2) + P(H_3) * P(E|H_3) + P(H_4) * P(E|H_4)$$

$$= 0.07 * \frac{0.008}{0.07} + 0.23 * \frac{0.002}{0.23} + 0.20 * \frac{0.0.09}{0.20} + 0.50 * \frac{0.033}{0.50} = 0.133$$

Teorema di Bayes

Le condizioni sperimentali non sempre consentono soluzione intuitive. Possiamo però dimostrare il seguente teorema

Data una partizione $H_1, H_2, \ldots, H_k$ dell'universo degli eventi la probabilità a posteriori rispetto all'evento "E" di H_i è data dalla formula

Formula

$$P(H_i|E) = \frac{P(H_i)*P(E|H_i)}{\sum_{j=1}^{k} P(H_j)*P(E|H_j)}$$

Logica

Si sa già il risultato della prova e si cerca quale ne sia la causa fra quelle possibili (PRINCIPIO DELLA PROBABILITA' INVERSA)

Da quanto detto sulle partizioni

$$P(E) = \sum_{j=1}^{k} P(H_j)*P(E|H_j); \qquad P(H_i \cap E) = P(H_i)*P(E|H_i)$$

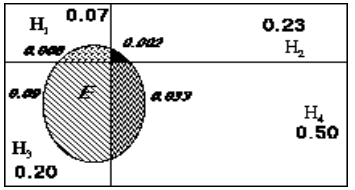
Per cui

$$P(H_i|E) = \frac{P(H_i)*P(E|H_i)}{\sum_{j=1}^{k} P(H_j)*P(E|H_j)} = \frac{P(H_i \cap E)}{P(E)}$$

Esempio

Ritorniamo al caso illustrato con i diagrammi di Venn e determiniamo la causa più probabile di "E"

$$P(H_1|E) = \frac{0.008}{0.133} = 0.0602 \quad P(H_3|E) = \frac{0.090}{0.133} = 0.6767$$
$$P(H_2|E) = \frac{0.002}{0.133} = 0.0150 \quad P(H_4|E) = \frac{0.033}{0.133} = 0.2481$$



La causa più probabile è allora "H3" come il diagramma mostra con chiarezza: se, in una scommessa, tutti gli eventi dessero luogo alla stessa vincita, la logica ci imporrebbe di scegliere H3.

Nella forula di Bayes il denominatore costante per cui spesso si scrive

$$P(H_i|E) = \frac{P(H_i)P(E|H_i)}{P(E)} \propto P(H_i)P(E|H_i)$$

Che esprime la probabilità posteriori come proporzionale a quella a priori con un fattore di proporzionalità noto come VEROSIMIGLIANZA cioè la probabilità, sotto H_i, che si verifichi E.

Esempio

Può un ciuccio superare un esame?

Dati:
1) il 75% di chi si presenta all'esame, supera l'esame.
2) il 70% di chi supera l'esame è bravo.
3) il 90% dei bocciati è ciuccio.

S="Superato", R="Respinto", B="Bravo", C="Ciuccio"

$$1)P(S) = 0.75, \qquad 2)P(B|S) = 0.70, \qquad 3)P(C|R) = 0.9$$

E' richiesto il calcolo di P(SIC).

$$P(S|C) = \frac{P(S \cap C)}{P(C)} = \frac{P(S)P(C|S)}{P(S)P(C|S) + P(R)P(C|R)} = \frac{P(S)\left[1 - P(B|S)\right]}{P(S)P(C|S) + P(R)P(C|R)}$$
$$= \frac{0.75*0.30}{0.75*0.30 + 0.25*0.90} = 0.5$$

Uso del teorema di Bayes

il 5% degli abitanti di un paese è affetto da una malattia.

Poniamo:

$$E_1 = \{ha \ \ la \ \ malattia\}, \ E_2 = \{Non \ \ ha \ \ la \ \ malattia\}$$

Si usa un test clinico la cui SENSITIVITA' (la probabilità che sia positivo (+) dato che la persona è ammalata, è: $P(+|E_1) = 0.90$

e con probabilità di FALSO POSITIVO (la persona è sana, ma il test indica il contrario) $P(+|E_2) = 0.15$

Scelta a caso una persona si effettua il test e questo risulta positivo, qual'è la probabilità che la persona sia ammalata?

$$P(E_1|+) = \frac{P(E_1)P(+|E_1)}{P(E_1)P(+|E_1) + P(E_2)P(+|E_2)} = \frac{(0.05)0.90}{(0.05)0.90 + (0.95)0.15} = 0.24$$

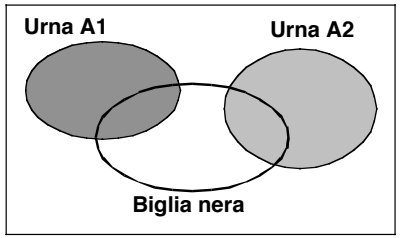
Probabilità a priori e a posteriori

1) si sceglie a caso l'urna;
2) Si scegli a caso la biglia.

La prova è stata effettuata ed è risultata "biglia nera". Da dove proviene?

La probabilità assegnata ad A_1 e A_2 prima dell'esperimento è detta A PRIORI.

Come si modifica alla luce del fatto è stata scelta una biglia nera (Evento N)?



il verificarsi di N limita l'attenzione alla sola intersezione di A_1 con N

$$P(A_1|N) = \frac{P(A_1 \cap N)}{P(N)} = \frac{0.35}{0.45} = 0.78$$

La probabilità dell'evento dopo il verificarsi di un altro è detta A POSTERIORI

Esempi

1.a) il lancio di due dadi non truccati ha prodotto almeno un "3". Qual'è la probabilità che la somma sia "7"?

S = La somma é "7"; E = é uscito almeno un "3"

$$S = \{(1,6);(2,5);(3,4);(4,3);(5,2);(6,1)\}$$
$$E = \{(3,1);(3,2);(3,3);(3,4);(3,5);(3,6);(1,3);(2,3);(4,3);(5,3);(6,3)\}$$

$$P(S|E) = \frac{P(S \cap E)}{P(E)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$$

2) Un mazzo di carte francesi ha 52 carte di cui 4 sono assi. Se si estraggono due carte senza che la prima estratta venga reimmessa prima della seconda estrazione. Qual'é la probabilità che siano entrambi degli assi?

A_1 = asso alla prima
A_2 = asso alla seconda

$$P(A_1 \cap A_2) = P(A_1) * P(A_2|A_1) = \frac{4}{52} * \frac{3}{51} = 0.0045$$

Indipendenza (in probabilità)

DUE EVENTI A E B SONO INDIPENDENTi SE IL VERIFICARSI DELL'UNO NON ALTERA LA PROBABILITA' DELL'ALTRO

$$P(E \mid F) = P(E)$$

Tale interpretazione è coerente con l definizione di probabilità condizionata

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} \Rightarrow \frac{P(E) * P(F)}{P(F)} = P(E)$$

L'indipendenza è una relazione BILATERALE: se "E" è indipendente da "F" allora è vero anche il viceversa purché "E" non sia impossibilie"

$$P(E|F) = P(E) \Rightarrow P(F|E) = \frac{P(E \cap F)}{P(E)} = \frac{P(E) * P(F)}{P(E)} = P(F)$$

Esempi

Si supponga che gli eventi "A" e "B" siano indipendenti e che si abbia

$$P(E) = 0.45, \quad P(F) = 0.80$$

a) Calcolare P(E U F)

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = P(E) + P(F) - P(E) * P(F) = 0.89$$

b) Calcolare P(E^c|F^c)

$$P(E^c \mid F^c) = \frac{P(E^c \cap F^c)}{P(F^c)} = \frac{P[(E \cup F)^c]}{P(F^c)} = \frac{1 - P(E \cup F)}{1 - P(F)} = \frac{0.11}{0.20} = 0.55$$

Soluzione di problemi con le probabilità

1) Individuare i dati del problema e tradurli in simboli.

2) Delimitare le richieste del problema ed esprimerle in simboli

3) applicare le regole del calcolo delle probabilità

In molti casi può essere utile la seguente formula

$$P(E|F) = \frac{P(E)}{P(F)} P(F|E)$$

Che consente di scambiare il ruolo degli eventi tra condizionato e condizionante

Esempi

Un controllo di qualità rivela:

1) il 20% delle componenti è difettoso.
2) Il 90% delle componenti passa il controllo.
3) Il prodotti privi di difetti passano il test nel 95% dei casi.

Qual è la probabilità che una componente non risulti difettosa una volta superato il controllo?

Poniamo E="La componente è difettosa"; F="La componente passa il test"

Il problema ci suggerisce

$$1) P(E) = 0.20, \qquad 2) P(F) = 0.90, \qquad 3) P\left(F|E^c\right) = 0.95$$

E' richiesto il calcolo di $P(E^c|F)$.

$$P\left(E^c|F\right) = \frac{P\left(E^c\right)}{P(F)} P\left(F|E^c\right) = \frac{\left[1-P(E)\right]}{P(F)} P\left(F|E^c\right) = \frac{0.8*0.9}{0.95} = 0.7579$$

Indipendenza di "n" eventi

Per evitare le difficoltà del concetto di indipendenza e per esaltarne la mera natura concettuale, diremo che una m-tupla è costituita da eventi indipendenti se:

$$P\left(E_{k_1} \cap E_{k_2} \cap \ldots \cap E_{k_m}\right) = \prod_{i=1}^{m} P\left(E_{k_i}\right)$$

per ogni permutazione degli indici: $2 \le k_1 < k_2 < \ldots < k_m \le m$.
Questo significa che tutte le possibili coppie di eventi sono indipendenti:

$$P(E_i \cap E_j) = P(E_i)\,P(E_j) \text{ per } i \neq j$$

e sono indipendenti anche tutte le combinazioni di tre eventi:

$$P(E_i \cap E_j \cap E_r) = P(E_i)P(E_j)P(E_r) \text{ per } i \neq j \neq r$$

e così via fino ad arrivare alla indipendenza della m-tupla.

Esempio di Bernstein

Supponiamo che S consista di 4 eventi: $S = \{E_1, E_2, E_3, E_4\}$

Assumiamo anche gli eventi siano equiprobabili: $P(E_i) = 1/4$

Definiamo ora gli eventi composti:

$$A_1 = \{E_1, E_2\} \Rightarrow P(A_1) = 1/2$$
$$A_2 = \{E_1, E_3\} \Rightarrow P(A_2) = 1/2$$
$$A_3 = \{E_1, E_4\} \Rightarrow P(A_3) = 1/2$$

Si verifica subito che

$$P(A_1 \cap A_2) = 1/4 = P(A_1) * P(A_2);$$
$$P(A_2 \cap A_3) = 1/4 = P(A_2) * P(A_3)$$
$$P(A_1 \cap A_3) = 1/4 = P(A_1) * P(A_3);$$

Indipendenti due a due

ma anche che: $P(A_1 \cap A_2 \cap A_3) = 1/4 \neq P(A_1) * P(A_2) * P(A_3) = 1/8$

l'indipendenza a due a due non implica che gli eventi siano indipendenti se considerati in terne. Tale risultato è generalizzabile a un numero di eventi qualsiasi

Problema del compleanno

Siete in una sala con "n" persone in un dato giorno. Qual'è la probabilità che almeno una delle persone presenti festeggi il compleanno quel giorno?

Ipotizziamo che l'anno sia di 365 giorni e che le nascite siano uniformi nel corso dell'anno.

Sia "A" l'evento "Una persona festeggia il compleanno".

$$P(\overline{A}) = P(\text{Nessuno festeggia il compleanno}) = P(\text{non festeggia la } 1^a) * P(\text{non festeggia la } 2^a) * \dots$$
$$= \left(1 - \frac{1}{365}\right) * \left(1 - \frac{1}{365}\right) * \dots * \left(1 - \frac{1}{365}\right) = \left(1 - \frac{1}{365}\right)^n$$

Se n=200 si ha $P(\overline{A}) = 58\% \Rightarrow P(A) = 42\%$

Se n=500 si ha $P(\overline{A}) = 25\% \Rightarrow P(A) = 75\%$

Incompletezza

Il sistema di Kolmogorov non dice che come scegliere le probabilità.

La teoria matematica interviene dopo l'assegnazione delle probabilità.

L'impegno maggiore nella trattazione matematica della probabilità si concentra su due questioni fondamentali:

1) Come determinare la probabilità di un evento qualsiasi a partire dalle probabilità già assegnate ai risultati elementari di un prova;

2) Come aggiornare tali probabilità allorché si rendano disponibili nuove informazioni rilevanti sulla prova

Sorte e indipendenza

L'indipendenza è una condizione forte che talvolta sembra porsi contro il senso comune.

Ciccillo è un affezionato del 12 sulla ruota di Napoli. Indichiamo con E_i l'evento "Esce il 12 nella estrazione i-esima". Non si ha motivo di dubitare della onestà delle estrazioni. Negli ultimi tempi il 12 non è uscito per 150 estrazioni.

Che probabilità ha di uscire alla 151ª?

$$P\left(E_{151} \middle| E_1^c \cap E_2^c \cap \dots \cap E_{150}^c\right) = \frac{P\left(E_1^c \cap E_2^c \cap \dots \cap E_{150}^c \cap E_{151}^c\right)}{P\left(E_1^c \cap E_2^c \cap \dots \cap E_{150}^c\right)} = P(E_{151})$$

E' evidente che la probabilità è la stessa non solo dopo 10, 100, 1000 estrazioni, ma che non c'è sequenza di ritardiche potrà mai provocare l'uscita del "12".

Attenzione! Questo non significa che il "12" non uscirà, ma solo l'assenza di raziocinio nell'idea che la propensione ad uscire aumenti con il ritardo.

Interpretazione classica

La probabilità è un modello numerico delle relazioni che intercorrono tra le possibili occorrenze degli eventi e le proprietà fisiche dell'esperimento

Principio della ragione insufficiente (gli eventi sono equiprobabili a meno che non si dimostri il contrario

Sia n(A) il numero di eventi elementari in A e sia n(S) il numero totale di eventi elementari.

La probabilità dell'evento composto è data da

$$P(A) = \frac{n(A)}{n(S)} = \frac{\text{casi favorevoli}}{\text{casi possibili}}$$

La presenza di "simmetrie" negli esperimenti consente una assegnazione oggettiva delle probabilità (almeno tra coloro che condividono le simmetrie)

Interpretazione classica/2

Una lotteria ha venduto 4750 biglietti. Ciccillo ne ha comprati 4.

Se tutti i numeri hanno la stessa probabilità di vincere allora:

$$P(Ciccillo\ \ vince) = \frac{4}{4750} = 0,000842 = 8.42\ \ per\ \ mille$$

PREGI **E' "naturale". Nel valutare il verificarsi di un evento eseguiamo a mente il rapporto tra le circostanze a favore e quelle contro**

DIFETTI **1) Include una tautologia: "ugualmente possibili" è già una definizione di probabilità.**

2) Non può essere richiamata se si ignora la struttura fisica della prova e come questa influenza gli eventi.

Interpretazione classica/3

Presenta delle contraddizioni:

Da due mazzi di carte francesi si sceglie una carta per ogni mazzo. Una di esse è di colore nero.

Qual'è la probabilità che l'altra sia di colore nero?

POISSON: i casi possibili sono: (N_1,N_2), (N_1,R_2), (R_1,N_2) e (R_1,R_2). Se una delle due è nera allora restano solo 3 casi di cui due a favore. Perciò:

$$P(N_2|N_1) = 2/3$$

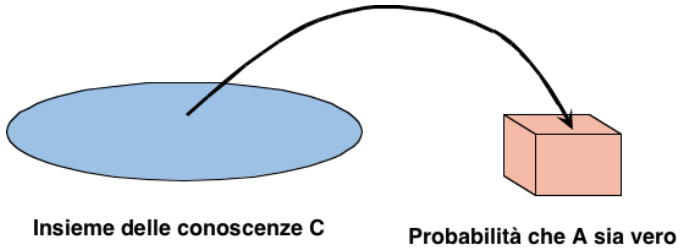
Il segno "I" va letto come : "dato che"

von Kries: le due scelte sono indipendenti per cui la scelta della 2^ carta può ignorare la scelta della 1^. Quindi:

$$P(N_2|N_1) = P(N_2) = \frac{26}{52} = \frac{1}{2}$$

Interpretazione logica (o keynesiana)

La probabilità esprime la relazione logica che sussiste tra la validità di una asserzione sul verificarsi di un evento e l'insieme delle conoscenze che si hanno sulla prova



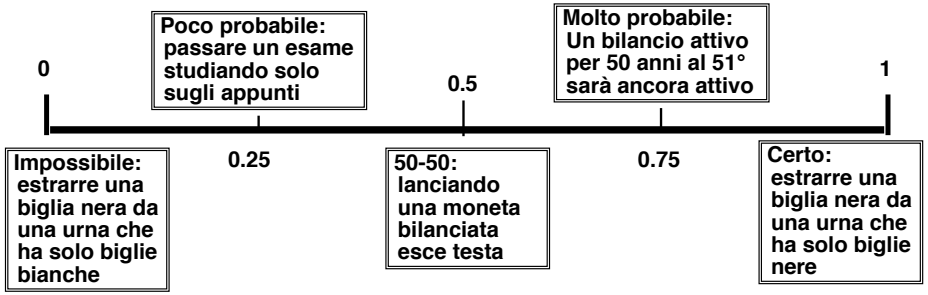
Dato un certo stato di informazione "C" esiste una ed una sola probabilità che esprime il grado di fiducia sulla validità dell'affermazione "A".

Due diverse persone, se danno due valori diversi a P(A), hanno due diversi stati di conoscenza "C"

Probabilità soggettiva (Bayesiana)

La probabilità è l'espressione numerica del grado di convinzione sulla verità di una certa asserzione .

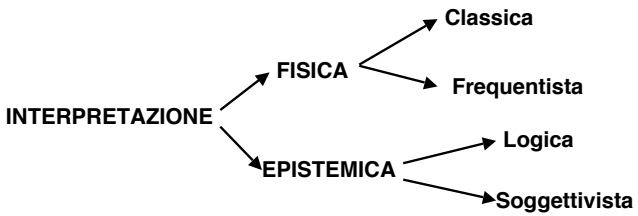
Per un dato insieme di conoscenze può esserci più di una probabilità



L"unico problema sono le regole di COERENZA cioè le opinioni di probabilità dovrebbero rispettare i postulato di additività e questo non è garantito

Significato della probabilità

L'interpretazione della probabilità, in sostanza, prescinde dalla sua rappresentazione matematica. I vari approcci alla probabilità si distinguono:



L'interpretazione "fisica o oggettivista" della probabilità si riferisce alle condizioni materiali di un esperimento.

L'interpretazione "epistemica o soggettivista" si riferisce alle idee di chi effettua l'esperimento.

Esempio

La probabilità che l'arciere centri il bersaglio è del 40%

Visione oggettivista

0.40 è una proprietà fisica dell'evento legata a: materiali, distanza da cui tira, tipo di bersaglioe, etc.

Può anche derivare da una lunga serie di esperienze analoghe.

Visione soggettivista

0.40 esprime il grado di fiducia, sul verificarsi dell'evento da parte di chi osserva l'esperimento.

Può essere unica o cambiare da persona a persona

Nuove rilevazioni

L'acquisizione di nuovi dati è dovuta al fatto che:

- **La base informativa di un problema non è soddisfacente**
- **E' utile e praticabile realizzarne una nuova o integrare quella esistente**

La rilevazione dei dati consiste nella annotazione sistematica, precisa e impersonale della modalità delle variabili riscontrate sull'unità

Le rilevazioni possono essere classificate in vario modo. Quella più rilevante è la distinzione tra TOTALI e PARZIALI:

TOTALI: coinvolgono tutti gli elementi di una popolazione

PARZIALI: la rilevazione è estesa solo ad una parte, comunque scelta, di popolazione

Le rilevazioni totali

Le RILEVAZIONI TOTALI (O CENSIMENTI) sono quelle in cui sono enumerate o misurate tutte ed indistintamente le unità della popolazione

All'interno delle totali si hanno:

RILEVAZIONI GENERALI: riguardano la rilevazione di tutte le unità rispetto alle variabili di interesse (POPOLAZIONE)

***Esempio:* un'indagine sul voto che si rivolga a tutti gli elettori di qualsiasi sesso e regione di residenza**

RILEVAZIONI SPECIALI: riguardano la rilevazione delle sole unità rispondenti a certe specifiche (SOTTOPOPOLAZIONE)

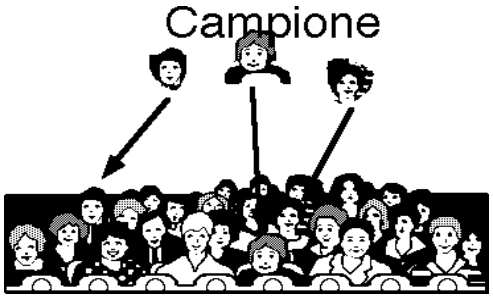
***Esempio:* un'indagine sul voto che si rivolga a tutti, ma i soli iscritti alle camere di commercio come "artigiani"**

Le rilevazioni parziali

Sono limitate solo ad una parte della popolazionescelta in base ad opportuni criteri. La parte esaminata si chiama CAMPIONE.

La riduzione delle unità è valida solo se permette il raggiungimento di risultati molto prossimi di quelli ottenibili con la rilevazione TOTALE.

TOTALE/PARZIALE NON E' UNA COTRAPPOSIZIONE, MA UNA COMPLEMENTARITA'



Esperienze consolidate in molti paesi e in molte discipline dimostrano che si può dare pieno affidamento ai campioni purché scelti con accuratezza.

Le ragioni del campione

Nel corso di un'indagine ci si può accorgere che la RILEVAZIONE TOTALE non è praticabile perché:

- **HA UN COSTO ECCESSIVO O RICHIEDE GRANDI ORGANIZZAZIONI**

  <u>Esempio:</u> il censimento generale si realizza ogni 10 anni

- **RICHIEDE TROPO TEMPO**

  <u>Esempio:</u> l'intervista di tutti i lavoratori dipendenti richiederebbe tanti anni che una volta finita la popolazione attuale sia molto diversa dalla censita

- **E' TEORICA: PARTE DELLE SUE UNITA' NON ESISTE ANCORA o NON ESISTE PIU'**

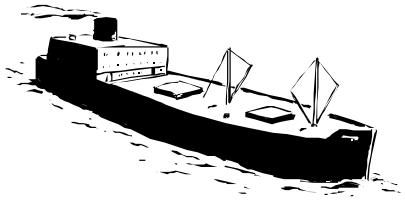
  <u>Esempi:</u> il controllo della qualità dovrebbe riguardare anche le unità non ancora prodotte.
  Le vestigia di antiche civiltà

Il campione

- **L'analisi del campione è meno costosa, più precisa, più asettica, più controllabile e più rapida dell'esame della rilevazione totale.**
- **I censimenti generali si limitano alle variabili fondamentali lasciando ai campioni il compito di scendere nei dettagli.**

La popolazione è la nave che quando naviga lascia vedere solo la parte che galleggia: il campione.

Osservando e analizzando la parte visibile si conoscerà anche la parte che è sotto l'acqua.



Applicazioni del campione

- **Sondaggi elettorali; gradimento delle amministrazioni locali; consenso alle scelte politiche governative.**
- **Ricerche di mercato: accettazione di un nuovo prodotto; apprezzamento della modifica di un prodotto conosciuto; desiderio di un nuovo servizio.**
- **Controllo della qualità: aderenza agli standard di un item; verifica della integrità di una fornitura; certificazione della composizione di un prodotto.**
- **Indagini di laboratorio: efficacia di un fertilizzante; pericolosità di un farmaco; validità di terapie comportamentali; tolleranza ad un prodotto.**
- **Imprenditoria: pagamento di *royalties*; diffusione di quotidiani e settimanali; *audience* televisiva; revisione dei conti.**

Campionamento

Come circoscrivere -grazie al calcolo delle probabilità- il numero di unità e come realizzare la loro scelta.

C'è un duplice aspetto:

Procedura di selezione

Insieme di regole ed operazioni con cui si realizza la scelta delle unità;

Procedura di stima

Calcolo delle statistiche e loro uso come valori presunti dei parametri incogniti.

Per ora ci interessa solo il primo

Campionamento/2

Operare con dei campioni significa agire con informazioni incomplete

Da un lato c'è la POPOLAZIONE: un collettivo di oggetti, persone, valori per la quale si cerca di conosceere un fatto

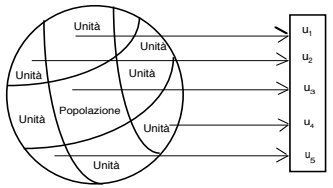
Dall'altro c'è il CAMPIONE: una parte della popolazione che viene esaminata perché possa fornire informazioni sulla popolazione

Dobbiamo perciò essere preparati ad un errore dovuto alla mancanza di dati relativi alle unità non campionate.

C'è un modo per minimizzare lo scarto tra valori campionari e valori nella popolazione?

Popolazione teorica ed effettiva

In ogni rilevazione di dati è necessario un sistema di riconoscimento delle unità



Non sempre è facile fare la lista delle unità: esistono popolazioni elusive od incerte in cui non sappiamo chi siano e quante siano le unità

Dobbiamo quindi distinguere tra

POPOLAZIONE TEORICA: quella che vorremmo esaminare

POPOLAZIONE EFFETTIVA: unità effettivamente raggiungibili

La *frame* o lista

Tra TEORICA ed EFFETTIVA si inserisce la *frame* o lista cioè un sistema di codici che identificano le unità

La lista è una sovrastruttura imposta alla popolazione.

Per essere utile deve risultare:

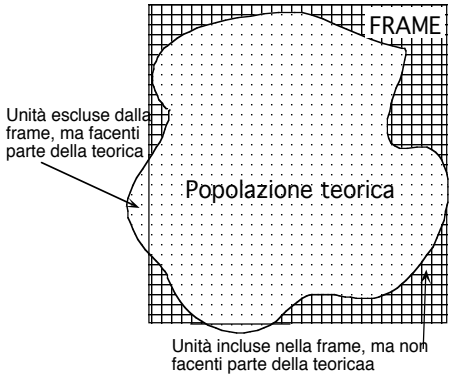
ESATTA

AGGIORANTA

COMPLETA

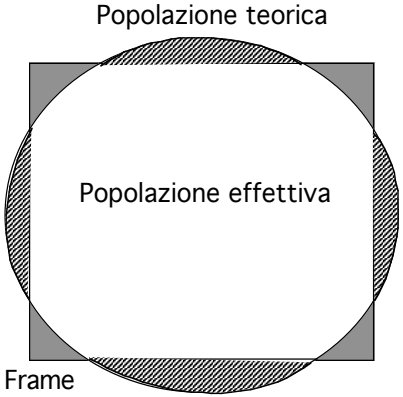
DOCUMENTATA

con regole note e disponibili



Le ragioni della scelta di una certa popolazione effettiva per una data popolazione teorica devono essere capillari ed accurate.

La *frame/2*



La popolazione effettiva è data dalle unità della popolazione teorica incluse nella *frame*.

La copertura assicurata dalla popolazione effettiva è soddisfacente se rende superflua la ricerca sulla teorica.

Una frame è perfetta se tutte le unità della popolazione vi sono elencate una ed una sola volta.

Errori di frame

UNDERCOVERAGE — **Unità della popolazione teorica non viste dalle *frame***

OVERCOVERAGE — **Unità viste dalle *frame* ma non facenti parte della teorica (include i doppioni)**

NONCOVERAGE — **Unità della popolazione teorica viste dalle *frame* ma dalle quali non è stato possibile avere alcun dato**

NON RESPONSE — **Per quanto si faccia non si riescono ad acquisire, dalla unità, le informazioni necessarie**

Tali errori impediscono un corretto collegamento (selection bias) tra teorica ed effettiva molto pericoloso per le indagini.

Errori non campionari

DISTORSIONI NELLA SCELTA DELLE UNITA'

Il meccanismo di estrazione delle unità agisce solo su alcune parti e ne esclude altre (undercoverage e over coverage)

ESEMPIO:

i sondaggi telefonici: in dipendenza dell'orario in cui si telefona si raggiungono unità diverse. Persone che non hanno telefono o il cui telefono non è in elenco non potranno mai essere raggiunte.

DISTORSIONI NELLA RILEVAZIONE DELLE UNITA'

Non sempre è possibile garantire l'accuratezza delle misurazioni ovvero non sempre le unità sono disposte a farsi rilevare o a rispondere con sincerità (noncoverage e nonresponse)

ESEMPIO:

Rilevazione degli affiliati a "Cosa Nostra". Vincitori di lotterie

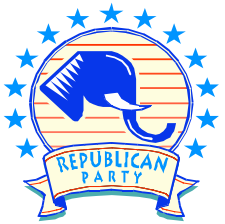
Il noto caso del Literary Digest

Nel 1936 tale rivista attivò un sondaggio postale su dieci milioni di votanti scelti da elenchi telefonici e registri di possessori di auto.

Lo scopo era di prevedere il risultato delle elezioni presidenziali: Roosvelt (democratici-progressisti) e Landon (repubblicani-conservatori).

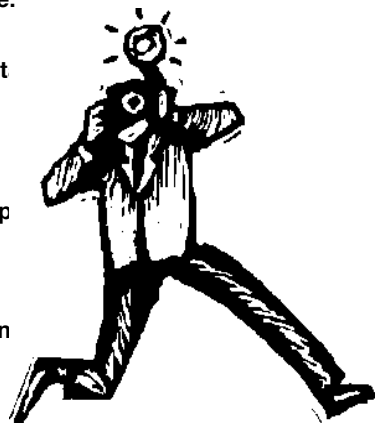
Si ottennero 2.4 milioni di riposte: il 57% avrebbe votato Landon ed il 38.5% Roosvelt. Vinse però Roosvelt con il 63%. Gran parte del fiasco è da attribuire ad una scelta inadeguata della lista.

Perché?



Le non risposte

In ogni indagine è presente una certa percentuale di unità che non collabora o che non è posta in grado di collaborare.



Sfuggono all'indagine per ragioni loro proprie (unit

Gli incaricati non riescono ad individuarle

Entrano in contatto tra di loro e concordano le risp

Regolano le loro risposte in base a chi le contatta

La risposta dipende dalla forma di contatto: telefon posta, rete, intervista, etc.

Per correggere queste anomalie bisogna eliminare delle unità dalla frame oppure -se possibile- ripetere l'acquisizione

Le non risposte/2

Possiamo catalogare le unità in tre categorie

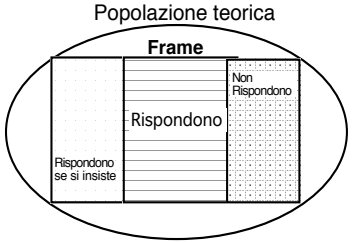
a) Unità che rispondono già al primo contatto

b) Unità che rispondono solo se subiscono dei call-back cioé se si insiste un certo numero di volte

c) Unità dalle quali non è possibile ottenere risposte

Le unità c) riducono l'ampiezza effettiva della popolazione o del campione.

Poiché sono spesso peculiari riducono anche la qualità dell'indagine dato che forse un particolare segmento di unità le sta sfuggendo



L'ampiezza del campione

il numero di unità "n" considerate nel capione è detto AMPIEZZA del campione. L'ampiezza della popolazione è indicata con "N"

Frazione di campionamento: $\frac{n}{N}$

Frazione di unità della popolazione "rappresentata"nel campione

Intervallo di campionamento: $\frac{N}{n}$

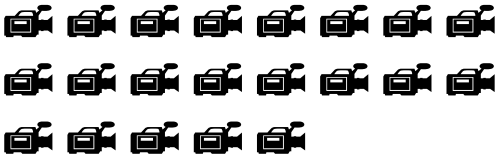
Ogni quante unità della popolazione si sceglie una unità del campione.

Anche, fattore di proporzionalità.

Se la popolazione è infinita questi due rapporti perdono di significato.

Errori campionari e variabilità

Se le unità fossero uguali basterebbe un campione di ampiezza n=1



Se le modalità sono due o più saremo certi che entrambe siano nel campione solo se n=N.

Ma la variabilità non è nota. Spesso è uno degli scopi di una ricerca.

Determinare "n"

E' un elemento fondamentale del piano di campionamento. Sulla scelta incidono...

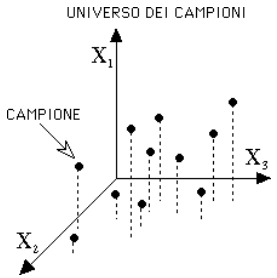
- **Obiettivo dell'indagine**
- **Variabilità attesa nella popolazione (controllo errori campionari)**
- **Costo dell'acquisizione**
- **Controllo errori non campionari**

La determinazione di "n" è molto complessa e verrà ripresa in un corso successivo

L'universo dei campioni

Fissata l'ampiezza campionaria "n" definiamo UNIVERSO DEI CAMPIONI (di ampiezza "n") l'insieme di tutti campioni di tale ampiezza che possono essere ottenuti da una data popolazione "P"

$$T_n(P) = \{C_1, C_2, \ldots, C_i, \ldots\}$$



L'universo dei campioni può anche essere considerato a sua volta una POPOLAZIONE le cui unità sono i campioni di ampiezza "n"

Cardinalità di $T_n(P)$

Dipende

- **Dalla possibilità di ripetere o no la stessa unità**
- **Se rileva o no l'ordine di comparizione nel campione.**

Se non c'è reimmissione ed i campioni sono considerati uguali purché formati dalle stesse unità allora la cardinalità è il coefficiente binomiale:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N*(N-1)*(N-2)*\ldots*(N-n+1)}{n!}$$

Esempio:
Ad un test sull'impatto visivo di un poster 6x3 metri sono stati invitati N=50 automobilisti che hanno dato la loro opinione. Di questi, n=7 dovrebbero essere sottoposti ad un altro test sulla leggibilità delle scritte.

Le scelte possibili sono: $\binom{50}{7} = \frac{50!}{7!43!} = 99'\,884'\,400$

Cardinalità/2

Se le unità possono ripetersi fermo restando che l'estrazione è senza reimmissione e che l'ordine non conta, la cardinalità è:

$$\binom{N+n-1}{n} = \frac{N*(N+1)*\ldots*(N+n-1)}{n!}$$

Esempio:
Si analizzano le N=70 sentenze emesse da un collegio giudicante con un'intervista di n=6 condannati. La presenza di recidivi può provocare la ripetizione delle unità.

I campioni possibili sono

$$\binom{70+6-1}{6} = \frac{70*71*\ldots*75}{6!} = 201'\,359'\,550$$

Cardinalità/3

Se è NON consentita la reimmissione e l'ordine diverso rende diversi due campioni con uguali unità allora i campioni possibili sono

$$\frac{N!}{(N-n)!}$$

Esempio:
La famosa scienziata ha intuito che un trattamento di 5 farmaci scelti tra 20 principi attivi e somministrati nel giusto ordine può curare una fastidiosa patologia.

Scegliendo a caso i principi attivi quanti campioni sono possibili?

$$\frac{20!}{15!} = 20*19*18*17*16 = 1'860'480$$

Cardinalità/4

Se è consentita la reimmissione e l'ordine diverso rende diversi due campioni con uguali unità allora l'universo dei campioni ha un numero di elementi pari a:

$$N^n$$

Nell'esempio del poster e del giudice si otterrebbe

$$50^7 = 781'250'000'000; \quad 70^6 = 117'649'000'000$$

il numero di elementi dell'universo dei campioni è quasi sempre troppo elevato - anche con i supercalcolatori- perché valga la pena di studiare il comportamento su tutti.

Effetti del Campionamento

Se si considerano tutte le unità di una popolazione, il problema della selezione delle unità non si pone.

Se non è possibile effettuare un'indagine completa ci saranno unità effettivamente esaminate ed altre no.

Il risultato è che ci si trova di fronte a dei dati che sono quelli, ma avrebbero potuto essere altri.

Cosa si può dire allora sui risultati ottenuti?

Errori campionari

L'uso del campione introduce un errore dovuto alle differenze tra risultati nel campione e risultati POTENZIALI ottenibili dall'esame di tutta la popolazione.

Gli errori sono dovuti a fluttuazioni in parte attribuibili alla naturale variabilità campionaria: i dati sono quelli, ma potevano essere altri

ESEMPIO
Vogliamo conoscere il totale dei valori della tabella (popolazione).

Si sceglie una riga o una colonna di cinque numeri (campione)

7	13	5	5	10
2	8	5	4	1
6	10	11	1	12
1	7	8	4	8
2	3	3	1	3

Riga	Stima	Errore camp.	Colonna	Stima	Errore camp.
40	200	60	18	90	-50
20	100	-40	41	205	65
40	200	60	32	160	20
28	140	0	15	75	-65
12	60	-80	34	170	30

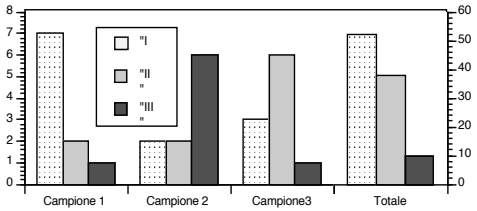
Solo se si sceglie la 4ª riga non c'è errore campionario

Errori campionari/2

Qualunque sia la conclusione raggiunta a mezzo del campione essa è dominata dall'incertezza.

il suo successo può corroborare la validità della procedura per il passato, ma ben poco può aggiungere sulla conoscenza del suo comportamento futuro.

ESEMPIO:
una variabile può assumere tre soli valori: 1,2,3.
Per stimare la sua distribuzione di frequenza: per campione si sceglie: prima colonna, ultima colonna e terza riga.



Popolazione dei valori

	1	2	3	4	5	6	7	8	8	9	10
1	1	1	1	3	1	1	2	3	1	1	2
2	1	1	1	3	2	3	1	1	2	2	1
3	2	2	2	2	2	2	2	1	1	1	3
4	3	3	2	1	3	2	1	2	1	2	3
5	3	1	3	1	1	2	1	1	1	3	3
6	1	1	1	2	2	1	2	2	1	1	2
7	1	3	2	3	1	1	2	2	2	1	1
8	1	1	1	1	2	2	1	1	2	1	3
9	1	2	3	1	2	2	1	3	3	2	3
10	1	2	2	1	2	1	1	1	3	2	3

Per ciascuno dei campioni si prenderà una decisione sbagliata

La rappresentatività

il campione deve essere RAPPRESENTATIVO della popolazione da cui è estratto, cioè assicurare che i risultati qui ottenuti si estendano a tutta la popolazione. Almeno in relazione alla caratteristica in esame

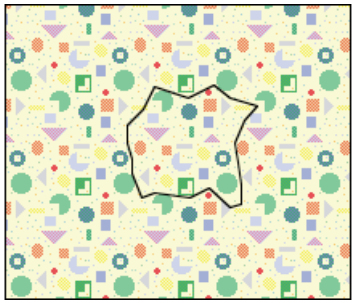
I giocatori di una squadra di basket non sono rappresentativi della popolazione per l'altezza, ma potrebbero esserlo per capacità di apprendimento.

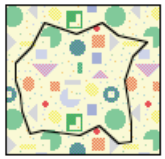
Solo la rilevazione totale è certamente rappresentativa ovvero la selezione di un numero qualsiasi di unità da una popolazione invariabile, ma così è inutile il campione.

La rappresentatività/2

La figlia vuole un vestito con il medesimo disegno di quello della madre. Che campione si dovrà portare al negozio di stoffe?

POPOLAZIONE





Campione

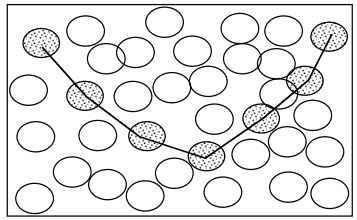
Il campione deve essere abbastanza piccolo per evitare di impacchettare l'intero vestito, ma deve anche essere abbastanza grande da includere il motivo ricorrente della stoffa.

La rappresentatività/3

Ci sono tanti campioni possibili

Alcuni sono rappresentativi cioè coincidono con l'intera popolazione per le variabili che ci interessano

Altri non sono rappresentativi e danno una idea distorta delle variabili di interesse.



Due fattori possono incidere sulla rappresentatività

L'ampiezza del campione "n"

La scelta casuale tra le unità della popolazione

Casualità e campionamento

E' casuale il meccanismo di scelta e non il campione ottenuto.

Perché scegliendo a caso dalla popolazione posso ottenere un campione rappresentativo?

Una popolazione è formata da tre tipi di unità: A, B, C di cui è nota la proporzione nella popolazione: p(A)=50%, p(B)=30%, p(C)=20%.

All'aumentare dell' ampiezza del campione, il meccanismo casuale di scelta porta a campioni che riproducono la popolazione

Ampiezza	A	B	C
n=10	0.6	0. 2	0.2
n=100	0.51	0.29	0.2
n=1000	0.512	0. 293	0.195
n=10000	0.5017	0.2983	0.2
n=100000	0.50047	0. 302	0.19573
n=1000000	0.500482	0.301929	0.197589
Popolazione	0.5	0.3	0.2

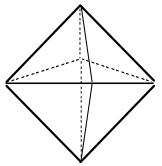
Questo è il postulato empirico del caso

Sorteggio delle unità

il modello della pura sorte può essere simulato in molti modi: monete, dadi.

Ogni processo fisico che nel suo funzionamento segua lo schema del sorteggio tra unità identiche può servire a formare il campione

ESEMPIO: l'ottaedro ha 8 facce uguali a forma di triangolo. Se fatto rotolare su di una superficie liscia finirà col poggiarsi su di una faccia. Se è ben costruito risulta imprevedibile la faccia su cui si poggerà



Accostando le uscite si ottiene il numero casuale in base ottale: 7201 convertibile in base decimale: $7*8^3+2*8^2+0*8+1= 3713$ o tra zero ed uno dividendo per 4095

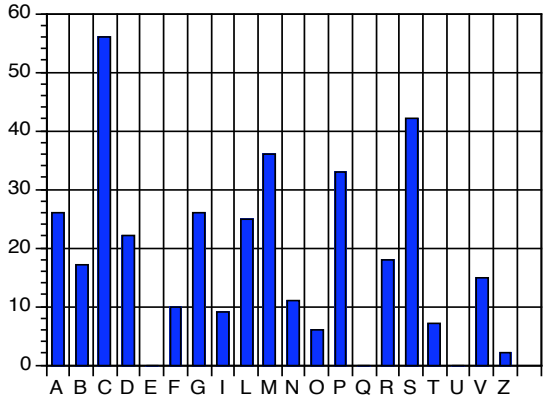
I giochi di sorte sono emblematici: le uscite sono casuali se nessun giocatore -per quanto furbo- riesce a determinare una regola che gli consenta di scommettere meglio che alla pari.

Esempio: la casualità dei cognomi

Ritenete casuale una selezione di unità che avvenga scegliendo a caso lettera dell'alfabeto italiano per includere tutte le unità che hanno un cognome che inizia con quella lettera?

Lettera	f.a.	f.r
A	26	7.18%
B	17	4.70%
C	56	15.47%
D	22	6.08%
E	0	0.00%
F	10	2.76%
G	26	7.18%
I	9	2.49%
L	25	6.91%
M	36	9.94%
N	11	3.04%
O	6	1.66%
P	33	9.12%
Q	0	0.00%
R	18	4.97%
S	42	11.60%
T	7	1.93%
U	0	0.00%
V	15	4.14%
Z	2	0.55%
	361	

Qual'è la lettera più diffusa per i vostri cognomi?



Estrazione con i numeri casuali

Nella pratica che precedette l'home computing il processo di estrazione dall'urna era spesso simulato con numeri casuali disponibili in tabelle

Le tavole dei numeri casuali sono delle raccolte di numeri da 0 a 9 variamente raggruppate e caratterizzate dall'assenza di una legge di successione o di ordinamento.

La costruzione è fatta in modo che le cifre da "0" a "9" hanno ciascuna frequenza 1/10 di ripetersi, le coppie da "00" a "99"frequenza 1/100, le terne "000"…"999" hanno 1/1000 ecc.

Per usare bene la tabella occorre selezionare, per sorteggio, il blocco riga/colonna da cui iniziare e poi prestabilire come continuare

Esempio di tabella di numeri casuali

	1	2	3	4
1	53 74 23 99 67 63 38 06 86 54 35 30 58 21 46 63 43 36 82 69 98 25 37 55 26	61 32 28 69 84 99 00 65 26 94 06 72 17 10 94 65 51 18 37 88 01 91 82 81 46	94 62 67 86 24 02 82 90 23 07 25 21 31 75 96 61 38 44 12 45 74 71 12 94 97	98 33 41 19 95 79 62 67 80 60 49 28 24 00 49 32 92 85 88 65 24 02 71 37 07
2	02 63 21 17 69 64 55 22 21 82 85 07 26 13 89 58 54 16 24 15 34 85 27 84 87	71 50 80 89 56 48 22 28 06 00 01 10 07 82 04 51 54 44 82 00 61 48 64 56 26	38 15 70 11 48 61 54 13 43 91 59 63 69 36 03 62 61 65 04 69 90 18 48 13 26	43 40 45 86 98 82 78 12 23 29 69 11 15 83 80 38 18 65 18 97 37 70 15 42 57
3	03 92 18 27 46 62 95 30 27 59 08 45 93 15 22 07 08 55 18 40 01 85 89 95 66	57 99 16 96 56 37 75 41 66 48 60 21 75 46 91 45 44 75 13 90 51 10 19 34 88	30 33 72 85 22 86 97 80 61 45 98 77 27 85 42 24 94 96 61 02 15 84 97 19 75	84 64 38 56 98 23 53 04 01 63 28 88 61 08 84 57 55 66 83 15 12 76 39 43 78
4	72 84 71 14 35 88 78 28 16 84 45 17 75 65 57 96 76 28 12 54 43 31 67 72 30	19 11 58 49 26 13 52 53 94 53 28 40 19 72 12 22 01 11 94 25 24 02 94 08 63	50 11 17 17 76 75 45 69 30 96 25 12 74 75 67 71 96 16 16 88 38 32 36 66 02	86 31 57 20 18 73 89 65 70 31 60 40 60 81 19 68 64 36 74 45 69 36 38 25 39
5	50 44 66 44 21 22 66 22 15 86 96 24 40 14 51 31 73 91 61 19 78 60 73 99 84	66 06 58 05 62 26 63 75 41 99 23 22 30 88 57 60 20 72 93 48 43 89 94 36 45	68 15 54 35 02 58 42 36 72 24 95 67 47 29 83 98 57 07 23 69 56 69 47 07 41	42 35 48 96 32 58 37 52 18 51 94 69 40 06 07 65 95 39 69 58 90 22 91 07 12
6	84 37 90 61 56 36 67 10 08 23 07 28 59 07 48 10 15 83 87 60 55 19 88 97 65	70 10 23 98 05 98 93 35 08 86 89 64 58 89 75 79 24 31 66 56 03 73 52 16 56	85 11 34 76 60 99 29 76 29 81 83 85 62 27 89 21 48 24 06 93 00 53 55 90 27	76 48 45 34 60 33 34 91 58 93 30 14 78 56 27 91 98 94 05 49 33 42 29 38 87
7	53 81 29 13 39 51 86 32 68 92 35 91 70 29 13 37 71 67 95 13 93 66 13 83 27	35 01 20 71 34 33 98 74 66 99 80 03 54 07 27 20 02 44 95 94 92 79 64 64 72	62 33 74 82 14 40 14 71 94 58 96 94 78 32 66 64 85 04 05 72 28 54 96 53 84	53 73 19 09 03 45 94 19 38 81 50 95 52 74 33 01 32 90 76 14 48 14 52 98 94
8	02 96 08 45 65 49 83 43 48 35 84 60 71 62 46 18 17 30 88 71 79 69 10 61 78	13 05 00 41 84 82 88 33 69 96 40 80 81 30 37 44 91 14 88 47 71 32 76 95 62	93 07 54 72 59 72 36 04 19 76 34 39 23 05 38 89 23 30 63 15 87 00 22 58 40	21 45 57 09 77 47 45 15 18 60 25 15 35 71 30 56 34 20 47 89 92 54 01 75 25

Estrazione con i numeri pseudo-casuali

Ci sono formule matematiche come i generatori congruenziali lineari, che, per opportune scelte dei parametri producono numeri tra 0 ed (m-1) aventi comportamento simile ai numeri casuali

$$X_{i+1} \equiv (aX_i + c) \text{Mod } m \qquad \text{con} \begin{cases} a = 293 \\ c = 1 \\ m = 1024 \\ X_1 = 68 \end{cases}$$

68	469	202	819	352	737	902	95	188	813	642	715	600	697	446
631	564	389	314	867	80	913	246	399	172	221	242	251	840	361
302	423	36	309	426	915	832	65	614	703	156	653	866	811	56
25	158	215	532	229	538	963	560	241	982	1007	140	61	466	347
296	713	14	7	4	149	650	1011	288	417	326	287	124	493	66
907	536	377	894	823	500	69	762	35	16	593	694	591	108	925

In molti computer si usa $m = 2^{64} = 1.8 \text{x} 10^{19}$

Se se ne usano un miliardo al secondo per finirli ci vorrebbero più di 200 mila anni

Numeri pseudo-casuali

I numeri pseudo-casuali sembrano prodotti dalla sorte, ma sono TUTTI noti a priori così come è nota la loro sequenza

$$X_k = Resto\left[aX_0 + \frac{(a^k - 1)c}{(a-1)}, m \right] \qquad [\ ]=\text{parte intera}$$

Basta conoscere il primo e tutti gli altri sono noti.

La sequenza è ciclica: dopo "m" valori i numero si ripetono nello stesso ordine

Il periodo "m" deve essere grande rispetto al campione da estrarre: Regola di Ripley

$$m \geq 200n^2 \Rightarrow se \quad n = 10'000 \quad m \geq 20'000'000'000 > 2^{31}$$

Campionamento casuale semplice

Un tipico esperimento che rientra nel calcolo combinatorio è la scelta casuale di "n" unità da una popolazione finita di "N".

L'evento elementare è la n-tupla di interi $C_i=(i_1,i_2,\ldots,i_n)$ corrispondenti alle posizioni di una lista univoca ed esaustiva delle unità della popolazione.

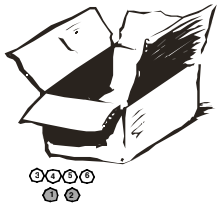
1	2	…	$i-1$	i	$i+1$	…	$n-2$	$n-1$	n
u	u		u	u	u		u	u	u

Se La popolazione è grande si usano tecniche di selezione computistiche.

Per ampiezze più piccole basta una scatola con delle biglie di uguale volume, superficie, temperatura, porosità, colore, lucidatura, etc.

Prima di ogni estrazione la scatola è ben scossa con moto sussultorio e ondulatorio.

Si mette in opera ogni accorgimento per assicurare la equiprobabilità delle biglie nella scatola.



Campione casuale semplice con reimmissione

Scelta di n=3 famiglie in una lista di N=100. Supponiamo che, dopo ogni estrazione, la biglia sia rimessa nell'urna che poi è adeguatamente scossa

La procedura descritta assicura che:

Ognuna delle N famiglie della popolazione ha la stessa probabilità di comparire in una qualsiasi delle n posizioni del campione:

$$P(Fam_i \ \ sia \ \ in \ \ posizione \ \ j) = \frac{100*100}{100*100*100} = \frac{1}{100} \qquad \frac{casi \ \ favorevoli}{casi \ \ possibili}$$

Ne consegue che ogni gruppo di n famiglie ha la stessa probabilità di costituire il campione:

$$P\left(Fam_{i_1},\ Fam_{i_2},\ Fam_{i_3}\right) = \frac{1}{100*100*100}$$

Campione casuale in blocco

Dopo ogni estrazione, la biglia NON è rimessa nell'urna.

La probabilità che la famiglia "i" compaia al 1° posto del campione è 1/(99*98)

Fam_i	2°	3°

il 1° posto e bloccato, il 2° può essere occupato dalle 99 restanti famiglie ed il 3° in 98 modi diversi dato che due famiglie impegnano già i primi due posti.

La stessa cosa succede per tutte le altre posizioni

1°	Fam_i	3°
99	1	98

1°	2°	Fam_i
99	98	1

Ognuna delle N famiglie della popolazione ha la stessa probabilità di comparire in una qualsiasi delle n posizioni del campione:

$$P(Fam_i \ \ sia \ \ in \ \ posizione \ \ j) = \frac{99*98}{100*99*98} = \frac{1}{100}$$

Campione casuale in blocco/2

Scelta la prima famiglia su N=100, la seconda è scelta su 99, la terza su 98.

Qualunque famiglia può essere la prima, la seconda o la terza. Ne consegue che:

Ogni gruppo di n famiglie ha la stessa probabilità di costituire il campione:

$$P\left(Fam_{i_1},\ Fam_{i_2},\ Fam_{i_3}\right) = \frac{1}{100*99*98}$$

E' chiaro che se una famiglia compare in una posizione non può comparire in un'altra dato che la scelta è senza reimmissione.

Ogni scelta, tranne la prima, dipende da quella e da quelle precedenti.

Proprietà del campione casuale

Il campione casuale ha le proprietà seguenti:

1) Ogni unità può comparire in una qualsiasi delle posizioni del campione.

2) Se le unità sono equiprobabili, ogni gruppo di unità ha la stessa probabilità di formare il campione

3) Se le unità sono scelte con reimmisione le singole scelte sono indipendenti

4) Se il campione è piccolo rispetto alla popolazione la differenza dovuta alla reimmissione/non reimmissione diventa trascurabile

Codice	Lista
Urto	77.08
Yama	78.56
Xilo	79.55
Wide	80.87
Zone	81.43
Dior	82.47
Idea	82.78
Urna	83.65
Mola	84.43
Ecos	85.31
Dato	86.42
Ceca	86.78
Alea	87.99
Tino	88.23
Erto	89.47
Mega	90.61
Doge	92.22
Alfa	92.23
Page	92.32
Foca	92.49
Jana	93.28
Tezo	93.50
Ossi	94.12
Dagà	94.25
Diga	94.75
Remo	95.28
Host	95.36
Sale	95.45
Biro	96.12
West	96.21
Riso	96.31
Cala	96.64
Indy	96.84
Posa	97.28
Demo	97.48
Duna	97.50
Quod	97.71
Wild	98.05
Yack	98.11
Nova	98.18
Vice	99.37
Nord	99.74
Sata	99.97
Home	100.10
Remi	100.69
Luce	101.24
Unno	101.52
Dole	101.69
Sito	101.88
Zeta	101.89
Vela	102.09
Beta	102.25
Paga	102.35
Tura	103.02
Nave	103.50
Xeno	103.85
Gelo	103.97
Arca	104.97
Raso	105.52
Zara	106.86
Sula	106.92
Oroo	107.24
Jena	108.17
Mina	109.27
Muta	110.13
Nero	111.65
Rada	112.81
Arpa	112.98
Lena	113.98
Giro	114.74
Hera	115.91
Jole	116.53
Oboe	116.96
Polo	118.61
Xmas	119.44

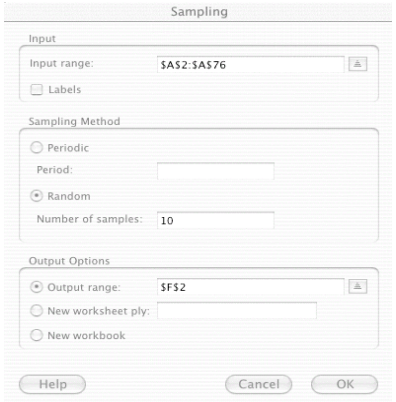
Excel: C.C.S. con Reimmissione

EXCEL ha una procedura per il campionamento con reimmissione

<u>Strumenti: Analisi dei dati: Campionamento</u>

Dopo aver indicato l'intervallo di input e il numero di unità da campionare Excel costruisce un campione casuale con reimmissione nella zona di output richiesta

Unità campionate	
Mola	84.43
Ceca	86.78
Jana	93.28
Remo	95.28
Remo	95.28
Remo	95.28
Demo	97.48
Home	100.1
Tura	103.02
Jole	116.53



Codice	Lista
Urto	77.08
Yama	78.56
Xilo	79.55
Wide	80.87
Zona	81.43
Dior	82.47
Idea	82.78
Urna	83.65
Mola	84.43
Ecos	85.31
Dato	86.42
Caca	86.78
Alea	87.99
Tino	88.23
Erto	89.47
Mega	90.61
Doge	92.22
Alfa	92.23
Page	92.32
Foce	92.49
Jena	93.28
Teco	93.50
Oasi	94.12
Daga	94.25
Diga	94.75
Remo	95.28
Host	95.36
Sale	95.45
Bits	96.12
West	96.21
Riso	96.31
Cala	96.84
Indy	96.84
Posa	97.28
Demo	97.48
Duna	97.50
Quod	97.71
Wild	98.05
Yack	98.11
Nora	98.18
Vice	99.37
Nord	99.74
Seta	99.97
Home	100.10
Remi	100.69
Luce	101.24
Unno	101.52
Dote	101.69
Sito	101.88
Zeta	101.89
Vela	102.09
Beta	102.25
Paga	102.35
Tura	103.02
Nave	103.50
Xeno	103.85
Gelo	103.97
Arca	104.97
Raso	105.52
Zara	106.86
Sula	106.92
Orso	107.24
Jena	108.17
Mina	109.27
Muta	110.13
Nero	111.65
Rada	112.81
Arpa	112.98
Lena	113.98
Giro	114.74
Hera	115.91
Jole	116.53
Oboe	118.06
Polo	118.61
Xmas	119.44

C.C.S. Senza Reimmissione

Si può procedere nel modo seguente

1)Si pone a fianco delle righe "popolazione" una nuova colonna SELETTORE in cui sia inserita la funzione CASUALE().

2) Si riordinano le righe in base al SELETTORE (ascendente o discendente non importa)

3) Le unità che si trovano nelle prime "n" posizioni sono il CCS-SR.

Codice	Lista	Selettore
Host	95.36	0.339044806
Demo	97.48	0.91048498
Beta	102.25	0.477187422
Orso	107.24	0.685980471
Page	92.32	0.256799318
Indy	96.84	0.309094032
Yack	98.11	0.239753839
Unno	101.52	0.469452099
Sula	106.92	0.554879814
Sito	101.88	0.721677718

E' poco probabile che due di questi valori siano uguali(se n<32768)

ESEMPIO

Indagine nell'ambito del corso sul grado di conoscenza della Legge sul lavoro *part-time* degli studenti nelle strutture universitarie

Estrazione di un campione casuale stratificato di ampiezza n=100

	Sesso		
Residenza	Donne	Uomini	
In sede	**9500**	**7000**	**17500**
Fuori sede	**1300**	**2200**	**3500**
	10800	**9200**	**20000**

POPOLAZIONE

campione

	Donne	Uomini	Totale
In sede	**48**	**35**	**83**
Fuori sede	**7**	**10**	**17**
TOTALE	**55**	**45**	**100**

RILEVAZIONE CAMPIONARIA

Grado di conoscenza	frequenza
Ottimo	**9**
Medio	**13**
Scarso o nullo	**78**
TOTALE	**100**

Selezione sistematica

Sia data una popolazione di "N" unità numerate da 1 a N dalla quale si deve estrarre un campione di ampiezza "n"

Si ricordi che l'intervallo (o passo) di campionamento è la frazione $f = \frac{N}{n}$

Caso di "f" intero

Si estrae un numero casuale "c" tra 1 e f "c". La 1^ unità estratta è quella in posizione r=c+1. La 2^ è in posizione r+f, la 3^ in r+2f.
In generale

$$\text{L'stratta } i-\text{esima è in posizione } r+(i-1)*f$$

Ad esempio se N=80 e n=8 (quindi con f=10) ed c=2 le unità estratte sono quelle nelle posizioni: {3, 13, 23, 33, 43, 53, 63, 73}

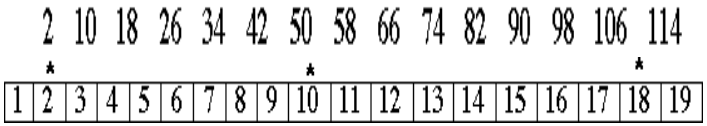
Da notare che il campionamento sistematico non va usato se il passo è legato all'ordine: se ogni 30 soldati c'è un sergente ed il passo è 30 il campione sarà formato da soli sergenti

Esempio

120 studenti di un corso sono elencati in ordine alfabetico e con numerazione progressiva. Si estrae un campione sistematico di 15 unità per partecipare ad un viaggio-studio

L'intervallo di campionamento è 120/15=8 unità. Scegliamo "r" nel blocco F-2: r=2

le unità prescelte sono



E' ovvio che se "f" è pari ed "r" è pari non sarà mai estratta una posizione dispari per cui la allocazione delle unità sulla lista non deve avere sistematicità almeno a livello di pari e dispari

Non si deve per forza avviare il conteggio dalla prima. Si può partire dalla N-esima ovvero numerare le unità da 1 ad N, ma da destra verso sinistra.

Caso di "f" non intero

Quando il passo risulta un numero frazionario ci sono varie possibilità

Si può approssimare f all'intero inferiore (difetto) ma le unità finali non hanno probabilità di essere scelte.

Esempio: **N=120, n=13** $f = \frac{120}{13} = 9.23 \cong 9$ $\{1,10,19,28,37,46,55,64,73,82,91,100,109\}$ $\{9,18,27,36,45,54,63,72,81,90,99,108,117\}$

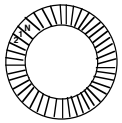
le unità 118,119,120 non usciranno

Si può approssimare all'intero superiore (eccesso), ma la numerosità del campione è inferiore a quella prefissata

Esempio: **N=120, n=13** $f = \frac{120}{13} = 9.23 \cong 10$ $\{1,11,21,31,41,51,61,71,81,91,101,111,?\}$ $\{9,19,29,39,49,59,69,79,89,99,109,119,?\}$

Per risolvere il problema si può selezionare a caso una unità tra tutte quelle non uscite e colmare la lacuna

Si può anche pensare ad una disposizione circolare



A che serve il campione?

Dalla rilevazione della variabile nelle unità del campione si esce con la serie di valori (data set o collettivo statistico)

$$\{X_i,\ i = 1,2,\ldots,n\}$$

che può essere utilizzata per stimare ...

Ammontare di una variabile $\hat{T} = \left(\frac{N}{n}\right)\sum_{X_i \in C} X_i \Rightarrow T = \sum_{i=1}^{N} X_i$

Valor medio di una variabile $\hat{\mu} = \frac{\sum_{X_i \in C} X_i}{n} \Rightarrow \mu = \frac{\sum_{i=1}^{N} X_i}{N}$

Varianza di una variabile $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Frazione di unità che possiede una certa caratteristica $\hat{\pi} = \frac{\sum_{X_i \in C} B_i}{n} \Rightarrow \pi = \frac{\sum_{i=1}^{N} B_i}{N}$; dove: $\begin{cases} B_i = 1 \text{ se presente} \\ B_i = 0 \text{ se assente} \end{cases}$

A che serve il campione?/2

N.I.P.

Differenze tra due valori medi $\hat{\mu}_1 - \hat{\mu}_2 = \frac{\sum_{X_i \in C_1} X_i}{n_1} - \frac{\sum_{Y_i \in C_2} Y_i}{n_2} \Rightarrow \mu_1 - \mu_2 = \frac{\sum_{i=1}^{N_1} X_i}{N_1} - \frac{\sum_{i=1}^{N_2} Y_i}{N_2}$

Rapporto tra due totali o due medie $\hat{R} = \frac{\sum_{X_i \in C} X_i}{\sum_{Y_i \in C} Y_i} \Rightarrow R = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} Y_i}$

Regressione lineare semplice $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Rightarrow y_i = \beta_0 + \beta_1 x_i$

Funzone di ripartizione di una variabile $\hat{F}(X) = \frac{\sum i}{\frac{x_{(i)} \le X}{n}}$