www.sce.unimore.it

Scienze della Comunicazione
e dell'Economia

ELEMENTI DI CAMPIONAMENTO

Luca La Rocca
luca.larocca@unimore.it

Si consideri la seguente notizia (Huff, 1954):

> *the average Yale man, Class of '24, makes $25111 a year.*

Cosa vuol dire? Fino a che punto fare studiare un figlio a Yale vuol dire garantirgli un futuro? (si tenga presente che all'epoca un reddito di 25111 dollari annui era molto elevato) È un'informazione affidabile?

Per rispondere conviene porsi alcune domande fondamentali:

1. chi lo dice?
2. come fa a saperlo?
3. cosa manca?

1. È una notizia riportata dalla stampa (non dalla Yale University): questo ci tranquillizza rispetto all'imparzialità di chi fornisce il dato (anche se dietro ogni giornale c'e' un editore) ma implica un interesse a enfatizzare il dato per "fare notizia" (publication bias).
2. È stato contattato un campione di laureati della Classe '24 ed è stato chiesto loro di riportare il reddito relativo all'anno trascorso: questo pone il problema della veridicità delle dichiarazioni rese (la tendenza a esagerare di alcuni compenserà la tendenza a minimizzare di altri?) e quello degli esclusi (chi saranno?)...
3. Manca una misura dell'incertezza associata al dato: non è plausibile che si conosca il reddito medio della Classe '24 "al dollaro"! (precisione $\neq$ accuratezza)

Chi resta escluso dal campione?

i. i laureati che non è stato possibile contattare (indirizzo sconosciuto): commessi, meccanici, artisti di scarso successo, disoccupati, barboni. . . non certo i dirigenti di grandi aziende o gli autori di bestseller!

ii. i laureati che è stato possibile contattare, ma che non hanno risposto: perché? Forse perché il loro reddito li imbarazza? Non abbiamo certezze sui non rispondenti, ma possiamo lo stesso provare a caratterizzarli. . .

. . . per fare un caso limite, se ponessimo la domanda “Ti piace compilare questionari?” con le stesse modalità, ci stupiremmo forse di trovare una larghissima maggioranza di risposte affermative?

Ci sono dunque buone ragioni per ritenere che il valore riportato come reddito medio dei laureati di Yale della Classe ’24 sia una sovrastima del valore vero e comunque il dato (che si presenta molto preciso) manca di un’adeguata indicazione della sua accuratezza. . .

. . . anche perché non ci viene detto quanti sono gli indirizzi sconosciuti e i non rispondenti (se chiedessimo “ti piace compilare questionari?” la proporzione di non rispondenti sarebbe il dato più interessante!)

Quanto a valutare se sia una buona idea far studiare un figlio a Yale, dipende anche da altre considerazioni: quello che vale per la Classe ’24 vale ancora al momento di prendere la decisione? Davvero ci interessa il reddito medio (o magari invece ci interessa il reddito mediano)?

Dobbiamo dunque disperare di potere avere un campione rappresentativo della popolazione da cui lo estraiamo?

No, dobbiamo “solo” essere

- consapevoli che la rappresentatività di un campione dipende dal meccanismo con cui esso è stato estratto;
- critici nei confronti di qualsiasi informazione statistica (always give the data a second look).

Nel seguito vedremo come prendere un campione casuale sia una buona strategia per proteggerci (in media) da possibili distorsioni.

Lasciando completamente al caso la scelta del campione:

- non abbiamo nessun controllo sulle unità statistiche che ne faranno parte (dunque perdiamo un'opportunità di scegliere per il meglio);
- sappiamo però descrivere efficacemente, in termini probabilistici, la variabilità dei risultati al variare del campione e possiamo sfruttare questa descrizione per valutare l'incertezza associata ai risultati effettivamente ottenuti.

Vedremo come la descrizione probabilistica della variabilità campionaria sia la chiave di volta della statistica inferenziale.

Se estraiamo a caso n aziende da una popolazione di numerosità N, prima dell'estrazione la quantità

$$X_i \quad = \quad \text{“fatturato dell'}i\text{-esima azienda estratta”}$$

è un numero aleatorio la cui distribuzione è determinata (classicamente) dalla distribuzione del fatturato nella popolazione; in particolare si avrà

$$\begin{aligned} \mathbb{E}[X_i] &= \mu \\ sd(X_i) &= \sigma \end{aligned}$$

dove μ e σ sono, rispettivamente, la media e la deviazione standard del fatturato nell'intera popolazione.

Per esempio (Borra & Di Ciaccio, 2008, Esempio 10.2.1) potremmo avere $N = 7$ grandi aziende operanti in un certo settore economico del paese i cui fatturati valgono

45, 49, 52, 56, 62, 65 e 74 milioni di euro.

Il fatturato medio di popolazione varrà

$$\mu = 57.57143 \quad \text{milioni di euro}$$

e la deviazione standard di popolazione

$$\sigma = 9.332847 \quad \text{milioni di euro;}$$

questi valori saranno il valore atteso e la deviazione standard...

...dei numeri aleatori $X_1 =$ "fatturato della prima azienda estratta" e $X_2 =$ "fatturato della seconda azienda estratta", definiti, quando si estraggano a caso $n = 2$ aziende, sullo spazio campionario

$$\begin{aligned}\Omega = \{&(45,49),(45,52),(45,56),(45,62),(45,65),(45,74),\\ &(49,45),(49,52),(49,56),(49,62),(49,65),(49,74),\\ &(52,45),(52,49),(52,56),(52,62),(52,65),(52,74),\\ &(56,45),(56,49),(56,52),(56,62),(56,65),(56,74),\\ &(62,45),(62,49),(62,52),(62,56),(62,65),(62,74),\\ &(65,45),(65,49),(65,52),(65,56),(65,62),(65,74),\\ &(74,45),(74,49),(74,52),(74,56),(74,62),(74,65)\}\end{aligned}$$

dove per semplicità le aziende sono identificate dai loro fatturati.

In pratica μ e σ sono quantità incognite: si tratta di parametri (caratteristiche di popolazione) e se potessimo conoscerle (con esattezza) non avremmo motivo di campionare...

...per ricordarci di questo fatto le denotiamo con lettere greche

`http://en.wikipedia.org/wiki/Greek_alphabet`

riservando quelle latine per le quantità calcolabili dai dati (statistiche).

Si noti il differente ruolo giocato, tipicamente, da μ e σ:

- μ è un parametro di interesse, nella misura in cui è sul suo valore che vogliamo trarre delle conlcusioni (fare inferenza);
- σ è un parametro di disturbo, nella misura in cui li suo valore ci interessa "solo" per descrivere la variabilità campionaria.

Dopo l'estrazione delle n aziende le quantità

$$x_i \quad = \quad \text{“fatturato dell'}i\text{-esima azienda estratta”}, \quad i = 1, \ldots, n,$$

sono note e vogliamo usarle per fare inferenza su μ...

...nonché su σ (per valutare la variabilità campionaria).

Si noti l'uso di lettere maiuscole e minuscole per distinguere tra numeri aleatori (prima dell'estrazione) e valori osservati (dopo l'estrazione).

Per esempio potremmo avere osservato $x_1 = 62$ e $x_2 = 45$, ottenendo $\bar{x} = (x_1 + x_2)/2 = (62 + 45)/2 = 53.5$; ci chiederemo allora:

- quanto potrà essere diverso $\bar{x}$ da μ?

La distribuzione campionaria del numero aleatorio

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

da cui proviene il valore osservato

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

è descritta dai suoi atomi e dalle loro probabilità. . .

. . . vale a dire dalla funzione di probabilità

$\bar{x}$	$\mathbb{P}\{\bar{X} = \bar{x}\}$	$\{\omega \in \Omega : \bar{X}(\omega) = \bar{x}\}$	$\bar{x}\mathbb{P}\{\bar{X} = \bar{x}\}$
47.0	0.04761905	{(45, 49), (49, 45)}	2.238095
48.5	0.04761905	{(45, 52), (52, 45)}	2.309524
50.5	0.09523810	{(45, 56), (49, 52), (52, 49), (56,45)}	4.809524
52.5	0.04761905	{(49, 56), (56, 49)}	2.500000
53.5	0.04761905	{(45, 62), (62, 45)}	2.547619
54.0	0.04761905	{(52, 56), (56, 52)}	2.571429
55.0	0.04761905	{(45, 65), (65, 45)}	2.619048
55.5	0.04761905	{(49, 62), (62, 49)}	2.642857
57.0	0.09523810	{(49, 65), (52, 62), (62, 52), (65,49)}	5.428571
58.5	0.04761905	{(52, 65), (65, 52)}	2.785714
59.0	0.04761905	{(56, 62), (62, 56)}	2.809524
59.5	0.04761905	{(45, 74), (74, 45)}	2.833333
60.5	0.04761905	{(56, 65), (65, 56)}	2.880952
61.5	0.04761905	{(49, 74), (74, 49)}	2.928571
63.0	0.04761905	{(52, 74), (74, 52)}	3.000000
63.5	0.04761905	{(62, 65), (65, 62)}	3.023810
65.0	0.04761905	{(56, 74), (74, 56)}	3.095238
68.0	0.04761905	{(62, 74), (74, 62)}	3.238095
69.5	0.04761905	{(65, 74), (74, 65)}	3.309524
Tot.	1.00000000	-	57.571428

La media campionaria ha valore atteso

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mu = \mu$$

e varianza

$$\mathcal{V}ar(\bar{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \quad \leq \quad \frac{\sigma^2}{n},$$

dove

- σ^2/n è il valore limite per popolazione infinita;
- $\frac{N-n}{N-1}$ è il fattore di correzione per popolazione finita.

Dunque lo scarto tra $\bar{x}$ e μ (nell'esempio pari a $53.5 - 57.6 = -4.1$) non avrà un segno privilegiato (in media non vi sarà distorsione) e il suo valore assoluto "tipicamente" varrà $sd(\bar{X}) = \sqrt{\mathcal{V}ar(\bar{X})}$...

...se rimpiazziamo σ^2 con la varianza campionaria osservata

$$sd_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

possiamo valutare tale scarto circa pari a

$$\sqrt{\frac{N-n}{N-1}}\,\frac{sd_X}{\sqrt{n}} = \sqrt{\frac{7-2}{7-1}}\,\frac{8.5}{\sqrt{2}} = 0.91 \times 6.01 = 5.5,$$

nell'esempio, avendo calcolato $sd_X = |x_2 - x_1|/2 = (62-45)/2 = 8.5$ (laddove $\sigma = 9.3$).

Il caso limite di popolazione infinita corrisponde a supporre di effettuare delle estrazioni con reinserimento:

- si basa su una teoria matematica più semplice, nella quale le osservazioni $X_1, \ldots X_n$ sono indipendenti, ed è un'ottima approssimazione della realtà nella misura in cui $n \ll N$;
- costituisce un approccio prudenziale, in quanto sovrastima la variabilità campionaria, visto che in pratica si cercherà di evitare l'eventualità di campionare due volte la stessa unità statistica.

Nel seguito supporremo senz'altro di essere in questo caso limite, nel quale si sviluppa "in prima approssimazione" l'inferenza statistica; di conseguenza il valore di N (numerosità di popolazione) non conterà, ma conterà solo il valore di n (numerosità campionaria).

Nel caso di popolazione finita (estrazioni senza reinserimento)
la media campionaria “alla fine” sarà uguale a quella di popolazione:
certamente $\bar{X} = \mu$ quando $n = N$.

La legge dei grandi numeri ci dice che qualcosa di analogo accade nel caso di popolazione infinita (estrazioni con reinserimento):

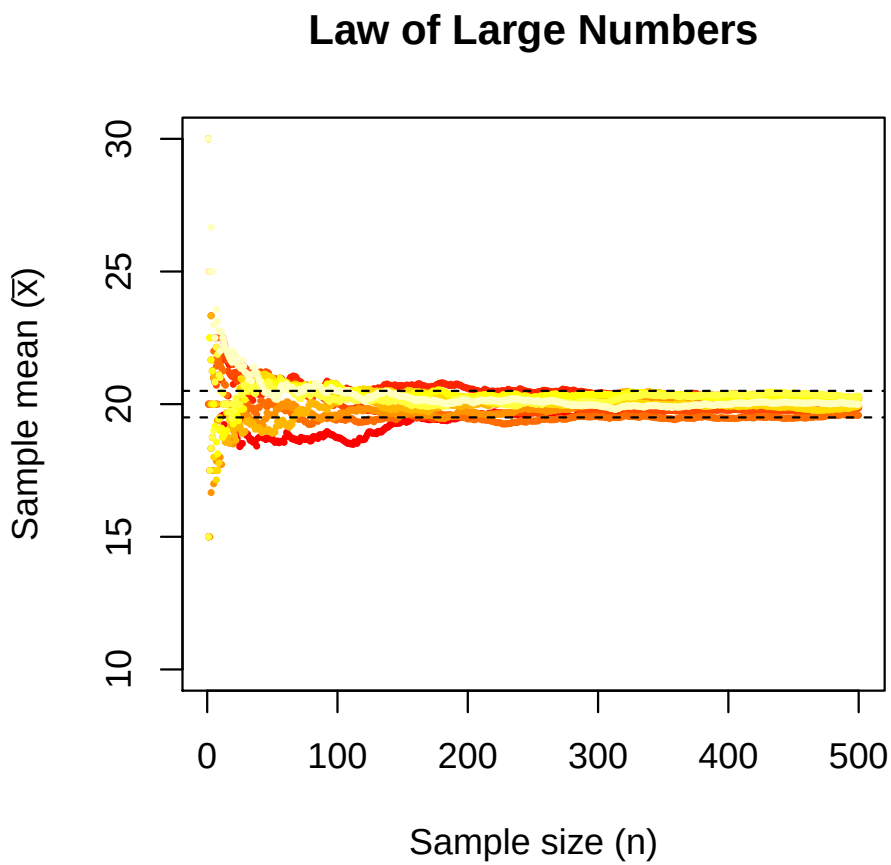
Theorem (Strong Law of Large Numbers)

Se $X_1, \ldots, X_n$ sono numeri aleatori indipendenti e identicamente distribuiti con $\mu = \mathbb{E}[X_1]$, allora quasi certamente (con probabilità uno) per $n \to \infty$ la loro media $\bar{X}$ converge a μ.

In pratica, come la figura seguente mostra, la media campionaria osservata $\bar{x}$ sarà "arbitrariamente" vicina alla media di popolazione μ, a patto di prendere n "abbastanza" grande...

...e in concreto questo sarà importante perché spesso n riesce a essere "abbastanza" grande anche quando $n \ll N$...

...i dati su cui la figura è basata sono stati ottenuti campionando ripetutamente con la funzione `sample()` di R da un'urna virtuale definita come `urn <- c(15, 15, 20, 20, 20, 30)`.


Law of Large Numbers
Sample mean ($\bar{x}$)
10
15
20
25
30
0
100
200
300
400
500
Sample size (n)

Un caso di particolare importanza è quello in cui si osservi un carattere dicotomico, per esempio chiedendo a un campione casuale semplice di cittadini se siano favorevoli o contrari a una certa proposta di legge.

In questo caso il parametro di interesse è la proporzione ψ di soggetti favorevoli alla proposta di legge nell'intera popolazione.

Se codifichiamo le osservazioni (prima dell'estrazione) come

$$X_i = \begin{cases} 1 & \text{se l'}i\text{-esimo intervistato è favorevole} \\ 0 & \text{se l'}i\text{-esimo intervistato è contrario} \end{cases}$$

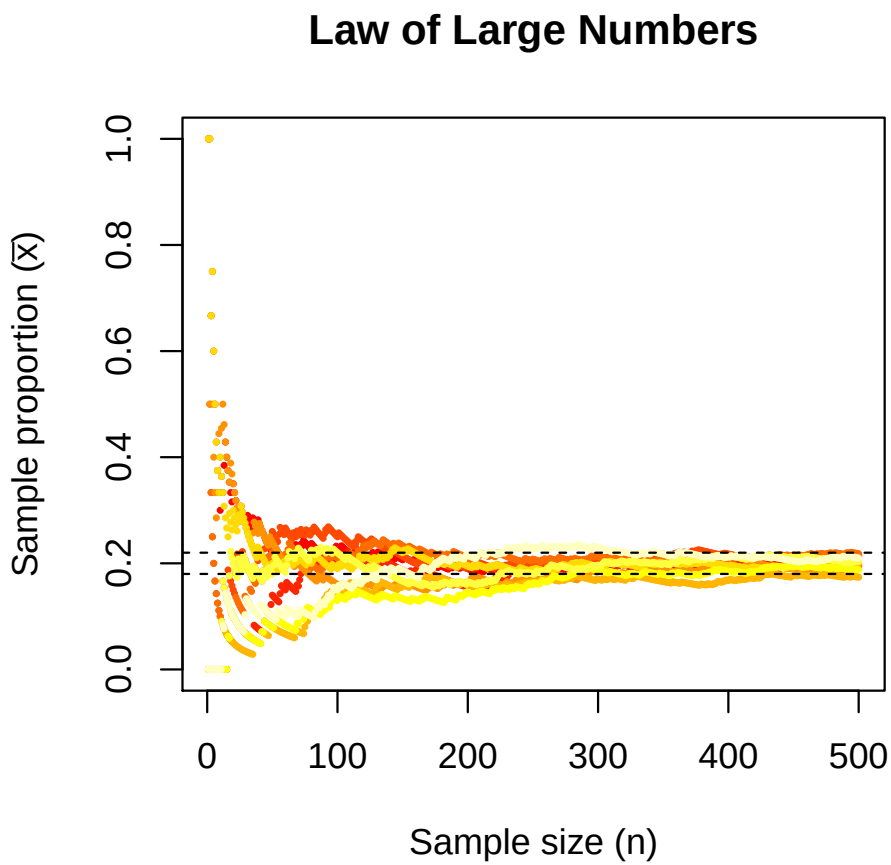
si verifica subito che $\mathbb{E}[X_i] = \psi \ldots$

...e $\mathcal{V}ar(X_i) = \psi(1-\psi)$, di modo che ci troviamo nella stessa situazione di prima, ma con la riparametrizzazione

$$\begin{aligned} \mu &= \psi \\ \sigma &= \sqrt{\psi(1-\psi)} \end{aligned}$$

che ha l'effetto di eliminare il parametro di disturbo.

In questo caso la legge dei grandi numeri, applicata alla proporzione campionaria $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, ci dice che, quasi certamente, per $n \to \infty$, la frequenza osservata $\bar{x}$ "coincide" con la probabilità ψ.


Law of Large Numbers
Sample proportion ($\bar{x}$)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
Sample size (n)

La legge dei grandi numeri è un teorema, vale a dire una conseguenza degli assiomi della probabilità e dell'ipotesi che $X_1, \ldots, X_n$ siano indipendenti e identicamente distribuiti con valore atteso $\mu(\psi)$;
di questo teorema si hanno riscontri empirici, quali ad esempio le figure precedenti, che confermano la bontà della teoria di cui esso fa parte in termini di applicabilità al mondo reale.

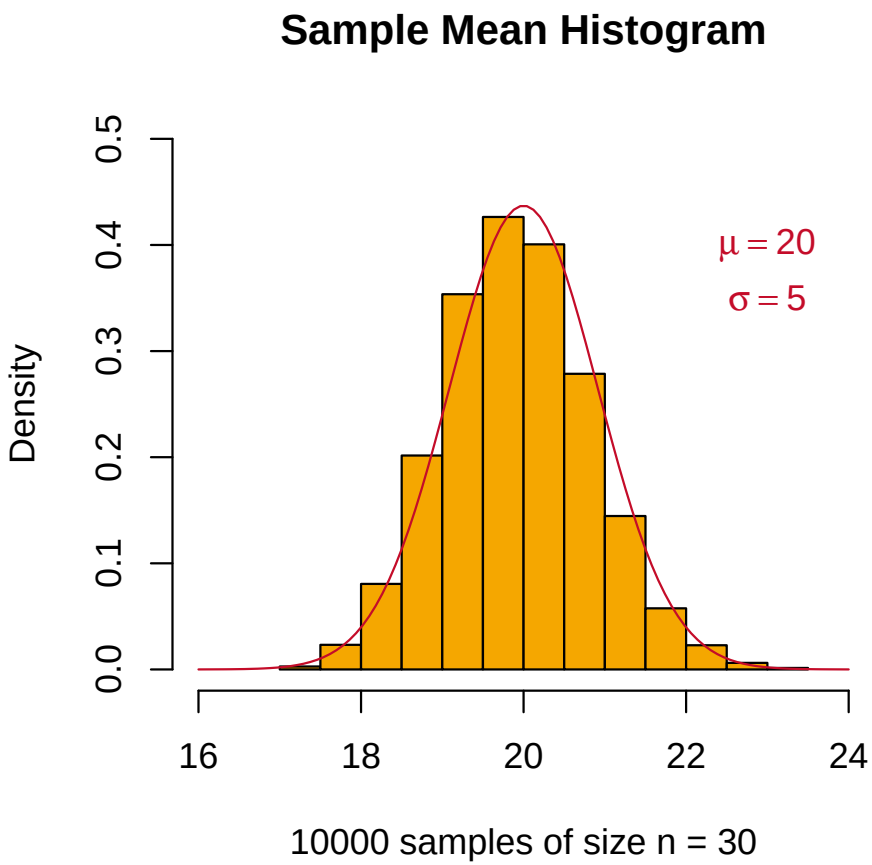
L'interpretazione frequentista della probabilità fa dei riscontri empirici un assioma (postulato empirico del caso) andando a definire la probabilità di $E_1, \ldots, E_n$ (dell'evento di cui $E_1, \ldots, E_n$ sono "prove" ripetute) come limite della proporzione $\bar{X}$, avendo posto $X_i = \mathbb{I}_{E_i}$ per ogni $i \geq 1$.

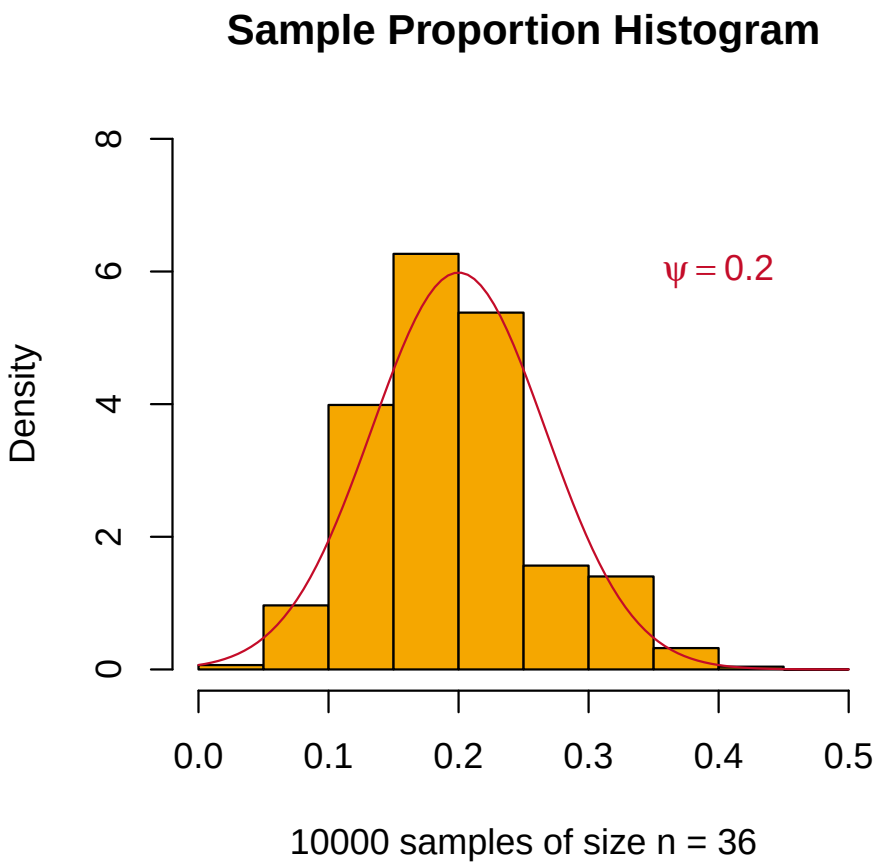
Il teorema del limite centrale ci dice che, per n "grande", la distribuzione campionaria di $\bar{X}$ è "ben" approssimata da una distribuzione normale (con media μ e deviazione standard $\sigma/\sqrt{n}$).

Quanto "grande" deve essere n perché l'approssimazione sia "buona"?

In teoria si tratta di un risultato asintotico (al limite per $n \to \infty$) in pratica dipende da quanto la distribuzione di X_1 (distribuzione di popolazione) differisce qualitativamente da una distribuzione normale (in particolare da quanto è asimmetrica): indicativamente, per fissare le idee, diciamo

- $n \geq 30$ nel caso generale
- $n\psi \geq 5$ e $n(1-\psi) \geq 5$ nel caso dicotomico

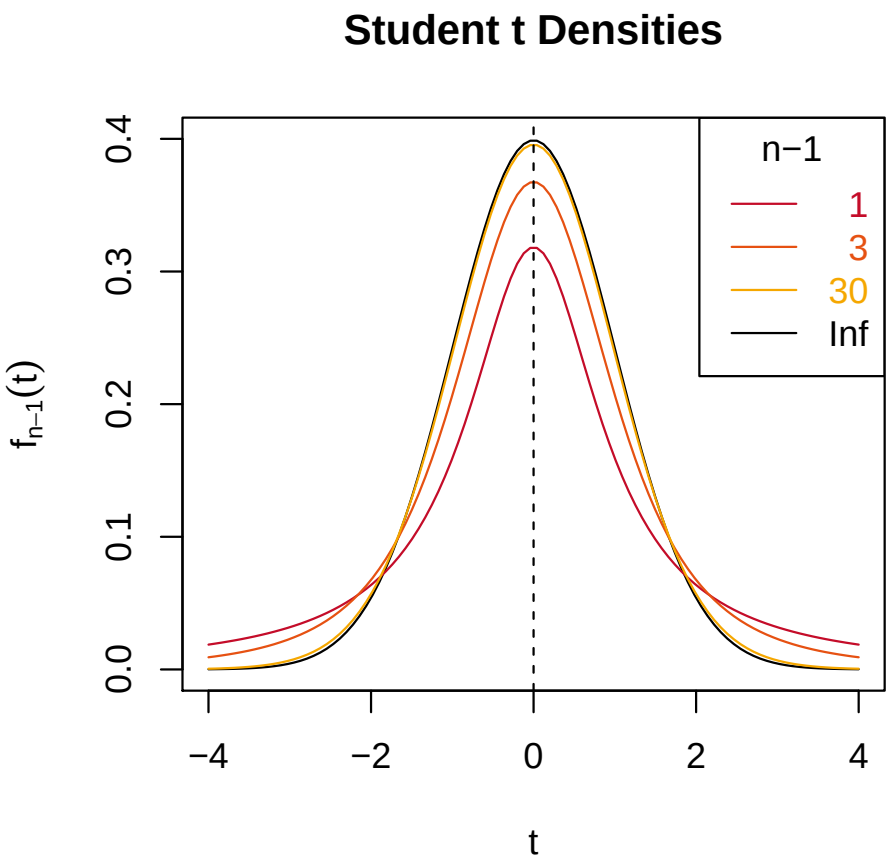

Sample Mean Histogram
Density
0.0
0.1
0.2
0.3
0.4
0.5
16
18
20
22
24
μ = 20
σ = 5
10000 samples of size n = 30


Sample Proportion Histogram
Density
0
2
4
6
8
ψ = 0.2
0.0
0.1
0.2
0.3
0.4
0.5
10000 samples of size n = 36

Se la distribuzione di popolazione ammette approssimazione normale (ipotesi di popolazione normale) la distribuzione campionaria di $\bar{X}$ può considerarsi normale anche in caso di "piccola" numerosità campionaria (diciamo per $n < 30$). In questo caso, tuttavia, è richiesta cautela nel rimpiazzare σ con sd_X (indispensabile ai fini pratici); si trova infatti che

$$T = \sqrt{n-1} \times \frac{\bar{X} - \mu}{sd_X}$$

segue una distribuzione *t* di Student con $n-1$ gradi di libertà (la cui densità è rappresentata nella figura seguente per $n-1 = 1, 3, 30$): questa ha code pesanti rispetto alla normale standard (in nero nella figura seguente) ma le si avvicina sempre più al crescere di $n-1$.


Student t Densities
n−1
1
3
30
Inf
$f_{n-1}(t)$
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
t

Per ottenere un campione casuale semplice,

> *dove l'aggettivo "semplice" indica che, fissata la numerosità campionaria n, tutti i possibili campioni sono equiprobabili,*

è necessario procedere con i seguenti passi:

1. individuare la popolazione di interesse e redigere una lista delle sue unità statistiche;
2. selezionare in modo completamente casualmente *n* unità statistiche dalla lista redatta al punto 1;
3. contattare le *n* unità statistiche selezionate al punto 2.

La risultante teoria matematica è decisamente semplice, ma possono sorgere alcuni problemi...

. . . nel seguito discutiamo informalmente alcune soluzioni, rinviando a Keeping (1962) per un approfondimento matematico.

In primo luogo potrebbe essere proibitivo attuare il passo 1 (elencare tutte le unità statistiche di interesse): si pensi per esempio al caso in cui interessino i clienti di un certo supermercato in un certa fascia oraria di un certo giorno. . .

. . . per certo non possiamo trattenerli tutti sino a sera prima di cominciare le interviste (e non è nemmeno pratico identificarli tutti e ricontattarli successivamente). . .

. . . però possiamo selezionare un cliente ogni k (es. $k = 20$) e ottenere in questo modo un campione sistematico (introdurremo un elemento di aleatorietà scegliendo a caso l'unità statistica da cui cominiciare).

Se la lista (implicita) da cui partiamo non presenta un ordinamento sistematico (es. la fascia oraria è breve) il campionamento sistematico si comporta come una buona approssimazione pratica del campionamento casuale semplice (avremo un campione di numerosità $n = N/k$); supponiamo in pratica che i clienti arrivino in ordine causale.

Se invece la lista presenta un ordinamento sistematico, la rappresentatività del campione può essere nulla:

> *Bailey (2006) riporta il caso di uno studio condotto durante la Seconda Guerra Mondiale in cui fu campionata sistematicamente una lista (esplicita) di soldati, prendendo* $k = 10$*; poiché i soldati erano raggruppati in plotoni da 10 unità e il primo soldato di ogni plotone era un sergente, non furono intervistati né caporali né soldati semplici.*

In secondo luogo, nel redigere la lista al punto 1, potremmo riuscire a raccogliere altre informazioni rilevanti sulle unità statistiche di interesse: per esempio (Borra & Di Ciaccio, 2008) se vogliamo stimare la spesa mensile media per generi alimentari delle famiglie residenti in un certa regione, potremmo conoscere la dimensione demografica dei centri abitati in cui vivono le famiglie di interesse...

...sapendo, da studi precedenti, che grandi centri abitati corrispondono, in media, a un minore consumo di generi alimentari, possiamo stratificare le famiglie a seconda che risiedano in un piccolo, medio o grande centro abitato e prendere un campione casuale semplice in ognuno dei tre strati...

...in questo modo otterremo un campione casuale stratificato.

Il campionamento casuale stratificato permette di

- migliorare l'accuratezza delle stime (se gli strati sono ben scelti);
- ottenere, oltre alla stima per l'intera popolazione, delle stime per i singoli strati (sottopopolazioni).

Gli strati saranno ben scelti quando quando la variabilità sarà:
i) bassa all'interno di ogni strato; ii) alta tra due strati distinti.

Il prezzo da pagare, ovviamente, è il reperimento delle altre informazioni rilevanti (oltre al lavoro da svolgere per la determinazione degli strati. . . strati mal scelti implicano che sia stato svolto del lavoro invano).

In terzo luogo potrebbe essere proibitivo attuare il passo 3 (contattare le n unità statistiche scelte in modo completamente casuale): se per esempio (Borra & Di Ciaccio, 2008) interessa studiare le caratteristiche delle abitazioni in una certa regione, potrebbe essere troppo costoso visitare n case senza alcun vincolo di prossimità...

...siccome però le case saranno elencate per comune (o addirittura potremmo avere solo l'elenco dei comuni) possiamo scegliere a caso un certo numero di comuni (grappoli) e rilevare le informazioni che ci interessano per tutte le case ubicate nei comuni scelti...

...otterremo in questo modo un campione casuale a grappoli.

Il campionamento casuale a grappoli permette di

risparmiare sui costi di campionamento

al prezzo di

peggiorare le stime (se i grappoli non sono buoni).

I grappoli saranno buoni nella misura in cui ognuno di essi sarà rappresentativo dell'intera popolazione (si noti la differenza con il campionamento casuale stratificato).

Se i grappoli sono buoni, un ulteriore risparmio può ottenersi mediante campionamento casuale semplice applicato ai grappoli scelti: otterremo in questo modo un campione casuale a stadi (due stadi).

- BAILEY, K.D. (2006).
  Metodi della Ricerca Sociale. Volume 2: L'Inchiesta.
  Il Mulino, Bologna.
- BORRA, S. & DI CIACCIO, A. (2008).
  Statistica: Metodologie per le Scienze Economiche e Sociali (Seconda Edizione).
  McGraw-Hill, Milano.
- HUFF, D. (1954).
  How to Lie with Statistics.
  Norton, New York.

KEEPING, E.S. (1962).
Introduction to Statistical Inference.
Van Nostrand, Princeton.