Università di Firenze - Corso di Laurea in Statistica
Statistica
Leonardo Grilli
A.A. 2010/2011

Distribuzioni campionarie

Cicchitelli: cap 15

Statistica descrittiva vs inferenziale

- **Statistica descrittiva**
  - Raccogliere, presentare, e descrivere i dati
- **Statistica inferenziale**
  - Trarre conclusioni e/o prendere decisioni riguardanti una popolazione sulla base dei dati campionari

Popolazione e Campione

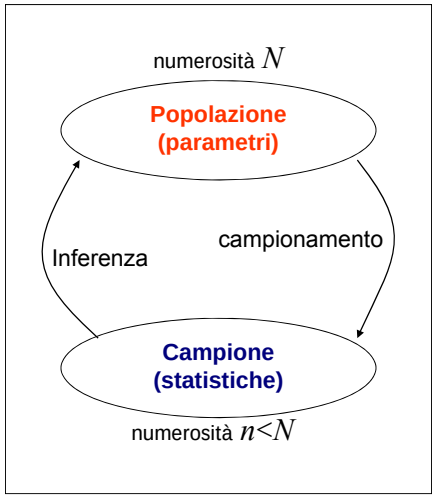
- Una **Popolazione** è l'insieme di tutte le unità o individui oggetto di studio
  - Tutti i potenziali votanti nelle prossime elezioni
  - Tutti i pezzi prodotti oggi
  - Tutti gli scontrini di novembre
- Un **Campione** è un sottoinsieme della popolazione
  - Alcuni votanti selezionati a caso per un'intervista
  - Alcuni pezzi selezionati per un test di distruzione
  - Alcuni scontrini selezionati a caso per una verifica

Perché usare un campione?

- Un campione consente di ottenere risultati statistici con precisione sufficientemente alta
- Presenta notevoli vantaggi rispetto ad un censimento
  - Organizzazione più semplice
  - Minore spesa
  - Tempi più brevi

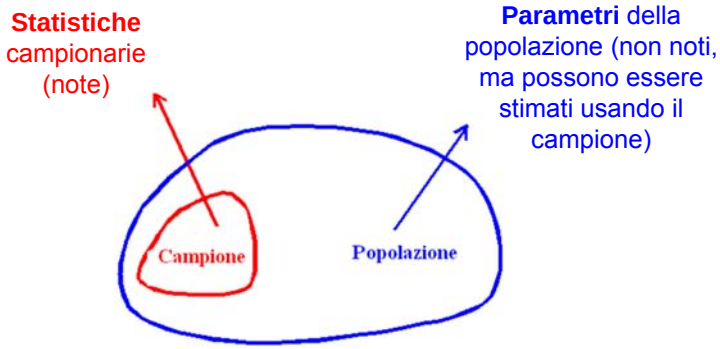
Campionamento e inferenza

- **Campionamento**: modalità di estrazione del campione dalla popolazione
- **Inferenza**: processo di generalizzazione per il quale i risultati ottenuti su un campione vengono estesi alla popolazione



Inferenza statistica

- Facciamo inferenza sulla popolazione esaminando i risultati campionari



Deduzione vs Induzione

- Deduzione
  - dal generale al particolare
  - tipica della logica e della matematica
  - conclusioni certe
- Induzione
  - dal particolare al generale
  - tipica delle discipline scientifiche
  - conclusioni incerte

Inferenza statistica e induzione

L'**INFERENZA STATISTICA** è un procedimento di induzione di tipo quantitativo, per cui l'incertezza del procedimento viene quantificata (si traduce in uno o più numeri)

- L'incertezza è dovuta a due fonti principali:
  - **Variabilità campionaria** (in principio, tutti i possibili campioni sono diversi e quindi la loro analisi produce risultati diversi; noi disponiamo di un solo campione)
  - **Errori di misurazione** (in molti casi ripetendo la misurazione della stessa entità si ottengono valori diversi; tipicamente noi disponiamo di una sola misurazione per ogni entità)

Logica del campionamento /1

- Popolazione di N unità, carattere X con distribuzione

Modalità	Freq. Rel.
x_1	f_1
.	.
x_i	f_i
.	.
x_k	f_k
Totale	1

- Si estrae un'unità a caso (con equiprobabilità) → il valore che si osserverà è una v.a. X con
  - supporto = modalità
  - probabilità = freq.rel.
- La distribuzione di probabilità della v.a. X coincide con la distribuzione delle frequenze relative del carattere X → tutti gli indici coincidono, ad es. la media (valore atteso) della v.a. X è uguale alla media del carattere X

Logica del campionamento /2

- **Estrazione a caso di 1 unità** (con equiprobabilità)
  → v.a. X (con distribuzione coincidente con quella del carattere X)
- **Estrazione a caso di n unità** (con equiprobabilità e con reimbussolamento)
  → n v.a. $X_1,X_2,\ldots,X_n$ *indipendenti e con identica distribuzione* (coincidente con quella del carattere X)

Natura del campione

- Si consideri l'esperimento aleatorio che consiste nel estrarre le unità e osservare i valori
- **Prima** dell'esperimento il campione è un vettore di *n* **variabili aleatorie** $X_1,X_2,\ldots,X_n$ (lettere maiuscole)
- **Dopo** l'esperimento il campione è un vettore di *n* **numeri** $x_1,x_2,\ldots,x_n$ (lettere minuscole)

Parametri

- Tipicamente l'inferenza riguarda alcuni parametri (indici relativi alla distribuzione del carattere di interesse nella popolazione, es. media, mediana, deviazione std)
- Poiché il campionamento casuale genera una v.a. X con la stessa distribuzione del carattere X, si chiamano parametri anche gli indici della distribuzione della v.a. X
- Se si assume che la v.a. X appartenga ad una certa famiglia parametrica (es. Normale, Poisson) i parametri della famiglia rappresentano gli aspetti ignoti della popolazione
  - Es. se X ha distribuzione Normale gli aspetti ignoti sono ricondotti a due soli parametri, la media e la deviazione std. (se si conoscono questi due parametri allora si conosce l'intera distribuzione)
- I parametri sono quantità
  - Fisse (cioè non sono v.a. – approccio frequentista)
  - Incognite

Campionamento da popolazione finita

- Popolazione **finita**: esiste un collettivo di N unità da cui se ne estraggono casualmente n<N
- Fasi:
  1. elencare e numerare preventivamente le N unità;
  2. estrarre a caso n numeri (es. n estrazioni da un'urna con palline numerate da 1 a N);
  3. registrare il valore del carattere (sottoporre a test i prodotti, intervistare le persone …)
- Varie strategie di campionamento
  - Semplice (= equiprobabilità) vs complesso (= diverse prob.)
  - Con ripetizione (reimbussolamento) o senza (in blocco)

Popolazioni infinite /1

- La popolazione è **infinita** tutte le volte che non è esattamente delimitata, cioè non è concettualmente possibile elencare i suoi membri
  - In una indagine di mercato si seleziona un campione di 200 consumatori da intervistare. Chiaramente non si è interessati ai quei 200 consumatori, ma ai consumatori in generale, quindi i dati relativi ai 200 intervistati devono essere generalizzati ad una più ampia popolazione. Ma quali sono i contorni di tale popolazione? Si tratta dei consumatori in astratto, quelli del Nord e del Sud, quelli di oggi e di domani … si tratta quindi di una popolazione non esattamente definita, quindi di numerosità infinita

Popolazioni infinite /2

- Esempio: si prelevano a caso 50 confezioni di pasta da un processo industriale automatizzato: la popolazione è costituita dalle confezioni potenzialmente producibili, quindi è infinita
  - Carattere di interesse: X = peso della confezione
  - Ipotesi: $X \sim N(\mu,\sigma^2)$
  - Peso *i*-ma confezione prelevata: $X_i \sim N(\mu,\sigma^2)$
  - Peso delle 50 confezioni prelevate: $X_1, X_2, \ldots X_{50}$ indipendenti $\sim N(\mu,\sigma^2)$

Popolazioni infinite /3

- La popolazione è **infinita** anche nei casi in cui non vi è alcuna estrazione di unità, ma (al fine di generalizzare i risultati) è opportuno considerare il valore del carattere X nelle unità sotto osservazione come realizzazione di un processo aleatorio:
  - una nuova terapia viene applicata a 20 soggetti "omogenei", cioè soggetti che per le loro caratteristiche (età, anamnesi,…) hanno identica probabilità di guarigione p (incognita):
    - esito della terapia per il soggetto *i* → $X_i \sim Be(p)$
    - esito per i 20 soggetti → $X_1, X_2, \ldots X_{20}$ indipendenti $\sim Be(p)$

Popolazioni infinite /4

- Spesso si assume che la distribuzione del carattere nella popolazione segua una distribuzione di tipo continuo (es. Normale)
- In tal caso implicitamente si assume che la popolazione sia infinita ($N=\infty$): infatti, la distribuzione di un carattere in una popolazione finita, per quanto numerosa, è necessariamente discreta e quindi non può essere esattamente Normale (al massimo può essere approssimativamente Normale se N è grande)

Popolazione generatrice

- Inferenza: procedimento induttivo che mira ad estendere ciò che si osserva su un caso particolare (il campione) ad una realtà più ampia (la popolazione)
- Inferenza statistica (o probabilistica): l'**incertezza** del procedimento induttivo viene **quantificata** in modo rigoroso
- Per fare inferenza statistica ciò che si osserva nel campione va considerato come realizzazione di una variabile aleatoria
- Che la popolazione sia finita o infinita, che vi sia o non vi sia un'estrazione, l'oggetto di studio è una v.a. X, detta **popolazione generatrice**

Campione casuale

- La teoria statistica di base si basa sulla nozione di **campione casuale**
- Indichiamo con X la v.a. che descrive la distribuzione del carattere nella popolazione e supponiamo di estrarre un campione di dimensione n
- Prima di effettuare l'estrazione, il valore che mostrerà la *i*-ma unità estratta è ignoto, è una v.a. che indichiamo con X_i
- Il campione è quindi un vettore di n v.a. $X_1, X_2, \ldots X_n$

Campione casuale

- Il campione si dice ***campione casuale*** quando le n v.a. $X_1, X_2, \ldots X_n$ sono iid-X, cioè **i**ndipendenti e **i**denticamente **d**istribuite come X
  - Indipendenti: la distribuzione di probabilità di un elemento campionario X_i non dipende dai valori assunti dagli altri elementi campionari. Questo accade se vi è indipendenza nella popolazione e il metodo di campionamento preserva tale indipendenza
  - Identicamente distribuite come X: tutti gli elementi campionari hanno la stessa distribuzione (sono dei cloni) e tale distribuzione è la stessa del carattere nella popolazione (= ogni X_i ha la stessa distribuzione di X). Questo accade se le probabilità di estrazione sono identiche per tutte le unità della popolazione e non vi sono problemi di mancata risposta o errore di misurazione

Campione casuale

- Dire che $X_1, X_2, \ldots X_n$ è un *campione casuale da una popolazione Normale* significa dire che
  - La distribuzione del carattere di interesse X nella popolazione è Normale (di solito media e varianza sono ignote)
  - Le v.a. $X_1, X_2, \ldots X_n$ sono indipendenti
  - Le v.a. $X_1, X_2, \ldots X_n$ hanno tutte la stessa distribuzione e tale distribuzione è Normale e coincide con quella del carattere nella popolazione. Ad es.

campione casuale n=3 da $X \sim N(6,2)$ significa
$X_1 \sim N(6,2)$ $X_2 \sim N(6,2)$ $X_3 \sim N(6,2)$ tra loro indipendenti

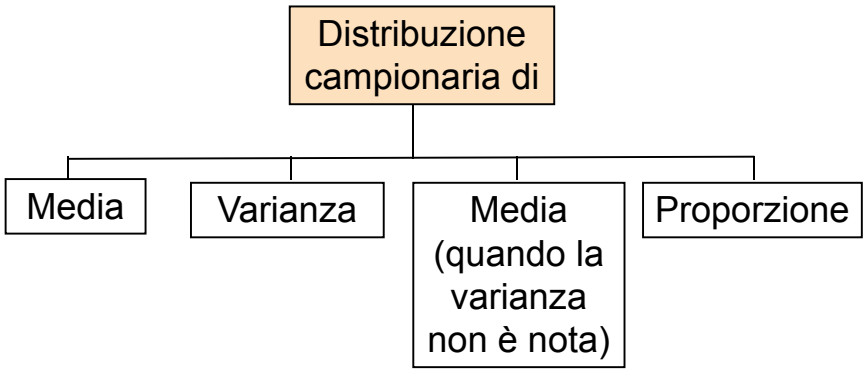
OSSERVAZIONE: quando si assume che il carattere abbia una distribuzione continua l'insieme dei possibili campioni è infinito. Ad es. i campioni di ampiezza 3 da una popolazione Normale sono tutte le possibili terne di numeri reali

Campione casuale

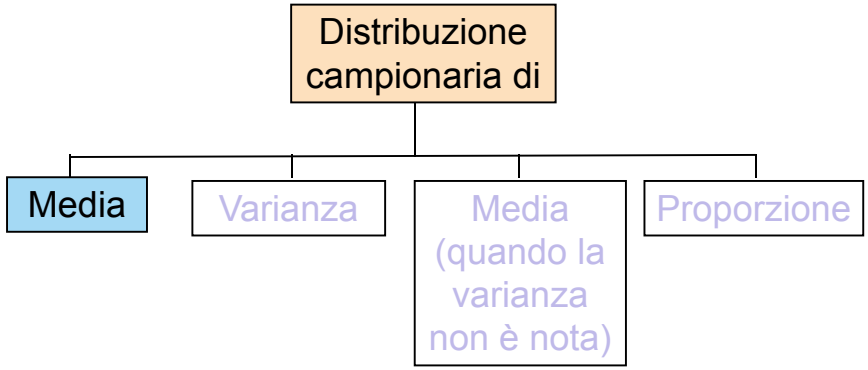
- Il metodo di campionamento "senza ripetizione" non può generare un "campione casuale" poiché induce correlazione tra gli elementi campionari
- Viceversa il metodo di campionamento "con ripetizione" può generare un "campione casuale", ma non è detto che ciò accada:
  - Infatti se i valori sono correlati nella popolazione, lo sono anche nel campione, qualunque metodo di campionamento si usi
  - Es. la serie storica del prezzo di scambio di un titolo azionario ad ogni singola transazione è caratterizzata da una forte correlazione (il prezzo ad una certa transazione dipende dai prezzi alle transazioni precedenti); comunque si effettui l'estrazione, le unità campionarie non sono indipendenti

Distribuzioni campionarie

Una **distribuzione campionaria** è una distribuzione di tutti i possibili valori di una statistica ottenuti da campioni della stessa ampiezza estratti dalla popolazione



Distribuzione campionaria della media



Stimatori

- POPOLAZIONE (numerosità N) → *parametri* (μ_X, σ_X, …)
- CAMPIONE (numerosità $n<N$) → *statistiche* ($\bar{X}$, S , …)

Ogni quantità della popolazione (parametro) ha un suo analogo nel campione (statistica). Ad es. al parametro

media del carattere nella popolazione, indicata con la lettera greca μ

corrisponde la statistica

media del carattere nel campione, detta "media campionaria" e indicata con la lettera latina $\bar{X}$

E' naturale quindi cercare di stimare un parametro di interesse (es. μ_X) con la corrispondente statistica (cioè $\bar{X}$). Quando una statistica viene usata a fini inferenziali per stimare un parametro viene detta ***stimatore*** (es. $\bar{X}$ è uno stimatore di μ_X).

Inferenza sulla media

- Tipicamente il parametro di interesse primario è la ***media della popolazione***
- Disponendo di un campione, lo *stimatore* naturale della media della popolazione è la ***media campionaria***

Parametro:	Stimatore:	Stima:
μ: media della popolazione	$\bar{X}$: variabile aleatoria media campionaria (stimatore di μ)	$\bar{x}$: valore assunto dalla media campionaria nel particolare campione osservato (stima di μ)

Esempio /1

Supponiamo che l'intera popolazione di interesse sia composta da $N = 4$ individui, sui quali si misura la variabile "numero di libri letti nell'ultimo mese":

Individuo	Libri letti
A	0
B	1
C	1
D	3

Esempio /2

La distribuzione della variabile di interesse nella popolazione è

Libri letti	Freq. Rel.
0	0.25
1	0.50
3	0.25
Totale	1

Questa è anche la distribuzione di probabilità della v.a. X = numero di libri letti da un individuo estratto a sorte da questa popolazione

I due principali indici di sintesi della popolazione (parametri) sono:

$$\mu_X = 1.25 \qquad \sigma_X^2 = 1.1875$$

Esempio /3

- Supponiamo di non poter osservare i dati relativi all'intera popolazione → si estrae un *campione*
  - *casuale* (= le unità della popolazione vengono estratte a caso)
  - *semplice* (= tutte le unità hanno la stessa probabilità di essere estratte)
  - di dimensione $n = 2$
- Il campionamento può essere
  - *con ripetizione* → i possibili campioni sono $N^n = 4^2 = 16$
  - *senza ripetizione* → i possibili campioni sono $4 \times 3 = 12$
- Poiché i possibili campioni sono pochi è possibile elencarli
- Una volta fatta l'estrazione si osserverà solo uno dei possibili campioni

Campionamento CON ripetizione

Campione	Individui selezionati	Risultato campionario	Media campionaria $\bar{X}$
1	A, A	0, 0	0
2	A, B	0, 1	0.5
3	A, C	0, 1	0.5
4	A, D	0, 3	1.5
5	B, A	1, 0	0.5
6	B, B	1, 1	1
7	B, C	1, 1	1
8	B, D	1, 3	2
9	C, A	1, 0	0.5
10	C, B	1, 1	1
11	C, C	1, 1	1
12	C, D	1, 3	2
13	D, A	3, 0	1.5
14	D, B	3, 1	2
15	D, C	3, 1	2
16	D, D	3, 3	3

Ogni campione ha probabilità 1/16 di essere estratto

Distribuzione campionaria della media

Campione	Media campionaria $\bar{X}$	Freq. Rel. (= probabilità)
1	0	1/16 = 0.0625
2,3,5,9	0.5	4/16 = 0.2500
6,7,10,11	1	4/16 = 0.2500
4,13	1.5	2/16 = 0.1250
8,12,14,15	2	4/16 = 0.2500
16	3	1/16 = 0.0625
		1.0000

L'estrazione di un campione e il calcolo della media sul campione estratto è un esperimento aleatorio: prima di estrarre il campione il valore della media che si otterrà è ignoto, ma è possibile determinare quali valori si potranno osservare e con quale probabilità → prima di estrarre il campione, la media campionaria è una variabile aleatoria. Nell'esempio è una v.a. discreta, che assume il valore 0 con probabilità 0.0625 (se viene estratto il campione n. 1), 0.5 con probabilità 0.2500 (se viene estratto il campione n. 2, oppure n. 3, oppure n. 5, oppure n. 9), ecc.

Ignorare le etichette: "campione casuale"

Un approccio equivalente consiste nell'ignorare le etichette A, B, … e considerare solo i valori. Poiché si effettuano estrazioni casuali con reimmissione, il risultato dell'esperimento campionario è una v.a. doppia (X_1, X_2) che ha le caratteristiche di "campione casuale", per cui

$$P(X_1 = x_1 \cap X_2 = x_2) = P(X_1 = x_1) \times P(X_2 = x_2)$$
$$P(X_i = 0) = 0.25, \quad P(X_i = 1) = 0.50, \quad P(X_i = 3) = 0.25$$

Campione	Probabilità	Media campionaria $\bar{X}$
0, 0	0.25×0.25 = 0.0625	0
0, 1	0.25×0.50 = 0.1250	0.5
0, 3	0.25×0.25 = 0.0625	1.5
1, 0	0.50×0.25 = 0.1250	0.5
1, 1	0.50×0.50 = 0.2500	1
1, 3	0.50×0.25 = 0.1250	2
3, 0	0.25×0.25 = 0.0625	1.5
3, 1	0.25×0.50 = 0.1250	2
3, 3	0.25×0.25 = 0.0625	3

La distribuzione campionaria della media è come prima

Sintetizzare la distribuzione campionaria

I due principali indici di sintesi di una v.a. sono il valore atteso e la varianza. Nell'esempio

$$\mu_{\bar{X}} = E(\bar{X}) = 0 \times 0.0625 + 0.5 \times 0.2500 + \ldots = 1.25$$

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = (0-1.25)^2 \times 0.0625 + (0.5-1.25)^2 \times 0.2500 + \ldots = 0.59375$$

$$\sigma_{\bar{X}} = \sqrt{0.59375} = 0.77055$$

La media campionaria è uno stimatore della media della popolazione.

Una volta estratto il campione, lo **stimatore** (che è una v.a.) produce una **stima** (che è un numero): ad es., se viene estratto il campione n. 6 lo stimatore $\bar{X}$ produce la stima $\bar{x} = 1$

Per convenzione: stimatore → lettera latina maiuscola
stima → corrispondente minuscola

La stima è un procedimento inferenziale (cioè induttivo) e quindi è soggetto ad errore → occorre quantificare l'errore

Errore di stima

Campione	Media camp. $\bar{X}$	Media popol. μ_X	Errore di stima $\bar{X} - \mu_X$	Freq. Rel. (= probabilità)
1	0	1.25	–1.25	1/16 = 0.0625
2,3,5,9	0.5	1.25	–0.75	4/16 = 0.2500
6,7,10,11	1	1.25	–0.25	4/16 = 0.2500
4,13	1.5	1.25	+0.25	2/16 = 0.1250
8,12,14,15	2	1.25	+0.75	4/16 = 0.2500
16	3	1.25	+1.75	1/16 = 0.0625
				1.0000

Ogni campione è caratterizzato da un **errore di stima**, ad es.

- se viene estratto il campione n. 3 la stima è 0.5 → sottostima di –0.75
- se viene estratto il campione n. 4 la stima è 1.5 → sovrastima di +0.25

Una volta estratto il campione la stima è nota, ma il valore del parametro di interesse no, per cui di fatto l'errore di stima è ignoto (non si può nemmeno sapere se l'errore è per eccesso o per difetto)

Stimatore non distorto /1

Quindi non si può valutare se una specifica **stima** è buona o no

Ma si possono valutare le proprietà dello **stimatore**: in generale, cioè considerando tutti i possibili campioni, lo stimatore come si comporta?

Definizione: uno stimatore si dice ***corretto*** o ***non distorto*** quando il valore atteso dell'errore di stima (= errore di stima medio nell'insieme dei possibili campioni) è nullo

Nell'esempio il valore atteso dell'errore di stima è

$$E(\bar{X} - \mu_X) = -1.25 \times 0.0625 - 0.75 \times 0.2500 + \ldots = 0$$

E' un caso fortunato? No, è vero in generale, *qualunque sia la distribuzione del carattere nella popolazione*, che la media campionaria è uno stimatore *non distorto* della media della popolazione → in alcuni campioni sovrastima, in altri sottostima, ma nell'insieme dei campioni sovrastime e sottostime si compensano, per cui lo stimatore non ha una tendenza sistematica né alla sovrastima né alla sottostima

Stimatore non distorto /2

Formalmente la proprietà di non distorsione si scrive come

$$E(\bar{X} - \mu_X) = 0 \qquad \text{qualunque sia il valore di } \mu_X$$

In alternativa, per le proprietà del valore atteso si può scrivere anche

$$E(\bar{X}) = \mu_X \qquad \text{qualunque sia il valore di } \mu_X$$

A parole: uno stimatore è *non distorto* quando il suo valore atteso coincide con il parametro di interesse (qualunque sia il suo valore)

Nell'es. $E(\bar{X}) = 1.25$, che coincide con la media della popolazione $\mu = 1.25$

Osservazione 1: $E(\ )$ è il valore atteso della distribuzione campionaria, cioè il valore medio nell'insieme dei possibili campioni

Osservazione 2: la precisazione "qualunque sia il valore del parametro" è cruciale perché in pratica il valore del parametro è ignoto

Variabilità dello stimatore

- Ammettiamo che lo stimatore in questione sia *non distorto*, per cui nell'insieme dei campioni sovrastime e sottostime si compensano. Questa proprietà è una buona base di partenza, ma *non garantisce una stima accurata*
- Infatti, in pratica si dispone di un solo campione, al quale è associato un errore di stima ignoto che potrebbe anche essere enorme
- Si pongono allora domande del tipo:
  - Qual è l'ordine di grandezza degli errori di stima?
  - Quanto è probabile incorrere in un errore di stima più grande di un certo valore prefissato?
- Occorre dunque ***quantificare il livello di incertezza*** associato allo stimatore, cioè quanto le stime (e quindi gli errori di stima) variano da campione a campione → ***varianza campionaria ed errore standard***

Errore standard della media /1

La ***varianza campionaria della media campionaria*** è

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2_X}{n} = \frac{\text{varianza di } X \text{ nella popolazione}}{\text{dimensione del campione}}$$

qualunque siano i valori di μ_X e σ^2_X

Nell'esempio $Var(\bar{X}) = 0.59375$

che effettivamente coincide con il rapporto tra la varianza della popolazione, 1.1875, e la numerosità campionaria, 2

La deviazione standard di $\bar{X}$ è detta ***errore standard della media campionaria*** e descrive la variabilità di $\bar{X}$ intorno a μ_X

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{\text{deviazione std. di } X \text{ nella popolazione}}{\text{radice quadrata della dimensione del campione}}$$

qualunque siano i valori di μ_X e σ_X

Errore standard della media /2

In generale, l'errore standard della media campionaria è

- direttamente proporzionale alla deviazione standard del carattere nella popolazione → quanto più il carattere varia nella popolazione, tanto più la media varia da campione a campione
- inversamente proporzionale alla radice quadrata della dimensione del campione → quanto più grande è il campione, tanto meno la media varia da campione a campione

- La media campionaria è una statistica relativa all'osservazione di un campione composto da *n* unità
- Nel calcolare la media i valori grandi e piccoli si compensano → la media è meno variabile delle singole osservazioni

Campionamento SENZA ripetizione ('in blocco')

Individui selezionati	Risultato campionario	Media campionaria $\bar{X}$
A, A	0, 0	0
A, B	0, 1	0.5
A, C	0, 1	0.5
A, D	0, 3	1.5
B, A	1, 0	0.5
B, B	1, 1	1
B, C	1, 1	1
B, D	1, 3	2
C, A	1, 0	0.5
C, B	1, 1	1
C, C	1, 1	1
C, D	1, 3	2
D, A	3, 0	1.5
D, B	3, 1	2
D, C	3, 1	2
D, D	3, 3	3

I campioni senza ripetizione sono 12 → ogni campione ha probabilità 1/12 di essere estratto

Il campionamento SENZA ripetizione non produce un "campione casuale"

- Supponiamo di ignorare le etichette A, B, … e considerare solo i valori
- Campionamento senza ripetizione → correlazione tra X_1 e X_2 → (X_1 , X_2) non è un "campione casuale"
- Infatti è ancora vero che

$$P(X_i = 0) = 0.25, \quad P(X_i = 1) = 0.50, \quad P(X_i = 3) = 0.25$$

- Tuttavia

$$P(X_1 = x_1 \cap X_2 = x_2) \neq P(X_1 = x_1) \times P(X_2 = x_2)$$

- Ad es. il campione (0, 3)
  - è uno dei 12 possibili → ha probabilità 1/12 = 0.0833
  - tale prob. è diversa da 0.25×0.25 = 0.0625

Distribuzione campionaria della media (campionamento SENZA ripetizione)

Rispetto al campionamento con ripetizione, come cambiano le caratteristiche della media campionaria?

La distribuzione campionaria diviene

Media campionaria $\bar{X}$	Freq. Rel. (= probabilità)
0.5	4/12 = 0.3333
1	2/12 = 0.1667
1.5	2/12 = 0.1667
2	4/12 = 0.3333
	1.0000

$$\mu_{\bar{X}} = 1.25$$
$$\sigma^2_{\bar{X}} = 0.39583$$
$$\sigma_{\bar{X}} = \sqrt{0.39583} = 0.62915$$

Distribuzione campionaria della media

	Campionamento casuale semplice	
Indice	Con ripetizione	Senza ripetizione
$\mu_{\bar{X}}$	1.25	1.25
$\sigma^2_{\bar{X}}$	0.59375	0.39583
$\sigma_{\bar{X}}$	0.77055	0.62915

- La media campionaria è uno stimatore *non distorto* in entrambi i casi
- Tuttavia la sua varianza è *inferiore* se il campionamento è senza ripetizione →rapporto tra le varianze = 0.39583/0.59375 = 2/3

Con il campionamento casuale semplice la media campionaria è uno stimatore *non distorto* sia nel caso "con ripetizione" che nel caso "senza ripetizione", ma nel secondo caso la varianza è ridotta di un fattore chiamato **fattore di correzione per popolazioni finite**

$$\frac{N-n}{N-1} \approx 1-f \qquad \text{dove } f = \frac{n}{N} = \frac{\text{numerosità del campione}}{\text{numerosità della popolazione}}$$

Nell'es. il fattore di riduzione della varianza vale (4-2)/(4-1)=2/3

Campionamento casuale semplice: con o senza ripetizione? /1

- Il buon senso suggerisce di evitare il campionamento con ripetizione perché ammette la possibilità che una unità compaia più volte → tenendo fissa l'ampiezza del campione questa ripetizione provoca una perdita di informazione
- Questa intuizione è confermata dalle proprietà della media campionaria come stimatore della media della popolazione: la proprietà di non distorsione vale in entrambi i casi, ma ***il campionamento con ripetizione causa un aumento della varianza campionaria*** (= perdita di efficienza)
- Di conseguenza nelle applicazioni la versione utilizzata è quella senza ripetizione
- Perché allora nella teoria statistica di base non si assume il campionamento senza ripetizione?

Campionamento casuale semplice: con o senza ripetizione? /2

- Quando il campione è un "campione casuale" (v.a. iid) le proprietà degli stimatori sono semplici e di facile dimostrazione
- **Il campionamento senza ripetizione genera dipendenza tra le osservazioni** → il campione non può essere un "campione casuale" → il quadro si complica
- Supponiamo di estrarre a caso 2 palline da un'urna contenente: 5 palline contrassegnate con il numero 1 + 5 palline contrassegnate con il numero 0. Indichiamo con X_1 il numero portato dalla pallina prima estratta e con X_2 il numero portato dalla pallina seconda estratta.
  - La distribuzione di probabilità di X_1 è la stessa sia che si estragga con ripetizione che senza: $P(X_1=1) = 5/10 = 0.5$.
  - Ma per X_2 le cose cambiano.

... per X_2 le cose cambiano perché

- CON RIPETIZIONE: rimettendo nell'urna la pallina prima estratta si ripristinano le condizioni iniziali e quindi X_2 è indipendente da X_1 , cioè conoscere il valore di X_1 non modifica le probabilità di X_2

  $P(X_2=1 \mid X_1=0) = P(X_2=1 \mid X_1=1) = P(X_2=1) = 5/10 = 0.5$

- SENZA RIPETIZIONE: le probabilità di X_2 cambiano a seconda di ciò che si è verificato nella prima estrazione, cioè del valore assunto da X_1 . Infatti al momento della seconda estrazione l'urna contiene 9 palline, ma la composizione in termini di 0 e 1 dipende dall'esito della prima estrazione

  $P(X_2=1 \mid X_1=0) = 5/9 = 0.56$ $P(X_2=1 \mid X_1=1) = 4/9 = 0.44$

  quindi X_2 non è indipendente da X_1

  La probabilità marginale di X_2 è

  $P(X_2=1) = P(X_2=1 \mid X_1=0)\,P(X_1=0) + P(X_2=1 \mid X_1=1)\,P(X_1=1)$

  $= (5/9)\times(5/10) + (4/9)\times(5/10) = 45/90 = 0.5$

  → X_1 e X_2 hanno identica distribuzione, ma non sono indipendenti

Campionamento casuale semplice: con o senza ripetizione? /3

- Il grado di dipendenza indotto dal campionamento senza ripetizione è funzione della ***frazione di campionamento***

$$f = \frac{n}{N} = \frac{\text{numerosità del campione}}{\text{numerosità della popolazione}}$$

- La dipendenza indotta è tanto più debole quanto più la frazione di campionamento *f* è piccola
- Nell'esempio precedente se si passa ad un urna di 1000 palline la frazione di campionamento passa da 2/10 a 2/1000. Assumendo che la composizione rimanga invariata (→ 500 palline con 1 e 500 con 0) le probabilità di X_1 e X_2 nel campionamento con ripetizione rimangono invariate, mentre nel campionamento senza ripetizione si ottiene

  $P(X_2=1 \mid X_1=0) = 500/999 = 0.5005$ $P(X_2=1 \mid X_1=1) = 499/999 = 0.4995$

  Questi valori sono molto vicini → la dipendenza è debole, ovvero la situazione è molto prossima all'indipendenza

Campionamento casuale semplice: con o senza ripetizione? /4

- Se la ***frazione di campionamento*** è prossima a 0 gli elementi campionari sono approssimativamente indipendenti e quindi i due tipi di estrazione, con e senza ripetizione, sono quasi del tutto equivalenti
- Per la maggior parte delle finalità **quando la frazione di campionamento è inferiore al 5% la dipendenza indotta dal campionamento senza ripetizione è trascurabile**
  - In tal caso se il campionamento è senza ripetizione si può comunque usare l'espressione dell'errore standard della media campionaria $\sigma_X/\sqrt{n}$ in quanto il fattore di correzione da applicare (cioè circa 1-*f*) è prossimo a 1
- Quando la popolazione è infinita ($N=\infty$, es. campionamento di pezzi realizzati da un processo produttivo continuo) la distinzione fra campionamento con e senza ripetizione svanisce del tutto (la frazione di campionamento è 0)

Correzione per popolazioni finite

- Applicare la **correzione per popolazioni finite** se:
  1. La popolazione è finita
  2. il campionamento è senza reintroduzione (in blocco)
  3. il campione è ampio rispetto alla popolazione
     (*n* è superiore al 5% di *N*)

- Allora

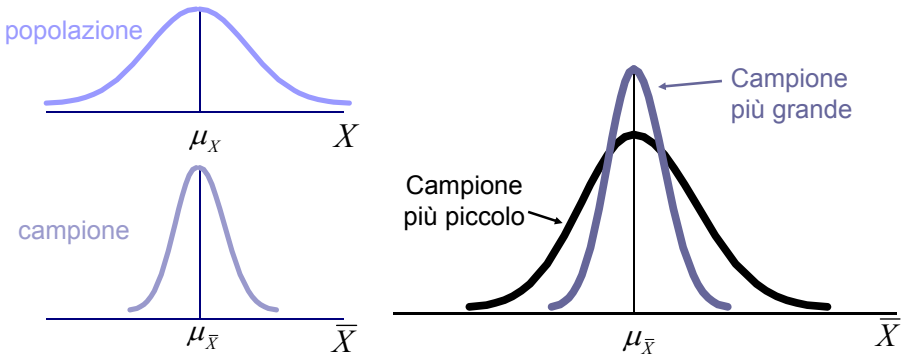
$$Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$ oppure $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

www.causeweb.org

Distribuzione campionaria della media /1

- La media campionaria è uno stimatore della media della popolazione. *Qualunque sia la distribuzione del carattere nella popolazione*, la media di un "campione casuale" è uno stimatore …

non distorto: $\mu_{\bar{X}} = \mu_X$ **con errore standard** $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$



Distribuzione campionaria della media /2

- Qualunque sia la distribuzione di X, la media di un campione casuale da X ha sempre valore atteso $\mu_{\bar{X}} = \mu_X$ e deviazione std $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$
- Questi due risultati sono molto importanti, ma di per sé non consentono di rispondere a quesiti del tipo: *qual è la probabilità che un campione presenti una media al di sotto di una certa soglia? Qual è l'intervallo di valori in cui cade il 95% delle medie campionarie?*
- Per rispondere a queste domande non basta conoscere valore atteso ed deviazione std, serve l'intera distribuzione, salvo che …

Distribuzione campionaria della media /3

- … salvo che la distribuzione appartenga ad una famiglia i cui parametri sono completamente identificati da valore atteso e deviazione std. In particolare, se la media campionaria ha distribuzione Normale, allora

$$\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$$

→ ricorrendo alle tavole della standardizzata si può rispondere alle domande precedentemente formulate

- In quali casi la media campionaria ha distribuzione esattamente o approssimativamente Normale?
  - Nell'esempio del numero di libri letti, la distribuzione della media campionaria è molto diversa da una Normale, innanzitutto perché ha solo 6 punti massa e comunque guardando il diagramma a bastoncini risulta chiaro tale distribuzione non può essere approssimata da una Normale

Distribuzione campionaria della media /4

- Nel caso di campione casuale da una popolazione Normale, la media campionaria ha distribuzione **esattamente** Normale, qualunque sia l'ampiezza campionaria

$$X \sim N(\mu_X, \sigma_X^2) \quad e \quad X_1, \ldots, X_n \sim iid - X \quad \Rightarrow \quad \bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$$

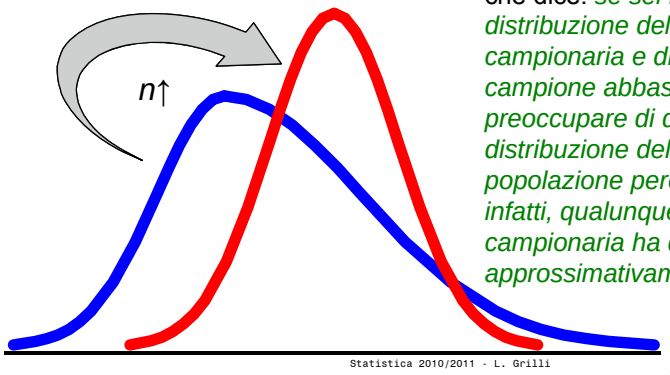
- Nel caso di campione casuale da una popolazione in cui la distribuzione del carattere non è Normale (es. carattere qualitativo o discreto) la media campionaria ha **approssimativamente** distribuzione Normale *se il campione è abbastanza numeroso*. Questa è una conseguenza del **Teorema Limite Centrale** (**TLC**)

$$X \sim ? \quad E(X) = \mu_X \quad Var(X) = \sigma_X^2$$

$$e \quad X_1, \ldots, X_n \sim iid - X \quad \Rightarrow \quad \bar{X} \overset{approx}{\sim} N(\mu_X, \frac{\sigma_X^2}{n})$$

Teorema limite centrale /1

- Il TLC è un risultato ***asintotico***, cioè indica quello che accade quando *n*, la dimensione del campione, tende all'infinito

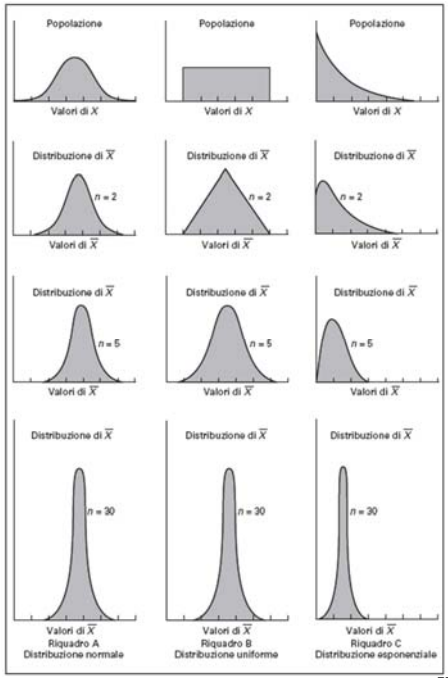


Il TLC è un risultato straordinario che dice: *se sei interessato alla distribuzione della media campionaria e disponi di un campione abbastanza ampio, non ti preoccupare di qual è la distribuzione del carattere nella popolazione perché ciò è irrilevante: infatti, qualunque essa sia, la media campionaria ha distribuzione approssimativamente Normale*

Teorema limite centrale /2

- Al crescere della dimensione campionaria *n* l'approssimazione diventa sempre migliore
- Problema pratico: quanto grande deve essere la dimensione campionaria *n* affinché l'approssimazione sia buona?
  - Infatti nelle applicazioni si dispone di un campione di una certa ampiezza *n* e si deve valutare se l'approssimazione è accettabile: in caso di risposta affermativa si usa l'approssimazione alla Normale, altrimenti occorre seguire altre strade (alquanto impervie, che noi non vedremo)

Distribuzione della media campionaria per campioni di diversa ampiezza (*n* =2, 5, 30) estratti da tre popolazioni con diversa distribuzione



Quanto più la distribuzione del carattere nella popolazione è simmetrica e campanulare tanto più bassa è la dimensione campionaria per la quale l'approssimazione alla Normale è buona (nei casi favorevoli *n*=5 è sufficiente)

Regola pratica prudenziale: un campione di ampiezza *n*=25 è sufficiente per una buona approssimazione nella maggior parte dei casi

Standardizzare la media campionaria

- Z per la distribuzione campionaria di $\bar{X}$:

$$Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

Se si usa la correzione per popolazioni finite allora

$$Z = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

dove: $\bar{X}$ = media campionaria
μ = media della popolazione
σ = deviazione std della popolazione
n = dimensione del campione

Media campionaria: probabilità di non superare una certa soglia

- Dunque in moltissimi casi la distribuzione della media campionaria è (almeno approssimativamente) Normale,

$$\bar{X} \overset{(approx)}{\sim} N(\mu_X, \frac{\sigma_X^2}{n})$$

- Assumiamo μ_X =368, σ_X =15 e n =25
- *Qual è la probabilità che un campione presenti una media al di sotto di una certa soglia, es. 365?*

$$P(\bar{X} < 365) = P\left(\frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} < \frac{365 - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}\right) = P\left(Z < \frac{365 - 368}{3}\right) = P(Z < -1) = 0.1587$$

$$\frac{\sigma_X}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

Media campionaria: intervalli di accettazione /1

- **Intervallo di accettazione al 95%:** intervallo centrato sulla media della popolazione che contiene il 95% delle medie campionarie

Dalla tavola della Normale standard l'intervallo centrato sulla media (cioè su 0) contenente il 95% della probabilità è [–1.96, 1.96]. Per passare dalla scala standard alla scala originale delle medie si effettua la trasformazione

$$\frac{\sigma_X}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

$$\mu_{\bar{X}} + \sigma_{\bar{X}} Z = \mu_X + \frac{\sigma_X}{\sqrt{n}} Z = 368 + 3Z$$

estremo inferiore $= 368 + 3 \times (-1.96) = 362.12$

estremo superiore $= 368 + 3 \times (+1.96) = 373.88$

Media campionaria: intervalli di accettazione /2

- ***Intervallo di accettazione al livello (1–α)%:***
  - E' un intervallo centrato sulla media della popolazione che contiene (1–α)% delle medie campionarie
  - Ovvero: se si estrae un campione casuale, si ha una probabilità (1–α) che la media del campione sia compresa in tale intervallo
- Sia $z_{\alpha/2}$ il valore di Z che lascia nella coda destra della distribuzione normale standard l'area $\alpha/2$ (cioè, l'intervallo da $-z_{\alpha/2}$ a $z_{\alpha/2}$ racchiude una probabilità $1-\alpha$)
- Allora l'intervallo

$$\mu_X \pm z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}$$

include la media campionaria con probabilità $1-\alpha$

Media campionaria: effetto della dimensione campionaria

- Le probabilità e gli intervalli relativi alla media campionaria dipendono dall'ampiezza campionaria *n*

	n=1	*n*=25	*n*=100
$\sigma_X/\sqrt{n}$	15	3	1.5
$P(\bar{X}<365)$	0.4207	0.1587	0.0228
Intervallo di accettazione 95%	[338.60, 397.40]	[362.12, 373.88]	[365.06, 370.94]

Nel caso *n*=1 la media coincide con la prima ed unica osservazione → la media campionaria ha la stessa distribuzione del carattere nella popolazione, ad es. 0.4207 è la probabilità che una singola osservazione sia inferiore a 365

Gli ingredienti necessari di una distribuzione campionaria

- Per studiare il comportamento della media campionaria nell'insieme dei possibili campioni (es. probabilità di non superare un certa soglia, intervallo che include il 95% delle medie …) è necessario disporre di *media e deviazione std della popolazione*
- Di solito tali valori non sono noti (certamente non sono noti nei problemi di inferenza statistica) → occorre ipotizzare i valori (si ragiona per scenari)

Esempio /1

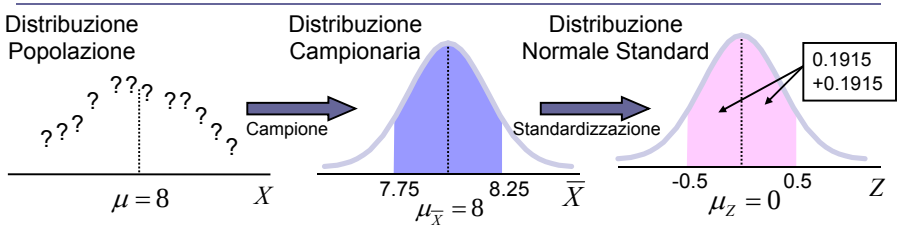
- Supponiamo che (un carattere X in) una popolazione abbia media μ = 8 e scarto quadratico medio σ = 3. Consideriamo un campione casuale di dimensione n = 36
- Qual è la probabilità che la media campionaria sia compresa fra 7.75 e 8.25?

Esempio /2

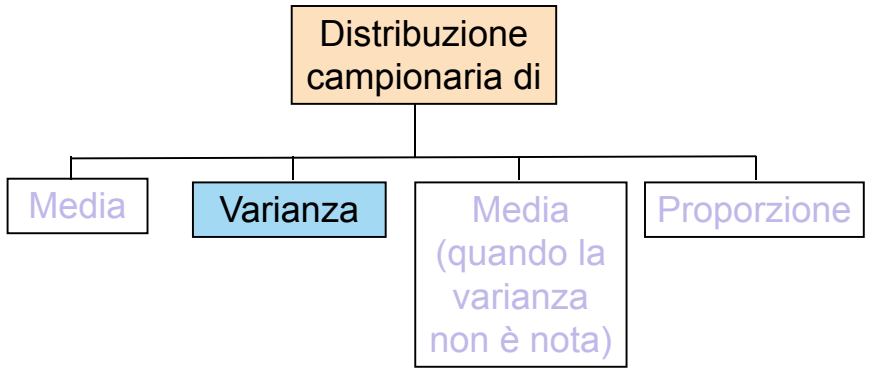
- Anche se la popolazione non ha distribuzione normale, il teorema del limite centrale può essere usato ($n > 25$)
- … quindi la distribuzione campionaria di $\overline{X}$ è approssimativamente Normale
- … con media $\mu_{\overline{X}} = 8$
- … e deviazione std $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

Esempio /3

$$P(7.75 < \overline{X} < 8.25) = P\left(\frac{7.75-8}{\frac{3}{\sqrt{36}}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{8.25-8}{\frac{3}{\sqrt{36}}}\right)$$
$$= P(-0.5 < Z < 0.5) = 0.3830$$



Distribuzione campionaria della varianza



Varianza campionaria

- Lo stimatore usuale della varianza della popolazione σ^2 è la **varianza campionaria**, quella con il divisore n-1

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

gradi di libertà (gdl)

- La varianza campionaria S^2 è uno stimatore non distorto poiché si dimostra che

$$E(S^2) = \sigma^2 \qquad \forall \sigma^2 \in (0, +\infty)$$

- Questo significa che in alcuni campioni S^2 sovrastima σ^2, in altri sottostima, ma nel complesso non vi è una tendenza sistematica né alla sovrastima né alla sottostima

Gradi di libertà (gdl)

Idea: Numero di osservazioni che sono libere di variare dopo che la media campionaria è stata calcolata

Esempio: Supponiamo la media di 3 numeri sia 4

$X_1 = 1$
$X_2 = 2$
$X_3 = ?$

→ X_3 **deve essere 9**
(cioè X_3 non è libero di variare)

Qui $n = 3$ → gradi di libertà $= n - 1 = 3 - 1 = 2$

(2 osservazioni possono assumere qualsiasi valore, ma dato il valore della media campionaria, la terza non è libera di variare)

Stima della varianza

- La proprietà di non distorsione è dovuta all'uso al denominatore di S^2 dei gradi di libertà (gdl) $n-1$ invece dell'ampiezza campionaria n
- In questo contesto gdl = numero di scarti dalla media "liberi" = $n-1$ (infatti, dati n numeri si calcolano n scarti dalla media aritmetica; tuttavia la somma degli scarti è 0 e quindi una volta noti $n-1$ scarti l'n-esimo è automaticamente determinato: è quel numero che aggiunto alla somma degli altri scarti dà 0)
- Lo stimatore con il divisore n (varianza descrittiva) è uno stimatore *distorto verso il basso*, poiché il suo valore atteso è pari a $\sigma^2(n-1)/n$
- *Al crescere di n la distorsione diviene trascurabile* per cui in pratica l'uso del denominatore corretto $n-1$ è importante solo in campioni di piccola ampiezza

Distribuzione campionaria della varianza quando X~Normale

- Se la popolazione generatrice è Normale, la varianza campionaria ha una distribuzione campionaria legata alla Chi-quadrato

$$X \sim N(\mu,\sigma^2) \quad X_1,\ldots,X_n \text{ iid-}X \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\Rightarrow \quad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$P\left(a < S^2 < b\right) = P\left(\frac{(n-1)a}{\sigma^2} < V < \frac{(n-1)b}{\sigma^2}\right) \quad \text{con } V \sim \chi^2_{n-1}$$

Esempio Chi-quadrato /1

- Un congelatore commerciale deve mantenere la temperatura selezionata con bassa variabilità. Si richiede una deviazione standard non superiore a 4 gradi (una varianza di 16 gradi^2).
- Un campione di 14 congelatori viene controllato rilevando la temperatura
- Assumendo che la deviazione standard della popolazione sia davvero 4, qual è il limite superiore (K) per la varianza campionaria che viene superato con probabilità 0.05?

$$P(S^2 > K) = 0.05$$

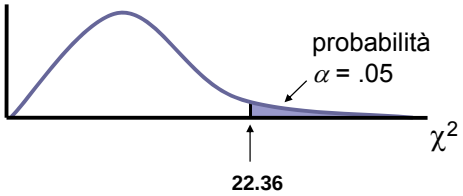
?

Esempio Chi-quadrato /2

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{13}$$

Distribuzione chi-quadrato con (n – 1) = 13 gradi di libertà

- Usiamo la distribuzione chi-quadrato con area 0.05 nella coda di destra:



$$P(\chi^2_{13} > 22.36) = 0.05$$

Esempio Chi-quadrato /3

Quantile superiore 0.05 della Chi-quadrato 13 gdl = 22.36

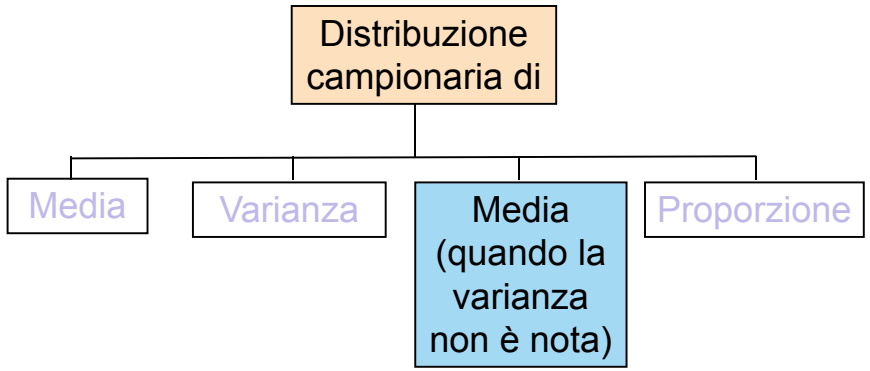
$$P(S^2 > K) = P\left(\frac{(n-1)S^2}{16} > \chi^2_{13}\right) = 0.05$$

oppure $\frac{(n-1)K}{16} = 22.36$ (dove n = 14)

allora $K = \frac{(22.36)(16)}{(14-1)} = 27.52$

Se, sulla base di un campione casuale di dimensione n = 14, si osservasse S^2 maggiore di 27.52, ci sarebbero buoni motivi per ipotizzare una varianza della popolazione superiore a 16.

Distribuzione campionaria della media – varianza non nota



Distribuzione campionaria della media – X~Normale con σ^2 ignota

- Se la popolazione generatrice è Normale, la standardizzata della media campionaria ha distribuzione
  - … Normale Standard se si usa la dev.std. della popolazione σ
  - … *t* di Student con *n*-1 gdl se si usa la dev.std. campionaria *S*

$$X \sim N(\mu, \sigma^2) \quad X_1, \ldots, X_n \; iid\text{-}X$$

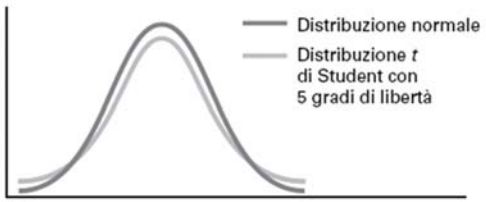
$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$$

La distribuzione *t* di Student /1

- La ***t* di Student** è una famiglia parametrica di v.a. continue che hanno come supporto l'intero asse dei numeri reali
- Il parametro della famiglia è un numero intero detto gradi di libertà (gdl)
- Ogni membro della famiglia (cioè, qualunque sia il numero di gdl) è una *distribuzione simmetrica* con media 0, varianza appena maggiore di 1 e *code più pesanti rispetto alla Normale Standard* (cioè i valori lontani dalla media hanno maggiore probabilità nella *t* che nella Normale Standard)



La *t* di Student è sostanzialmente diversa dalla Normale standard quando il numero di gdl è piccolo (meno di 20); al crescere del numero di gdl la *t* diviene sempre più simile alla Normale standard, tanto che per gdl>120 le due distribuzioni presentano differenze trascurabili

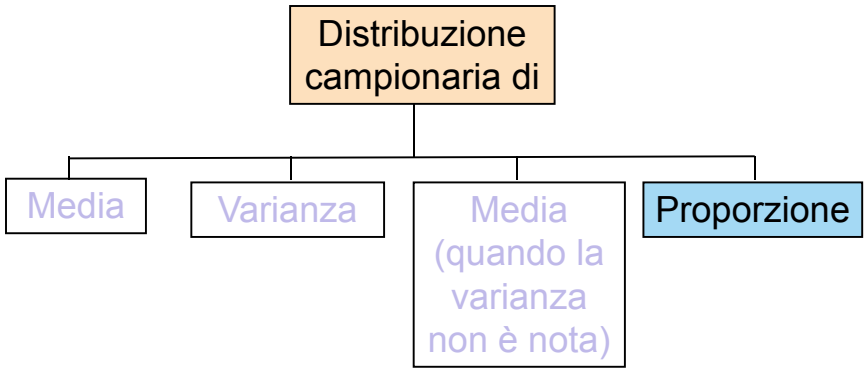
La distribuzione *t* di Student /2

- La ***t di Student*** ha code più pesanti della Normale standard → per ogni data probabilità α da lasciare sulla coda destra il valore della *t* è più grande (= spostato verso destra) rispetto alla Normale Standard
- In altre parole, per ogni $\alpha < 0.5$, il quantile superiore della *t* è maggiore di quello della Normale Standard
- La differenza nei quantili superiori è rilevante quando il numero di gdl è piccolo e tende a zero al crescere del numero di gdl

Esempio: quantili superiori α=0.025

n	gdl	*t*	*z*	*t*/*z*
3	2	4.30	1.96	2.20
5	4	2.78	1.96	1.42
10	9	2.26	1.96	1.15
20	19	2.09	1.96	1.07
30	29	2.05	1.96	1.04
40	39	2.02	1.96	1.03
50	49	2.01	1.96	1.03
75	74	1.99	1.96	1.02
100	99	1.98	1.96	1.01
120	119	1.98	1.96	1.01
200	199	1.97	1.96	1.01
500	499	1.96	1.96	1.00

Distribuz. campionaria della proporzione



Proporzione campionaria /1

- In molte applicazioni il carattere di interesse è *qualitativo con due modalità* (sì/no, conforme/non conforme, soddisfatto/insoddisfatto, acquista/non acquista …). Si dice anche che i dati sono *binari* o *dicotomici*
- In tal caso la distribuzione del carattere nella popolazione è necessariamente Bernoulli (successo/insuccesso)
  - successo = presenza della caratteristica di interesse (sì, conforme, soddisfatto …)
- L'unico parametro è p = ***probabilità di successo*** = "*probabilità che un'unità a caso della popolazione presenti la caratteristica di interesse*".
- Popolazione finita → p = ***proporzione di successi*** = "*proporzione di unità della popolazione che presentano la caratteristica di interesse*"

Proporzione campionaria /2

- Lo stimatore naturale della proporzione nella popolazione, p , è il corrispondente nel campione, cioè la ***proporzione campionaria***

$$\hat{P} = \frac{X}{n} = \frac{\text{numero di successi (=numero di casi che presentano la caratteristica di interesse)}}{\text{numero di prove (=ampiezza campionaria)}}$$

- Codificando il successo con 1 e l'insuccesso con 0 il campione $X_1, X_2, \ldots X_n$ è una sequenza di numeri 0 e 1
- Allora la proporzione campionaria coincide con la media campionaria calcolata sugli elementi $X_1, X_2, \ldots X_n$

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

Proporzione campionaria /3

- La proporzione campionaria è un tipo di media campionaria → *valgono tutte le proprietà viste in generale per la media campionaria*
  - La proporzione campionaria è uno stimatore ***non distorto*** della proporzione nella popolazione

$$\mu_{\hat{P}} = E(\hat{P}) = p \qquad \text{qualunque sia il valore di } p$$

  - La proporzione campionaria ha varianza campionaria σ_X^2/n , dove $\sigma_X^2 = p(1-p)$ [varianza Bernoulli], quindi

$$\sigma_{\hat{P}}^2 = Var(\hat{P}) = \frac{p(1-p)}{n} \qquad \text{qualunque sia il valore di } p$$

Errore standard della proporzione campionaria

$$\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$$

Proporzione campionaria: limite superiore dell'errore standard

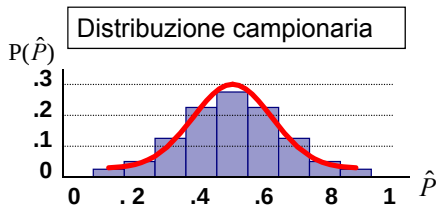
- La deviazione std Bernoulli è limitata superiormente:
  - quando p=0.5 la deviazione std è max: $\sqrt{p(1-p)} = \sqrt{0.5(1-0.5)} = 0.5$
- Quindi l'errore standard della proporzione campionaria è limitato superiormente:
  - quando p=0.5 l'errore standard è max: $\sigma_{\hat{P}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}} = \frac{1}{\sqrt{4n}}$
- Esempio: in un sondaggio d'opinione si intervistano n persone e la risposta è binaria (es. favorevole/contrario)
  - n=25 → l'errore standard è al più 0.1 (ovvero 10%)
  - n=100 → l'errore standard è al più 0.05 (ovvero 5%)
  - n=2500 → l'errore standard è al più 0.01 (ovvero 1%)

Proporzione campionaria: approssimazione alla Normale /1

- La proporzione campionaria ha una distribuzione (detta "binomiale relativa") le cui probabilità si calcolano facilmente da quelle binomiali
- Quando l'ampiezza campionaria n (che corrisponde al numero di prove) è grande il calcolo delle probabilità binomiali è complesso (ad es. 50! è un numero di 65 cifre)
- Tuttavia *quando n è grande, per il TLC la distribuzione della proporzione campionaria è ben approssimata dalla Normale*

Per la proporzione campionaria il criterio per giudicare se l'approssimazione Normale è accettabile non è quello generale che considera sufficiente un campione di ampiezza >25; si adotta invece il criterio $np(1-p) > 9$ e poiché p non è noto in pratica il criterio diviene

$$n\hat{P}(1-\hat{P}) > 9$$



Proporzione campionaria: approssimazione alla Normale /2

- Criterio per giudicare se la distribuzione della proporzione campionaria è ben approssimata dalla Normale:

$$\underbrace{n}_{\text{Ampiezza del campione}} \times \underbrace{\hat{P}(1-\hat{P})}_{\text{Grado di simmetria della Bernoulli}} > 9$$

Proporzione vicina a 0 o 1 ↓ distribuzione fortemente asimmetrica ↓ occorre un campione molto grande

Valore del criterio al variare di n e $\hat{P}$

	proporzione campionaria				
n	5.00%	25.00%	50.00%	75.00%	95.00%
30	1.43	5.63	7.50	5.63	1.43
40	1.90	7.50	**10.00**	7.50	1.90
50	2.38	**9.38**	**12.50**	**9.38**	2.38
100	4.75	**18.75**	**25.00**	**18.75**	4.75
200	**9.50**	**37.50**	**50.00**	**37.50**	**9.50**

Standardizzare la proporzione campionaria

- La proporzione campionaria standardizzata è

$$Z_{\hat{P}} = \frac{\hat{P} - \text{valore atteso}}{\text{errore standard}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- Pertanto, se il criterio per l'approssimazione alla Normale è soddisfatto

$$\hat{P} \overset{approx}{\sim} N\left(p, \frac{p(1-p)}{n}\right) \qquad Z_{\hat{P}} \overset{approx}{\sim} N(0,1)$$

- Ad esempio, la probabilità di osservare un campione con una proporzione di successi non superiore al 60% è

$$P(\hat{P} \le 0.60) = P\left(Z_{\hat{P}} \le \frac{0.60 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) \approx \Phi\left(\frac{0.60 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

Il segno di approssimazione ≈ è dovuto al fatto che Z_P è solo approssimativamente Normale Standard

Esempio

- Supponiamo che il 75% dei clienti sia soddisfatto del servizio
- La popolazione è infinita: si tratta dei clienti in astratto, quelli effettivi e quelli potenziali, quelli di ieri e quelli di domani
- Il carattere di interesse è dicotomico; ponendo "successo"="cliente soddisfatto" la distribuzione del carattere nella popolazione è Bernoulli con probabilità di successo (= di cliente soddisfatto) p =0.75 : $X \sim Be(0.75)$
- Supponiamo di intervistare n=200 clienti; in tal caso l'errore standard della proporzione campionaria è

$$\sigma_{\hat{P}} = \sqrt{\frac{0.75(1-0.75)}{200}} = 0.0306$$

Inoltre la distribuzione è ben approssimata dalla Normale (infatti 200×0.75×0.25 > 9)

- La proporzione di clienti soddisfatti cambia da campione a campione: in alcuni è superiore a quella vera del 75%, in altri è inferiore. Qual è la probabilità di osservare un campione in cui i clienti soddisfatti sono non più del 70%?

$$P(\hat{P} \le 0.70) = P\left(\frac{\hat{P} - 0.75}{0.0306} \le \frac{0.70 - 0.75}{0.0306}\right) = P(Z_{\hat{P}} \le -1.6330) \approx \Phi(-1.6330) = 0.0512$$