Stima puntuale e intervallare
Cap.6 Luccio & Caudek

In statistica molto spesso non si può ricorrere ad intere popolazioni, ma occorre lavorar su campioni, e i risultati che si ottengono vanno poi *generalizzati* alle popolazioni di provenienza.

Per descrivere i campioni di variabili aleatorie (ma anche le popolazioni) si usano indici riassuntivi riferiti alla *tendenza centrale*, alla *variabilità* e alla *forma*.

Questi indici vengono chiamati

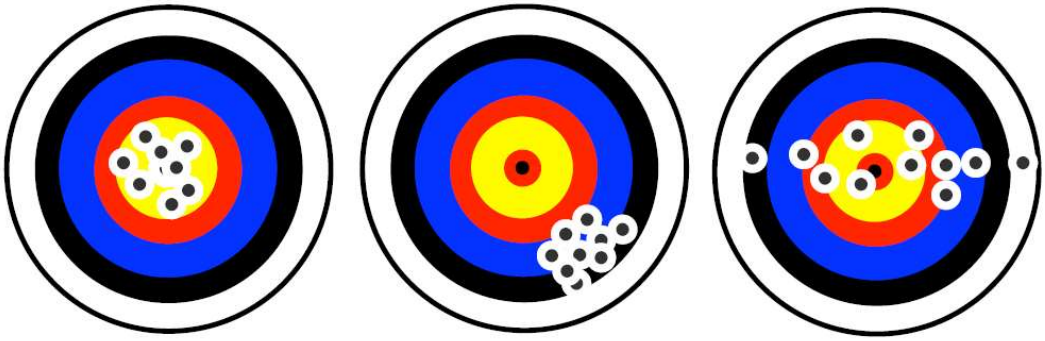
Statistiche	**Parametri**
Campione	**Popolazione**
Media aritmetica $\bar{X} = \sum_{i=1}^{n} x_i / n$	**Media (Valore atteso)** in variabile aleatoria discreta (pag 67) $\mu = \sum_{i=1}^{n} x_i p(x_i)$ in variabile aleatoria continua (pag 77) $\mu = \int_{-\infty}^{+\infty} x f(x) dx$ dove la quantità $f(x)dx$ nella funzione di *densità di probabilità* corrisponde alla quantità $p(x_i)$ nel caso discreto e l'integrazione $\int_{-\infty}^{+\infty}$ è analoga ad una somma.
Deviazione standard $S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}$	**Radice della varianza** (della popolazione) (pag 78) $\sigma = [V(X)]^{1/2} = \sqrt{E(X^2) - \mu^2}$

- Per statistica intenderemo una qualunque funzione delle *n* variabili aleatorie che costituiscono il campione.
- Quando le statistiche vengono usate per stimare i parametri della popolazione da cui il campione considerato è stato tratto, esse vengono chiamate **stimatori**.

Stima puntuale. Singolo valore che "mira" al parametro della popolazione.

Stima intervallare. Specifica un intervallo di valori che contiene il parametro della popolazione, con un certo valore di probabilità.

Stima puntuale



Stimatore non distorto poco variabile

Stimatore distorto poco variabile

Stimatore non distorto molto variabile

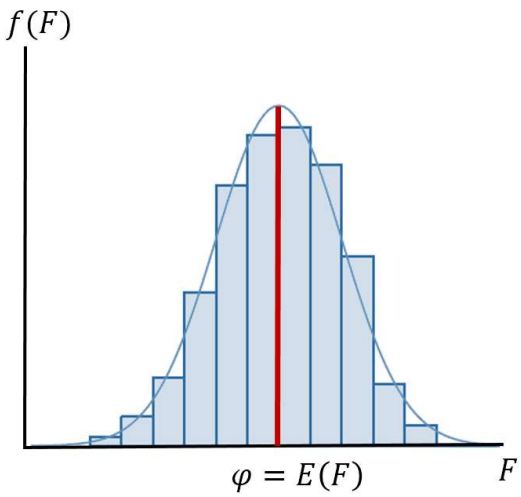
Criteri che consentano di distinguere l'adeguatezza di uno stimatore puntuale rispetto ad un altro.

1. Correttezza ed errore quadratico medio

L'estrazione di un campione e il calcolo di uno stimatore sono simili ad un singolo tiro al bersaglio. Per valutare la bontà di uno stimatore è necessario considerare le stime ottenute ripetendo un grande numero di volte il processo impiegato per eseguire una stima.

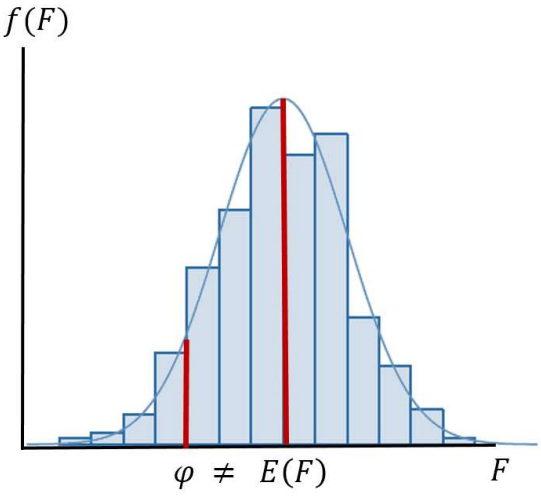
La bontà di uno stimatore puntuale viene dunque valutata (pg. 104)

- (*lo strumento grafico*) mediante la distribuzione di frequenze (istogramma, valore atteso, varianza) generata dalle stime ottenute estraendo molteplici campioni (distribuzione campionaria dello stimatore (!))
- (*la procedura*) calcolando una stima del parametro in ciascuno di essi
- (*il criterio*) stabilendo quanto precisamente la distribuzione delle stime si addensa intorno al parametro che si vuole stimare.



Distribuzione di frequenze campionarie di uno stimatore F centrato sul parametro.

Predittore *equilibrato* (unbiased)



Distribuzione di frequenze campionarie di uno stimatore F dotato di errore sistematico.

L'errore sistematico B è dato da

$$B = E(F) - \varphi$$

Oltre che dall'errore sistematico, la bontà di uno stimatore dipende anche dalla varianza della sua distribuzione campionaria (Errore/Scarto Quadratico Medio)

$$E(F-\varphi)^2 = V(F) + B^2$$

La tabella 6.1 sul teso di Luccio & Caudek, a pag. 105, illustra alcuni degli stimatori *bilanciati* di più corrente uso.

- La *media campionaria* ha come *valore atteso* (…la media delle medie di tutti i campioni di grandezza n che possono essere estratti da una popolazione) il parametro μ della popolazione.
  Viene detta predittore equilibrato del valore atteso della popolazione.

$$E(\bar{X}) = E\left(\sum_{i=1}^{n} x_i /n\right) = \sum_{i=1}^{n} E(x_i)/n = \frac{1}{n} n\mu = \mu$$

- Predittore equilibrato della *proporzione* di casi dotati di una certa caratteristica nella popolazione è:

$$\hat{p} = \frac{Y}{n},$$

  dove $Y = (Bern_1 + Bern_2 + \ldots + Bern_n)$ è la somma di casi nel campione che presentano una certa caratteristica (variabile aleatoria binomiale).
  Analogamente alla media campionaria, avremo che

$$E(\hat{p}) = E\left(\frac{Y}{n}\right) = E\left(\sum_{i=1}^{n} Bern_i\,/n\right) = \frac{1}{n}\sum_{i=1}^{n} E(Be\quad_i) = \frac{1}{n}np = p.$$

Che forma ha li distribuzione campionaria delle medie (e delle proporzioni) campionarie?

Rivediamo la binomiale e osserviamo cosa succede al crescere di *n*

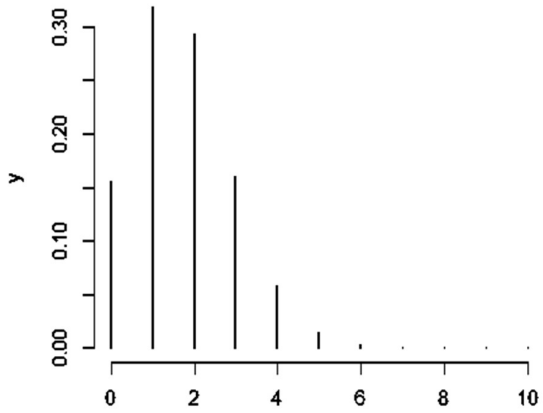


Figura 1 Distribuzione di probabilità discreta binomiale per n = 10, p = 0.17. Si noti l'asimmetria della distribuzione.

Vediamo come si modifica la binomiale osservando n = 30 ed n = 100 prove indipendenti di Bernoulli, sempre con p = 0.17.

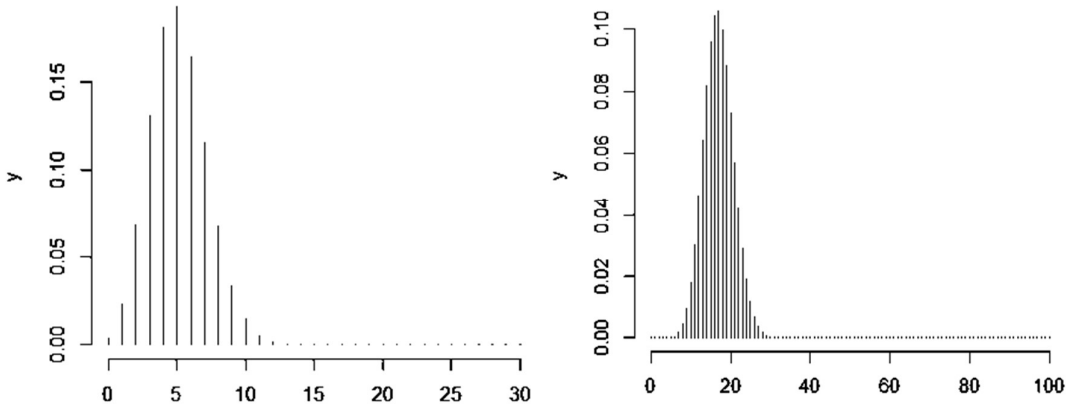


Figura 2 N =30 e N = 100. La distribuzione tende alla normalità.

La distribuzione campionaria degli stimatori nella tabella 6.1 a pag 105 è approssimativamente normale per campioni sufficientemente grandi.

Capitolo 7 -> TLC in grandi campioni o assunzione di normalità

Cosa hanno in comune la media campionaria e la variabile aleatoria binomiale?

SOMMATORIA DEI VALORI DEL CAMPIONE

Non tutti gli stimatori sono privi di errore sistematico: il caso della varianza campionaria (corretta) e il concetto di gradi di libertà.

Abbiamo visto che

$$E(\bar{X}) = \mu$$

$$E(\hat{p}) = p$$

Mentre

$$E(S^2) \neq \sigma^2$$

Dimostrazione (pg. 106, mediante la forma alternativa della varianza campionaria)

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n} = \sum \frac{(X_i)^2}{n} - \bar{X}^2$$

Consideriamo il numeratore, lasciando stare il fattore 1/n

$$nS^2 = \sum(X_i - \bar{X})^2 = \sum(X_i)^2 - n\bar{X}^2$$

Prendiamo il valore atteso $E(nS^2)$

$$E\left[\sum(X_i)^2 - n\bar{X}^2\right] = E\left(\sum(X_i)^2\right) - nE(\bar{X}^2)$$

$$= \sum E(X_i)^2 - n\,E(\bar{X}^2)$$

Ricordiamo che (dal Cap 3.):

$$V(X) = \sigma^2 = E(X_i{}^2) - \mu^2$$

dalla quale otteniamo che

$$E(X_i{}^2) = \sigma^2 + \mu^2$$

Anticipiamo che (Cap 7.):

$$\boldsymbol{V(\bar{X}) = E(\bar{X}^2) - \mu^2} = V\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \boldsymbol{\frac{\sigma^2}{n}}$$

dalla quale otteniamo che

$$E(\bar{X}^2) = {\sigma^2}/{n} + \mu^2$$

Ciò detto riscriviamo il tutto

$$E\left[\sum (X_i)^2 - n\bar{X}^2\right] = \sum (\sigma^2 + \mu^2) - n\left(\sigma^2/n + \mu^2\right) =$$

$$n(\sigma^2 + \mu^2) - n\left(\sigma^2/n + \mu^2\right) =$$

$$n\sigma^2 - \sigma^2 + n\mu^2 - n\,\mu^2 =$$

$$\sigma^2(n-1)$$

Quindi il valore atteso della varianza campionaria (non corretta)

$$E(S^2) = \frac{1}{n}E\left[\sum (X_i)^2 - n\bar{X}^2\right] = \sigma^2\frac{1}{n}(n-1)$$

$$= \sigma^2 - \frac{\sigma^2}{n} \quad (< \sigma^2)$$

(Ricordiamo quanto detto a riguardo di un generico stimatore puntuale distorto F: $E(F) = B \pm \varphi$)

Fornisce sempre una stima più piccola della vera varianza della popolazione.

n	$\frac{1}{n}(n-1)$
5	0.8
20	0.95
100	0.99
500	0.998
1000	0.999

Per correggere tale errore sistematico definiremo la varianza campionaria corretta (stimatore puntuale bilanciato) come:

$$s^2 = \frac{n}{n-1}S^2 = \frac{\sum(\mathrm{X_i} - \bar{\mathrm{X}})^2}{\mathrm{n}-1}$$

poiché

$$E(s^2) = \frac{1}{n-1}E(nS^2) = \frac{1}{n-1}E\left[\sum (X_i)^2 - n\bar{X}^2\right] = \frac{1}{n-1}\,\boldsymbol{\sigma^2}(n-1)$$

La quantità al denominatore $n-1$ è chiamata (numero di) *gradi di libertà* (gdl).

L'operazione di aggiustamento del valore atteso di una statistica campionaria incorretta (biased) attraverso l'uso dei *gdl*, in luogo di *n*, è detta aggiustamento per i gradi di libertà.

Nel calcolo della varianza campionaria quanti sono i valori del campione che possono falsificare una stima?

X={1,2,3}; Media = 2

$$\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3}=0.7$$

X={4,8,12}; Media = 8

$$\frac{(4-8)^2+(8-8)^2+(12-8)^2}{3}=10.7$$

Al numeratore non ci sono tre *scarti* indipendenti (gdl = 3), ma soltanto due:

$$\sum_{i=1}^{3}(X_i-\bar{X})=0$$

quindi uno scarto può essere recuperato per somma degli altri $n-1$.

$$(1-2)+(2-2)+(3-2)=0 \quad \rightarrow \quad (1-2)+(2-2)=-(3-2)$$
$$(4-8)+(8-8)+(12-8)=0 \quad \rightarrow \quad (4-8)+(8-8)=-(12-8)$$

Usando una statistica campionaria basata sulla somma degli scarti dalla media si utilizzano $n-1$ dati genuini, essendo che una qualsiasi delle differenze dalla media può essere recuperata, sebbene con segno negativo, dalla somma delle altre.

Di qui la sovrastima della varianza campionaria non corretta, che utilizza n al denominatore invece dei gradi di libertà (n-1).

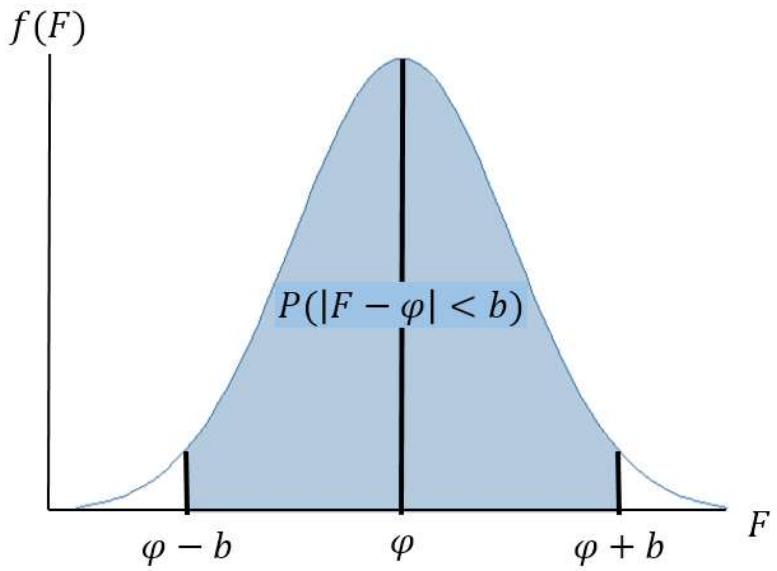
2. Errore della stima

Dato uno stimatore campionario *F* ***privo di errore sistematico***, sarà possibile fare solo delle affermazioni probabilistiche circa la distanza di ogni possibile F_i dal parametro ignoto φ.

La quantità $F_i - \varphi$ è essa stessa una variabile aleatoria di cui, in ciascuna singola istanza, non è possibile conoscere la grandezza.

La probabilità che l'errore della stima sia minore di una certa quantità, che qui chiameremo genericamente *b*, corrisponde all'area evidenziata in figura.

$$P(|F - \varphi| < b) = P(\varphi - b < F < \varphi + b)$$



Possiamo quantificare (*al minimo*) questa probabilità (*diseguaglianza di Cebicev*), con la seguente equazione (pg. 108)

$$P(|F - \varphi| < k\sigma_F) \geq 1 - \frac{1}{k^2}$$

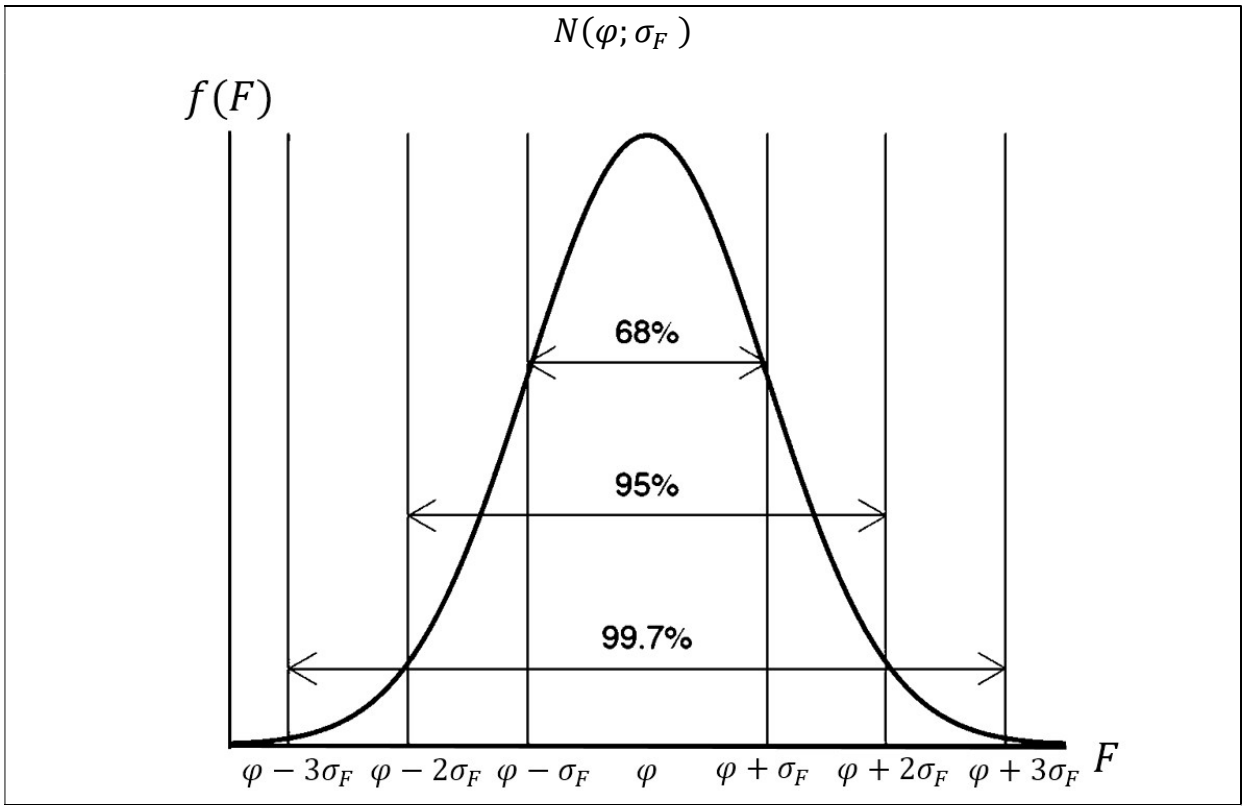
dove σ_F è la deviazione standard (cfr. *errore standard*, Tabella 6.1) della variabile aleatoria *F*.

Così, se fossimo interessati alla probabilità che l'errore della stima non superi le due deviazioni standard attorno a φ, potremo stimare tale probabilità in un valore maggiore o uguale a 0.75.

$$P(|F - \varphi| < 2\sigma_F) \geq 1 - \frac{1}{4}.$$

Se le nostre conoscenze circa la distribuzione campionaria dello stimatore *F* dovessero essere più precise, allora potremo rifinire questa stima.

Ad esempio, se la distribuzione campionaria di uno stimatore *F* fosse *normale*,



per $b = 2\sigma_F$ avremo che:

$$P(|F-\varphi| < b) = P(\varphi - 2\sigma_F < F < \varphi + 2\sigma_F) \sim 0.95.$$

Potremmo anche trasformare la *normale* degli indici statistici *F* in una *normale standard* mediante la nota trasformazione

$$z_i = \frac{F_i - \varphi}{\sigma_F} \sim N(0;1).$$

Consultando la Tabella 1 a pag. 286, troviamo il valore più preciso:

$$P(\varphi - 2\sigma_F < F < \varphi + 2\sigma_F) =$$

$$= P(-2 < z < +2)$$

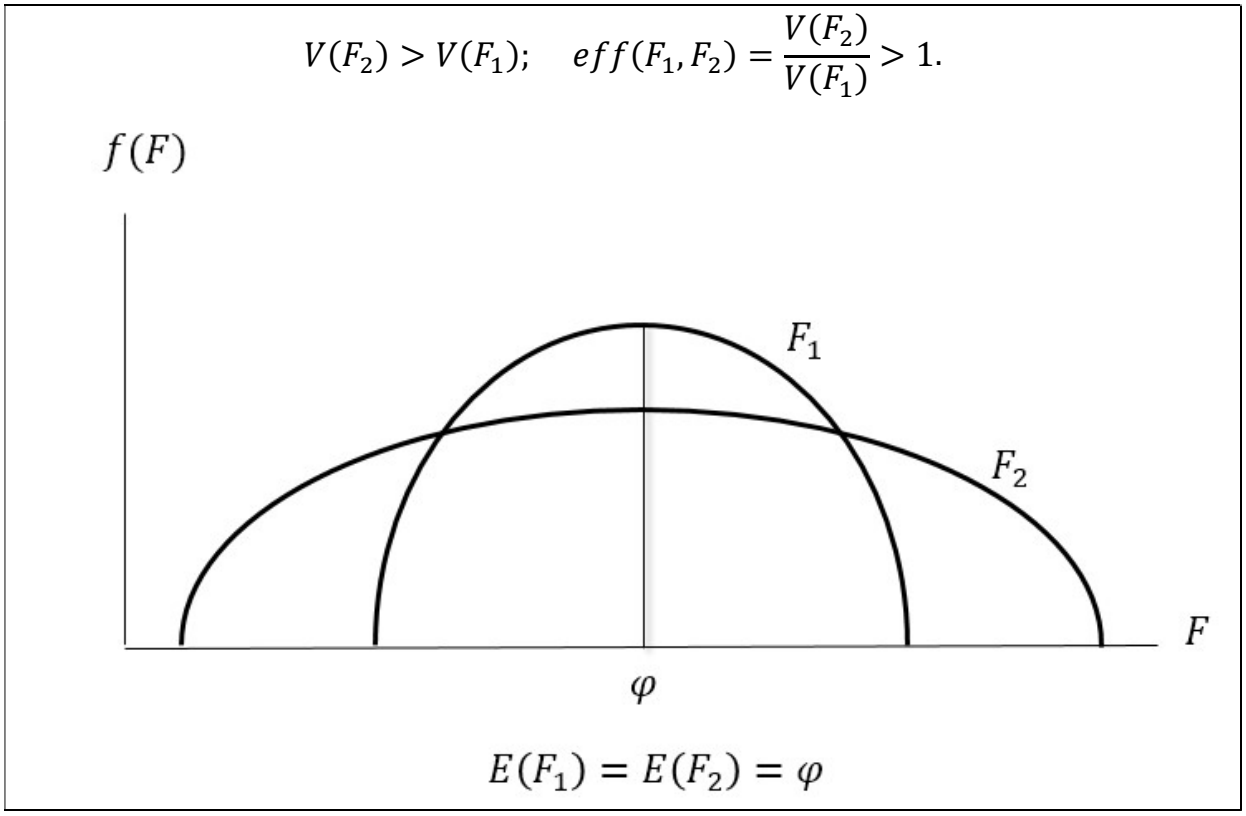
$$= 1 - [P(z < -2) + P(z > +2)] = 1 - (0.0228 + 0.0228) = 0.9544.$$

3. Efficienza

L'assenza di distorsione non è l'unica proprietà desiderabile di uno stimatore.

Il valore medio di una statistica non è l'unica proprietà di una statistica dato che, in campioni diversi, una statistica varia attorno alla sua media.

Una statistica potrebbe, infatti, essere centrata sul parametro ignoto perché errori molto grandi vengono bilanciati da errori molto piccoli:



Distribuzione campionaria di due stimatori puntuali bilanciati sul parametro φ.

L'efficienza relativa di uno stimatore bilanciato rispetto ad un altro è definita dal **rapporto delle loro varianze campionarie** (il quadrato dei rispettivi *errori standard*, cfr. Tabella 6.1).

Nella circostanza descritta in figura risulta che F_1 è migliore di F_2: le misure ottenute in campioni diversi sono più vicine al valore reale della popolazione, rispetto a quelle dello stimatore F_2.

Fra più stimatori corretti, il migliore è quello che ha la varianza campionaria minore, così da aumentare la probabilità di piccoli errori della stima:

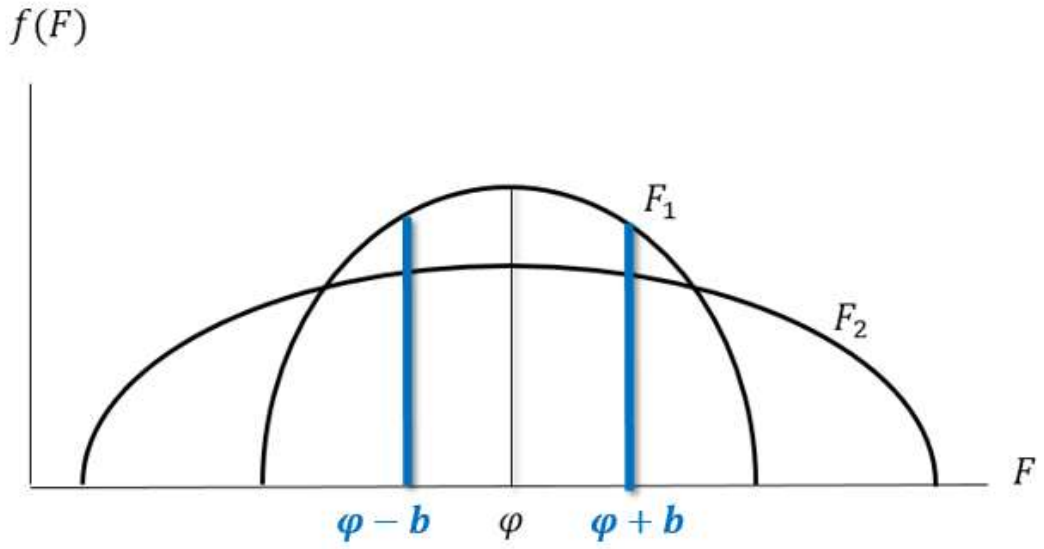
$$\boldsymbol{P(|F_1 - \varphi| < b)} = ?$$

VS

$$\boldsymbol{P(|F_2 - \varphi| < b)} = ?$$

$$P(\varphi - b < F_1 < \varphi + b) > P(\varphi - b < F_2 < \varphi + b)$$

L'area sotto la curva di F_1 nell'intervallo di errore della stima +/- b è maggiore.



Esempio 1. Supponiamo di stimare la media di una popolazione gaussiana utilizzando la **media** campionaria ($\boldsymbol{F_1}$) e la **mediana** ($\boldsymbol{F_2}$). Sappiamo infatti che nella distribuzione gaussiana media e mediana coincidono.

Abbiamo già anticipato che la varianza campionaria della media è:

$$V(F_1) = \frac{\sigma^2}{n}$$

Assumendo una distribuzione normale per la popolazione dalla quale vengono estratti grandi campioni su cui si calcola la mediana, può essere dimostrato che:

$$V(F_2) = 1.570796\frac{\sigma^2}{n}$$

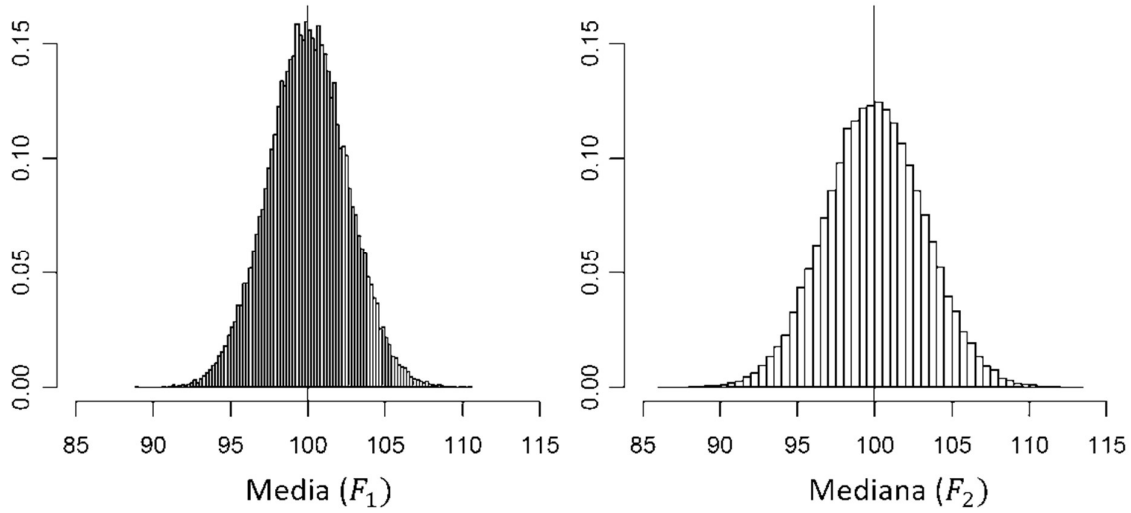
Ne segue che l'efficienza della media relativamente alla mediana come stimatore di μ è:

$$eff(F_1, F_2) = \frac{V(F_2)}{V(F_1)} = \frac{1.570796\sigma^2/n}{\sigma^2/n} = 1.570796$$

Detto in altri termini, la variabilità campionaria della media è il 64% circa della variabilità della mediana

$$\frac{V(F_1)}{V(F_2)} = \frac{1}{1.570796} = 0.6366.$$

Esaminiamo i risultati di una simulazione: calcoliamo empiricamente la varianza della distribuzione campionaria di media e mediana di un campione casuale di n = 200 osservazioni estratto (50000 volte) da una popolazione normale con media μ = 100 e deviazione standard σ = 36.



$$V(F_1) = 6.499387 \cong \frac{36^2}{200} = 6.48$$

la varianza della distribuzione campionaria della media è uguale al rapporto tra la varianza della popolazione e la numerosità del campione.

$$V(F_2) = 10.06176 \cong 1.570796\frac{36^2}{200} = 10.17853$$

$$\frac{V(F_1)}{V(F_2)} = \frac{6.499387}{10.06176} = 0.6459 \cong 64\%.$$

La simulazione precedente rivela che la mediana è meno efficiente della media quale operatore di tendenza centrale (ha una varianza campionaria maggiore).

Esempio 2.

- Si può dimostrare che

$$Var\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x}_j)^2}{n}\right] < Var\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x}_j)^2}{n-1}\right]$$

- La varianza campionaria corretta, che usa $n-1$ al denominatore, è quindi uno stimatore meno efficiente della varianza campionaria non corretta.
- La sua varianza campionaria è superiore a quella della statistica non corretta, che usa n al denominatore.
- Sarà comunque da preferire come *stima puntale non distorta* del parametro σ.

3. Consistenza

Consistenza. Uno stimatore si dice consistente quando la differenza tra la stima ed il valore vero del parametro della popolazione diminuisce all'aumentare delle dimensioni del campione.

Se guardiamo all'ultima colonna della tabella 6.1 noteremo che l'errore standard (la radice della varianza campionaria) degli stimatori puntuali *media*, *differenza tra medie* e *proporzione campionaria* hanno *n* al denominatore.

Quando n tende all'infinito allora la varianza campionaria tende a 0. In altre parole, al crescere della grandezza campionaria lo stimatore converge sempre più al parametro ignoto (*legge dei grandi numeri*).

4. Sufficienza

Si ha sufficienza quando uno stimatore sintetizza tutte le informazioni presenti nel campione che sono importanti per la stima del parametro.