

Indici e formule varie di statistica e probabilità

Indici di posizione: descrivono la tendenza centrale di un insieme di dati.

- **Media campionaria:** utilizzabile solo se i dati sono valori numerici, si calcola come:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{j=1}^k f_j^{rel} x_j$$

Si tratta di un operatore lineare, ovvero se per ogni dato $y = ax + b$, allora $\bar{y} = a\bar{x} + b$.
Trattando direttamente con i dati grezzi è purtroppo *non robusta agli outlier*.

- **Mediana campionaria:** utilizzabile solo se i dati sono numerici o categorici ordinali, si tratta del valore intermedio della lista ordinata ottenuta mettendo le osservazioni in fila dalla più piccola alla più grande; se il numero di osservazioni è pari, la mediana si ottiene come la media aritmetica dei due valori centrali. *Robusta agli outlier*, è un tipo particolare di **quantile**, ovvero “il valore del campione maggiore o uguale di almeno nq valori e minore o uguale di almeno $n(1 - q)$ valori.”
- **Moda campionaria:** utilizzabile con qualunque tipo di dato (numerico o categorico, ordinale o nominale), è il valore che compare con frequenza maggiore nel campione.

Indici di dispersione: descrivono quanto i valori sono differenti l'uno dall'altro in un insieme di dati.

- **Range e range interquartile:** rispettivamente la differenza tra il valore massimo e il valore minimo, che descrive dove si trovano complessivamente i dati, e tra il 0.75-esimo quantile e il 0.25-esimo, che indica dove si trova il 50% centrale dei dati.
- **Varianza campionaria:** la “media” degli scarti quadratici dei valori dalla media campionaria:

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

Non si tratta di un operatore lineare e invece vale che se $y = ax + b$, allora $\sigma_y^2 = a^2 \sigma_x^2$.

- **Deviazione standard campionaria:** la radice quadrata positiva della varianza campionaria, introdotta per risolvere alcuni problemi riguardanti l'unità di misura:

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

Nemmeno questo è un operatore lineare, e infatti se $y = ax + b$, allora $\sigma_y = |a| \sigma_x$.

- **Coefficiente di variazione:** una sorta di deviazione standard normalizzata, utilizzata per confrontare la dispersione di campioni aventi valori medi molto distanti:

$$\sigma^* = \frac{\sigma}{\bar{x}}$$

Indici di correlazione: indicano quanto i grandi valori di un attributo x del campione corrispondono a grandi valori di un altro attributo y , e quanto i piccoli valori del primo ai piccoli valori del secondo; si utilizzano cioè per scoprire se esiste un qualche tipo di relazione tra due attributi delle osservazioni. In tal senso, si indicano come “grandi” i valori superiori alla media campionaria e “piccoli” quelli inferiori.

- **Covarianza campionaria:** moltiplicando gli scarti dei due attributi otteniamo in media valori positivi se questi sono legati da una correlazione *diretta* (grande-grande, piccolo-piccolo) e negativi se sono legati da una correlazione *indiretta* (piccolo-grande, grande-piccolo). Facendo la media di tali prodotti per comprendere quale sia la relazione tendenziale si ottiene allora:

$$cov_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

- **Coefficiente di correlazione lineare campionario:** si divide la covarianza campionaria per il prodotto delle deviazioni standard dei due attributi per far scomparire le unità di misura:

$$r = \frac{cov_{x,y}}{\sigma_x \sigma_y}$$

L'indice così ottenuto varia da -1 a $+1$, dove il primo indica una completa relazione lineare inversa e il secondo una completa relazione lineare diretta, mentre più i valori sono vicini allo zero minore è la correlazione tra i due attributi. L'indice è indifferente alle trasformazioni lineari $x' = ax + b$ e $y' = cy + d$ condizione che a e c abbiano segno concorde: se ciò non accade, l'indice cambia semplicemente di segno.

Indici di eterogeneità: non potendo calcolare la dispersione su attributi qualitativi nominali tramite il normale strumento della varianza e degli altri indici da essa derivati, per ottenere un risultato simile si misura in questi casi un indice della dispersione della distribuzione delle frequenze, detta *eterogeneità*. In particolare, diremo che un attributo si distribuisce in modo massimamente eterogeneo se ogni suo valore si presenta con la stessa frequenza (a patto però che ci siano almeno due valori distinti).

- **Indice di eterogeneità di Gini:** indicando con f_i la frequenza relativa del valore v_i del campione e detto s il numero di valori unici, questo indice si calcola come:

$$I = 1 - \sum_{i=1}^s f_i^2$$

Si dimostra che $0 \leq I \leq \frac{s-1}{s}$. Per poter confrontare l'eterogeneità di campioni con un numero s di valori distinti diversi si crea poi un **indice di Gini normalizzato** come segue:

$$I' = \frac{s}{s-1} I$$

Si dimostra che $0 \leq I' \leq 1$, e che 1 indica la massima eterogeneità (tutti stessa frequenza rel. $\frac{1}{s}$), mentre 0 rappresenta la massima omogeneità (un unico valore con frequenza 1).

- **Entropia:** altra misura dell'eterogeneità di un campione, stando le stesse definizioni di prima si ha:

$$H = \sum_{i=1}^s f_i \log \frac{1}{f_i} = - \sum_{i=1}^s f_i \log f_i$$

Si dimostra che $0 \leq H \leq \log s$; si definisce perciò un'**entropia normalizzata**:

$$H' = \frac{H}{\log s}$$

i cui valori variano tra 0 e 1, dove 0 corrisponde alla minima eterogeneità e 1 alla massima.

Indici di concentrazione: descrivono quanto una variabile numerica che può essere vista come una risorsa (es. ricchezza) sia o meno distribuita equamente tra gli individui all'interno della popolazione.

- **Indice di concentrazione di Gini:** indichiamo con a_1, \dots, a_n le osservazioni una volta che queste sono state ordinate in modo non decrescente, e diciamo poi:

- $F_i = \frac{i}{n}$ la frequenza relativa cumulata fino all' i -esima osservazione

- $Q_i = \frac{\sum_{k=1}^i a_k}{n\bar{a}}$ la quantità relativa cumulata fino all' i -esima osservazione

Essendo che le osservazioni sono state ordinate in modo non decrescente si ha $Q_i \leq F_i$, e in particolare in caso di concentrazione minima si avrà che $Q_i = F_i, \forall i$, in quanto questo rappresenta il

caso in cui la quantità è ripartita in modo perfettamente omogeneo, mentre il caso opposto è quanto $Q_i = 0 \forall i \leq n - 1$. Per calcolare quanto si è vicini a questa ultima situazione si calcola così:

$$G = \frac{\sum_{i=1}^{n-1} F_i - Q_i}{\sum_{i=1}^{n-1} F_i}$$

Analisi dei classificatori: prendendo in considerazione unicamente i classificatori binari, i principali assi per misurare la loro efficacia sono:

- **Sensibilità:** quanto è in grado di rilevare i casi positivi: $sensibilità = \frac{VP}{VP+FN}$
- **Specificità:** quanto è in grado di rilevare i casi negativi: $specificità = \frac{VN}{VN+FP}$

Utilizzando queste due dimensioni è possibile tracciare un punto $(1 - specificità, sensibilità)$ in un piano cartesiano dove il punto $(0,1)$ rappresenta il classificatore perfetto: nel caso di classificatori a soglia, quest'ultima si può allora aggiustare progressivamente per ottenere una curva ROC in modo da scegliere il valore della soglia che più avvicina al classificatore ideale. In generale un classificatore è buono quanto più quella curva si allontana dalla bisettrice $(0,0) - (1,1)$, dove risiedono i classificatori del tutto casuali: un buon indice della bontà di un classificatore è allora la AUC, Area Under (ROC) Curve.

Analisi della varianza: se si vuole testare l'ipotesi che i valori medi di un certo attributo x siano sensibilmente diversi in una serie di G gruppi presenti in un campione, ognuno con cardinalità n_g , è possibile applicare un metodo detto ANOVA: l'idea alla base è che se non ci sono grandi differenze calcolare la varianza in uno specifico gruppo non dovrebbe dare un risultato tanto distante dalla varianza calcolata su tutti i dati a disposizione. Definite allora \bar{x} la media campionaria su tutto il campione e \bar{x}^g quella nei singoli gruppi, e dette poi:

- s_T^2 la varianza campionaria su tutte le osservazioni: $s_T^2 = \frac{1}{n-1} \sum_{g=1}^G \sum_{i=1}^{n_g} (x_i^g - \bar{x})^2$
- s_W^2 la varianza campionaria dei valori entro i gruppi: $s_W^2 = \frac{1}{n-G} \sum_{g=1}^G \sum_{i=1}^{n_g} (x_i^g - \bar{x}^g)^2$
- s_B^2 la varianza campionaria tra le medie dei gruppi: $s_B^2 = \frac{1}{G-1} \sum_{g=1}^G (\bar{x}^g - \bar{x})^2$

si può dimostrare che:

$$s_T^2 = \frac{n-G}{n-1} s_W^2 + \frac{G-1}{n-1} s_B^2$$

Dunque, se la s_T^2 e $\frac{n-G}{n-1} s_W^2$ assumono valori molto vicini ciò significa che s_B^2 è trascurabile, e dunque non vi è grande differenza tra i valori medi nei gruppi; al contrario, se ciò non accade confermiamo l'ipotesi.

Calcolo combinatorio: insieme di formule che permettono di calcolare in quanti modi si può estrarre un certo numero di oggetti da un'urna a seconda che continuo o meno certi fattori.

- **Permutazioni semplici:** sequenze ordinate in cui compaiono tutti gli n oggetti, si calcolano come:

$$P_n = n!$$

- **Permutazioni di oggetti distinguibili a gruppi:** se si individuano k gruppi di oggetti indistinguibili tra di loro, il numero di sequenze ordinate di n oggetti effettivamente distinguibili dalle altre (cioè in cui non si scambiano semplicemente di posto oggetti da uno stesso gruppo) si calcola come:

$$P_{n;n_1, \dots, n_k} = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$$

- **Disposizioni senza ripetizione:** sequenze ordinate di k oggetti tra gli n disponibili in cui uno stesso oggetto non può comparire due volte, si calcolano come:

$$d_{n,k} = \frac{n!}{(n-k)!}$$

- **Disposizioni con ripetizione:** sequenze ordinate di k oggetti tra gli n disponibili in cui uno stesso oggetto può comparire più volte, si calcolano come:

$$D_{n,k} = n^k$$

- **Combinazioni senza ripetizione:** insiemi di k oggetti estratti tra gli n disponibili in cui uno stesso oggetto non può comparire più volte (essendo insiemi non conta l'ordine), si calcolano come:

$$c_{n,k} = \frac{n!}{k!(n-k)!}$$

- **Combinazioni con ripetizione:** insiemi di k oggetti estratti tra gli n disponibili in cui uno stesso oggetto può comparire più volte (essendo insiemi non conta l'ordine), si calcolano come:

$$C_{n,k} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Teoremi di probabilità: detto Ω lo spazio degli esiti di un esperimento casuale, un evento è un qualcosa che può verificarsi in base all'esito di un esperimento, ovvero un insieme di esiti ($E \subseteq \Omega$). Scaturiscono dunque una serie di teoremi sulla probabilità:

- **Assiomi di Kolmogorov:** definiscono il concetto di funzione di probabilità. Fissato uno spazio misurabile, ovvero uno spazio degli esiti Ω e un'algebra degli eventi \mathcal{A} (insieme di tutti gli eventi costruibili dagli esiti di Ω), la funzione $P: \mathcal{A} \rightarrow \mathbb{R}$ è una funzione di probabilità se e solo se:
 - 1) $\forall E \in \mathcal{A}, 0 \leq P(E) \leq 1$
 - 2) $P(\Omega) = 1$
 - 3) Dati eventi disgiunti, ovvero t.c. $\forall i \neq j \ E_i \cap E_j = \emptyset$, allora $P(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i)$
- **Corollari degli assiomi di Kolmogorov:**
 - 1) $P(E) + P(\bar{E}) = 1$
 - 2) $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- **Definizione di probabilità condizionata:** si tratta della probabilità che un evento E si verifichi dato che si è verificato un altro evento F . Si definisce come:

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

Si ricava poi facilmente che $P(E|F) + P(\bar{E}|F) = 1$.

- **Teorema delle probabilità totali e Formula di Bayes:** Dato un qualunque evento E è una partizione di Ω formata dagli eventi F_1, \dots, F_n , ovvero tale che $\forall i \neq j, F_i \cap F_j = \emptyset$ (disgiunti) e $\bigcup_{i=1}^n F_i = \Omega$ (ricoprenti), si può scrivere la probabilità di E come la media pesata della sua probabilità condizionata dagli eventi F_j , dove il peso è dato dalla probabilità degli F_j stessi:

$$P(E) = \sum_{i=1}^n P(E|F_i)P(F_i)$$

Immaginando poi di sapere che l'evento E si è effettivamente verificato, possiamo modificare la nostra credenza sulle probabilità dei vari F_j utilizzando la seguente formula di Bayes:

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^n P(E|F_i)P(F_i)} \quad \left(= \frac{P(E|F_j)P(F_j)}{P(E)} \right)$$

- **Indipendenza tra eventi:** Due eventi si dicono indipendenti se e solo se:

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

Si dimostra poi che se E ed F sono indipendenti lo sono anche E e \bar{F} , mentre perché tre o più eventi siano indipendenti deve accadere che per *ogni* loro sottogruppo vale la formula precedente opportunamente generalizzata per accogliere più eventi.

Teoremi sulle variabili aleatorie

- **Disuguaglianze di Markov e di Chebyshev:** Se $X > 0$ è una variabile aleatoria non negativa, allora per ogni $a > 0$ valgono le seguenti disuguaglianze di Markov:

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad P(X < a) \geq 1 - \frac{\mathbb{E}[X]}{a}$$

Questa disuguaglianza fornisce un bound sulla probabilità dell'evento $X \geq a$; se invece si vuole porre un bound sulla probabilità che una variabile aleatoria con media μ e varianza σ^2 disti almeno o al più $r > 0$ dal suo valore atteso si utilizzano le disuguaglianze di Chebyshev:

$$P(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2} \quad P(|X - \mu| < r) \geq 1 - \frac{\sigma^2}{r^2}$$

Se in quest'ultima disuguaglianza si pone $r = k\sigma$ si ottiene un bound sulla probabilità che una variabile aleatoria disti dalla sua media di almeno o al più k deviazioni standard.

- **Leggi dei grandi numeri:** esistono due leggi dei grandi numeri, una debole e una forte. La prima afferma che, data una successione di n variabili aleatorie i.i.d. X_1, \dots, X_n tutte con valore atteso μ allora per ogni $\epsilon > 0$ la probabilità che la loro media disti almeno ϵ dal valore atteso tende a 0 quando il numero di variabili è molto alto, ovvero:

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \epsilon\right) \rightarrow 0 \quad \text{quando } n \rightarrow \infty$$

La legge forte dei grandi numeri afferma invece che data la media campionaria \bar{X}_n di un campione la probabilità che essa assuma valore pari al valore atteso della popolazione quando n tende a $+\infty$ equivale a quella dell'evento certo:

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mathbb{E}[X]\right) = 1$$

- **Teorema centrale del limite:** date n variabili aleatorie i.i.d. X_1, \dots, X_n con valore atteso $\mathbb{E}[X]$ e deviazione standard σ_X , la loro somma ha una distribuzione che può essere approssimata tanto meglio quanto più n è grande con una distribuzione normale di parametri $n\mathbb{E}[X]$ e $\sqrt{n}\sigma_X$:

$$\sum_{i=1}^n X_i \approx N(n\mathbb{E}[X], \sqrt{n}\sigma_X)$$

- **Definizione di quantile di una variabile aleatoria:** chiamiamo quantile q -esimo di una qualunque variabile aleatoria X la specificazione $\chi_q \in D_X$ tale per cui:

$$P(X \leq \chi_q) = q, \text{ ovvero } F_X(\chi_q) = q$$

Se la variabile aleatoria è continua, si può inoltre ricavare il valore del quantile come:

$$\chi_q = F_X^{-1}(q)$$

Vale inoltre la seguente coimplicazione, utile nel caso si vogliano confrontare le distribuzioni di due variabili aleatorie attraverso un diagramma QQ:

$$\forall q \in [0,1] \quad \chi_q^X = \chi_q^Y \Leftrightarrow X \text{ e } Y \text{ sono identicamente distribuite}$$

Definizione di buon stimatore

- **Definizione di stimatore:** uno stimatore è una funzione di variabili aleatorie che preso in input un campione X_1, \dots, X_n , ovvero un numero di variabili aleatorie estratte da un'unica popolazione X , restituisce in output una stima di una quantità $\tau(\theta)$ dipendente dal parametro θ di X .
- **Definizione di stimatore non deviato e bias:** uno stimatore T si dice non deviato rispetto ad una quantità $\tau(\theta)$ se e solo se il suo valore atteso è pari a tale quantità, ovvero:

$$\mathbb{E}[T(X_1, \dots, X_n)] = \tau(\theta)$$

In caso contrario, il valore della statistica oscilla attorno a valore diverso da quello da stimare, ovvero sta sovrastimando o sottostimando la quantità corretta; si dice allora che è presente un *bias*, così definito:

$$b_{\tau(\theta)}(T) := \mathbb{E}[T] - \tau(\theta)$$

- **Definizione di stimatore consistente in media quadratica e MSE:** perché una statistica sia buona non solo essa deve oscillare attorno alla quantità da stimare, ma non deve neanche oscillare troppo! Definito errore quadratico medio della statistica rispetto a $\tau(\theta)$ il valore atteso del loro scarto quadratico, così:

$$MSE_{\tau(\theta)}(T) = \mathbb{E}[(T - \tau(\theta))^2]$$

si dice che la statistica è consistente in media quadratica rispetto a $\tau(\theta)$ quando tale errore quadratico medio tende a zero quando la taglia del campione tende a $+\infty$, ovvero quando:

$$\lim_{n \rightarrow +\infty} MSE_{\tau(\theta)}(T_n) = 0$$

Esiste poi il concetto di consistenza debole, per cui uno stimatore è debolmente consistente rispetto a una quantità $\tau(\theta)$ quando per un ϵ abbastanza piccolo vale che:

$$\lim_{n \rightarrow +\infty} P(|T - \tau(\theta)| < \epsilon) = 1$$

- **Relazione tra MSE, bias e varianza:** vale la seguente interessante formula:

$$MSE_{\tau(\theta)}(T) = Var(T) + (b_{\tau(\theta)}(T))^2$$

Come conseguenza, per gli stimatori non devianti l'errore quadratico medio è pari alla varianza.

- **Media campionaria come stimatore del valore atteso:** la media campionaria $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ è uno stimatore non deviato e consistente in media quadratica rispetto al valore atteso μ di qualunque popolazione X in quanto:

$$\begin{aligned} \mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{n} = \frac{\sum_{i=1}^n \mathbb{E}[X]}{n} = \frac{n\mu}{n} = \mu \\ MSE_{\mu}(\bar{X}) &= Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{1}{n^2} \sum_{i=1}^n Var(X) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \\ &\rightarrow \lim_{n \rightarrow +\infty} MSE_{\mu}(\bar{X}) = 0 \end{aligned}$$

Si noti che alcune delle semplificazioni fatte (varianza della somma come somma delle varianze) derivano dal fatto che le X_i sono indipendenti e identicamente distribuite come la popolazione X .

- **Varianza campionaria come stimatore della varianza:** la varianza campionaria $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ è uno stimatore non deviato della varianza σ^2 di qualunque popolazione X in quanto:

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) \rightarrow (n-1)S^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 \\ \mathbb{E}[S^2] &= \mathbb{E}\left[\frac{(n-1)S^2}{n-1}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1}\right] = \frac{1}{n-1} (n\mathbb{E}[X^2] - n\mathbb{E}[\bar{X}^2]) \\ &= \frac{1}{n-1} (nVar(X) + n\mathbb{E}[X]^2 - nVar(\bar{X}) - n\mathbb{E}[\bar{X}]^2) \\ &= \frac{1}{n-1} \left(n\sigma^2 + n\mu^2 - \frac{n\sigma^2}{n} - n\mu^2 \right) = \frac{1}{n-1} (\sigma^2(n-1)) = \sigma^2 \end{aligned}$$

Si noti che la sostituzione centrale è dovuta al fatto che $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Calcolo della confidenza

- **Calcolo col teorema centrale del limite:** ci si potrebbe chiedere a partire da quale taglia n del campione la media campionaria inizia ad approssimare il valore atteso di una distribuzione con un errore di al più ϵ con una confidenza di almeno $1 - \delta$ %, ovvero:

$$P(|\bar{X} - \mu| < \epsilon) \geq 1 - \delta$$

Ricordando che $\mathbb{E}[\bar{X}] = \mu$ e $Var(\bar{X}) = \frac{\sigma^2}{n}$ si può standardizzare la variabile e risolvere il valore assoluto, ottenendo la stessa probabilità ma rispetto a una variabile standardizzata:

$$P(|\bar{X} - \mu| \leq \epsilon) = P(-\epsilon \leq \bar{X} - \mu \leq \epsilon) = P\left(-\frac{\epsilon}{\sigma} \sqrt{n} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\epsilon}{\sigma} \sqrt{n}\right)$$

Ora, per la chiusura delle variabili normali se sapessimo che X è normale anche \bar{X} lo sarebbe, ma così non è purtroppo; tuttavia, per il teorema centrale del limite ricordando che \bar{X} è a conti fatti la somma di n variabili aleatorie la sua distribuzione può essere approssimata con quella di una normale, in questo caso una standard vista l'operazione di standardizzazione applicatagli:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1) \rightarrow P\left(-\frac{\epsilon}{\sigma} \sqrt{n} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\epsilon}{\sigma} \sqrt{n}\right) = P\left(-\frac{\epsilon}{\sigma} \sqrt{n} \leq Z \leq \frac{\epsilon}{\sigma} \sqrt{n}\right) = 2\Phi\left(\frac{\epsilon}{\sigma} \sqrt{n}\right) - 1$$

Ecco dunque che la disequazione originale si trasforma in:

$$2\Phi\left(\frac{\epsilon}{\sigma} \sqrt{n}\right) - 1 \geq 1 - \delta \rightarrow n \geq \left(\frac{\sigma}{\epsilon} \Phi^{-1}\left(1 - \frac{\delta}{2}\right)\right)^2$$

- **Calcolo con la disuguaglianza di Chebyshev:** sappiamo che su \bar{X} vale necessariamente la disuguaglianza di Chebyshev, ovvero che:

$$P(|\bar{X} - \mathbb{E}[\bar{X}]| < \epsilon) \geq 1 - \frac{Var(\bar{X})}{\epsilon^2}$$

Ricordando che $\mathbb{E}[\bar{X}] = \mu$, questa è proprio ciò che cercavamo nel punto precedente; per porre che sia maggiore di $1 - \delta$ basta allora porre che:

$$1 - \frac{\sigma^2}{n\epsilon^2} \geq 1 - \delta \rightarrow n \geq \frac{\sigma^2}{\delta\epsilon^2}$$

Tuttavia, questa formula, non introducendo alcuna approssimazione fornisce sempre una taglia necessaria del campione più grande, ovvero una stima troppo conservativa.