# CAMPIONAMENTO E STIMA

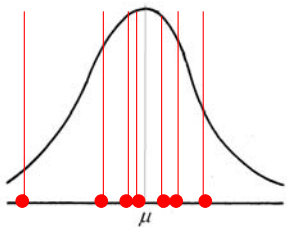


## CAMPIONAMENTO E STIMA

- Si parla di *campionamento* quando invece di osservare tutte le unità di una popolazione, se ne rilevano solo un sottoinsieme, detto campione
- *Tutta* la conoscenza empirica è effettivamente di tipo *campionario*
  - anche la misura di un oggetto non è altro che un campionamento di una osservazione tratta dalla distribuzione che caratterizza il processo di misura
  - il processo di misura può essere visto come il meccanismo generatore del dato osservato x:

$$x = \mu + \varepsilon$$

  - x è una stima campionaria di $\mu$

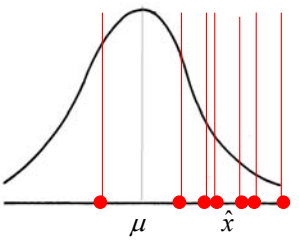



  - $\mu$ è il valore vero ed ignoto che vorremmo conoscere
  - $\varepsilon = x - \mu$ è un *errore* casuale, distribuito normalmente
- Per migliorare la *stima* di $\mu$, possiamo operare un maggior numero di osservazioni, ovvero di campionamenti da questa distribuzione, e utilizzare come stima la loro media aritmetica

# CAMPIONAMENTO E STIMA

- In generale, indichiamo la stima campionaria di $\mu$ con : $\hat{\mu} = \hat{x}$
  - è una funzione dei dati campionari: può essere una singola osservazione, nella maggior parte dei casi sarà la media di un campione di n osservazioni
- Le misure prodotte dallo strumento possono essere affette da errore:
  - errore **casuale**: somma di molteplici fattori indipendenti ciascuno di minima entità -> si distribuisce normalmente
  - errore **sistematico**: effetto di qualche aspetto rilevante del processo di misura -> sposta *sistematicamente* la misura in una direzione, provocando una *distorsione* della stima, cioè della misura prodotta

    se accade questo, significa che in pratica la distribuzione da cui stiamo campionando non è più centrata sul valore vero $\mu$






# CAMPIONAMENTO E STIMA

E[ ] indica il valore atteso: media calcolata sulle infinite prove

- **Principio del Campionamento Ripetuto**
- Quanto si avvicina la nostra stima al vero ed ignoto valore del carattere da misurare ?
- Pensiamo di ripetere il campionamento più volte, al limite *infinite* volte:
  - l'errore effettivamente commesso sarà diverso per ciascun campione estratto

    $\varepsilon_j = \hat{x}_j - \mu \qquad j = 1, \ldots, \infty$
  - l'errore commesso è *sconosciuto*, per ciascuna ripetizione del campionamento, perché non conosciamo il valore vero $\mu$ ...
- Possiamo valutare la stima sulla base dell'errore che si commette *in media*, ripetendo il campionamento all'infinito, cioè $E[\varepsilon]$
  - la stima si dice **corretta** (o **non distorta**) se l'errore è in media pari a 0, cioè errori positivi e negativi, nel corso delle infinite prove, si compensano:

    $E[\varepsilon] = 0$
  - quindi la distribuzione della stima sarà centrata sul valore vero $\mu$ :

    $E[\hat{x}] = \mu \qquad infatti: \quad E[\hat{x}] = E[\mu + \varepsilon] = \mu + E[\varepsilon]$
  - se viceversa la stima non *centra* il valore vero *nemmeno* in media, si dice che è **distorta** (o **non corretta**)

# CAMPIONAMENTO E STIMA

- **L'Errore Quadratico Medio di Stima (MSE)**
- Per valutare l'errore complessivo che si commette in media, in base il principio del campionamento ripetuto, utilizzando una certa stima, si definisce l'errore quadratico medio MSE :

$$MSE(\hat{x}) = E[\varepsilon^2] = E[(\hat{x} - \mu)^2]$$

- L'errore quadratico medio MSE può essere scomposto in due componenti:

$$E[\varepsilon^2] = E[(\hat{x} - \mu)^2] = E[(\hat{x} - \underbrace{E[\hat{x}] + E[\hat{x}]}_{\text{sommo e sottraggo la quantità } E[\hat{x}] \text{ costante}} - \mu)^2] =$$

$$= E[(\hat{x} - E[\hat{x}])^2 + (E[\hat{x}] - \mu)^2 + 2\,(\hat{x} - E[\hat{x}])\,(E[\hat{x}] - \mu)] =$$

$$= \underbrace{E[(\hat{x} - E[\hat{x}])^2]}_{V(\hat{x})} + E[\underbrace{(E[\hat{x}] - \mu)^2}_{\text{costante rispetto all'operatore } E[\,]}] + E[2\,(\hat{x} - E[\hat{x}])\,\underbrace{(E[\hat{x}] - \mu)}_{\text{costante rispetto all'operatore } E[\,]}] =$$

$$= V(\hat{x}) + (E[\hat{x}] - \mu)^2 + 2\,(E[\hat{x}] - \mu)\,\underbrace{E[(\hat{x} - E[\hat{x}])]}_{=0} = V(\hat{x}) + \Delta^2$$

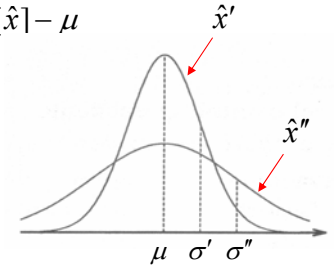


# CAMPIONAMENTO E STIMA

- Dunque l'errore complessivo di stima può essere scomposto nelle due componenti:

$$MSE(\hat{x}) = E[(\hat{x} - \mu)^2] = V[\hat{x}] + \Delta^2$$

  - la **distorsione** (***bias***), al quadrato : $\Delta = E[\hat{x}] - \mu$
  - la **precisione** (varianza della stima) : $V[\hat{x}]$




- Se lo stimatore è non distorto, la sua varianza misura la precisione della stima prodotta: minore è tale varianza e maggiore è la precisione della stima
- Come si determina se una stima è distorta e la sua precisione ?
- Vediamo il caso dello stimatore più diffuso: la media campionaria
  Pensiamo di effettuare un campionamento di n osservazioni tratte dalla distribuzione che caratterizza il processo di misura:

$$x_i = \mu + \varepsilon_i \quad \varepsilon_i \sim N(\delta, \sigma) \quad \forall i = 1, \ldots, n$$

In genere si ipotizza che l'errore si distribuisca normalmente, con media $\delta$ e scarto quadratico medio $\sigma$

# CAMPIONAMENTO E STIMA

- **Lo stimatore Media Campionaria**

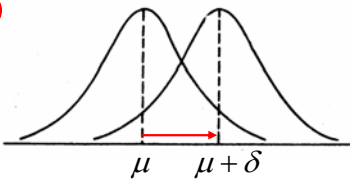
$$\hat{\mu} = \hat{x} = \bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

- Pensiamo di ripetere infinite volte il campionamento:
  - in corrispondenza di ciascun campione otterremo una stima
  - in definitiva avremo un numero infinito di stime campionarie, in corrispondenza degli infiniti campioni estratti
- In media, sugli infiniti campioni, riusciamo a centrare il valore vero $\mu$ ?

$$x_i = \mu + \varepsilon_i \quad \varepsilon_i \sim N(\delta, \sigma)$$

$$E[\bar{x}] = E\left[\frac{\sum x_i}{n}\right] = \frac{1}{n}\sum E[x_i] = \frac{1}{n}\sum E[\mu + \varepsilon_i] =$$

$$= \frac{1}{n}\left(\sum \mu + \sum E[\varepsilon_i]\right) = \frac{1}{n} n\,\mu + \frac{1}{n} n\,\delta = \mu + \delta$$




- *Bias* : da *quale* distribuzione stiamo campionando ?!



# CAMPIONAMENTO E STIMA

- Determiniamo la precisione dello stimatore media campionaria:

$$V[\bar{x}] = V\left[\frac{\sum x_i}{n}\right] = \frac{1}{n^2} V\left[\sum x_i\right] = \frac{1}{n^2}\sum \underbrace{V[x_i]}_{\substack{\text{tutte uguali} \\ \text{a } \sigma^2}} = \frac{1}{n^2}\sum \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$x_i = \mu + \varepsilon_i \quad \varepsilon_i \sim N(\delta, \sigma)$$

*infatti le singole osservazioni $x_i$ sono indipendenti e provengono tutte dalla stessa distribuzione (si dice che sono i.i.d.) con varianza:* $V[x_i] = V[\varepsilon_i] = \sigma^2$

*Ricordiamo che, per la somma di n variabili indipendenti, vale la proprietà:*

$$V\left[X_1 + X_2 + \ldots + X_n\right] = \left(V[X_1] + V[X_2] + \ldots + V[X_n]\right)$$

- La varianza dello stimatore media campionaria dipende da :
  - la varianza del processo di misura, cioè la $V[\varepsilon] = \sigma^2$
  - la dimensione del campione

# CAMPIONAMENTO E STIMA

- Come si può *controllare* la precisione della stima ?
- Proprio in virtù di questa semplice proprietà della Media Campionaria, è possibile *controllare* a piacimento la precisione della stima, infatti osserviamo che :

$$V[\bar{x}] = \frac{\sigma^2}{n}$$

  - la precisione è funzione inversa di n, cioè diminuisce al crescere di n
  - è quindi sempre possibile garantire una precisione desiderata, dimensionando adeguatamente il campione: il problema diventa determinare n
  - sarà possibile determinare il numero di misurazioni da eseguire (o di soggetti da selezionare) tale da garantire di ottenere una stima con la precisione richiesta

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- Come si può *controllare* la distorsione della stima ?
  - La distorsione non dipende tanto dalla *dimensione* del campione, cioè dal fatto di effettuare solo un numero limitato di osservazioni, quanto dalle *modalità di scelta* delle unità che vengono "estratte" per fare parte del campione
  - La distorsione dipende inoltre dallo strumento di rilevazione, dal processo di misura, e da un gran numero di altri aspetti collegati alle modalità di rilevazione dei dati osservati, che determinano in ultima analisi la distribuzione campionaria
  - Questo significa che il problema della distorsione non si risolve semplicemente aumentando la dimensione del campione: anche eseguendo infinite estrazioni-misurazioni, produrremo sempre una stima distorta se il problema risiede ad es. nello strumento di misura (es. bilancia, formulazione domanda questionario)
  - Per evitare che la stima sia distorta, è necessario garantire che il campione estratto provenga ovvero riproduca fedelmente la distribuzione del "fenomeno" di interesse, e non un'altra
  - A tal fine, occorre prestare la massima attenzione a tutte le fasi del processo di rilevazione, per evitare di commettere, in particolare, quel tipo di errori che possono introdurre una distorsione sulle stime prodotte

# CAMPIONAMENTO E STIMA

- **Fonti di distorsione dovute al processo di misura**:
  - Metodo di selezione del campione: un *bias* si verifica quando il campione selezionato non riproduce la distribuzione del fenomeno di interesse, ma presenta caratteristiche diverse rispetto alla popolazione obiettivo, per la quale si vogliono ottenere delle stime:
    - utilizzo di soggetti facilmente accessibili (es. studenti), non rappresentativi della popolazione obiettivo
    - auto-selezione dei soggetti (es. soggetti "volontari" / inaccessibili)
    - problematiche di "copertura" della popolazione, in relazione alla tecnica di indagine utilizzata (es. interviste telefoniche, internet)
    - incidenza delle mancate risposte / rifiuti ("mortalità")
  - Strumento di rilevazione non adeguato: ad es. in un questionario, una domanda può essere interpretata in modo diverso da quello desiderato; la formulazione della domanda o delle modalità di risposta previste possono indurre a fornire una particolare risposta (magari non sono state previste tutte le possibilità); anche l'ordine delle modalità può influenzare la risposta
  - Vari errori che si possono commettere *sistematicamente* in tutto il processo di rilevazione, sia nel caso di esperimenti che di indagini sociali: errori in fase di registrazione dei dati, di codifica, di trascrizione delle risposte su supporti informatici, ecc … (se invece gli stessi errori si verificano in modo casuale, non introducono una distorsione)

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- **Fonti di distorsione dovute all'osservatore** (rilevatore/sperimentatore/intervistatore)
  - Deviazione dalle direttive del protocollo di rilevazione: es. in una intervista parafrasare le domande, usare "parole proprie"
  - Le aspettative del rilevatore: un ricercatore che si aspetta di trovare delle differenze tra gruppi di soggetti, può inconsapevolmente assumere atteggiamenti e comportamenti che possono:
    - influenzare i soggetti studiati nella direzione attesa
    - rendere selettive o distorcere le sue stesse percezioni, provocando errori sistematici di osservazione e persino di registrazione dei dati
  - L'interazione tra rilevatore e soggetto: si instaura una situazione sociale asimmetrica, nella quale il ruolo del rilevatore non risulta indifferente a chi è *sottoposto* all'osservazione
  - Le caratteristiche del rilevatore: caratteristiche fisiche, il sesso, l'età, la personalità, l'esperienza, possono influenzare in vari modi i risultati
  - Errori di comportamento inintenzionale del rilevatore: mancanza di uniformità nella presentazione delle istruzioni, variazione del tono della voce nel sottolineare determinate risposte, reazioni fisiche in coincidenza con particolari risposte dei soggetti(cambiamenti di tensione del corpo, dello sguardo, del sorriso, movimenti degli occhi, dilatazione della pupilla)

# CAMPIONAMENTO E STIMA

- **Fonti di distorsione dovute al soggetto osservato**
  - Direzione delle risposte: l'atteggiamento dei soggetti può variare molto, dalla condiscendenza al boicottaggio; dalla semplice collaborazione, al desiderio di contribuire al progresso della conoscenza, fino al soggetto che vuole fare buona impressione...
  - Desiderabilità sociale: induce risposte "ideali" in soggetti con alto grado di desiderabilità sociale, ad es. si verifica che i soggetti, avendo intuito quello che a loro parere è lo scopo della ricerca, cercano di rispondere in modo da avvalorarla; può essere necessario ricorrere all'inganno per sviare i soggetti dal vero obiettivo della ricerca
  - Percezione di sé: la consapevolezza di essere osservati, per scopi scientifici o anche per scopi non noti, può alterare le risposte o le prestazioni dei soggetti e addirittura le reazioni ai trattamenti (es. effetto placebo, effetto Hawthorne)
  - L'errore del soggetto in un campione casuale è generalmente incorrelato, cioè non sempre introduce una distorsione, ma aumenta la variabilità osservata

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- **Campionamento di popolazioni**
- Quando si parla di campionamento nelle scienze sociali, si intende generalmente campionamento di popolazioni (individui, famiglie, imprese, eventi, ...)
- Si definisce popolazione **obiettivo** (o **universo**) quella per la quale si vogliono produrre le stime, che deve essere definita nel *contenuto*, nello *spazio* e nel *tempo* : (es. pop. italiana ≡ esseri umani residenti sul territorio Italiano il 12/04/2005 ...)
- Il ricorso al campionamento è di fatto obbligato quando la rilevazione esaustiva di tutte le unità della popolazione non è possibile, ad esempio quando:
  - la popolazione è costituita da un numero virtualmente infinito di unità
  - l'osservazione dell'unità ne comporta la distruzione (es. durata lampadina)
- Quando la popolazione obiettivo è reale e finita, l'indagine può essere estesa a tutte le unità (es. censimento): si tratta di valutare se ciò sia conveniente in relazione a costi, tempi e obiettivi.
- Si procede solitamente alla rilevazione esaustiva quando:
  - le unità che costituiscono la popolazione obiettivo sono rare o comunque relativamente poco numerose
  - è necessario un livello di dettaglio, ad es. territoriale, molto elevato
  - il costo di reperimento dell'informazione è trascurabile (es. dati ricavabili da archivi amministrativi informatizzati)

# CAMPIONAMENTO E STIMA

- Il ricorso ad una rilevazione campionaria è in generale preferibile per ragioni economiche :
  - quando la popolazione è molto numerosa
  - quando le unità sono difficilmente raggiungibili (es. disperse su ampio territorio)
  - il costo del contatto è elevato (es. intervista faccia a faccia, esame medico o di laboratorio, esperimenti, ...)
- Si preferisce l'indagine campionaria a quella esaustiva anche per altre ragioni:
  - i risultati sono disponibili con maggiore tempestività
  - è possibile realizzare indagini più approfondite e mirate
  - il questionario può essere più complesso
  - è possibile attuare un maggiore controllo dell'errore di rilevazione (extra-campionario) rispetto alle indagini di vaste dimensioni, che :
    - richiedono l'impiego massivo di personale non sempre adeguatamente addestrato,
    - e sono gravate da enormi volumi di lavoro (e di dati) con scarse possibilità di controllo

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- **Il campionamento probabilistico**
- Il campionamento probabilistico (o casuale, o statistico) si caratterizza per il metodo di selezione delle unità della popolazione che vengono incluse nel campione per essere osservate: la selezione delle unità avviene in modo *casuale*
- Selezionare le unità *casualmente* non significa sceglierle "a casaccio", senza un criterio, ma al contrario impiegare tecniche specifiche per garantire una probabilità di selezione prestabilita ad ogni unità della popolazione (es. uguale per tutti)
- La metafora che possiamo utilizzare per descrivere la selezione casuale è quella dell'estrazione delle palline da un'urna: l'estrazione casuale di un numero *sufficiente* di unità permette di riprodurre nel campione la distribuzione della popolazione => è questo il concetto di **campione *rappresentativo***
- Il "*caso*" è il massimo garante che la distribuzione di una variabile nel campione riproduca fedelmente quella dell'universo: è improbabile che in un campione di numerosità adeguata, per effetto del caso si trovino ad es. solo soggetti giovani, sarà molto più probabile ritrovarvi la stessa distribuzione d'età della popolazione di provenienza
- La selezione casuale è quindi la migliore garanzia che le stime campionarie non risultino affette da distorsione per effetto del campionamento e siano dunque *generalizzabili* all'universo da cui il campione è stato estratto; tutte le altre fonti di distorsione di natura extra-campionaria ovviamente permangono
- E' possibile determinare, ed anzi prefissare, la precisione delle stime

# CAMPIONAMENTO E STIMA

- **Campioni non probabilistici**
- Nel campionamento non probabilistico vengono compresi tutti gli altri criteri di formazione di un campione che non garantiscono la *casualità* della selezione delle unità da osservare
- I campioni le cui unità non vengono selezionate casualmente riflettono, nel bene e nel male, le idee e gli orientamenti di chi li costruisce, e saranno quindi sempre caratterizzati da un elevato grado di soggettività
- Il *controllo* della distorsione non è più garantito dal caso, ma risulta affidato a considerazioni e valutazioni soggettive
- Tra i campioni non probabilistici troviamo:
  - campione a scelta ragionata: le unità sono selezionate in modo da somigliare nell'insieme, per alcuni caratteri strutturali (età, sesso, ...), alla popolazione obiettivo
  - campione per quote: la dimensione del campione è prefissata ma la scelta delle unità è lasciata agli intervistatori, che devono rispettare delle "quote" di soggetti che presentano determinate caratteristiche prestabilite
  - campioni "volontari", basati sull'adesione volontaria dei rispondenti: portano a risultati quasi certamente distorti a causa dell'*autoselezione* dei soggetti (es. lettori di un giornale, televoto di Biscardi, ...)

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- **Tecniche di campionamento statistiche**
- Nell'ambito del campionamento casuale sono stati sviluppati molti diverse tecniche (o "disegni") per la formazione del campione, per risolvere in modo *efficiente* le diverse situazioni che ci si trova a dover affrontare nella realtà:
  - **casuale semplice**
  - su due (o più) stadi
  - stratificato
  - a grappoli
  - ...
- Nel seguito ci limiteremo al primo e più semplice metodo, detto **Campionamento Casuale Semplice**: direttamente assimilabile all'estrazione da un'urna che contiene tutte le unità della popolazione, è il punto di riferimento per tutto il campionamento
- Il CCS fornisce le maggiori garanzie di controllo, eventualmente al prezzo di una minore *efficienza*, rispetto ad altri schemi, cioè richiede un campione più numeroso
- Il CCS è applicabile anche in assenza di informazioni sulla struttura della popolazione, che permetterebbero di progettare un campione più strutturato (ad es. stratificato), con maggiore efficienza.
- I limiti del CCS sono rappresentati dalla non applicabilità in determinate situazioni: quando la lista di tutte le unità che formano la popolazione non è disponibile, o per ragioni di costo, in relazione alla tecnica di rilevazione che si intende adottare

# CAMPIONAMENTO E STIMA

- **Fasi del Campionamento Casuale (semplice)**
- 1. LISTA: formazione della lista di tutte le unità che compongono la popolazione (cosa non sempre agevole, né sempre possibile ...)
- 2. PROBABILIZZAZIONE: attribuzione ad ogni unità della lista di una probabilità di selezione maggiore di zero (nel CCS è uguale per tutti)
- 3. NUMEROSITÀ: si determina il numero di unità da selezionare, sufficiente a garantire una precisione prestabilita della stima risultante
- 4. SELEZIONE: scelta delle unità campionarie dalla lista della popolazione con metodi che garantiscono la casualità; in pratica, non si mettono materialmente delle palline in un urna... ci sono diverse tecniche, ma due sono quelle principali:
  - selezione **pseudo-casuale**: ad ogni unità della lista vengono assegnati uno o più numeri, proporzionalmente alla probabilità di selezione assegnata, e poi con un computer si generano n numeri pseudo-casuali (*random*) che simulano l'estrazione
  - selezione **sistematica**: si ordinano le unità in modo opportuno (a volte va già bene l'ordine *naturale*) e se ne prende una ogni k, dove:
    k = N / n è detto **passo di campionamento**
    N = numero di unità statistiche che costituiscono la (lista della) popolazione
    n = numero di unità da selezionare, cioè la dimensione del campione

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- Esempio: Vogliamo stimare il numero di studenti che si presenteranno al primo appello. Supponiamo di avere la brillante idea di prendere come campione gli studenti della prima fila: questo è un criterio soggettivo, non una selezione casuale, quindi la stima potrebbe risultare distorta
- Per esempio perché gli studenti della prima potrebbero essere quelli che:
  - studiano di più
  - si svegliano prima
  - abitano più vicino
  - ci vedono meno bene ...
- Il punto cruciale è la relazione (eventuale) che questa diversità dei soggetti selezionati può avere con la variabile oggetto di stima:
  - se gli studenti della prima fila si differenziano, ad es. solo per l'ultimo aspetto, plausibilmente indipendente dalla variabile "voglia/tempo di studiare" -> non avremo alcuna distorsione, cioè la stima prodotta sulla base delle loro risposte sarà attendibile e rappresentativa dell'intera classe
  - se invece in prima fila troviamo quelli che studiano di più, chiedendo *solo* a loro se intendono sostenere l'esame in preappello, avremo una *sovrastima* del numero di studenti realmente intenzionati a parteciparvi

# CAMPIONAMENTO E STIMA

- **Teorema del limite centrale**
- Supponiamo di ripetere infinite volte l'estrazione di un campione di n unità da una distribuzione normale con media $\mu$ e varianza $\sigma^2$:

$$x_i = \mu + \varepsilon_i \quad \varepsilon_i \sim N(0,\sigma) \quad \forall i = 1, n$$

- Il teorema del limite centrale afferma che la media campionaria, *al crescere di n*, *tende* a distribuirsi normalmente con media $\mu$ e varianza $\sigma^2/n$

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \xrightarrow{n} N(\mu, \frac{\sigma}{\sqrt{n}})$$

- Il teorema del limite centrale, in effetti, riassume i risultati sulla distribuzione della media campionaria che avevamo già ricavato :

$$E[\bar{x}] = \mu, \; data \; E[\varepsilon] = 0, \; e \quad V[\bar{x}] = \frac{\sigma^2}{n}$$

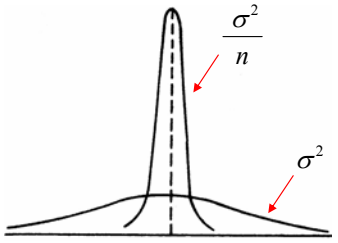
E aggiunge che la *forma* della sua distribuzione tende ad una normale:
la media, considerata come trasformazione di variabile, è la somma di n variabili i.i.d. indipendenti e identicamente distribuite (diviso una costante n)



# CAMPIONAMENTO E STIMA

- La distribuzione della media campionaria dunque *tende* alla normale al crescere di n (cioè all'infinito): ma per n finito, e relativamente piccolo ?
  - in generale, più la distribuzione di partenza è lontana dalla normalità e più la convergenza della media alla normale è lenta, e quindi il campione dovrà essere più numeroso
  - se la distribuzione da cui si campionano le osservazioni X(i) è normale o almeno simmetrica, la distribuzione della media *converge* alla normale molto rapidamente
  - se invece la distribuzione delle X(i) è asimmetrica, il campione dovrà essere di almeno 30 unità;
  - per n > 100 la convergenza è praticamente assicurata (salvo casi anomali)
- E' importante sottolineare ancora come la media campionaria permetta di migliorare la precisione della stima di un fattore pari a 1/n
- Lo scarto quadratico medio della media campionaria, detto anche ***Standard Error,*** si riduce in funzione della radice quadrata di n :

$$SE = \sqrt{V(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$



# CAMPIONAMENTO E STIMA

- **Intervallo fiduciario della stima**
- Sulla base del teorema del limite centrale, che ci fornisce la distribuzione della media campionaria, possiamo capire esattamente che cosa significa stimare $\mu$ con il dato campionario

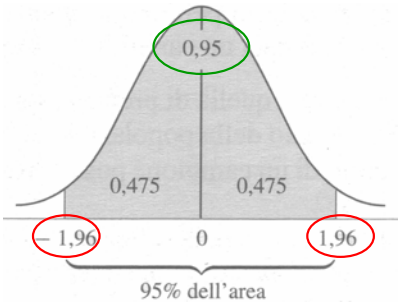
$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- Consideriamo la seguente trasformazione della media campionaria:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- Affermazione probabilistica:

$$P\{-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\} = 1 - \alpha$$




Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- Esplicitando rispetto a $\bar{x}$ otteniamo :

$$P\{-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\} = 1 - \alpha$$

$$P\{-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq (\bar{x} - \mu) \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1 - \alpha$$

$$P\{\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1 - \alpha$$

- Dunque si può affermare che la probabilità che la media campionaria cada in un intervallo centrato su $\mu$ di dimensione $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ è pari a $1 - \alpha$

# CAMPIONAMENTO E STIMA

- Capiamo il significato di questo risultato:

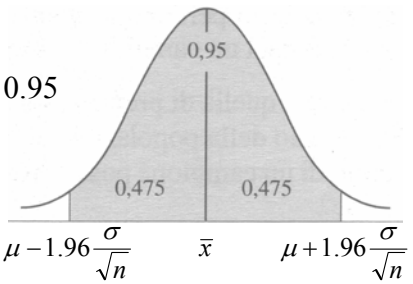
$$P\{\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1-\alpha$$

- Significato di $\alpha$ : se ad es. scegliamo (1-$\alpha$) = 0.95 (95%), allora

$$z_{\alpha/2} = z_{0.95/2} = 1.96$$

significa che ripetendo il campionamento 100 volte, in 95 campioni la media campionaria cadrà in un intorno di $\mu$ di ampiezza, cioè non più distante dal valore vero:

$$P\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\} = 0.95$$






# CAMPIONAMENTO E STIMA

- **Stima Intervallare**
- La conoscenza della distribuzione della media campionaria, fornitaci dal teorema del limite centrale, permette di dare una stima intervallare per $\mu$ , cioè di individuare un intervallo in cui il valore vero cadrà con probabilità prefissata
- Tale intervallo è detto **intervallo di confidenza**
- Il punto di partenza è sempre il teorema del limite centrale, ma questa volta esplicitando rispetto a $\mu$ si ottiene:

$$P\{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1-\alpha$$

- L'intervallo di confidenza è centrato sul valore della media campionaria (**stima puntuale**) e ha dimensione ...
- L'intervallo di confidenza contiene il valore vero di $\mu$ con probabilità (1-$\alpha$): questa probabilità è detta **livello di significatività** dell'intervallo di confidenza
- Cosa significa fissare (1-$\alpha$) = 0.95 ? Ipotizzando di ripetere 100 volte il campionamento, significa che per ben 95 volte il vero valore cadrà nell'intervallo di confidenza:

$$P\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\} = 0.95$$

# CAMPIONAMENTO E STIMA

- Come si arriva a determinare l'intervallo di confidenza: partendo dal teorema del limite centrale, esplicitiamo la diseguaglianza rispetto a $\mu$ :

$$P\{-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\alpha/2}\} = 1-\alpha$$

$$P\{-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le (\bar{x}-\mu) \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1-\alpha$$

$$P\{-\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1-\alpha$$

*raccogliendo* $-1$ *si ottiene* :

$$P\{-(\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \le -(\mu) \le -(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}})\} = 1-\alpha$$

*Ricordando una proprietà delle diseguaglianze* : $-a \le -b \le -c \Rightarrow c \le b \le a$

*ad esempio* : $-3 \le -2 \le -1 \Rightarrow 1 \le 2 \le 3$ , *si ha* :

$$P\{\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1-\alpha$$



# CAMPIONAMENTO E STIMA

- Esercizio: Una società di indagini di mercato esegue un sondaggio sulle intenzioni di voto nel prossimo referendum ottenendo, su 1746 risposte, il 39% di SI. Determinare l'intervallo di confidenza al 95% di significatività.
- Intanto osserviamo che la variabile rilevata è dicotomica (si, no): il valore da stimare è quindi una percentuale p, e la varianza è data da p (1 – p)
- L'intervallo di confidenza, in generale, è quello che garantisce che:

$$P\{\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1-\alpha$$

- Vogliamo una stima significativa al 95%, quindi l'intervallo richiesto diventa:

$$P\{\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\} = 0.95$$

- A posteriori possiamo stimare la varianza della popolazione con quella campionaria, che per una variabile dicotomica è data da: 0,39 (1 – 0,39)

$$1.96\frac{\sqrt{0.39\,(1-0.39)}}{\sqrt{1746}} = 1.96\sqrt{\frac{0.2379}{1746}} = 1.96 \cdot 0{,}0117 = 0.0229$$

$$P\{0.39-0.0229 \le \mu \le 0.39+0.0229\} = 0.95$$

*L'intervallo di confidenza al 95% è dato da* : $0.3671 \le \mu \le 0.4129$

# CAMPIONAMENTO E STIMA

- **Quanto deve essere grande il campione ?**
- Errore molto comune è pensare che il campione debba essere proporzionato alla dimensione della popolazione: invece non c'è nessuna ragione perché il campione debba essere più numeroso all'aumentare della popolazione obiettivo
- Il punto fondamentale è invece la *variabilità* presente nella popolazione: se la variabilità che caratterizza la distribuzione della variabile obiettivo nella popolazione è piccola, cioè la popolazione è molto omogenea, sarà sufficiente un campione piccolo per ottenere una stima anche molto precisa
- Caso limite: se una popolazione fosse composta di 1.000.000 di soggetti, tutti *identici* (almeno per la variabile di interesse), basterebbe un campione di *una* sola unità per avere una stima esatta
- La numerosità *ottimale* del campione è quella che permette di garantire gli obiettivi dell'indagine, in termini di precisione della stima prodotta, con il minimo costo: è cioè il numero *minimo* di unità necessario per assicurare l'obiettivo informativo
- Per determinare la numerosità ottimale occorre avere già un'idea, o fare delle ipotesi, sulla variabilità della distribuzione del carattere nella popolazione

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- **Determinazione della numerosità ottimale del campione**
- L'obiettivo informativo è garantire una determinata precisione della stima, che può essere esplicitata in due modi diversi:
  - prefissando direttamente la precisione, cioè la varianza della stima: ma non è facile ragionare direttamente in termini di varianza ...
  - prefissando la dimensione (ampiezza) dell'intervallo di confidenza, in cui la stima dovrà cadere *con probabilità prefissata* (1-α), usualmente il 95%
- 1° Metodo - Prefissare direttamente la precisione della stima :

$$V[\bar{x}] = \frac{\sigma^2}{n} \quad \Rightarrow \quad \boxed{n = \frac{\sigma^2}{V[\bar{x}]}}$$

- 2° Metodo - Prefissare la dimensione desiderata dell'intervallo di confidenza e il **livello di significatività**, cioè la probabilità (1-α) :

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le d \;\Rightarrow\; z_{\alpha/2}\,\sigma \le d\sqrt{n} \;\Rightarrow\; \sqrt{n} \ge \frac{z_{\alpha/2}\sigma}{d} \;\Rightarrow\; \boxed{n \ge \frac{z_{\alpha/2}^2\,\sigma^2}{d^2}}$$

$$Es.\ Fissato\ \ 1-\alpha = 0{,}95\ \ abbiamo: \quad n \ge \frac{z_{0{,}025}^2\,\sigma^2}{d^2} = \frac{(1{,}96)^2\,\sigma^2}{d^2}$$

# CAMPIONAMENTO E STIMA

- Bene: ora sappiamo calcolare la numerosità campionaria... o quasi
- C'è un piccolo problema: guardiamo meglio le espressioni per determinare n ... Dipendono entrambe dalla varianza della popolazione, che purtroppo è ignota
- Per procedere al calcolo di n è necessario avere un'idea di qual è la variabilità naturale del fenomeno. Abbiamo due strade:
  - avanzare delle ipotesi (sulla base delle conoscenze a priori sul fenomeno) sul livello di variabilità che ci aspettiamo di trovare, e basare su di esse il calcolo di n
  - stimare la varianza della popolazione con dati eventualmente disponibili, per esempio risultati di precedenti ricerche, o di una indagine pilota
  - in casi particolari (es. variabile dicotomica), in assenza di informazioni, si può assumere la situazione peggiore (massima variabilità)
- *A posteriori*, cioè dopo aver portato a termine l'indagine, potremo stimare la varianza della popolazione con i dati del campione, e verificare quale precisione siamo riusciti a garantire *effettivamente*

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- **Stima campionaria della Varianza della popolazione**
- La stima campionaria della varianza elementare

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \qquad \text{è data da:} \qquad \hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- La stima $s^2$ è *quasi* intuitiva : ci può stupire però la stranezza di dividere per (n-1). Perché dividere per (n-1) e non per n ?
- La risposta è che $s^2$ è uno stimatore **corretto** (**non distorto**) per $\sigma^2$, mentre

$$\frac{\sum (x_i - \bar{x})^2}{n}$$ invece non lo è

- Si può dimostrare infatti che : $E\left[\frac{\sum (x_i - \bar{x})^2}{n}\right] = \sigma^2 \frac{n-1}{n} \neq \sigma^2$

  quindi la varianza campionaria non è una buona stima di $\sigma^2$: è una stima distorta, che sottostima sistematicamente la vera varianza
- Per questo si usa invece $s^2$ che fornisce una stima non distorta per $\sigma^2$

# CAMPIONAMENTO E STIMA

- Dimostrazione :

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

*ricordiamo che* : $V[y]=M[y^2]-(M[y])^2$ *e quindi* : $M[y^2]=V[y]+(M[y])^2$

$$=\frac{1}{n}\sum E[X_i^2]-E[\bar{X}^2]=\frac{1}{n}\sum(V[X_i]+E[X_i]^2)-(V[\bar{X}]+E[\bar{X}]^2)=$$

$$=\frac{1}{n}\sum(\sigma^2+\mu^2)-\frac{\sigma^2}{n}-\mu^2=\frac{1}{n}n\sigma^2+\frac{1}{n}n\mu^2-\frac{\sigma^2}{n}-\mu^2=$$

$$=\sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2=\sigma^2-\frac{\sigma^2}{n}=\frac{n\sigma^2-\sigma^2}{n}=\sigma^2\frac{n-1}{n}$$

*Ecco dunque come si arriva a* $s^2$ :

$$s^2=\frac{\sum(x_i-\bar{x})^2}{n}\frac{n}{n-1}=\frac{\sum(x_i-\bar{x})^2}{n-1}\quad\Rightarrow\quad E[s^2]=\sigma^2$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2006/07

# CAMPIONAMENTO E STIMA

- **Stima della varianza di stima**
- Dopo aver portato a termine l'indagine, potremo stimare la varianza della popolazione con i dati del campione, per verificare quale precisione siamo riusciti a garantire *effettivamente*
- La varianza di stima effettiva dell'indagine potrà essere *stimata*, a posteriori, nel modo seguente :

$$V[\bar{x}]=\frac{\sigma^2}{n}\quad\Rightarrow\quad\hat{V}[\bar{x}]=\frac{\hat{\sigma}^2}{n}=\frac{s^2}{n}$$

# CAMPIONAMENTO E STIMA

- Esercizio. La società di indagini di mercato che deve eseguire il sondaggio sul referendum, deve determinare la numerosità campionaria in grado di garantire una precisione del 2%, cioè che la stima cadrà, con il 95% di probabilità, in un intervallo di ampiezza 4% (+2% a destra e -2% a sinistra della stima puntuale)
- La dimensione del campione da determinare è la numerosità minima in grado di garantire che:

$$P\{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} = 1-\alpha$$

- Abbiamo visto che n si determina come:

$$n \geq \frac{z_{\alpha/2}^2\,\sigma^2}{d^2} \qquad \textit{fissato } 1-\alpha = 0{,}95: \quad n \geq \frac{(1{,}96)^2\,\sigma^2}{d^2}$$

- A priori non abbiamo una stima della varianza della popolazione: se riteniamo il risultato incerto, o in assenza di ipotesi su tale variabilità, assumiamo la variabilità massima possibile, che per una variabile dicotomica si registra quando p=0.5

$$n \geq \frac{(1{,}96)^2\,0{,}5(1-0{,}5)}{(0.02)^2} = \frac{(1{,}96)^2\,0{,}25}{(0.02)^2} = \frac{3{,}8415\cdot 0{,}25}{(2/100)^2} = \frac{0{,}9604}{4/10000} = \frac{9604}{4} = 2401$$



# CAMPIONAMENTO E STIMA

- Esercizi. Variazioni sul tema ...
- Se invece avessimo voluto garantire una precisione dell'1% (in più e in meno), cioè una dimensione dell'intervallo di confidenza di 2 punti percentuali ?
- La dimensione del campione diventa:

$$n \geq \frac{(1{,}96)^2\,0{,}5(1-0{,}5)}{(0.01)^2} = \frac{(1{,}96)^2\,0{,}25}{(0.01)^2} \cong \frac{3{,}8415\cdot 0{,}25}{(1/100)^2} = \frac{0{,}9604}{1/10000} = 9604$$

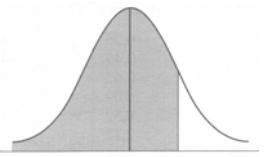
- Osserviamo che per avere una precisione *doppia* dell'intervallo di confidenza (cioè con ampiezza dimezzata), la numerosità del campione deve *quadruplicare*

# CAMPIONAMENTO E STIMA

- Esercizi. Variazioni sul tema … Torniamo ad un intervallo di ampiezza +/- 2%
- Se volessimo essere ancora più sicuri che l'intervallo di confidenza contenga la stima, invece del 95% di significatività potremmo fissare il 99%. Cosa cambia ?

$$n \geq \frac{z_{\alpha/2}^2\ \sigma^2}{d^2} \qquad \textit{fissato}\ \ 1-\alpha = 0{,}99: \qquad n \geq \frac{(z_{0{,}005})^2\ \sigma^2}{d^2}$$

- Se abbiamo una tavola con la Funzione di ripartizione F(z) della Normale standard, oppure usiamo Excel, cercheremo il punto z(a/2) = z(0,005) = -2,5758, da cui:



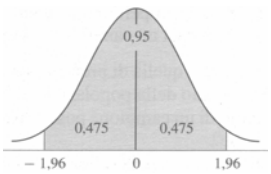
$$n \geq \frac{(-2{,}5758)^2\ \sigma^2}{d^2} = \frac{(2{,}5758)^2 \cdot 0{,}25}{(0{,}02)^2} = \frac{6{,}6349 \cdot 0{,}25}{4/10000} = 4147$$

- Sulla tavola della normale che abbiamo usato finora, che ci fornisce invece l'area compresa in [0, z] , dovremo cercare l'area 0,99/2 = 0,495, che sembra esattamente (per l'imprecisione della tavola) a metà tra 2,57 e 2,58, per cui concluderemmo z = 2,575
- La precisione assoluta non è poi così importante nella determinazione di n, quindi possiamo limitarci a due cifre decimale e in definitiva usare, per (1-a) = 99% : z = 2,58



# CAMPIONAMENTO E STIMA

- **Valori di z(α/2) utilizzati nei problemi di campionamento:**




- $1-\alpha = 0{,}95$

  $z_{\alpha/2} = z_{0{,}025} = 1{,}96$

  $n \geq \dfrac{(1{,}96)^2\ \sigma^2}{d^2}$

- $1-\alpha = 0{,}99$

  $z_{\alpha/2} = z_{0{,}005} = 2{,}5758$

  $n \geq \dfrac{(2{,}58)^2\ \sigma^2}{d^2}$

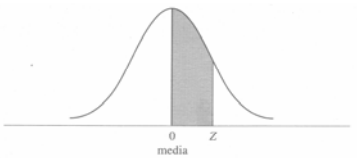
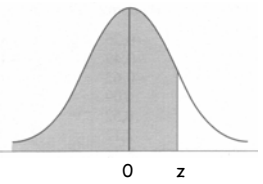



**Tavola A.1** La distribuzione normale.

| z | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,0 | 0,0000 | 0,0040 | 0,0080 | 0,0120 | 0,0160 | 0,0199 | 0,0239 | 0,0279 | 0,0319 | 0,0359 |
| 0,1 | 0,0398 | 0,0438 | 0,0478 | 0,0517 | 0,0557 | 0,0596 | 0,0636 | 0,0675 | 0,0714 | 0,0753 |
| 0,2 | 0,0793 | 0,0832 | 0,0871 | 0,0910 | 0,0948 | 0,0987 | 0,1026 | 0,1064 | 0,1103 | 0,1141 |
| 0,3 | 0,1179 | 0,1217 | 0,1255 | 0,1293 | 0,1331 | 0,1368 | 0,1406 | 0,1443 | 0,1480 | 0,1517 |
| 0,4 | 0,1554 | 0,1591 | 0,1628 | 0,1664 | 0,1700 | 0,1736 | 0,1772 | 0,1808 | 0,1844 | 0,1879 |
| 0,5 | 0,1915 | 0,1950 | 0,1985 | 0,2019 | 0,2054 | 0,2088 | 0,2123 | 0,2157 | 0,2190 | 0,2224 |
| 0,6 | 0,2257 | 0,2291 | 0,2324 | 0,2357 | 0,2389 | 0,2422 | 0,2454 | 0,2486 | 0,2517 | 0,2549 |
| 0,7 | 0,2580 | 0,2611 | 0,2642 | 0,2673 | 0,2703 | 0,2734 | 0,2764 | 0,2794 | 0,2823 | 0,2852 |
| 0,8 | 0,2881 | 0,2910 | 0,2939 | 0,2967 | 0,2995 | 0,3023 | 0,3051 | 0,3078 | 0,3106 | 0,3133 |
| 0,9 | 0,3159 | 0,3186 | 0,3212 | 0,3238 | 0,3264 | 0,3289 | 0,3315 | 0,3340 | 0,3365 | 0,3389 |
| 1,0 | 0,3413 | 0,3438 | 0,3461 | 0,3485 | 0,3508 | 0,3531 | 0,3554 | 0,3577 | 0,3599 | 0,3621 |
| 1,1 | 0,3643 | 0,3665 | 0,3686 | 0,3708 | 0,3729 | 0,3749 | 0,3770 | 0,3790 | 0,3810 | 0,3830 |
| 1,2 | 0,3849 | 0,3869 | 0,3888 | 0,3907 | 0,3925 | 0,3944 | 0,3962 | 0,3980 | 0,3997 | 0,4015 |
| 1,3 | 0,4032 | 0,4049 | 0,4066 | 0,4082 | 0,4099 | 0,4115 | 0,4131 | 0,4147 | 0,4162 | 0,4177 |
| 1,4 | 0,4192 | 0,4207 | 0,4222 | 0,4236 | 0,4251 | 0,4265 | 0,4279 | 0,4292 | 0,4306 | 0,4319 |
| 1,5 | 0,4332 | 0,4345 | 0,4357 | 0,4370 | 0,4382 | 0,4394 | 0,4406 | 0,4418 | 0,4429 | 0,4441 |
| 1,6 | 0,4452 | 0,4463 | 0,4474 | 0,4484 | 0,4495 | 0,4505 | 0,4515 | 0,4525 | 0,4535 | 0,4545 |
| 1,7 | 0,4554 | 0,4564 | 0,4573 | 0,4582 | 0,4591 | 0,4599 | 0,4608 | 0,4616 | 0,4625 | 0,4633 |
| 1,8 | 0,4641 | 0,4649 | 0,4656 | 0,4664 | 0,4671 | 0,4678 | 0,4686 | 0,4693 | 0,4699 | 0,4706 |
| 1,9 | 0,4713 | 0,4719 | 0,4726 | 0,4732 | 0,4738 | 0,4744 | 0,4750 | 0,4756 | 0,4761 | 0,4767 |
| 2,0 | 0,4772 | 0,4778 | 0,4783 | 0,4788 | 0,4793 | 0,4798 | 0,4803 | 0,4808 | 0,4812 | 0,4817 |
| 2,1 | 0,4821 | 0,4826 | 0,4830 | 0,4834 | 0,4838 | 0,4842 | 0,4846 | 0,4850 | 0,4854 | 0,4857 |
| 2,2 | 0,4861 | 0,4864 | 0,4868 | 0,4871 | 0,4875 | 0,4878 | 0,4881 | 0,4884 | 0,4887 | 0,4890 |
| 2,3 | 0,4893 | 0,4896 | 0,4898 | 0,4901 | 0,4904 | 0,4906 | 0,4909 | 0,4911 | 0,4913 | 0,4916 |
| 2,4 | 0,4918 | 0,4920 | 0,4022 | 0,4925 | 0,4927 | 0,4929 | 0,4931 | 0,4932 | 0,4934 | 0,4936 |
| 2,5 | 0,4938 | 0,4940 | 0,4941 | 0,4943 | 0,4945 | 0,4946 | 0,4948 | 0,4949 | 0,4951 | 0,4952 |
| 2,6 | 0,4953 | 0,4955 | 0,4956 | 0,4957 | 0,4959 | 0,4960 | 0,4961 | 0,4962 | 0,4963 | 0,4964 |
| 2,7 | 0,4965 | 0,4966 | 0,4967 | 0,4968 | 0,4969 | 0,4970 | 0,4971 | 0,4972 | 0,4973 | 0,4974 |
| 2,8 | 0,4974 | 0,4975 | 0,4976 | 0,4977 | 0,4977 | 0,4978 | 0,4979 | 0,4979 | 0,4980 | 0,4981 |
| 2,9 | 0,4981 | 0,4982 | 0,4982 | 0,4983 | 0,4984 | 0,4984 | 0,4985 | 0,4985 | 0,4986 | 0,4986 |
| 3,0 | 0,4987 | 0,4987 | 0,4987 | 0,4988 | 0,4988 | 0,4989 | 0,4989 | 0,4989 | 0,4990 | 0,4990 |

# CAMPIONAMENTO E STIMA




- Esempio. Calcolo di n con Excel
- Excel fornisce la funzione DISTRIB.NORM.ST(z) che riporta la Funzione di ripartizione F(z) della Normale standard.
- Verifichiamo con questa funzione i dati della tavola usata finora:

| z | F(z) | F(z) - F(0) |
|---|---|---|
| 2,57 | 0,994915 | 0,494915 |
| 2,58 | 0,995060 | -0,005085 |

- Excel fornisce inoltre la funzione INV.NORM.ST(area), che fornisce il punto z corrispondente ad una F(z) = area specificata
- Con questa funzione è facile determinare il punto z(a/2) = z(0,005) = -2,5758 e poi calcolare la dimensione ottimale del campione:

$$n = \frac{z_{\alpha/2}^2 \, \sigma^2}{d^2}$$

| $\sigma^2$ | 1-a | d | z(a/2) | $z(a/2)^2$ | => n |
|---|---|---|---|---|---|
| 0,25 | 0,95 | 0,02 | -1,9600 | 3,8415 | 2401 |
| 0,25 | 0,99 | 0,02 | -2,5758 | 6,6349 | 4147 |
| 0,25 | 0,95 | 0,01 | -1,9600 | 3,8415 | 9604 |