# METODI STATISTICI DI BASE AVANZATI per le scienze sociali – AGRESTI, FINLAY

Psicometria

Università degli studi di Trieste

65 pag.

**METODI STATISTICI DI BASE AVANZATI per le scienze sociali – AGRESTI, FINLAY**

CAPITOLO 1: **INTRODUZIONE**

Ruolo fondamentale della statistica nella ricerca sociale: raccolta di info ed elaborazione attraverso metodi statistici.

La raccolta di info è il CUORE della scienza, attraverso essa vengono ottenute le OSSERVAZIONE impiegate per l'analisi statistica.

- DATI = OSSERVAZIONI raccolte su determinate caratteristiche di interesse

Esaminare i dati in modo opportuno è utile per comprendere come alcune caratteristiche sono tra loro in relazione. Per ottenere i dati si utilizzano diversi metodi tra cui INDAGINI, ESPERIMENTI, OSSERVAZIONI DAI COMPORTAMENTI. Vengono analizzati spesso anche dati che sono stati raccolti per altre finalità

Le raccolte di dati contenute in archivi sono detti DATABASE (numerosi disponibili in Internet)

Es.: General Social Survey (GSS): l'indagine che coinvolge circa 200 soggetti adulti che permette di ottenere info sulle opinioni e sui comportamenti della popolazione statunitense

STATISTICA = l'insieme delle metodologie finalizzate alla raccolta e all'analisi dei dati. È l'insieme dei dati per:

- PROGETTARE: pianificazione di COME dovranno essere OTTENUTI i dati (cioè la modalità di selezione degli individui e strutturazione del questionario)
- DESCRIVERE: sono i modi in cui dovranno essere sintetizzati i dati per meglio comprendere le info che sono in essi contenute (grafici, tabelle e sintesi numeriche sono definite come STATISTICHE DEFINITIVE)
- INFERIRE: formulare previsioni basate sui dati raccolti

  INFERENZE STATISTICHE = previsioni fatte utilizzando i dati raccolti

  È importante l'ACCURATEZZA delle statistiche campionarie che stimano i parametri della popolazione

Quelli DESCRITTIVI e quelli INFERENZIALI sono 2 metodi per l'analisi statistica dei dati.

La STATISTICA è un complesso di metodi per PROGETTARE RICERCE e ANALIZZARE i dati raccolti per le stesse.

1. METODI DESCRITTIVI: strumenti utilizzati per le sintesi – analisi statistica descrittiva
2. METODI INFERENZIALI: strumenti utilizzati per formulare previsioni – analisi statistica inferenziale

SOGGETTI di studio = le entità che vengono osservate in una ricerca (solitamente persone fisiche)

POPOLAZIONE = è costituita dal TOTALE DEI SOGGETTI di interesse di uno studio

CAMPIONE = sottoinsieme della popolazione di riferimento dello studio

L'*obiettivo principale* di qualunque ricerca è quello di acquisire conoscenze riferibili alle popolazioni. Spesso è necessario osservare campioni piuttosto che popolazioni.

**STATISTICA DESCRITTIVA**: viene impiegata per sintetizzare le info raccolte in un'indagine

• strumenti: grafici, tabelle e numeri come medie e percentuali

• obiettivo: sintetizzare i dati in formati semplici e facilmente leggibili (minima perdita di info)

• sono utili quando i dati sono stati raccolti per l'intera popolazione (es. censimenti)

**STATISTICA INFERENZIALE**: viene impiegata per ottenere previsioni su una popolazione sulla base di info raccolte su un campione selezionato da essa

- impiegata quando sono disponibili SOLO dati campionari e si vogliono fare previsioni per l'intera popolazione

- *ci serve per formulare previsioni riferite all'intera popolazione facendo uso dei dati campionari*

- sono in grado di prevedere valori caratteristici di grandi popolazioni attraverso analisi condotte su campioni di dimensioni relativamente ridotte

PARAMETRO = è una sintesi numerica delle caratteristiche della popolazione

STATISTICA = è una sintesi numerica dei dati campionari

Nelle applicazioni, il principale obiettivo è quello di conoscere il VALORE ASSUNTO DAI PARAMETRI (non il valore che assumono le statistiche in determinati campioni selezionati)

La popolazione a cui si fa riferimento nell'applicazione dei metodi inferenziali è un insieme reale di soggetti, tuttavia le generazioni spesso vengono riferite a una popolazione CONCETTUALE (cioè ipotetica)

FILE DI DATI: dati che vengono organizzati per poi essere analizzati da un software statistico

SOFTWARE STATISTICO: ha la forma di un foglio elettronico

• riga: contiene info riferite a un soggetto del campione

• colonna: osservazioni raccolte nel campione per ognuna delle caratteristiche esaminate

## CAPITOLO 2: **CAMPIONAMENTO E MISURAZIONE**

Metodi descrittivi sintetizzano i dati, mentre quelli inferenziali formulano previsioni sulle popolazioni in base ai dati raccolti.

AFFIDABILITA: una misura è affidabile se un soggetto darà un'identica risposta a una domanda che gli è già stata posta in precedenza. -> al contrario produrranno statistiche prive di significato

### 2.1 LE VARIABILI E LA LORO MISURAZIONE

L'impiego di metodi statistici ci aiuta a determinare i fattori che influiscono sulla VARIABILITÀ osservata tra i soggetti dello studio.

VARIABILI

**VARIABILE**: è una caratteristica che assume diversi valori tra i soggetti di un campione o di una popolazione. - Differenti soggetti possono avere differenti valori di una variabile (es: reddito, num fratelli)

- I valori della variabile formano una SCALA DI MISURA

La validità di un det metodo statistico è legata alla scala di misurazione del carattere a cui lo stesso è applicato.

VARIABILI QUANTITATIVE E VARIABILI CATEGORIALI

**VARIABILE QUANTITATIVA**: quando assume valori che sono numeri.

- valori assumono diversi ordini di grandezza assunti dalla variabile ( es: età, reddito annuo di un soggetto)

**VARIABILE CATEGORIALE**: quando i valori da essa assunti sono insiemi di categorie ( es: stato civile, musica preferita…)

- le diverse categorie sono da considerarsi differenziate in termini qualitativi e non di grandezza numerica: spesso vengono chiamate QUALITATIVE

SI DISTINGUONO perché vengono usati differenti metodi statistici per la loro sintesi

VARIABILI NOMINALI, ORDINALI E SCALE DI INTERVALLI

**SCALA DI INTERVALLI**: viene formata dai possibili valori numerici assunti da una **variabile quantitativa**. - le scale di intervalli sono caratterizzate da specifiche

distanze numeriche (intervalli) fra ciascuna coppia formata con i diversi valori assunti dalla variabile: possiamo fare confronti fra osservazioni considerando QUANTO una di esse è più grande o più piccola dell'altra

- presenta un'origine NON fissa MA convenzionale

Le variabili Categoriali formano due tipi di scale:

Variabili categoriali

**SCALA NOMINALE**: le scale non hanno un estremità alta o bassa e non rispettano alcun ordinamento (ne sono prive)

- spesso le diverse categorie sono chiamate LIVELLI DELLA SCALA, nel caso delle variabili NOMINALI non possiamo dire che un certo liv è più alto di un altro

- ogni valore di una variabile QUANTITATIVA è più grande o piccolo di ogni altro possibile valore. I vari livelli sono rappresentati da espressioni verbali e denotano variazioni di tipo QUALITATIVO.

**SCALA ORDINALE**: è una via di mezzo tra le prime due scale

- ha un ordinamento naturale dei suoi valori (es: la classe sociale, ideologia politica..)

- non sono NOMINALI: perché esiste un ordinamento fra le categorie

Non sono DI INTERVALLI: perché non è definibile una distanza tra i diversi livelli della scala

- possiedono un importante caratteristica delle scale quantitative: ciascun livello (categoria) è una grandezza che è minore o maggiore di quella di un altro livello

DIFFERENZA TRA NOMINALE E ORDINALE:

- ordinali: categorie hanno ordinamento naturale

- nominale: categorie sono prive di ordinamento

ASPETTI QUALITATIVI DEI DATI ORDINALI

- livelli della SCALA NOMINALE sono rappresentati da ESPRESSIONI VERBALI e denotano variazioni in termini QUALITATIVI e NON quantitativi

- livelli delle SCALE DI INTERVALLI sono QUANTITÀ e denotano variazioni in grandezza

Classificazione delle SCALE ORDINALI nella classificazione quantitativo-qualitativo è più sfumata:

- sono spesso analizzate con metodi simili a quelli adottate per le nominali, nonostante siano più simili alle intervalli

- le variabili ORDINALI infatti possiedono un'importante caratteristica delle scale quantitative: CIASCUN LIVELLO è una grandezza che è minore o maggiore di quella di un altro livello

- spesso è utile analizzare le scale ordinali assegnando valori numerici alle diverse categorie

VARIABILI DISCRETE E VARIABILI CONTINUE

**VARIABILE DISCRETA**: se i suoi valori formano un insieme di numeri distinti come 0,1,2,3.. (es: num di fratelli)

- è possibile elencare tutti i suoi possibili valori

- esse sono espresse con una unità di misura di base che non è divisibile (es. 2 e 3 è un valore possibile per un numero di fratelli; 2.5716 non lo è)

**VARIABILE CONTINUA**: se può assumere come valori ogni possibile numero reale incluso in un continuum infinito ( es: altezza, tempo trascorso x leggere un libro, l'età)

- risulta impossibile elencare tutti i possibili valori che una variabile assume in quanto costituiscono un intervallo infinito di numeri reali

- infinito insieme di valori compresi tra due estremi di un intervallo (es. età è continua perché non assume valori con cadenze discrete: in un certo istante compreso tra l'età 21 e 22 un individuo assume l'età di 21.3581 anni e così via per tutti i valori che è possibile immaginare nell'intervallo)

Qualunque variabile con un numero finito di possibili valori è discreta -> TUTTE le variabili categoriali, nominali o ordinali sono DISCRETE essendo costituite da un INSIEME FINITO di categorie

- le variabili qualitative possono essere o discrete o continue (età=variabile continua; numero di fratelli=discreta)

In sintesi:

Variabili sono di tipo QUANTITATIVO o CATEGORIALE:

- variabili categoriali con categorie NON ORDINATE sono NOMINALI

- variabili categoriale con categorie ORDINATE sono ORDINALI

Variabili CATEGORIALI (ordinali o nominali) sono DISCRETE

- le variabili QUANTITATIVE possono essere discrete o continue, MA nella pratica le variabili DISCRETE che assumono molti valori sono trattate come CONTINUE

VEDI SCHEMA PAG 15

## 2.2 CASUALIZZAZIONE

I metodi statistici inferenziali fanno uso delle statistiche campionarie per fare previsioni sui parametri delle popolazioni.

CASUALIZZAZIONE (randomization) = metodo di campionamento

- serve per avere una BUONA RAPPRESENTATIVITÀ del campione

CAMPIONE CASUALE SEMPLICE

Un CAMPIONE è un **CAMPIONE CASUALE SEMPLICE** (CCS) quando è ottenuto attraverso un metodo in grado di assicurare a ogni possibile campione che può essere selezionato dalla popolazione abbia una identica probabilità di selezione.

"n" = numero dei soggetti del campione

- tale grandezza è detta DIMENSIONE CAMPIONARIA

**Campione casuale semplice:** campione casuale semplice di n soggetti estratti da una popolazione è tale SE a ogni possibile campione di pari numerosità che poteva essere estratto ha un'uguale probabilità di selezione

- campione casuale semplice è spesso chiamato campione casuale: l'aggettivo semplice è impiegato per distinguerlo da schemi di campionamento casuale più complessi

Perché è importante selezionare campioni casuali?

Per l'**equiprobabilità** di selezione di ciascuno dei possibili campioni

- così viene ridotta la probabilità di selezionare campioni che per le loro caratteristiche distorcono in maniera importante la loro rappresentatività della popolazione e QUINDI portano a errate conclusioni inferenziali sui valori dei parametri

La maggior parte dei metodi dell'inferenza statistica assume la selezione di campioni basati sulla casualizzazione.

COME SELEZIONARE UN CAMPIONE CASUALE SEMPLICE

Per selezionare un campione casuale è necessario disporre di una **LISTA DI CAMPIONAMENTO** (SAMPLING FRAME) che elenchi tutti i soggetti della popolazione.

METODO PER LA SELEZIONE di un campione casuale semplice:

1) Assegnare un numero a ciascuno dei soggetti della lista di campionamento
2) Ottenere in modo casuale un insieme di questi numeri
3) Includere nel campione tutti i soggetti della popolazione i cui numeri sono stati ottenuti in (2)

➔ l'impiego di numeri casuali o *random numbers* = garantisce che ogni sogg della popolazione abbia uguale probabilità di essere selezionato

- numeri casuali: sono numeri generati attraverso un calcolatore elettronico secondo uno schema COMPUTAZIONALE in grado di assicurare che ciascuna delle cifre 0,1,2…9 abbia la STESSA PROBABILITÀ di essere generate e che NON dipenda dalle altre cifre generate precedentemente a essa

LA RACCOLTA DI DATI NELLE INDAGINI CAMPIONARIE

**INDAGINE CAMPIONARIA**: negli studi viene selezionato e intervistato un campione di individui da una popolazione

- intervista faccia a faccia, telefonica o autocompilata dall'intervistato

Esiste una pluralità di cause in grado di influire, distorcerle, sulle risposte date dai soggetti campionati ai quesiti sottoposti col questionario: se ciò si verifica, i risultati NON sono generalizzabili all'intera popolazione di riferimento dell'indagine

LA RACCOLTA DI DATI NELLE INDAGINI CAMPIONARIE

In alcuni studi i dati vengono raccolti attraverso un **ESPERIMENTO PIANIFICATO**

- obiettivo: confrontare le risposte date da alcuni soggetti a domande poste in differenti condizioni (cioè i valori assunti da una variabile che si ritiene possa influire sulle stesse risposte)

- condizioni: sono dette "**TRATTAMENTI**"

il loro processo di assegnazione: è detto **PIANO SPERIMENTALE**

Un corretto piano sperimentale prevede la casualizzazione per l'assegnazione dei soggetti dell'indagine ai diversi trattamenti. (esempio aspirina)

LA RACCOLTA DI DATI ATTRAVERSO UNO STUDIO OSSERVAZIONALE

**STUDI OSSERVAZIONALI:** studi in cui si OSSERVANO le manifestazioni delle variabili in esame sui diversi soggetti senza alcuna manipolazione sperimentale degli stessi

- il ricercatore registra i valori osservati delle variabili di interesse sui soggetti ma non ha possibilità di controllo sperimentale sugli stessi

In questi studi è difficile il confronto tra i gruppi perché gli stessi gruppi possono risultare non bilanciati rispetto al possesso di caratteristiche di interesse (vero anche quando campioni casuali).

- determinazione di rapporti causa-effetto è cruciale nell'ambito scientifico, MA non è possibile determinare in maniera definita i rapporti di causa effetto in studi di tipo non sperimentale anche se gli stessi vengono realizzati con un campione casuale

Che uno studio sia sperimentale o meno, è SEMPRE **fondamentale inserire** nello studio dei **meccanismi di casualizzazione**, soprattutto se l'obbiettivo è quello di fare INFERENZA.

## 2.3 VARIABILITÀ CAMPIONARIA E POTENZIALI FATTORI DI DISTRIBUZIONE

I risultati in uno studio dipendono SEMPRE E COMUNQUE dal campione che è stato selezionato.

- le unità che entreranno a far parte del campione selezionato da un ricercatore non saranno selezionate anche dall'altro

ERRORE CAMPIONARIO

**ERRORE CAMPIONARIO**: è l'errore che viene commesso quando viene impiegata una statistica campionaria per prevedere il valore di un parametro della popolazione

- cioè quanto si allontanano dal valore del parametro che esse stesse prevedono

Questo allontanamento è mostrato dal modo in cui i valori delle stesse statistiche variano da campione a campione.

Nella pratica l'errore campionario è sempre sconosciuto, perché NON SI CONOSCE il valore del parametro nella popolazione.

Il PRINCIPIO DI CAMPIONAMENTO CASUALE è un assicurazione rispetto alla possibilità di distorsioni nei risultati inferenziali: l'errore campionaria oscilla intorno allo 0, assumendo, di volta in volta, valori positivi e valori negativi

- il campionamento casuale ci permette di prevedere la probabile ampiezza dell'errore di campionario

**MARGINE DI ERRORE** = per un campione di ampiezza circa pari a 1000 unità, l'errore campionario che si commette per stimare percentuali è di circa più o meno 3%; questo intervallo (+ e poi sotto c'è il -) si chiama margine di errore

- anche quando si selezionano dei campioni non casuali c'è una variabilità nei risultati calcolati delle statistiche, MA in questo caso non è possibile determinare l'ampiezza dell'errore campionario

DISTORSIONE CAMPIONARIA: CAMPIONAMENTO NON-PROBABILISTICO

Altri fattori possono contribuire a determinare variabilità nei valori assunti dalle statistiche nei differenti campioni. Tali fattori a loro volta possono determinare DISTORSIONI.

Vedremo 3 tipi di distorsioni, la prima è la distorsione campionaria

**CAMPIONAMENTO PROBABILISTICO**: Per campioni casuali semplici, ogni possibile campione di ampiezza n che può essere selezionato dalla popolazione ha la stessa probabilità di essere estratto

- la probabilità di selezione di un particolare campione è nota ancor prima che venga fatta la selezione

Distorsione campionari

**CAMPIONAMENTO NON-PROBABILISTICO**: non è possibile determinare la probabilità di selezione dei possibili campioni

- l'inferenza fatta utilizzando questo tipo di campioni non è affidabile IN QUANTO i risultati a cui si pervieni subiscono una DISTORSIONE CAMPIONARIA

Il metodo più diffuso metodo non probabilistico è quello del **CAMPIONAMENTO VOLONTARIO:** i soggetti decidono volontariamente di essere inclusi nel campione

- il campione può rappresentare in maniera inappropriata la popolazione e può condurre a conclusioni ERRATE

Esempi di campioni volontari si ricavano quotidianamente su molti siti internet (opinioni in merito a un qualche cosa con un voto): I rispondenti non rappresentano l'intera popolazione!!!

- Le ampiezze campionarie elevate non servono a molto se abbiamo a che fare con il campionamento volontario: le distorsioni continuano a sussistere

- ricorso al campionamento volontario è spesso necessario e ciò accade soprattutto in ambito medico

Anche nel caso di **campionamenti casuali** si possono avere distorsioni campionarie:

QUANDO le liste da cui selezionare il campione non coprono per intero la popolazione

- nelle liste non vengono elencati tutti i soggetti campionabili facendo si, in tal modo, che non tutti i gruppi presenti nella popolazione siano rappresentati dal campione

DISTORSIONI DOVUTE ALLE RISPOSTE

I quesiti mal formulati producono le **DISTORSIONI DOVUTE ALLE RISPOSTE**. Anche l'ordine in cui vengono formulate le domande id un questionario è in grado di influire in maniera sensibile sui risultati di un'indagine.

- in un'intervista anche le caratteristiche dell'intervistatore possono provocare distorsioni nelle risposte

- gli intervistati potrebbero mentire se fossero convinti che le loro opinioni siano socialmente inaccettabili -> vengono prodotte risposte che più facilmente sono gradite all'intervistatore

DISTORSIONI DOVUTE ALLE NON-RISPOSTE: I DATI MANCANTI

Alcuni soggetti inclusi in un campione possono rifiutarsi di partecipar all'indagine oppure può risultare impossibile contattarli

- questo determina **la DISTORSIONE DOVUTA ALLE NON-RISPOSTE** (i risultati di un indagine devono essere valutati con cautela se si hanno percentuali di non-risposte superiori al 20%)

Il problema dei **dati mancanti** è rilevante in tutti gli studi che coinvolgono grandi numerosità campionarie.

- la maggior parte dei software statistici non considera nell'analisi i record nei quali si registrano dati mancanti per almeno una delle variabili prese in esame dalla studio. Questo comporta una notevole perdita di informazioni e può indurre distorsioni nei risultati.

SINTESI SUI VARI TIPI DI DISTORSIONI

Le indagini CAMPIONARIE sono soggette a diversi tipi di DISTORSIONI:

- **DISTORSIONI CAMPIONARIE:** si hanno quando vengono impiegati campioni non probabilistici o campioni sotto-rappresentativi della popolazione
- **DISTORSIONI DOVUTE A RISPOSTE**: si hanno quando o i soggetti campionati danno risposte non corrette ai quesiti che vengono loro presentati o, quando, le domande sono formulate in modo confuso o sono mal riportate dall'intervistatore
- **DISTORSIONI DOVUTE A NON-RISPOSTE**: si hanno quando alcuni dei soggetti campionati non vengono intervistati oppure quando gli stessi soggetti rifiutano di partecipare o di rispondere alle domande del questionario di rilevazione

Occorre tenere presente che ogni studio in cui viene fatta inferenza su una vasta popolazione è tanto più affidabile quanto meglio è stato selezionato il campione.

## 2.4 ALTRI METODI DI CAMPIONAMENTO PROBABILISTICI

È stato presentato il campione casuale semplice ed è stato spiegato perché è così importante per l'inferenza statistica. Nella pratica, altri modelli probabilistici basati su sistemi di casualizzazione sono preferibili al campione casuale semplice e sono anche più semplici da ottenere

CAMPIONE SISTEMATICO

**CAMPIONE**: costituito selezionando un determinato soggetto in prossimità dell'inizio della lista di campionamento e dopo aver saltato un certo numero di soggetti ne viene selezionato un altro da includere nel campione e così via

- PASSO DI ESTRAZIONE: il numero di soggetti che viene saltato

**CAMPIONE SISTEMATICO**: indichiamo l'ampiezza campionaria con "n" e la dimensione della popolazione con N

- con K=N/n indichiamo la dimensione della popolazione divisa per l'ampiezza campionaria

Nel campionamento sistematico:

1. viene selezionato un soggetto tra i primi K contenuti nella lista di campionamento

2. viene selezionato ogni soggetto che si trova nella K-esima posizione della lista dopo il primo soggetto selezionato

- il numero K è detto PASSO DI ESTRAZIONE

Quando viene fatto un campionamento a partire da una lista, la procedura di selezione del campione sistematico è molto più semplice di quella del campione casuale semplice

- il campionamento sistematico ci permette di ottenere una buona rappresentatività della popolazione in quanto, per una lista di campionamento formata da nomi in ordine alfabetico, i valori della maggior parte delle variabili osservate variano casualmente

- il campione sistematico è diverso da un campione casuale semplice perché non tutti i campioni di ampiezza n hanno la stessa possibilità di essere selezionati

Nel campionamento sistematico 2 soggetti che sono contigui nella lista di campionamento non potranno mai fare parte entrambi del campione

CAMPIONE STRATIFICATO

Un tipo di campione probabilistico usato nella ricerca di ambito sociale e negli studi in cui si confrontano gruppi

**CAMPIONE STRATIFICATO**: (campione probabilistico) per la selezione di un campione stratificato la popolazione viene suddivisa in gruppi distinti detti STRATI

- successivamente, da ogni strato viene selezionato un campiona casuale semplice

Esso prende in considerazione per l'analisi ogni strato

***CAMPIONE STRATIFICATO PROPORZIONALE***: se la proporzione di osservazioni contenute in ciascuno strato del campione è uguale alla proporzione di popolazione appartenente al corrispondente gruppo.

***CAMPIONE STRATIFICATO NON-PROPORZIONALE***: se le proporzioni campionarie dei vari strati non sono uguali a quelle osservate nella popolazione

Ciò è talvolta utile quando è relativamente ridotta la dimensione di un particolare strato nella popolazione

Per realizzare la stratificazione è necessario conoscere a quale strato appartiene ciascun soggetto della lista di campionamento.

Possiamo stratificare solo rispetto a variabili che sono facilmente identificabili.

CAMPIONE A GRAPPOLI

I campioni casuali semplici, sistematici e stratificati sono difficili da selezionare perché richiedono la disponibilità di una lista di campionamento completa.

QUANDO NON si dispone di liste complete delle popolazioni di interesse, è utile ottenere un campione a grappoli

**CAMPIONE A GRAPPOLI**: per selezionare un campione a grappoli la popolazione deve essere suddivisa in un gran numero di grappoli

- si seleziona poi un campione casuale di questi grappoli e si impiegano nello studio tutti i soggetti contenuti nei grappoli

Esso prende in considerazione per l'analisi un campione di grappoli piuttosto che la loro totalità

Differenza tra CAMPIONE STRATIFICATO e un CAMPIONE A GRAPPOLI:

CAMPIONE STRATIFICATO: prende in considerazione per l'analisi OGNI STRATO

- strati= sono i differenti gruppi che vogliamo confrontare rispetto ad alcune loro caratteristiche

CAMPIONE A GRAPPOLI: prende in considerazione un CAMPIONE di grappoli piuttosto che la loro totalità

- grappoli = sono il mezzo che permette di identificare più agevolmente gruppi di soggetti

- SCOPO: NON confrontare tra loro i grappoli, MA usare i grappoli per ottenere un campione

Esso necessita di ampiezze maggiori per raggiungere il livello di precisione delle stime cui si previene con il campione casuale semplice.

Le osservazioni all'interno dei grappoli tenderanno a essere simili-> vi è una tendenza all'aggregazione fra soggetti che hanno condizioni socio-economiche e demografiche analoghe e condividono opinioni comuni.

Per questo motivo per rappresentare bene la variabilità di una popolazione è necessario disporre di un elevato numero di osservazioni.

CAMPIONE A PIÙ STADI

**CAMPIONE A PIÙ STADI**: è la combinazione degli altri metodi di campionamento che abbiamo visto fino ad ora; ottenuto attraverso la combinazione del campione stratificato e a grappoli

- sono molto utilizzati nella ricerca sociale

- sono più facili da realizzare del campione casuale semplice e allo stesso tempo permettono di ottenere un campionamento della popolazione più esteso di quanto possa essere ottenuto impiegando un solo metodo quale il campionamento a grappoli

Il campione a grappoli, necessita di ampiezze campionarie maggiori per raggiungere il livello di precisione delle stime cui si perviene con il campione casuale semplice

- le osservazioni all'interno dei grappoli tenderanno ad essere simili: c'è infatti una tendenza all'aggregazione fra soggetti che hanno condizioni ad es socio economiche e demografiche analoghe e condividono opinioni comuni

I soggetti appartenenti ad uno stesso grappolo, assumono valori simili in variabili di fondamentale importanza per la ricerca sociale (età, reddito, razza)

- per rappresentare adeguatamente la variabilità di una popolazione è necessario disporre di un elevato numero di osservazioni

- di contro, i risultati inferenziali che si ottengono attraverso campioni stratificati possono essere ben più precisi di quelli presentati in questo libro

CAPITOLO 3: **STATISTICHE DESCRITTIVE**

**3.1 DESCRIVERE I DATI CON TABELLE E GRAFICI**
FREQUENZE RELATIVE: DATI CATEGORIALI
Per i dati categoriali, la rappresentazione tabellare prevede un elenco delle categorie della variabile con accanto a ciascuna di esse la FREQUENZA
- FREQUENZA ASSOLUTA: numero di osservazioni che hanno presentato quel valore della variabile nel collettivo
PROPORZIONI = frequenze relative

PERCENTUALI = frequenze percentuali

**FREQUENZA RELATIVA**: è la proporzione o percentuale di osservazioni che ricade in una determinata categoria

PROPORZIONE: è pari al numero di osservazioni in una categoria diviso il numero totale di osservazioni

- è compresa tra 0 e 1

- la PERCENTUALE è la proporzione moltiplicata per 100

VEDI ESEMPIO 3.1 PAG 34

DISTRIBUZIONI DI FREQUENZE E GRAFICI A BARRE: DATI CATEGORIALI

In una tabella è necessario riportare solo le percentuali e il numero totale di osservazioni prese in esame.

**DISTRIBUZIONE DI FREQUENZA**: è una lista di tutti i possibili valori di una variabile a ciascuno dei quali è associato un numero che rappresenta quante volte quel valore viene osservato nei dati in esame

- Una DISTRIBUZIONE DI FREQUENZA RELATIVA elenca i possibili valori della variabile con le loro rispettive proporzioni o percentuali

Per costruire una DISTRIBUZIONE DI FREQUENZE ASSOLUTE per una variabile categoriale, si crea una lista delle categorie della variabile e a ciascuna categoria viene affiancato il corrispondente conteggio. Per ottenere una più completa percezione delle info QUANTITATIVE di una distribuzione, è utile il grafico:

- **GRAFICO A BARRE** è costruito rappresentando con barre rettangolari ciascuna categoria; l'altezza della barra rappresenta la frequenza relativa.

- **GRAFICO A TORTA** è costruito utilizzando un cerchio suddiviso in vari spicchi, ciascuno dei quali rappresenta una categoria

DISTRIBUZIONI DI FREQUENZA: DATI QUANTITATIVI

VEDI ESEMPIO 3.2 PAG 35

Gli intervalli di valori nelle distribuzioni di frequenze spesso sono di uguale ampiezza. Essi devono includere tutti i possibili valori della variabile e ogni possibile valore osservato deve essere incluso in un singolo intervallo

- gli intervalli devono essere MUTUALMENTE ESCLUSIVI

ISTOGRAMMI

**ISTOGRAMMA**: è la rappresentazione grafica di una distribuzione di frequenze per una variabile quantitativa

- Ogni intervallo è rappresentato da una BARRA, la cui area rappresenta il numero di osservazioni nell'intervallo.

La scelta dell'ampiezza degli intervalli x distribuzioni di frequenze e istogrammi è basata sul buon senso.

SE impiegati pochi intervalli = perdita di info

SE troppi intervalli= difficoltà di lettura, irregolarità dell'istogramma

Principio a cui uniformarsi per la determinazione dell'ampiezza degli intervalli è basato sulla considerazione che:

- due osservazioni incluse in uno stesso intervallo devono essere pressoché uguali dal punto di vista pratico

Per le distribuzioni **discrete** che hanno un numero limitato di valori: l'istogramma riporta una barra per ciascun valore della distribuzione

Per le distribuzioni **continue** o **discrete** con un elevato numero di valori, è necessario dividere i possibili valori in intervalli

DIAGRAMMA RAMO E FOGLIE

**DIAGRAMMA RAMO E FOGLIE**: ogni valore registrato è rappresentato da una o più "principali" e da una cifra "finale"

- RAMO: formato dalle cifre principali (a sinistra)

- FOGLIA: quelle finali (a destra)

Le cifre che rappresentano le FOGLIE sono ordinate in maniera CRESCENTE

- se rovesciato su un lato, assume la forma di un istogramma

- fornisce un'immediata schematizzazione dei dati

CONFRONTI TRA GRUPPI

Confronti tra gruppi rispetto a valori posseduti da un'altra variabile sono molto utilizzati

I diagrammi ramo e foglie utili per confrontare visivamente i valori assunti da una variabile quantitativa in due collettivi di non elevata dimensione.

I due diagrammi vengono rappresentati a "schiena contro schiena".

DISTRIBUZIONE DI DATI DI POPOLAZIONE e DISTRIBUZIONI DI DATI CAMPIONARI

La costruzione di distribuzioni di frequenze o di istogrammi può essere fatta sia per i dati di popolazione, sia per quelli campionari

1. DISTRIBUZIONI DI DATI DI POPOLAZIONE

2. DISTRIBUZIONI DI DATI CAMPIONARI: è una "foto sfuocata" della distribuzione dei dati di popolazione

- mano a mano che l'ampiezza campionaria aumenta, le proporzioni campionarie in ciascun intervallo si approssimano sempre di più alle vere proporzioni della popolazione: la distribuzione dei dati campionari diventa sempre più simile alla distribuzione di popolazione

La forma dell'istogramma con cui rappresentiamo la distribuzione, diventerò sempre più simile a una curva liscia

LA FORMA DELLA DISTRIBUZIONE

Per sintetizzare le info di una distribuzione per un campione o una popolazione è efficace **descrivere la FORMA**.

**Distribuzione a U:** i punti più elevati (frequenze elevate) sono riferiti ai valori più bassi e più alti assunti dalle osservazioni della distribuzione

- indica POLARIZZAZIONE DELLA VARIABILE RISPETTO A DUE INSIEMI DI SOGGETTI

**Distribuzione CAMPANULARE**: la frequenza più elevata si registra per i valori centrali

- indica che MAGGIOR PARTE DEI SOGGETTI TENDE AD ASSUMERE VALORI PROSSIMI AL VALORE CENTRALE

La maggior parte delle distribuzioni vengono osservate nell'ambito della ricerca sociale NON sono SIMMETRICHE

- distribuzione viene definita

**ASSIMETRICA POSITIVA**: se la coda più lunga è quella destra

O

**ASSIMETRICA NEGATIVA**: se la coda più lunga è quella a sinistra

a seconda di quella che è la sua CODA più lunga

- CODA: zona in corrispondenza dei valori alti e bassi

### 3.2 DESCRIVERE IL "CENTRO" DEI DATI

**MISURE DI TENDENZA CENTRALE**: statistiche che misurano la TIPICITA della distribuzione e descrivono il centro di una distribuzione di frequenza definita per una variabile quantitativa.

LA MEDIA

- MEDIA (mean, average): è la somma dei valori assunti dalle osservazioni divisa per il totale delle osservazioni

VEDI ESEMPIO 3.4 PAG 41

NOTAZIONE PER LA NUMEROSITÀ DELLE OSSERVAZIONI E PER LA MEDIA CAMPIONARIA

n = **NUMEROSITÀ CAMPIONARIA**

y = generica variabile indicata

$y_1$(prima osservazione),$y_2$,$y_3$....$y_n$ = le n osservazioni della variabile y

y lineetta = media della variabile

- una barra sopra una lettera rappresenta la MEDIA CAMPIONARIA per la variabile.

Sigma = somma

(vedi formule pag 42)

***PROPRIETÀ DELLA MEDIA***:

1. Calcolo della media è appropriato solo per le variabili **quantitative** (valori numerici)

   - no variabili ordinali e nominali

2. Il valore assoluto della media può essere notevolmente influenzato da un'osservazione che assume un valore medio al di sotto o al di sopra di quello assunto dalla maggioranza delle restanti osservazioni

   - tale valore è detto OUTLIER (*valore anomalo*)

   - quando abbiamo a che fare con piccoli campioni dove è sufficiente la presenza di una sola osservazione molto più piccola o più grande delle altre per avere dei valori medi non effettivamente "tipici" della distribuzione

3. La media tende a spostarsi nella direzione della coda più lunga della distribuzione

   - più ASIMMETRICA è la distribuzione, MINORE è la tipicità dei dati

4. La media è il punto di equilibrio nella linea in cui sono riportati tutti i valori di una distribuzione

   - è il centro di gravità (baricentro) delle osservazioni -> la somma delle distanze dalla media delle osservazioni al di sopra delle medie è pari alla somma delle distanze della madia delle osservazioni al di sotto della media

5. MEDIA PONDERATA: $y = \frac{n_1 y_1 + n_2 y_2}{n_1 + n_2}$

   dove $y_1$ e $y_2$ sono le medie campionarie per due insiemi di dati con numerosità $n_1$ e $n_2$.

   Il numeratore è la somma di tutte le osservazioni in quanto $ny=\sum y$

LA MEDIANA

Media è una semplice misura del centro della distribuzione

MEDIANA: divide in due parti il campione, ORDINATO IN MANIERA CRESCENTE

- ciascuna parte contiene un identico numero di osservazioni

- è il valore della variabile assunto dall'osservazione centrale del campione ordinato

- quando l'ampiezza n è dispari, abbiamo una sola osservazione centrale

- quando l'ampiezza campionaria è pari, si hanno due osservazioni centrali: la mediana è il valore centrale tra i due.

Quando una distribuzione è **fortemente asimmetrica** la mediana descrive la sua tipicità molto meglio della media

***PROPRIETÀ DELLA MEDIANA***:

1. La mediana è un'appropriata sintesi per riassumere sia le **variabili quantitative sia le variabili ordinali**

   - le osservazioni devono essere ordinate rispetto al valore della variabile posseduto da ciascuna di esse. No x scale nominali.

2. Per le distribuzioni simmetriche la media e la mediana sono identiche
3. Per le distribuzioni asimmetriche la media è spostata, rispetto la mediana, nella direzione dell'asimmetria (cioè verso la coda più lunga)
4. La mediana non è sensibile alle distanze tra i valori della distribuzione e la loro media
5. Non è influenza dagli Outlier

CONFRONTO FRA MEDIA E MEDIANA

MEDIANA: è più appropriata della media nelle distribuzioni asimmetriche

- la mediana può essere determinata sia per scale QUANTITATIVE che per scale ORDINALI.

MEDIA: al contrario della mediana, è molto influenzata dagli Outlier.

- per il calcolo della media è necessario avere dati QUANTITATIVI

Per calcolare la media con i dati ordinali, è necessario assegnare dei numeri alle categoria
Anche l'impiego della mediana da degli svantaggi.
Per dati discreti che assumono un numero limitato di valori, possiamo avere distribuzioni molto diverse tra loro che però presentano la stessa mediana.
- la mediana è pari al valore che viene osservato con maggior frequenza MA non ci fornisce info sulla frequenza relativa che ricade in ciascuno dei due livelli della variabile

(0 e 1 <- dati binari che cioè possono assumere due valori nella variabile)

Una situazione estrema si determina quando vengono sintetizzati DATI BINARI

Quando le osservazioni assumono solo i valori 0 e 1, la **media** è pari alla proporzione di osservazioni che assumono valore 1.

Per la distribuzione BINARIA, la media è più informativa della mediana.

- SE distribuzione è FORTEMENTE ASIMMETRICA: MEDIANA rappresenta la tipicità dei dati

- SE distribuzione è QUASI SIMMETRICA o solo LEGGERMENTE ASUMMETRICA o DISCRETA con un numero molto limitato di valori: MEDIA è preferibile perché utilizza i valori numerici di tutte le osservazioni

MODA

MODA: è il valore assoluto nella distribuzione con maggior frequenza

- utilizzata quando si devono sintetizzare distribuzioni estremamente discrete come quelle che si hanno con i dati categoriali

***PROPRIETÀ DELLA MODA***:

1. è applicabile a dati misurati in qualunque scala
2. Una distribuzione di frequenza è chiamata BIMODALE se osserviamo due "picchi" nell'istogramma delle distribuzioni

   - vengono frequentemente osservate per le variabili che sono riferite a misure attitudinali
3. La media, la moda e la mediana sono identiche se la distribuzione è unimodale e simmetrica come la distribuzione campanulare

La media, la mediana e la moda sono misure complementari: ciascuna di esse è utile per descrivere differenti aspetti dei dati

- una misura numerica centrale fornisce info su un valore tipico della distribuzione ma non da info sulla dispersione delle osservazioni intorno a tale valore.

### 3.3 DESCRIVERE LA VARIABILITÀ DEI DATI

CAMPO DI VARIAZIONE (RANGE)

Per descrivere la variabilità di una VARIABILE QUANTITATIVA

CAMPO DI VARIAZIONE (range): è la differenza tra il valore più alto è quello più basso della distribuzione. Non è sensibile ad altre caratteristiche della variabile dei dati.

DEVIAZIONE STANDARD (STANDARD DEVIATION)

Basata sugli SCARTI dei dati da una misura di tendenza centrale come la media

## DEVIAZIONE

DEVIAZIONE di un'osservazione $y_i$ della media campionaria y lineetta: è la differenza tra i due valori ( $y_i$ - y)

Ogni osservazione ha una deviazione

- deviazione POSITIVA: quando l'osservazione ha un valore al di sopra della media

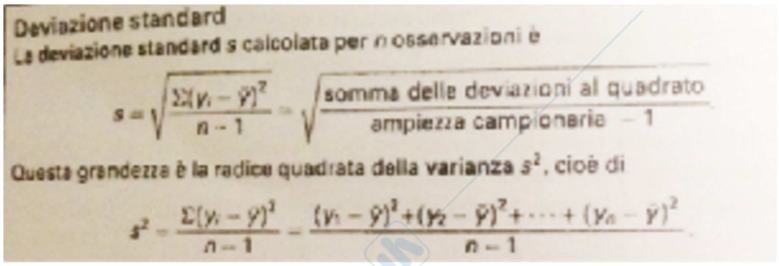
- NEGATIVA: quando ha un valore al di sotto

Interpretazione di y lineetta come centro di gravità dei dati, implica che: la somma delle deviazioni positive sia uguale alla somma delle deviazioni negative

-> la somma di tutte le deviazioni dalla media è 0

Per questo motivo le misure di variabilità fanno uso o dei valori assoluti delle deviazioni o dei loro quadrati: la più importante misura di variabilità fa uso dei quadrati

## DEVIAZIONE STANDARD




VARIANZA: è approssimativamente una media delle deviazioni al quadrato

- unità di misura: il quadrato dell'unita di misura con cui sono espressi i dati originali

Viene utilizzata la radice quadrata della varianza, cioè la deviazione standard

SOMMA DEI QUADRATI: è costituita dalla somma dei quadrati di ciascuna deviazione

- non bisogna sommare le deviazioni e poi elevare al quadrato il risultato ottenuto

s = distanza tipica delle osservazioni dalla media

- maggiore è la deviazione standard s, MAGGIORE è la dispersione delle osservazioni

- più grandi sono le deviazioni, maggiore è la somma dei quadrati e più grande è s

I software statistici e molte calcolatrici tascabili sono programmati per calcolare velocemente le deviazioni standard.

I risultati ottenuti procedendo manualmente possono essere sensibilmente differenti da quelli calcolati dai software, causa arrotondamenti.

***PROPRIETÀ DELLA DEVIAZIONE STANDARD***:

1. $s \geq 0$
2. $s = 0$ solo quando tutte le osservazioni hanno lo stesso valore
3. più grande è la variabilità intorno alla media, maggiore è il valore di s.
4. la ragione per cui si utilizza ( n – 1) piuttosto che n nel denominatore di s ( e di $s^2$) è di tipo tecnico e riguarda l'inferenza per i parametri della popolazione.
5. Se sui dati viene effettuata una trasformazione di scala, anche le deviazioni standard vengono trasformate

INTERPRETARE LA GRADEZZA DI s

s è interpretabile come: distanza tipica delle osservazioni dalla media

**REGOLA EMPIRICA per interpretare la grandezza di s**: se l'istogramma della distribuzione ha una forma approssimativamente campanulare, allora:

1. Circa il 68% delle osservazioni assume valori compresi tra ylineetta – s e y + s
2. Circa il 95% delle osservazioni assume valori compresi tra y – 2s e y + 2s
3. La quasi totalità delle osservazioni assume valori compresi tra y – 3s e y + 3s

- detta EMPIRICA poiché molte delle distribuzioni osservate nella pratica sono grossomodo *campanulari*

VEDI ESEMPIO 3.8 PAG 52

Quando le distribuzioni hanno altre forme, la percentuale di osservazioni che ricadono entro due deviazioni standard dalla media può non essere del 95%.

- la regola empirica non è appropriata QUANDO la distribuzione è molto asimmetrica oppure molto discreta con la variabile che pin assume pochi valori

Quando il più elevato o il più piccolo valore della distribuzione è a meno di una deviazione standard dalla media, possiamo presupporre che la distribuzione sia fortemente asimmetrica.

- le deviazioni standard al pari della media possono essere fortemente influenzate dalle osservazione OUTLIER, soprattutto se il dataset è di dimensioni ridotte.

### 3.4 MISURE DI POSIZIONE

MISURE DI POSIZIONE: altro modo per descrivere una distribuzione

- forniscono info sul valore al di sotto del quale ricade una certa percentuale di osservazioni della distribuzione. Alcune individuano il centro altre la variabilità della distribuzione

QUARTILI E ALTRI PERCENTILI

CAMPO DI VARIAZIONE è DEFINITO USANDO DUE MISURE DI POSIZIONE, il MASSIMO e il MINIMO dei valori osservati nella distribuzione.

MEDIANA è una misura di posizione e metà delle osservazioni ricadono al di sotto o al di sopra di essa

- mediana: è una grandezza che appartiene a un insieme di misure di posizione chiamate PERCENTILI

**PERCENTILI**: il *p*-esimo percentile è il valore nelle distribuzioni al di sotto del quale ricade il *p*% delle osservazioni e al di sopra del quale ricade il ( 100- *p*)% delle osservazioni.

Altri due percentili di uso comune sono il PRIMO QUARTILE e il TERZO QUARTILE:

- un quarto delle osservazioni ricade al di SOTTO del primo quartile, e un quarto al di SOPRA del terzo

**PRIMO QUARTILE** = 25-esimo percentile

- è la mediana per le osservazioni che ricadono al di sotto della mediana (cioè per la prima metà delle osservazioni)

**TERZO QUARTILE** = 75-esimo

- è la mediana per le osservazioni che ricadono al di sopra della mediana (cioè per la seconda metà delle osservazioni)

Insieme alla mediana, i quartili dividono la distribuzione in 4 parti ciascuna della quali contiene un quarto delle osservazioni.

MISURE DI VARIABILITÀ: LO SCARTO INTERQUARTILE

**SCARTO INTERQUARTILE**: è la differenza tra il primo e il terzo quartile, viene indicata con IQR

- descrive le dispersioni delle osservazioni “centrali” della distribuzione

L’impiego di questo scarto presenta un sostanziale vantaggio rispetto al campo di variazione

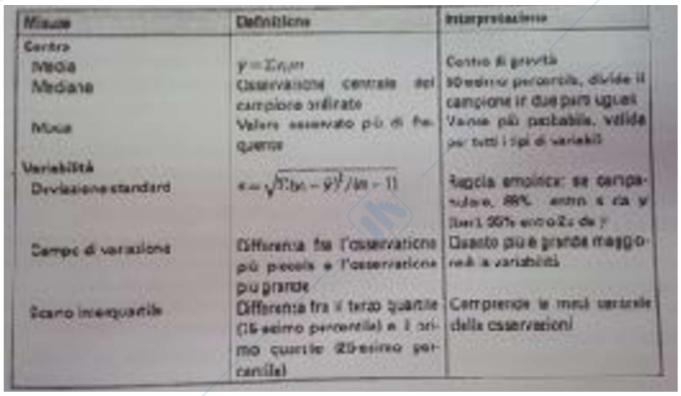
- NON è influenzato dalla presenza di OUTLIER.

Per le distribuzioni campanulari:

- la distanza dalla media di ciascun quartile è circa 2/3 della deviazione standard

- lo scarto interquartile è approssimativamente (4/3)s

SCHEMA RIASSUNTIVO DELLE MISURE DI CENTRALITÀ E DI VARIABILITÀ



| Misure | Definizione | Interpretazione |
|---|---|---|
| Centro | | |
| Media | $\bar{y} = \Sigma y_i/n$ | Centro di gravità |
| Mediana | Osservazione centrale del campione ordinato | 50esimo percentile, divide il campione in due parti uguali |
| Moda | Valore osservato più di frequente | Valore più probabile, valido per tutti i tipi di variabili |
| Variabilità | | |
| Deviazione standard | $s = \sqrt{\Sigma(y_i - \bar{y})^2/(n-1)}$ | Regola empirica: se campione [illegible] 68% entro 1s da y [illegible] 95% entro 2s da y |
| Campo di variazione | Differenza fra l'osservazione più piccola e l'osservazione più grande | Quanto più è grande maggiore è la variabilità |
| Scarto interquartile | Differenza fra il terzo quartile (75esimo percentile) e il primo quartile (25esimo percentile) | Comprende le metà centrale delle osservazioni |

BOX PLOT: RAPPRESENTAZIONE GRAFICA DI CINQUE MISURE DI POSIZIONE

La mediana, i quartili, il massimo e il minimo sono 5 misure di posizione spesso impiegate congiuntamente per descrivere la centralità e la variabilità di unna distribuzione.

Questi 5 numeri forniscono una semplice descrizione dei dati e li chiameremo

**SINTESI-A-CINQUE-NUMERI**

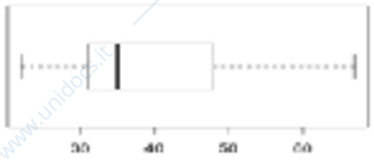
- SINTESI A CINQUE NUMERI: sono anche gli elementi base di una rappresentazione grafica detta **BOX PLOT** = sintetizza sia il centro sia la variabilità di una distribuzione

BOX= contiene il 50% centrale della distribuzione, dal primo al terzo quartile.

- MEDIANA: rappresentata da una linea che attraversa il box

BAFFI = linee che si estendono a partire dalla scatola

- esse si estendono fino al massimo e fino al minimo, a meno che nella distribuzione ci siano osservazioni outlier rappresentate in maniera differente dal grafico




I box plot affiancati sono utili per confrontare distribuzioni.

- sono degli strumenti efficaci anche per identificare le osservazione outlier

OUTLIER

**OUTLIER**: è un osservazione che ricade a più di 1.5x(IQR) al di sopra del terzo quartile

oppure a più di 1.5x(IQR) al di sotto del primo quartile.

SE nella distribuzione NON sono presenti osservazioni outlier: ALLORA i baffi del box plot si estendono fino alle osservazioni massima e minima

- in tal caso il minimo non assume valori al di sotto di Q1-[1.5X(IQR)] e il massimo non è oltre [Q3+1.5x(IQR)]

SE CI SONO outlier: i baffi si estendono fino ai valori osservati più piccoli e più grandi interni all'intervallo Q1-[1.5X(IQR):Q3+[1.5X(IQR)], invece, fino a Q1-[1.5X(IQR)] e fino a [Q3+1.5X(IQR)]

Perché è importante identificare gli outlier?

Possono darci preziose info sui dati: possono permetterci di evidenziare osservazioni in cui i valori siano stati trascritti in maniera NON corretta

- è importante evidenziare separatamente gli outlier nel box plot, perché le osservazioni OUTLIER non sono molto informative sulla forma della distribuzione, specie nei dateset di grandi dimensioni

Il criterio del 1.5x(IQR)è da considerarsi ARBITRARIO: è sempre meglio considerare un'osservazione che sulla base di esso viene definita outlier COME UN POTENZIALE OUTLIER

- quando una distribuzione ha una forte asimmetria positiva (lunga coda destra), alcune osservazioni possono ricadere a più di 1.5 IQR oltre il terzo quartile senza discostarsi dal generale andamento dei dati

QUANTE DEVIAZIONI STANDARD DALLA MEDIA? LO Z-SCORE

Un altro modo per misurare una posizione in una distribuzione è quello di contare a quante deviazioni standard dalla media ricade una certa osservazione.

**Z-SCORE**: è il numero di deviazioni standard che separano un'osservazione di una distribuzione dalla media della stessa.

z = osservazione – media / deviazione standard

Secondo la regola empirica, per una distribuzione campanulare è molto difficile che un'osservazione ricada a più di 3 deviazioni standard dalla media

- SE un'osservazione ha uno z-core > 3 essa è un OUTLIER

### 3.5 STATISTICHE DESCRITTIVE BIVARIATE

Il valore di un parametro la maggior parte delle volte è incognito e le statistiche campionarie vengono calcolate per stimarlo.

STATISTICHE BIVARIATE: impiegate per analizzare congiuntamente due variabili.

ASSOCIAZIONE FRA VARIABILI ESPLICATIVE E VARIABILI RISPOSTA

Obiettivo dell'analisi multivariata: studiare l'ASSOCIAZIONE fra variabili

- c'è ASSOCIAZIONE TRA 2 VARIABILI, SE una variabile tende ad assumere certi valori allorchè l'altra variabile cambia il suo valore

ANALISI BIVARIATA = analisi dell'associazione tra due variabili

- coinvolge due variabili (studia come il valore della variabile risposta DIPENDA DA o SIA SPIEGATA DA i valori assunti dalla variabile esplicativa):

1. **VARIABILE RISPOSTA o variabile dipendente**: i suoi valori vengono confrontati per i diversi valori assunti dall'altra variabile (variabile esplicativa)

2. **VARIABILE ESPLICATIVA o variabile indipendente**: definiscono i gruppi che mettiamo a confronto per valutare le differenze che esistono tra loro rispetto ai valori assunti dalla variabile risposta.

IL CONFRONTO FRA DUE GRUPPI COME ANALISI BIVARIATA

BOX PLOT: confronta le medie della variabile risposta rispetto alle categorie della variabile esplicativa

DATI CATEGORIALI BIVARIATI

Per due variabili categoriali, una **TABELLA DI CONTINGENZA** mostra quante osservazioni vengono registrate per le diverse combinazioni dei valori delle variabili

- mostra il numero di persone osservate nel campione rispetto alle possibili combinazioni delle risposte date ai due quesiti

- la tabella permette di osservare COME i valori di una variabile risposta siano CONTINGENTI alle categorie della variabili esplicativa

DATI QUANTITATIVI BIVARIATI

Quando entrambe le variabili sono variabili quantitative, uno **SCATTERPLOT** rappresenta mediante punti i valori assunti dalle osservazioni

- i valori della variabile di risposta sono rappresentati sull'asse y, mentre quelli della variabile esplicativa sull'asse x

- i valori delle osservazioni sono rappresentati da dei punti nel grafico

Per due variabili quantitative, la **CORRELAZIONE** = è una misura della forza del legame di associazione lineare fra le variabili

- essa assume valori tra -1 e +1 e indica se i valori della variabile risposta tendono ad aumentare (correlazione positiva) o a diminuire (correlazione negativa) quando aumentano i valori della variabile esplicativa

- lo studio della correlazione è finalizzato a: descrivere la forza dell'associazione che lega le due variabili nei termini di quanto i dati seguano una tendenza descritta da una linea retta

ANALISI DELLA REGRESSIONE: ci permette di ottenere l'equazione di una retta per prevedere i valori della variabile risposta sulla base dei valori assunti da una variabile esplicativa

- impiega la formula matematica di una linea retta per prevedere il valore della variabile allorchè la variabile esplicativa assuma un certo valore

### 3.6 STATISTICHE CAMPIONARIE E PARAMETRI DELLA POPOLAZIONE

- media y lineetta: sintesi di tendenza centrale più diffusa

- s: più diffusa misura di variabilità

Poiché questi due valori dipendono da quello che è il campione effettivamente selezionato, essi assumono valori diversi da campione a campione

- essi stessi SONO DELLE VARIABILI: il loro valore è incognito prima che il campione sia selezionato; una volta selezionato il campione, essi diventano delle STATISTICHE CAMPIONARIE

STATISTICA = descrive una caratteristica del campione

PARAMETRO = descrive una caratteristica della popolazione da cui quel campione è stato estratto

- lettere minuscole greche: indicano i PARAMETRI di una popolazione

- lettere latine: indicano le STATISTICHE campionarie

NOTAZIONE PER I PARAMETRI

Le lettere greche μ (mi) e σ (sigma) indicano la media e la deviazione standard di una variabile nella popolazione.

σ -> DEVIAZIONE STANDARD DELLA POPOLAZIONE (è costante)

μ -> MEDIA DELLA POPOLAZIONE (è costante)

- la media e la deviazione standard sono VARIABILI perché dipendono dal CAMPIONE SELEZIONATO

- sono costanti perché si riferiscono a un solo gruppo di osservazioni, un gruppo che comprende l'intera popolazione.

- CAPITOLO 4: **DISTRIBUZIONI DI PROBABILITÀ**

### 4.1 INTRODUZIONE ALLA PROBABILITÀ

Condizione importante per una corretta rilevazione dei dati statistici: CASUALITÀ DELLA SELEZIONE DELLE OSSERVAZIONI

LA PROBABILITÀ COME UNA LUNGA SERIE DI FREQUENZE RELATIVE

PROBABILITÀ = per un campione casuale o esperimento casuale, la probabilità di una osservazione è la PROPORZIONE di volte in cui essa dovrebbe verificarsi in una lunghissima sequenza di osservazioni

Es. intervistiamo un campione casuale di elettori californiani: chiediamo a ciascuno di essi di indicare per chi ha votato nella consultazione elettorale

- SE intervistati MOLTI elettori, allora la proporzione campionaria id chi dichiarerò di aver votato per Schwarzenegger sarà vicina alla popolazione

- proporzione popolazione: 0.56; probabilità che una persona selezionata a caso abbia votato: circa 0.56

PROBABILITÀ si riferisce a una lunga serie dell'esperimento casuale: perché c'è bisogno di un GRAN NUMERO DI OSSERVAZIONI per stimare con accuratezza una probabilità

- probabilità = è una proporzione, è un numero compreso tra 0 e 1

Nella pratica è anche espressa in percentuali

PERÒ la probabilità come una lunga serie di osservazioni non è sempre utile: ci si deve fidare A VOLTE di informazioni SOGGETTIVE

- approccio soggettivo: probabilità di un risultato è definita come il GRADO di fiducia che si ripone, sulla base dell'info disponibile, nel verificarsi di quel risultato

(Statistica bayesiana: usa la probabilità soggettiva come proprio fondamento)

LEGGI PROBABILISTICHE DI BASE

P (A) = probabilità di un possibile evento O di un insieme di eventi

A = evento o insieme di eventi

4 leggi sono:

1. **P(non A) = 1 – P(A)**

   - se è nota la probabilità del verificarsi di un evento, ALLORA la probabilità che esso NON si verifichi è pari a 1 MENO quella probabilità

   Es. Se P(A) = 0.56, allora 1 – 0.56 = 0.44

2. Se A e B sono 2 possibili risultati (che NON si sovrappongono), allora

   **P(A o B) = P(A) + P(B)**

   - es. indagine per stimare persone che ritengono che la ricerca sulle cellule embrionali sia vietata dal governo:

   A = valore stimato per la proporzione campionaria (al di sotto di 0.10)

   B = valore stimato della proporzione (al di sopra di 0.10)

   Allora

   P(A) = P(B) = 0.03

   La probabilità COMPLESSIVA che la proporzione campionaria sia sbagliata di più di 0.10 è:

   P(A o B) = P(A) + P(B) = 0.3 + 0.3 = 0.06

3. Se A e B sono possibili risultati, allora

   **P(A e B) = P(A) x P(B dato a)**

   - es. probabilità che adulto americano sia sposato = 0.56

   Probabilità persona sia molto felice = 0.40

   P(coniugato e molto felice) = P(coniugato) x P(molto felice) = 0.56 x 0.40 = 0.22

Circa il 22% della popolazione è sia coniugato che molto felice

- in alcuni casi A e B sono indipendenti, cioè SE uno degli eventi si verifica ciò non dipende dal fatto che si sia verificato l'altro: questo significa che

P(B dato A) = P(B)

La regola si semplifica nella regola 4

4. Se A e B sono indipendenti, allora

**P(A e B) = P(A) x P(B)**

- probabilità di effettuare una corretta inferenza = 0.95

A = corretta ipotesi sui maschi della popolazione

B = corretta ipotesi sulle femmine

Sono campioni e ipotesi indipendenti: la probabilità che ENTRAMBE le ipotesi siano corrette è

P(A e B) = P(A) x P(B) = 0.95 x 0.95 = 0.90

## 4.2 DISTRIBUZIONI DI PROBABILITÀ PER VARIABILI DISCRETE E CONTINUE

Variabile deve assumere almeno 2 valori

VARIABILE CASUALE = il risultato varia da osservazione a osservazione secondo una variazione casuale che può essere spiegata dalle probabilità

VARIABILE **discreta** = SE i possibili risultati appartengono a un insieme di valori discreti (es. variabile come "il numero di …" con possibili valori 0,1,2…)

VARIABILE **continua** = SE i possibili valori appartengono a un insieme di valori infinito; valori in continuum infiniti di possibili valori

DISTRIBUZIONE DI PROBABILITÀ = l'insieme dei POSSIBILI risultati e le corrispondenti probabilità

Vediamo come questo si realizza per variabili discrete e continue

<u>DISTRIBUZIONI DI PROBABILITÀ PER VARIABILI DISCRETE</u>

DISTRIBUZIONE DI PROBABILITÀ di una variabile discreta = assegna una probabilità a ciascun possibile valore della variabile

- ciascuna probabilità è un numero compreso tra 0 e 1

- la somma delle probabilità di TUTTI i possibili valori della variabile è 1

P(y) = probabilità di un possibile risultato per una variabile y

$0 \leq P(y) \leq 1$ e $\Sigma_{\text{tutti } y} P(y) = 1$ VEDI ESEMPIO 4.1 PAG 78

DISTRIBUZIONI DI PROBABILITÀ PER VARIABILI CONTINUE

DISTRIBUZIONI SI PROBABILITÀ DI VARIABILI CONTINUE: assegnano, probabilità intervalli di numeri

- probabilità che un valore di una variabile cada in ciascun particolare intervallo è compresa tra 0 e 1

- probabilità che intervallo contenga TUTTI i possibili valori è 1

GRAFICO della distribuzione di probabilità di una variabile continua: è una CURVA LISCIA, CONTINUA

- area sottesa alla curva per un intervallo di valori rappresenta: PROBABILITÀ che la variabile assuma un valore all'interno di quell'intervallo

VEDI ESEMPIO 4.2 PAG 79

PARAMETRI PER DESCRIVERE LE DISTRIBUZIONI DI PROBABILITÀ

Per la maggior parte delle distribuzioni di probabilità, impieghiamo delle formule matematiche

PER ALTRE distribuzioni, grafici o tabelle sono utili

DISTRIBUZIONE NELLA POPOLAZIONE di una variabile = è la distribuzione di probabilità di una variabile per un soggetto estratto casualmente dalla popolazione

DISTRIBUZIONE DI PROBABILITÀ ha dei **parametri** che descrivono le misure di tendenza centrale e la variabilità

MEDIA = descrive la tendenza centrale

DEVIAZIONE STANDARD = descrive la variabilità

I valori dei parametri sono UGUALI a quelli che si rileverebbero, in una lunga serie di osservazioni, se venissero selezionate casualmente osservazioni di una variabile y che ha una determinata distribuzione di probabilità

Es. 1% dei casi y = 0

3% dei casi y = 1

E così via..

Su 100 osservazioni, si avrà all'incirca:

uno 0, 3 volte 1, 60 volte 2, 23 volte 3, 12 volte 4, una volta 5

MEDIA = totale delle osservazioni diviso la dimensione campionaria; sarà pari a:

(1)0 + (3)1 + (60)2 + (23)3 + 12(4) + 1(5) / 100 = 245/100 = 2.45

Questa operazione assume la forma: 0(0.01) + 1(0.03) + 2(0.60) + 3(0.23) + 4(0.12) + 5(0.01)

Cioè la SOMMA dei possibili risultati MOLTIPLICATI per le corrispondenti probabilità

Infatti PER CIASCUNA variabile discreta y, questa è l'espressione per la MEDIA della distribuzione di probabilità

MEDIA DI UNA DISTRIBUZIONE DI PROBABILITÀ (valore atteso) = la media di una distribuzione di probabilità per una variabile discreta y è

$$\mu = \sum yP(y)$$

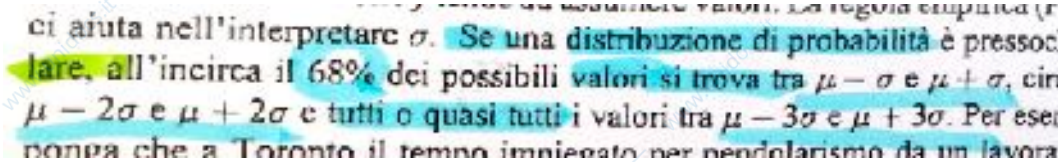
- la somma è calcolata su TUTTI i possibili valori della variabile

- questo parametro è anche chiamato VALORE ATTESO DI Y E(y)

La deviazione standard di una distribuzione di probabilità (σ): MISURA LA VARIABILITÀ

+ grande il valore di σ , + dispersione della distribuzione

- Se una distribuzione di probabilità è CAMPANULARE:


ci aiuta nell'interpretarc $\sigma$. Se una distribuzione di probabilità è pressoc
lare, all'incirca il 68% dei possibili valori si trova tra $\mu - \sigma$ e $\mu + \sigma$, cir
$\mu - 2\sigma$ e $\mu + 2\sigma$ e tutti o quasi tutti i valori tra $\mu - 3\sigma$ e $\mu + 3\sigma$. Per ese
ponga che a Toronto il tempo impiegato per pendolarismo da un lavora

DEVIAZIONE STANDARD = radice quadrata della VARIANZA di una distribuzione di probabilità

- VARIANZA = misura la media quadratica dello scarto di un'osservazione dalla media, cioè il valori atteso di $(y-\mu)^2$

**4.3 LA DISTRIBUZIONE DI PROBABILITÀ NORMALE**

Alcune distribuzioni teoriche della probabilità sono particolarmente importanti in relazione all'utilizzo che di esse di fa nella statistica inferenziale

DISTRIBUZIONE DI PROBABILITÀ NORMALE: curva campanulare, descrive molti istogrammi di dati per variabili che sono continue (o che, se discrete, assumono un gran numero di valori)

- è la più importante distribuzione per la statistica inferenziale!!!!!!!!!!!!!!

(sarà utile anche con distribuzioni di dati campionari che non presentano andamento campanulare

DISTRIBUZIONE NORMALE:

- è simmetrica, campanulare e caratterizzata da una media μ e da una deviazione standard σ

- probabilità che un'osservazione ricada all'interno di un intervallo definito dalla media μ più o meno un certo numero di deviazioni standard, è la stessa per tutte le distribuzioni normale

Tale probabilità è pari a

0.68 entro 1 deviazione standard

0.95 entro 2 deviazioni standard

0.997 entro 3 deviazioni standard

Ciascuna distribuzione normale è specificata da due parametri: la media $\mu$ e la deviazione standard $\sigma$. Per ciascun valore reale di $\mu$ e per ciascun valore non negativo di $\sigma$, esiste una distribuzione normale avente quella media e quella deviazione standard. La Figura 4.3 illustra tutto ciò. Sostanzialmente l'intera distribuzione è compresa tra $\mu - 3\sigma$ e $\mu + 3\sigma$.

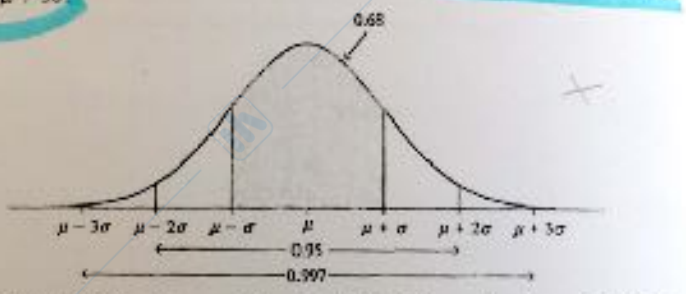



Figura 4.3 Per ogni distribuzione normale, la probabilità è pari (approssimativamente) a 0.68 entro $\sigma$ da $\mu$, a 0.95 entro $2\sigma$ da $\mu$ e a 0.997 entro $3\sigma$ da $\mu$.

ESEMPIO

Ad esempio, le altezze delle donne adulte nel Nord America hanno, all'incirca, una distribuzione normale con media $\mu = 65.0$ pollici[3] e deviazione standard $\sigma = 3.5$. La probabilità che una donna scelta a caso abbia un'altezza compresa tra $\mu - 3\sigma = 65.0 - 3(3.5) = 54.5$ pollici e $\mu + 3\sigma = 65.0 + 3(3.5) = 75.5$ pollici è vi-

cina a 1.0. L'altezza degli uomini adulti ha una distribuzione normale con $\mu = 70.0$ e $\sigma = 4.0$ pollici. Quindi, la probabilità che un uomo scelto a caso abbia un'altezza compresa tra $\mu - 3\sigma = 70.0 - 3(4.0) = 58$ pollici e $\mu + 3\sigma = 70.0 + 3(4.0) = 82$ pollici è quasi 1.0. Si osservi la Figura 4.4.

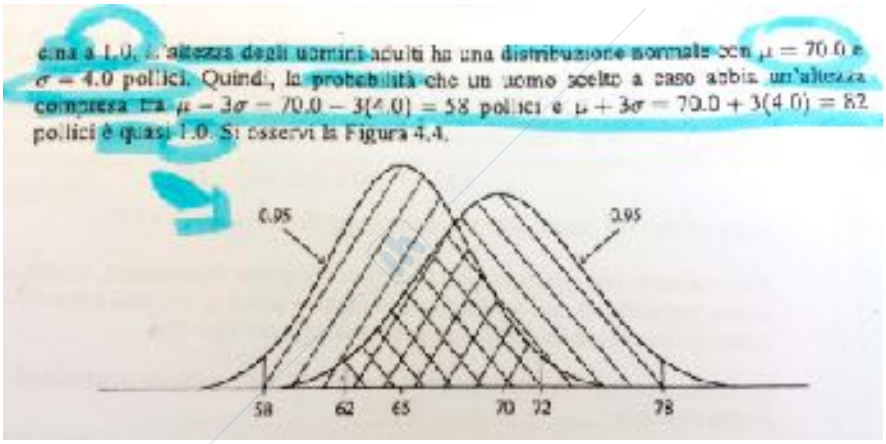



TAVOLA DELLE PROBABILITÀ SOTTESE ALLE CODE DI UNA DISTRIBUZIONE NORMALE

In una distribuzione normale, per ciascun fissato valore z: probabilità che osservazione ricada intorno alla media più o meno intorno alla media più o meno z volte la deviazione standard = dipende dal VALORE di z

- tale probabilità è data dall'area sottesa alla curva normale compresa tra $\mu - z\sigma$ e $\mu + z\sigma$

Per una distribuzione normale, la probabilità concentrata entro $z\sigma$ da $\mu$ è la STESSA per tutte le cure normale ANCHE SE z non è un numero intero

TAVOLA A dell'appendice A: riporta la probabilità per valori che cadono nella coda destra, almeno z volte la deviazione standard oltre la media

- colonna sul margine sinistro: elenca i valori per una z a una cifra decimale

- la seconda cifra decimale di z è nell'INTESTAZIONE delle colonne

Tabella 4.2 Riproduzione di parte della Tavola A in cui vengono mostrate le probabilità sottese alla coda destra dalla normale.

| $z$ | Seconda cifra decimale di $z$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
| 0.0 | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0722 | 0.0708 | 0.0694 | 0.0681 |
| 1.5 | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |

La Tabella 4.2 mostra un piccolo estratto della Tavola A. La probabilità per $z = 1.43$ si individua all'incrocio fra la riga corrispondente al valore 1.4 e alla colonna corrispondente al valore .03. Questa probabilità è pari a 0.0764. Ciò significa che per ogni distribuzione normale, la probabilità sottesa alla coda destra oltre $\mu + 1.43\sigma$ (cioè, più di 1.43 volte la deviazione standard oltre la media) vale 0.0764.

Poiché i valori all'interno della Tavola A sono le probabilità osservate per la metà destra della distribuzione normale (oltre $\mu + z\sigma$), esse assumono valori compresi tra 0 e 0.50. Data la simmetria della curva normale, queste probabilità sono valide anche per la coda sinistra per valori inferiori a $\mu - z\sigma$. Ad esempio, la probabilità sottesa alla curva prima del valore di $\mu - 1.43\sigma$ è sempre 0.0764. Le probabilità sottese alla coda sinistra, chiamate *probabilità cumulate*, sono fornite in automatico da molte calcolatrici tascabili e dai software statistici.

PROBABILITÀ DELLA NORMALE E REGOLA EMPIRICA

- regola empirica: per gli istogrammi a forma campanulare, circa il 68% dei dati cade in un intervallo entro 1 volta la deviazione standard dalla media; il 95& entro 2 volte la deviazione standard; TUTTE o quasi tutte le osservazioni entro 3 volte la deviazione standard

Ad esempio, il valore due deviazioni standard oltre la media ha un valore $z$ pari a 2.00. La probabilità della curva normale riportata nella Tavola A oltre $z = 2.00$ è 0.0228. La probabilità sulla coda destra oltre $\mu + 2\sigma$ vale 0.0228 per qualsiasi distribuzione normale. Dato la simmetria della distribuzione (Figura 4.5), la probabilità sottesa alla coda sinistra prima di $\mu - 2\sigma$ è, ugualmente, 0.0228. La probabilità complessiva oltre due volte la deviazione standard dalla media è $2(0.0228) = 0.0456$. Poiché tale probabilità è 0.0456, quella riferita all'intervallo compreso tra $\mu - 2\sigma$ e $\mu + 2\sigma$ è $1 - 0.0456 = 0.9544$ (si è utilizzata la regola (1) delle leggi sulle probabilità riportate alla fine del Paragrafo 4.1: $P(\text{non } A) = 1 - P(A)$.) Quando una variabile si distribuisce normalmente, il 95% delle osservazioni cade all'interno dell'intervallo dato dalla media più o meno due volte la deviazione standard.

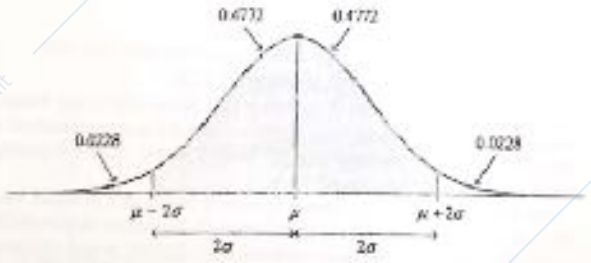



Figura 4.5 Per una distribuzione normale la probabilità corrispondente all'intervallo definito dalla media più o meno due volte la deviazione standard è pari a $1 - 2(0.0228)$, in pratica = 0.95.

Le percentuali approssimate attraverso l'applicazione della regola e arrotondare alla seconda cifra decimale sono: le percentuali effettivamente osservate per una distribuzione normale

REGOLA EMPIRICA: riporta le percentuale in modo APPROSSIMATO, piuttosto che esatto

Perché si applica a tutte le distribuzioni approssimativamente campanulari, non solo alla normale

- non tutte le distribuzioni campanulari sono normali: lo sono SOLO quelle descritte dalla formula matematica riportata nel Problema 4.56 alla fine del capitolo

TROVARE I VALORI Z PER DETERMINATE PROBABILITÀ IN PROSSIMITÀ DELLE CODE

Nelle metodologie inferenziali utilizzano i valori z di probabilità della curva normale: partendo dalla PROBABILITÀ (riportata nella parte interna della Tavola A, si trova il valore z che fornisce lo scostamento dalla media, in termini di deviazione standard, di quel valore

VEDI ESEMPIO 4.3 PAG 84

LO Z-SCORE CORRISPONDE AL NUMERO DI DEVIAZIONI STANDARD DALLA MEDIA

z = distanza, espressa in termini di deviazioni standard, tra un possibile valore y di una variabili E la media μ della sua distribuzione di probabilità

VEDI ESEMPIO 4.4 PAG 85

*z*-score

Lo *z*-score per un valore *y* di una variabile è il *numero di deviazioni standard* tra *y* e $\mu$. Tale valore è pari a

$$z = \frac{\text{osservazione} - \text{media}}{\text{deviazione standard}} = \frac{y - \mu}{\sigma}$$

Si avranno z-score POSITIVI quando: valore di y MAGGIORE della media

NEGATIVI: valore di y INFERIORE a quello della media

La tavola A riporta SOLO valori positivi per z

- poiché la distribuzione normale è simmetrica intorno alla media:

la probabilità sottesa alla coda sinistra prima di -z è uguale alla probabilità sottesa alla coda destra dopo +z

z-score sono UTILI per confrontare diverse distribuzioni normali

Quadro riassuntivo su come deve essere utilizzato uno z-score:

**Utilizzo degli *z*-score per trovare le probabilità o i valori di *y***

- Se è noto un valore di *y* e si ha bisogno di trovare la probabilità a esso corrispondente, si trasforma *y* in uno *z*-score mediante la formula $z = (y - \mu)/\sigma$ e si utilizza una tavola di probabilità della distribuzione normale per determinare la probabilità di interesse.
- Se è nota la probabilità e si ha bisogno di trovare un valore per *y*, si trasforma la probabilità in una probabilità sottesa alla coda della distribuzione e si trova lo *z*-score utilizzando la tavola della normale, quindi, si calcola $y = \mu + z\sigma$.

LA DISTRIBUZIONE NORMALE STANDARDIZZATA

- distribuzione normale standardizzata: è una DISTRIBUZIONE NORMALE con media μ = 0 e deviazione standard σ = 1

- i valori che ricadono z deviazioni standard OLTRE la media, è pari a μ – zσ = 0 + z(1)  CIOÈ semplicemente lo z-score

- QUINDI i valori della distribuzione normale standardizzata COINCIDONO con gli z-score

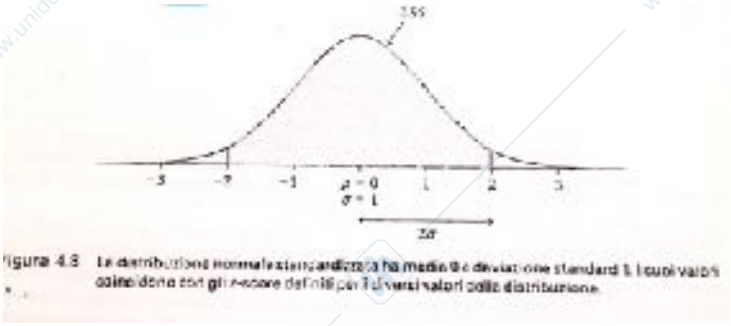


Figura 4.8 La distribuzione normale standardizzata ha media 0 e deviazione standard 1. I suoi valori coincidono con gli z-score definiti per i diversi valori della distribuzione.

Quando i valori di una qualsiasi distribuzione normale sono trasformati in z-score, VENGONO DISTRIBUITI intorno alla media 0 e hanno una deviazione standard = 1

- gli z-score ottenuti hanno una distribuzione normale standardizzata

SE una variabile ha distribuzione normale + i suoi valori sono trasformati in z-score: sottraendo la MEDIA e dividendo per la DEVIAZIONE STANDARD, ALLORA gli z-score avranno una distribuzione normale standardizzata

## 4.4 LE DISTRIBUZIONI CAMPIONARIE DESCRIVONO COME VARIANO LE STATISTICHE

Distribuzioni campionarie riportano le probabilità dei possibili valori per una variabile

- nella realtà sono raramente note

PER QUESTO si fa ricordo ai dati campionari per fare inferenza sui parametri di queste distribuzioni

Vediamo il legame fra inferenza statistica e calcolo delle probabilità in riferimento a valori dei parametri noti

VEDI ESEMPIO 4.6 PAG 88

SIMULAZIONE DEL PROCESSO DI STIMA

Simulazione = ci dice quanto un risultato di un exit pool, ad esempio, ben approssimi la proporzione della popolazione dei votanti per un candidato

Vedi esempio a pag 88-89

In sintesi: se metà degli elettori della popolazione avesse votato per Schwarzenegger, ci si sarebbe dovuti attendere di osservare una proporzione di voti a suo favore compresa tra circa il 47% e il 53% dei votanti

RAPPRESENTAZIONE DELLA VARIABILTÀ CAMPIONARIA ATTRAVERSO LA DISTRIBUZIONE CAMPIONARIA

Preferenza al voto: è una variabile che varia al variare degli elettori

Allo stesso modo, lo è anche la proporzione campionaria dei votanti per un certo candidato

DISTRIBUZIONE CAMPIONARIA: fornisce anche i possibili valori di una statistica per un SINGOLO campione di dimensione n

- distribuzione campionaria di una statistica: è la distribuzione di probabilità che fornisce le probabilità per i possibili valori che la statistica può assumere

Ciascuna statistica campionaria ha una DISTRIBUZIONE CAMPIONARIA: è un tipo di distribuzione di probabilità

- fornisce la probabilità NON per osservazioni individuali MA per possibili valori di una statistica calcolata attraverso le osservazioni

- aiuta a prevedere QUANTO precisa sia la stima del parametro di interesse

VEDI ESEMPIO 4.7 PAG 90

INTERPRETAZIONE DELLA DISTRIBUZIONE CAMPIONARIA NEL CASO DI CAMPIONAMENTO RIPETUTO

Distribuzioni campionarie descrivono la variabilità campionaria che si registra nella selezione dei dati e nell'utilizzo delle statistiche campionarie per stimare i parametri

- SE diverse società di sondaggi realizzassero ciascuna un proprio exit poll e stimassero la proporzione della popolazione che ha votato per il candidato Repubblicano, otterrebbero delle stime differenti, in quanto i campioni sarebbero costituiti da persone diverse

La distribuzione campionaria di una statistica basata su n osservazioni, è la DISTRIBUZIONE DELLE FREQUENZE REALTIVE per quella statistica ottenuta da estrazioni ripetute di campioni di dimensione n, avendo calcolato ogni volta il valore della statistica

- la forma delle distribuzioni campionarie è spesso nota teoricamente

- si possono allora trovare le probabilità per il valore della statistica campionaria per un campione di una data dimensione n

## 4.5 DISTRIBUZIONI CAMPIONARIE DELLE MEDIE CAMPIONARIE

Media campionaria (y con sopra la lineetta) viene spesso usata, ed è molto importate analizzare la sua distribuzione campionaria

- quando si analizzano dati e di trova y (con sopra lineetta), NON si conosce di quanto essa si discosta dalla media della popolazione μ, perché non si conosce il valore di μ

- utilizzando l'info sull'ampiezza della distribuzione campionaria, però, si può prevedere quanto essa sia precisa

2 aspetti principali della distribuzione campionaria della media campionaria:

- formule per la misura di tendenza centrale e l'ampiezza della distribuzione campionaria

- la sua forma

MEDIA ED ERRORE STANDARD DELLA DISTRIBUZIONE CAMPIONARIA DI y (lineetta)

Media campionaria è una variabile: assume valori diversi al variare dei campioni

- la media della distribuzione campionaria di y (lineetta) è pari a μ

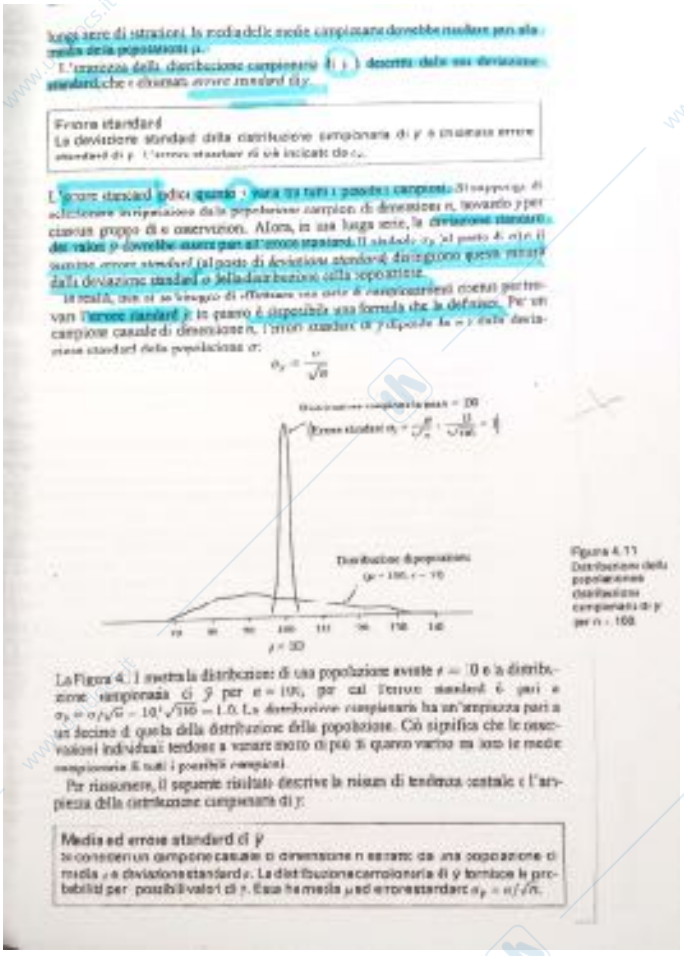
- SE si estraggono campioni con ripetizione, allora in una lunga serie di estrazioni, la media delle medie campionarie dovrebbe risultare pari alla media della popolazione μ

- AMPIEZZA della distribuzione campionaria di y (lineetta) è descritta dalla sua deviazione standard, chiamata ERRORE STANDARD DI y (lineetta)

ERRORE STANDARD: la deviazione standard della distribuzione campionaria di y (lineetta) è chiamata errore standard di y(lineetta)

ERRORE STANDARD indica quanto y(lineetta) VARIA tra tutti i possibili campioni

- deviazione standard dei valori y(lineetta) dovrebbe essere pari all'errore standard



Errore standard
La deviazione standard della distribuzione campionaria di $\bar{y}$ è chiamata errore standard di $\bar{y}$. L'errore standard di $\bar{y}$ è indicato da $\sigma_{\bar{y}}$.

L'errore standard indica quanto $\bar{y}$ varia tra tutti i possibili campioni. Si supponga di selezionare in ripetizione da la popolazione campioni di dimensione $n$, trovando $\bar{y}$ per ciascun gruppo di $n$ osservazioni. Allora, in una lunga serie, la deviazione standard dei valori $\bar{y}$ dovrebbe essere pari all'errore standard. Il simbolo $\sigma_{\bar{y}}$ al posto di $\sigma(n)$ il termine errore standard (al posto di deviazione standard) distinguono questa misura dalla deviazione standard o della distribuzione della popolazione.

In realtà, non ci si limita di ripetere un certo numero di campionamenti casuali per trovare l'errore standard $\bar{y}$ in quanto è disponibile una formula che la definisce. Per un campione casuale di dimensione $n$, l'errore standard di $\bar{y}$ è pari alla deviazione standard della popolazione $\sigma$:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

Distribuzione della popolazione ($\mu = 100$, $\sigma = 10$)

Figura 4.11 Distribuzione della popolazione e distribuzione campionaria di $\bar{y}$ per $n = 100$.

La Figura 4.11 mostra la distribuzione di una popolazione avente $\mu = 100$ e la distribuzione campionaria di $\bar{y}$ per $n = 100$, per cui l'errore standard è pari a $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 10/\sqrt{100} = 1.0$. La distribuzione campionaria ha un'ampiezza pari a un decimo di quella della distribuzione della popolazione. Ciò significa che le osservazioni individuali tendono a variare molto di più di quanto varino tra loro le medie campionarie di tutti i possibili campioni.

Per riassumere, il seguente risultato descrive la misura di tendenza centrale e l'ampiezza della distribuzione campionaria di $\bar{y}$:

**Media ed errore standard di $\bar{y}$**
Se consideriamo un campione casuale di dimensione $n$ estratto da una popolazione di media $\mu$ e deviazione standard $\sigma$. La distribuzione campionaria di $\bar{y}$ fornisce le probabilità per possibili valori di $\bar{y}$. Essa ha media $\mu$ ed errore standard $\sigma_{\bar{y}} = \sigma/\sqrt{n}$.

VEDI ESEMPIO 4.8 PAG 94

EFFETTO DELLA DIMENSIONE CAMPIONARIA SULLA DISTRIBUZIONE CAMPIONARIA E SULLA PRECISIONE DELLE STIME

Errore standard diminuisce all'AUMENTARE dell'AMPIEZZA n del campione

- questo perché il denominatore (radice di n) della formula dell'errore standard $\sigma_{\bar{y}} = \sigma/\sqrt{n}$

cresce al crescere di n

es. la deviazione standard della popolazione è pari a σ=0.50

l'errore standard è pari a 0.01 SE n=2705

SE n=100, una dimensione campionaria tipicamente bassa per un sondaggio, l'errore standard è pari a

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{n}} = \frac{0.50}{\sqrt{100}} = 0.05$$

Quando n=100, poiché tre volte l'errore standard equivale a 3(0.05)=0.15

l'intervallo che contiene la proporzione campionaria è molto grande, entro 0.15 da 0.50, quindi tra 0.35 e 0.65

La figura 4.13 mostra le distribuzioni campionarie della proporzione campionaria quando n=100 e quando n=2705

AL CRESCERE DI n, L'ERRORE STANDARD diminuisce e la distribuzione campionaria si restringe

- significa che: la proporzione campionaria tende ad assumere valori prossimi alla proporzione della popolazione

- è quindi più probabile che la proporzione campionaria stimi con maggior precisione una proporzione della popolazione non nota nel caso in cui n=2705 piuttosto che n=100

Tutto ciò è coerente con la nozione che campioni più ampi forniscono stime più precise delle caratteristiche della popolazione

(vedi esempio e grafico pag 93)

IN SINTESI

**ERRORE CAMPIONARIO**: stimare μ per mezzo di y(lineetta) produce un ERRORE, perchè si è osservata solo una parte della popolazione

- questo errore campionario tende a DIMINUIRE all'AUMENTARE della dimensione campionaria n

- errore standard è importante per le tecniche inferenziali perché fornisce una previsione dell'errore di campionamento che si commette nell'utilizzare y(lineetta) per stimare μ

<u>LA DISTRIBUZIONE CAMPIONARIA DELLA MEDIA CAMPIONARIA È APPROSSIMATIVAMENTE NORMALE</u>

FORMA della distribuzione campionaria y(lineetta)

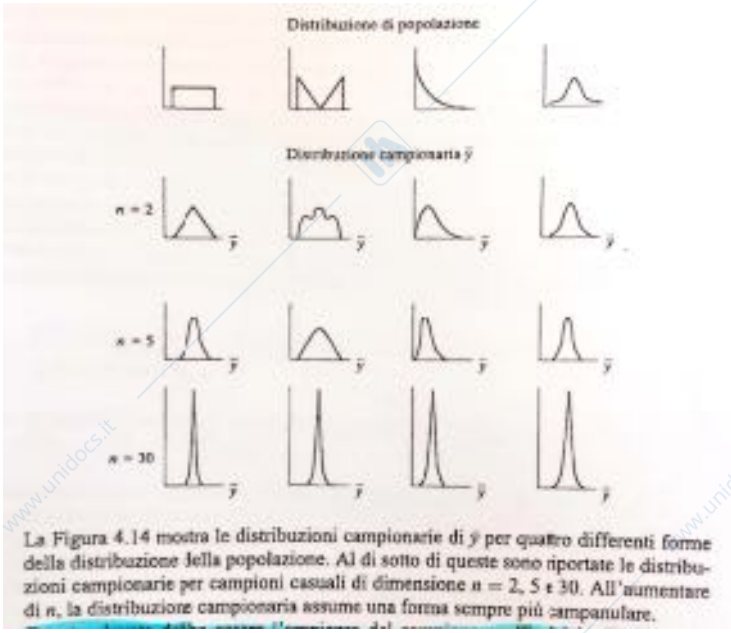
- è un risultato che può essere provato anche matematicamente e che viene ricondotto al TEOREMA DEL LIMITE CENTRALE

**TEOREMA DEL LIMITE CENTRALE**: per campioni casuali di elevata ampiezza n, la distribuzione campionaria della media campionaria y(lineetta) è approssimativamente una distribuzione normale

VEDIAMO delle interpretazioni di questo risultato:

1. l'approssimazione alla NORMALE della distribuzione campionaria, si applica QUALUNQUE SIA LA FORMA della distribuzione della popolazione

- per GRANDI CAMPIONI CASUALI: distribuzione campionaria y(lineetta) è all'incirca normale, anche se la distribuzione della popolazione è ASIMMETRICA (a forma di U) o fortemente discreta come la distribuzione binaria




La Figura 4.14 mostra le distribuzioni campionarie di $\bar{y}$ per quattro differenti forme della distribuzione della popolazione. Al di sotto di queste sono riportate le distribuzioni campionarie per campioni casuali di dimensione $n = 2, 5$ e 30. All'aumentare di $n$, la distribuzione campionaria assume una forma sempre più campanulare.

2. l'elevazione dell'ampiezza del campione n, affichè la distribuzione campionaria sia il più possibili a forma campanulare, DIPENDE DA gradi ASIMMETRIA della distribuzione della popolazione

   - SE la distribuzione della popolazione è di forma CAMPANUARE, allora la distribuzione campionaria è anch'essa campanulare, QUALUNQUE sia la dimensione campionaria (vedi parte destra della figura 4.14)

   - distribuzioni + asimmetriche necessitano di ampiezze campionarie maggiori

   - per campioni casuali, la distribuzione campionaria è quasi sempre approssimativamente campanulare

3. verificare empiricamente il Teorema del Limite Centrale: si estraggono campioni casuali con selezione ripetuta, calcolando y(lineetta) per ciascun campione di n osservazioni

   - vazioni. L'istogramma dei valori $\bar{y}$ si approssimerebbe a una curva normale di media $\mu$ ed errore standard pari a $\sigma/\sqrt{n}$ (cioè, la deviazione standard della popolazione

   Cioè la deviazione standard della popolazione divisa la radice quadrata della dimensione campionaria

4. sapere che la distribuzione campionaria di y(lineetta) è approssimativamente normale, aiuta a trovare le probabilità dei possibili valori di y(lineetta)

   trovare le probabilità dei possibili valori di $\bar{y}$. Ad esempio, $\bar{y}$ quasi certamente cade entro $3\sigma_{\bar{y}} = 3\sigma/\sqrt{n}$ da $\mu$. Vedremo in seguito che un'impostazione di questo genere è di fondamentale importanza per i metodi della statistica inferenziale.

VEDI ESEMPIO 4.9 PAG 97

Il valore della deviazione standard della popolazione σ è un valore che in realtà non è noto

- per fare inferenza, si può stimare σ attraverso la DEVIAZIONE STANDARD campionaria s

## 4.6 RICHIAMI SU: POPOLAZIONE, DATI CAMPIONARI E DISTRIBUZIONI CAMPIONARIE

Descriviamo i 3 tipi di distribuzioni:

- ***DISTRIBUZIONE DELLA POPOLAZIONE***: distribuzione dalla quale si seleziona il CAMPIONE

  - caratteristiche NON note: si fanno inferenze sulle sue caratt, come i parametri μ e σ che misurano la tendenza centrale e l'ampiezza della distribuzione

- ***DISTRIBUZIONE DEI DATI CAMPIONARI***: distribuzione dei DATI che si osservano, cioè le osservazioni campionarie y1,y2…yn

  - si descrive attraverso: statistiche campionarie quali la media y(lineetta) e la deviazione standard s

  + MAGGIORE dimensione del campione n + distribuzione dei dati campionari assomiglia alla distribuzione della popolazione e tanto + le statistiche campionarie (come y lineetta) prendono valori prossimi ai parametri della popolazione (come μ)

- ***DISTRIBUZIONE CAMPIONARIA di una statistica***: distribuzione di PROBABILITÀ per i possibili valori di una statistica campionaria, come y lineetta

- descrive la VARIABILITÀ tra i valori della statistica in tutti i possibili campioni di una determinata ampiezza

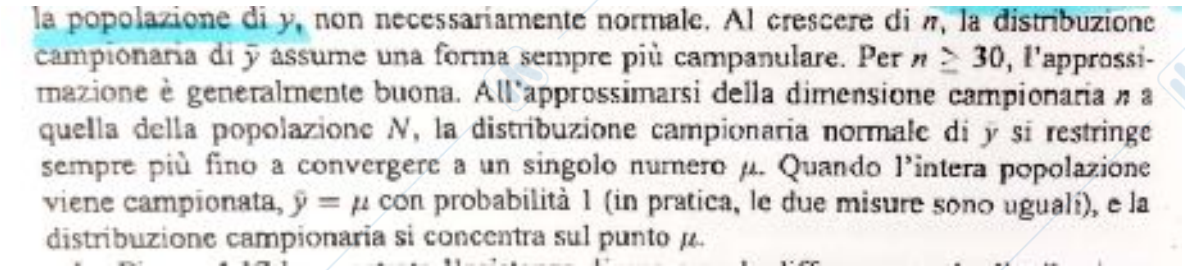
- questa distribuzione determina la probabilità che la statistica prenda valori entro una determinata distanza dal valore del parametro della popolazione che viene stimato

VEDI ESEMPIO 4.10 e 4.11 PAG 99

EFFETTO DELLA DIMENSIONE CAMPIONARIA SULLE DISTRIBUZIONI DEI DATI CAMPIONARI E SULLA DISTRIBUZIONE CAMPIONARIA

Forma della distribuzione della campionaria: è prossima a quella della NORMALE nella forma per valori crescenti di n

- DISTRIBUZIONE DELLA POPOLAZIONE DI y (non necessariamente normale): per campioni con UNA SOLA osservazione (n=1), ylineetta=y1, SI HA CHE la distribuzione campionaria di y lineetta è uguale alla distribuzione di probabilità per una osservazione su y



la popolazione di $y$, non necessariamente normale. Al crescere di $n$, la distribuzione campionaria di $\bar{y}$ assume una forma sempre più campanulare. Per $n \geq 30$, l'approssimazione è generalmente buona. All'approssimarsi della dimensione campionaria $n$ a quella della popolazione $N$, la distribuzione campionaria normale di $\bar{y}$ si restringe sempre più fino a convergere a un singolo numero $\mu$. Quando l'intera popolazione viene campionata, $\bar{y} = \mu$ con probabilità 1 (in pratica, le due misure sono uguali), e la distribuzione campionaria si concentra sul punto $\mu$.

- al crescere della dimensione campionaria, la **distribuzione dei dati campionari** assomiglia molto alla distribuzione della popolazione

- la **distribuzione campionaria**, al contrario, si presta con una forma campanulare e tende a restringersi al crescere si n

- i valori campionari di y possono essere solo 0 o 1

- al contrario i valori della media campionaria prendono valori nell'intervallo TRA 0 e 1

IL RUOLO CHIAVE DELLE DISTRIBUZIONI CAMPIONARIE NELL'INFERENZA STATISTICA

Per il Teorema del Limite Centrale, si può usare la distribuzione normale per trovare le probabilità y lineetta

- il RISULTATO riferito a grandi campioni casuali e riguardante le medie campionarie che tendono a distribuirsi come una normale, è importante PERCHÉ vale anche per molte altre statistiche

Es. per grandi campioni casuali, la maggior parte delle statistiche campionarie usate per stimare parametri della popolazione hanno distribuzioni campionarie approssimativamente normali

Questo è il motivo per cui si dà grande importanza allo studio della distribuzione normale

CAPITOLO 5: **INFERENZA STATISTICA – LA STIMA**

Come utilizzare i dati campionari per stimare i parametri della popolazione

## 5.1 STIMA PUNTUALE E INTERVALLARE

2 tipi di stima dei parametri:

1. STIMA PUNTUALE: SINGOLO numero che rappresenta la MIGLIORE previsione del valore assunto dal parametro
2. STIMA INTERVALLARE: è un INTERVALLO di numeri intorno alla stima puntuale, all'interno del quale si ritiene ricada il valore del parametro

- STIMA = da solo è un'abbreviazione di stima puntuale

- STIMATORE = particolare tipo di statistica impiegato per stimare un parametro;

STIMA = valore ottenuto applicando lo stimatore a uno specifico campione

Es. proporzione campionaria è uno stimatore della proporzione della popolazione

- valore 0.73 è la STIMA per la proporzione di chi crede nella vita dopo la morte

STIMA PUNTUALE DEI PARAMETRI

Ciascun parametro può avere diversi possibili stimatori

- distribuzione della popolazione normale: il centro è rappresentato SIA dalla media SIA dalla mediana perché la distribuzione è SIMMETRICA

- così in presenza di dati campionari, 2 possibili STIMATORI del centro della distribuzione sono: media campionaria e mediana campionaria

STIMATORI PUNTUALI CORRETTI ED EFFICIENTI

Buon STIMATORE = ha distribuzione campionaria che:

1. è centrata intorno al parametro

2. ha l'errore standard più piccolo possibile

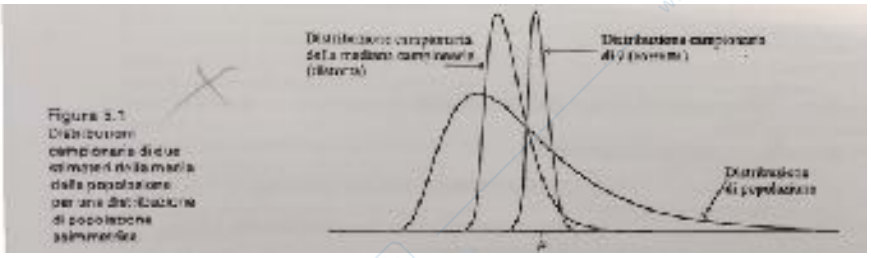
Uno STIMATORE è **CORRETTO**: SE la sua distribuzione campionaria è centrata intorno al parametro

- PARAMETRO: è la media della distribuzione campionaria

- per un campione casuale, la media delle distribuzione campionaria della media campionaria y lineetta, coincide con la media della popolazione μ

- per CIASCUN CAMPIONE: la media campionaria può sottostimare o sovrastimare μ

TUTTAVIA se si calcolasse ripetutamente la media campionaria su campioni diversi, le SOVRASTIME tenderebbero a controbilanciare le SOTTOSTIME



STIMATORE è **DISTORTO**: tende a SOTTOSTIMARE il parametro o a SOVRASTIMRLO

Es. il campo di variazione campionaria è tipicamente più piccolo di quello della popolazione e no può essere più grande

Infatti

- minimo e massimo campionario non possono essere quasi mai coincidenti con quelli della popolazione

- il campo di variazione campionario tende a SOTTOSTIMARE quello della popolazione

- si tratta quindi di uno STIMATORE DISTORTO del campo di variazione della popolazione

Es. distribuzione della popolazione asimmetrica positiva

Vogliamo stimare la media della popolazione

- se si ritiene che gli outlier possano influire sulla STIMA, si dovrebbe scegliere di stimarla utilizzando la MEDIANA CAMPIONARIA piuttosto che la media campionaria

TUTTAVIA

Mediana della popolazione è, in questo caso, più piccola della media della popolazione

QUINDI anche la mediana campionari tende a essere inferiore alla media della popolazione μ

Così la mediana campionaria è uno STIMATORE DISTORTO di μ, tendendo in media a sottostimare questa grandezza

- risulta più conveniente utilizzare la media campionaria, calcolandola dopo aver eliminato gli outlier

2. STIMATORE deve avere un **ERRORE STANDARD PICCOLO**

STIMATORE EFFICIENTE: errore standard più piccolo di quello di altri stimatori

- determina stime del parametro più vicine al vero valore

Es. distribuzione è normale: errore standard della media campionaria è il 25% più grande dell'errore standard della media campionaria

- media campionaria: più vicina della mediana campionaria al centro della popolazione

- media campionaria = stimatore efficiente; mediana campionaria non lo è

IN SINTESI

- buon stimatore di un parametro deve essere

CORRETTO (o comunque avere una dispersione trascurabile)

EFFICIENTE

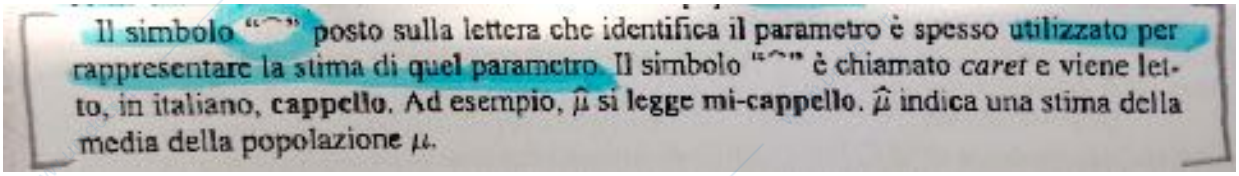
STIMATORI DELLA MEDIA, DELLA DEVIAZIONE STANDARD E DELLA PROPORZIONE

È comune usare la formula con cui si calcola il valore del parametro della popolazione COME STIMATORE applicato ai dati campionari

Es. per stimare una proporzione della popolazione, la proporzione campionaria è uno stimatore corretto ed efficiente

- per stimare una media della popolazione μ, la media campionaria y lineetta è corretta

- allo stesso modo, si usa la deviazione standard campionaria s come STIMATORE della deviazione standard della popolazione σ


Il simbolo "^" posto sulla lettera che identifica il parametro è spesso utilizzato per rappresentare la stima di quel parametro. Il simbolo "^" è chiamato *caret* e viene letto, in italiano, **cappello**. Ad esempio, $\hat{\mu}$ si legge **mi-cappello**. $\hat{\mu}$ indica una stima della media della popolazione $\mu$.

METODO DI STIMA DELLA MASSIMA VEROSIMIGLIANZA

STIMA PUNTUALE:

**TEORIA DELLA STIMA DI MASSIMA VEROSIMIGLIANZA** = è il valore del parametro che risulta più verosimile in relazione ai dati osservati

- STIMA si basa sul fornire una risposta a questa domanda:

SE il parametro fosse uguale a un certo valore (es. valore della stima)

QUALE sarebbe stata la probabilità di osservare i fati che sono stati raccolti con l'indagine campionaria?

Es. su uno studio di 1000 adulti statunitensi, è stato riportato che la STIMA di massima verosimiglianza della proporzione della popolazione che crede nell'astrologia è 0.37

- allora il campione osservato aveva una elevata probabilità di essere osservato SE la proporzione della popolazione fosse stata proprio pari a 0.37, piuttosto che essa fosse stata uguale a un qualsiasi altro possibile valore

Per molte distribuzione della popolazione, come la normale:

lo STIMATORE di massima verosimiglianza della media della popolazione è la MEDIA CAMPIONARIA

Per grandi campioni, gli STIMATORI DI VEROSIMIGLIANZA godono di 3 proprietà desiderabili:

1. Sono EFFICIENTI per campioni relativamente ampi: altri stimatori non hanno errori standard più piccoli e non tendono a prendere valori prossimi a quelli del parametro
2. Sono AFFETTI da ERRORE: diminuisce al CRESCERE dell'ampiezza campionaria
3. Hanno distribuzione campionaria approssimativamente normale

L'INTERVALLO DI CONFIDENZA COME STIMA PUNTUALE +O- UN MARGINE DI ERRORE

Per fornire una corretta informazione, l'inferenza su un certo parametro DEVE BASARSI SU:

STIMA PUNTUALE

+

STIMA rispetto al VERO VALORE del parametro

Le indicazioni sulla precisione della stima puntuale sono basate sull'ampiezza della STIMA INTERVALLARE di un parametro

STIMA INTERVALLARE: intervallo di valori intorno alla stima puntuale

- è progettato per contenere una prefissata probabilità prossima a 1, il vero valore del parametro

Poiché le STIME intervallari contengono il parametro con un certo livello di fiducia, essi vengono indicati come INTERVALLI DI CONFIDENZA

INTERVALLI DI CONFIDENZA: intervallo di confidenza per un parametro è un intervallo di valori entro cui si ritiene ricada il valore di detto parametro

- LIVELLO DI FIDUCIA = la probabilità associata al fatto che l'intervallo contenga il parametro: è un numero prossimo a 1, come 0.95 o 0.99

Aspetto cruciale nella costruzione di un INTERVALLO DI CONFIDENZA: è insito nella distribuzione campionaria dello stimatore puntuale

- spesso la distribuzione campionaria è approssimativamente normale

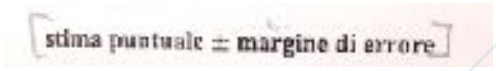
- la distribuzione normale, poi, determina la probabilità che lo stimatore produca valori entro una certa distanza dal parametro

PIÙ PICCOLO è L'ERRORE STANDARD, TANTO PIÙ PRECISO TENDE A ESSERE LO STIMATORE

Nella realtà, la distribuzione è solo approssimativamente normale: per costruire un intervallo di confidenza, si aggiunge e si sottrae DALLA stima puntuale, qualche multiplo (uno z-score) del suo errore standard

MARGINE D'ERRORE = multiplo dell'errore standard

Un intervallo di confidenza assume allora la forma:




Per costruire un INTERVALLO DI CONFIDENZA che ha il 95% livello di confidenza, si prende:

STIMA PUNTUALE + e – MARGINE DI ERRORE pari circa a 2 errori standard

## 5.2 INTERVALLO DI CONFIDENZA PER UNA PROPORZIONE $\hat{\pi}$

In presenza di DATI CATEGORIALI, un'osservazione può assumere valori che sono una delle categorie della variabilitàQuesto tipo di misura si realizza quando la variabili è NOMINALE

Inoltre si verifica quando variabili intrinsecamente continue sono misurate con scale di misura categoriali (es. reddito annuale che si presenta nelle categorie

Per sintetizzare dati categoriali, calcoliamo le PROPORZIONI (O PERCENTUALI) di osservazioni che ricadono in una categoria

LA PROPORZIONE CAMPIONARIA E IL SUO ERRORE STANDARD

Pi greco = Proporzione della popolazione

- assume valore compresi tra 0 e 1

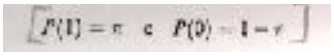
- la sua STIMA PUNTUALE è la PROPORZIONE CAMPIONARIA, indicata con

Proporzione campionaria: è una media QUANDO

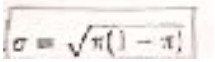
y = 1 per un'osservazione nella categoria di interesse

y = 0 in tutti gli altri casi

La proporzione della popolazione pi greco è: la MEDIA μ della distribuzione di probabilità, le cui probabilità sono pari a




La deviazione standard di questa distribuzione di probabilità è




Dato che la formula per l'errore standard della media campionaria $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, 

l'ERRORE STANDARD $\sigma_{\bar{\pi}}$ della PROPORZIONE CAMPIONARIA è:

$$\sigma_{\bar{\pi}} = \sigma/\sqrt{n} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

AL CRESCERE DELL'AMPIEZZA CAMPIONARIA, l'errore standard DIMINUISCE

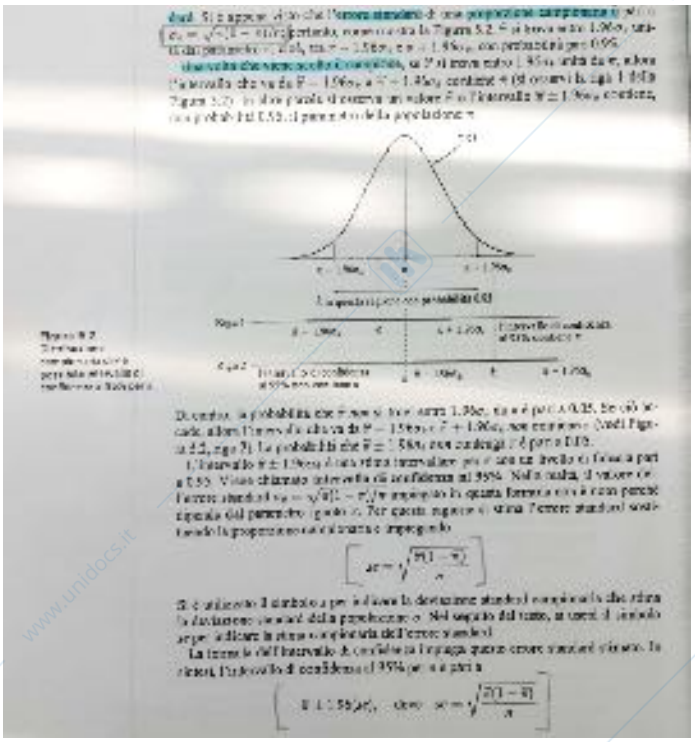
INTERVALLO DI CONFIDENZA PER UNA PROPORZIONE NEL CASO DI GRANDI CAMPIONI

Dato che una Proporzione campionaria $\bar{\pi}$ è: una media campionaria

ALLORA trova applicazione il Teorema del Limite Centrale = per campioni casuali di ampiezza

ELEVATA, la distribuzione campionaria di $\bar{\pi}$ è approssimativamente normale intorno al parametro pi greco oggetto di stima

- 95% delle osservazioni di una distribuzione normale è compreso entro 2 deviazioni standard dalla media o 1.96 deviazioni standard



VEDI ESEMPIO 5.1 e 5.2 PAG 116-117

SE l'intervallo di confidenza viene costruito facendo ricordo a una calcolatrice tascabile, NON BISOGNA ARROTODARE i calcoli, in modo tale da ridurre l'errore:

- ARROTONDAMENTO va fatto solo quando viene riportato il risultato finale

Es. è meglio riportare un intervallo di confidenza (0.30, 0.36), piuttosto che 0.303395, 0.356606

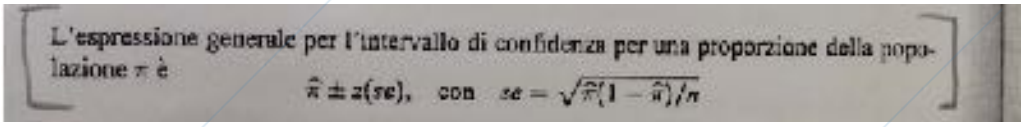
Le cifre DECIMALI in più tendono a distrarre e non sono particolarmente utili: non aggiungono alcun tipo di info

CONTROLLO DEL LIVELLO DI FIDUCIA

Con un livello di fiducia di 0.95, cioè il 95% di fiducia, c'è uno 0.05 di probabilità che il metodo produca un intervallo di confidenza che NON contiene il valore de parametro.

In alcuni casi pratici, un 5% di possibilità di un'inferenza non corretta è un rischio di errore inaccettabile: **per incrementare le possibilità di una corretta inferenza, si incrementa allora il livello di fiduci**a (es. fino al valore di 0.99)

VEDI ESEMPIO 5.3 PAG 118


L'espressione generale per l'intervallo di confidenza per una proporzione della popolazione $\pi$ è

$$\hat{\pi} \pm z(se), \quad \text{con} \quad se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

z dipende dal livello di fiducia

- più è alto il livello di fiducia, tanto maggiore sarà la possibilità che l'intervallo di confidenza contenga il parametro

- livelli di fiducia scelti più comunemente sono: 0.95 e 0.99

MARGINE D'ERRORE prodotto tra z-score e se

- con un livello di fiducia elevato, l'intervallo di confidenza si allarga, perché CRESCE il valore di z

Perché ci accontentiamo di un livello di fiducia inferiore al 100%?

Per una corretta inferenza, l'intervallo deve contenere tutti i possibili valori pi greco

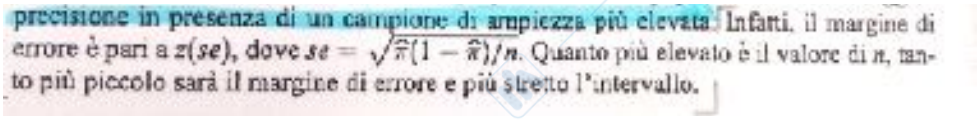
- ci si accontenta di essere meno precisi nella stima del valore del parametro: si accetta un compromesso tra il livello di fiducia desiderato che l'inferenza sia corretta e una desiderata precisione della stima

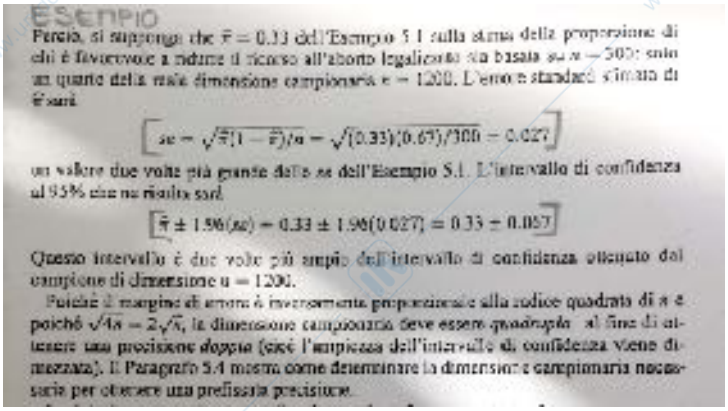
- AL MIGLIORARE DI UN ASPETTO, L'ALTRO PEGGIORA

- non si osservano intervalli di confidenza al 99.999999%: sarebbe troppo grande per poter fornire info utili su quale possa essere il valore del parametro della popolazione

CAMPIONI DI MAGGIORE AMPIEZZA FORNISCONO INTERVALLI PIÙ STRETTI

È lecito attendersi di poter stimare una proporzione della popolazione pi greco con maggior precisione in presenza di un campione di ampiezza più elevata


precisione in presenza di un campione di ampiezza più elevata. Infatti, il margine di errore è pari a $z(se)$, dove $se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$. Quanto più elevato è il valore di $n$, tanto più piccolo sarà il margine di errore e più stretto l'intervallo.



ESEMPIO

Perciò, si supponga che $\bar{p} = 0.33$ dell'Esempio 5.1 sulla stima della proporzione di elettori favorevoli a ridurre il ricorso all'aborto legalizzato sia basato su $n = 500$: sulla quarta della reale dimensione campionaria è $n = 1200$. L'errore standard stimato di $\bar{p}$ sarà

$$\left[ se = \sqrt{\bar{p}(1-\bar{p})/n} = \sqrt{(0.33)(0.67)/500} = 0.027 \right]$$

un valore due volte più grande della *se* dell'Esempio 5.1. L'intervallo di confidenza al 95% che ne risulta sarà

$$\left[ \bar{p} \pm 1.96(se) = 0.33 \pm 1.96(0.027) = 0.33 \pm 0.053 \right]$$

Questo intervallo è due volte più ampio dell'intervallo di confidenza ottenuto dal campione di dimensione $n = 1200$.

Poiché il margine di errore è inversamente proporzionale alla radice quadrata di $n$ e poiché $\sqrt{4n} = 2\sqrt{n}$, la dimensione campionaria deve essere *quadruplicata* al fine di ottenere una precisione *doppia* (cioè l'ampiezza dell'intervallo di confidenza viene dimezzata). Il Paragrafo 5.4 mostra come determinare la dimensione campionaria necessaria per ottenere una prefissata precisione.

In sintesi, questo paragrafo e quello precedente hanno dimostrato che

**Ampiezza di un intervallo di confidenza:**

1. cresce al crescere del livello di fiducia
2. decresce al crescere della dimensione campionaria

Queste proprietà si applicano a tutti gli intervalli di confidenza, non solo a quelli per la proporzione

LIVELLO DI SIGNIFICATIVITÀ = 1 – LIVELLO DI FIDUCIA

**LIVELLO DI SIGNIFICATIVITÀ =** probabilità che un metodo di stima intervallare fornisca un intervallo di confidenza che NON contenga un certo parametro

1 – livello di fiducia

Es. livello di fiducia = 0.95

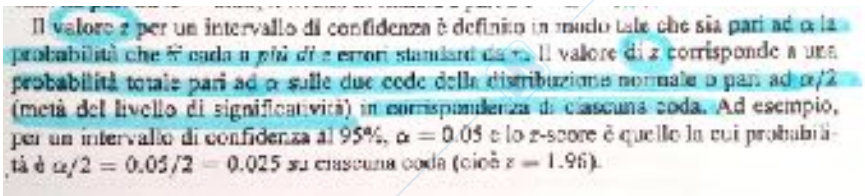
Livello di significatività = 1 – 0.95 = 0.05

Alfa: livello di significatività

1 – alfa: livello di fiducia

Es. livello di significatività alfa = 0.05

Livello di fiducia 1 – alfa = 0.95



Il valore $z$ per un intervallo di confidenza è definito in modo tale che sia pari ad $\alpha$ la probabilità che $\bar{p}$ cada a *più di* $z$ errori standard da $\pi$. Il valore di $z$ corrisponde a una probabilità totale pari ad $\alpha$ sulle due code della distribuzione normale o pari ad $\alpha/2$ (metà del livello di significatività) in corrispondenza di ciascuna coda. Ad esempio, per un intervallo di confidenza al 95%, $\alpha = 0.05$ e lo $z$-score è quello la cui probabilità è $\alpha/2 = 0.05/2 = 0.025$ su ciascuna coda (cioè $z = 1.96$).

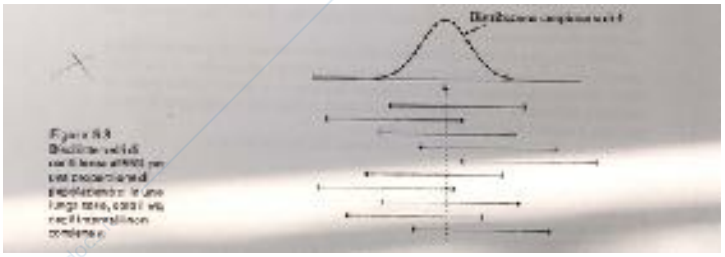
IL LIVELLO DI FIDUCIA È UNA PROPORZIONE CORRETTA PER UNA LUNGA SERIE DI PROVE

Livello di fiducia per un intervallo di confidenza DESCRIVE COME si comporta il metodo di stima scelto quando viene utilizzato più e più volte con molti differenti campioni casuali

- la proporzione della popolazione NON NOTA pi greco, è un NUMERO FISSO: un intervallo di confidenza determinato in corrispondenza di campioni di differente dimensione può contenere oppure no pi greco

Affermare che un particolare intervallo contiene pi greco il 95% di fiducia = il 95% delle volte l'inferenza è corretta

Gli intervalli di confidenza assumono posizioni differenti in quanto stima di pi greco varia da campione a campione: all'incirca solo una volta su 20, un intervallo di confidenza al 95% non contiene il vero valore del parametro della popolazione



Nella realtà: viene scelto solo un campione di una prestabilita dimensione n e si costruisce un intervallo di confidenza usando le osservazioni di quell'unico campione: non si può sapere se quell'intervallo di confidenza contiene realmente pi greco

La fiducia in quell'intervallo si basa sulle proprietà della procedura di stima

- si può solo tenere sotto controllo la probabilità che l'intervallo contenga pi greco

NECESSITÀ DI UN'ELEVATA AMPIEZZA CAMPIONARIA PER LA VALIDITÀ DEL METODO DI STIMA

La probabilità che l'intervallo di confidenza contenga pi greco è approssimativamente pari al livello di fiducia prescelto

- l'approssimazione è migliore per grandi campioni

Per il Teorema del Limite Centrale, al CRESCERE di n, la distribuzione campionaria di stima di pi greco è quasi normale

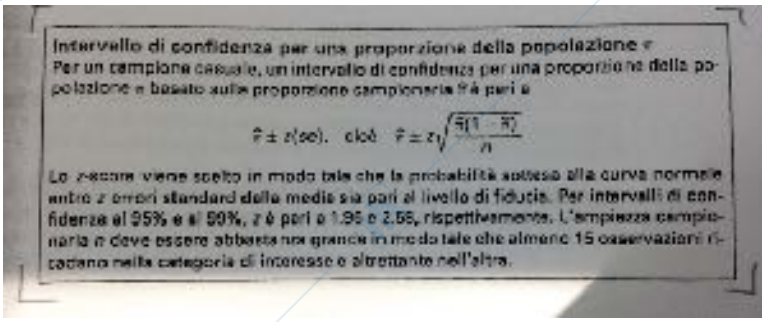
AL CRESCERE DI N, L'ERRORE STANDARD STIMATO $se = \sqrt{\bar{\pi}(1-\bar{\pi})/n}$

assume valori prossimi al VERO errore standard $\sigma_{\bar{\pi}} = \sqrt{\pi(1-\pi)/n}$

La formula dell'intervallo di confidenza si applica a grandi campioni casuali: almeno 15 osservazioni sia per la categoria di interesse sia per le altre

Procedura di determinazione dell'intervallo di confidenza per una proporzione

INTERVALLO DI CONFIDENZA PER UNA PROPORZIONE DELLA POPOLAZIONE PI GRECO



Intervallo di confidenza per una proporzione della popolazione *p*

Per un campione casuale, un intervallo di confidenza per una proporzione della popolazione è basato sulla proporzione campionaria $\hat{p}$ è pari a

$$\hat{p} \pm z(se), \text{ cioè } \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Lo *z*-score viene scelto in modo tale che la probabilità sottesa alla curva normale entro *z* errori standard della media sia pari al livello di fiducia. Per intervalli di confidenza del 95% e del 99%, *z* è pari a 1.96 e 2.58, rispettivamente. L'ampiezza campionaria *n* deve essere abbastanza grande in modo tale che almeno 15 osservazioni ricadano nella categoria di interesse e altrettante nell'altra.

## 5.3 INTERVALLO DI CONFIDENZA PER LA MEDIA

Abbiamo visto come costruire un intervallo di confidenza, ORA vediamo come costruirne uno per la media della popolazione nel caso di dati quantitativi

ERRORE STANDARD STIMATO PER IL MARGINE DI ERRORE

L'intervallo di confidenza per una media è basato sulla forma

[stima puntuale ± margine di errore]

Margine di errore: multiplo dell'errore standard

Stima puntuale della media della popolazione μ è la media campionaria y lineetta

- per grandi campioni casuali, secondo il Teorema del Limite Centrale, la distribuzione campionaria di y lineetta è all'incirca normale

- il margine d'errore può essere ottenuto moltiplicando

z-score della distribuzione normale x l'errore standard

Errore standard per la media campionaria è

$$\sigma_y = \frac{\sigma}{\sqrt{n}}$$

σ = deviazione standard della popolazione

- errore standard di una proporzione campionaria, dipende da σ

- in realtà si stima σ attraverso la deviazione standard campionaria s

Pertanto gli intervalli di confidenza utilizzano l'errore standard STIMATO

$$se = s/\sqrt{n}$$

VEDI ESEMPIO 5.4 PAG 120

Vediamo come trovare il margine di errore con precisione, moltiplicando

se x valore molto simile a z-score (a meno che n non si eccessivamente piccolo)

LA DISTRIBUZIONE T

Come costruire un intervallo di confidenza che si applica a campioni casuali di dimensione QUALSIASI

- necessario assumere che la distribuzione della popolazione sia normale: la distribuzione campionaria di y lineetta è normale anche per campioni di piccole dimensioni

Es. si supponga di conoscere l'esatto errore standard della media campionaria

$$\sigma_{\bar{y}} = \sigma/\sqrt{n}$$

Con l'ulteriore assunzione che la popolazione sia normale, per qualsiasi n potremo utilizzare la formula

$$\bar{y} \pm z\sigma_{\bar{y}}, \quad \text{cioè} \quad \bar{y} \pm z\sigma/\sqrt{n}$$

Non si conosce la deviazione standard della popolazione σ

Non si conosce l'esatto errore standard

SE si sostituisce la deviazione standard campionaria s al posto di σ per ottenere l'errore standard

STIMATO $se = s/\sqrt{n}$, si introduce una FONTE DI ERRORE

Tale errore può essere NOTEVOLE  n è PICCOLO

Per calcolare questo incremento di errore, dobbiamo:

Sostituire lo Z-SCORE con un punteggio più grande, cioè T-SCORE

- l'intervallo di confidenza diventa più ampio

T-SCORE = simile ad uno z-score, MA è riferito a una distribuzione campanulare più dispersa della distribuzione normale, chiamata DISTRIBUZIONE T

PROPRIETÀ DELLA DISTRIBUZIONE T

Proprietà della distribuzione t sono:

1. Campanulare e SIMMETRICA intorno alla media 0
2. Deviazione standard è poco più grande di 1: il valore esatto dipende da GRADI DI LIBERTÀ indicati con gdl

   - l'ampiezza è un po' diversa per ciascun differente valore dei gdl e si applicano differenti t-score per ciascun valore degli stessi

3. Nel caso di inferenza sulla media della popolazione: i gdl sono pari a gdl = n -1

   cioè un numero in meno rispetto all'ampiezza campionaria

4. Presenta aree sulle code più grandi + è più dispersa rispetto alla distribuzione normale standardizzata

   - più elevato è il valore dei gdl, più la distribuzione tenderà ad assomigliare a una normale standardizzata

   - quando i gdl sono PARI o MAGGIORI a 30, le 2 distribuzioni sono sostanzialmente identiche (vedi figura 5.4)

5. t-score moltiplicato per l'errore standard stimato = margine di errore per un intervallo di confidenza per la media

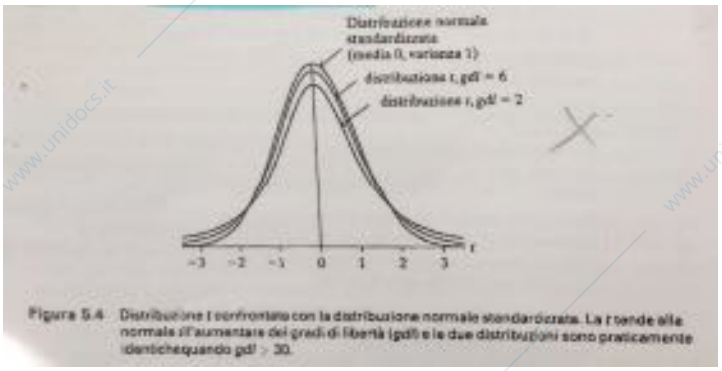
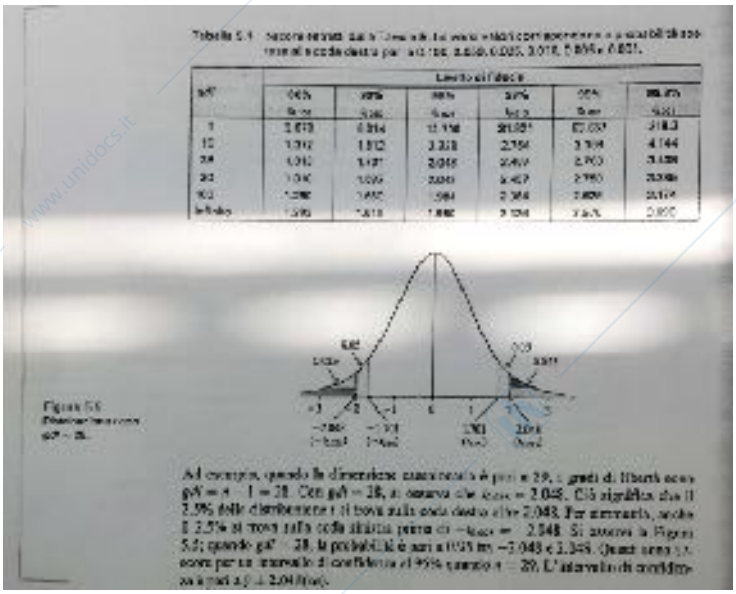



Figura 5.4 Distribuzione t confrontata con la distribuzione normale standardizzata. La t tende alla normale all'aumentare dei gradi di libertà (gdl) e le due distribuzioni sono praticamente identichequando gdl > 30.

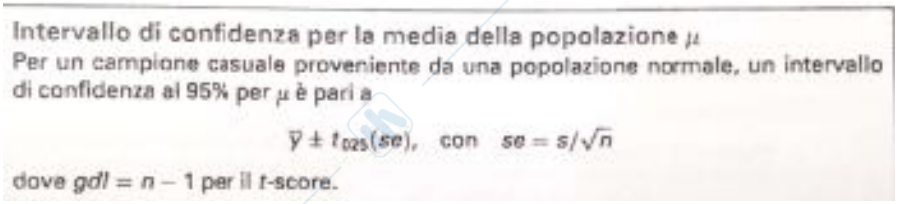
Tavola B: riporta t-score della distribuzione t per diverse probabilità sottese alla coda destra

Tabella 5.1 è un estratto di quella tavola: colonna nell'intestazione riporta i t-score corrispondenti a una probabilità di 0.025 sottesa alla coda destra, cioè i t-score utilizzati per il calcolo degli intervalli di confidenza al 95%



I T-SCORE NELL'INTERVALLO DI CONFIDENZA PER UNA MEDIA

Gli intervalli di confidenza per una media assomigliano a quelli costruiti per le proporzioni: fanno ricorso alla distribuzione t piuttosto che alla normale standardizzata



**Intervallo di confidenza per la media della popolazione $\mu$**

Per un campione casuale proveniente da una popolazione normale, un intervallo di confidenza al 95% per $\mu$ è pari a

$$\bar{y} \pm t_{.025}(se), \quad \text{con} \quad se = s/\sqrt{n}$$

dove $gdl = n - 1$ per il *t*-score.

Intervallo di confidenza ha un margine di errore pari al t-score moltiplicato per l'errore standard stimato

- impiego di t si basa sull'assunzione che la popolazione di distribuisca normalmente VEDI ESEMPIO 5.5 PAG 125

EFFETTO DELLA DIMENSIONE CAMPIONARIA SUL LIVELLO DI FIDUCIA

Utilizzata per la distribuzione t per trovare un intervallo di confidenza al 95%

Per altri livelli di fiducia si usa la stessa formula ma con un differente t-score

σ = livello di significatività

- rappresenta la probabilità che il metodo di calcolo produca un intervallo di confidenza che non contiene μ

L'intervallo di confidenza usa il t-score con una probabilità pari a alfa/2 su CIASCUNA coda

Es. intervallo di confidenza al 99%

Alfa = 1 – 0.99 = 0.01

Si ha alfa/2 = 0.005

Il t-score per uno specifico valore dei gdl è t0.005

A un livello di fiducia del 99%, il cambiamento medio nella popolazione è 0: nessun cambiamento

Anche nel caso dell'intervallo di confidenza per la media, l'ampiezza dello stesso intervallo dipende dalla dimensione campionaria n

A campioni di maggiore ampiezza corrispondono intervalli più stretti

ROBUSTEZZA RISPETTO A VIOLAZIONI DEL REQUISITO DELLA POOPLAZIONE NORMALE

I requisiti per la determinazione di un intervallo di confidenza per una media:

1. casualità della selezione delle osservazioni campionarie

2. l'assunzione di normalità per la popolazione

- la distribuzione campionaria di y lineetta è normale anche per piccoli valori di n

- lo z-score che misura il numero di errori standard con cui y lineetta si differenzia da μ HA UNA DISTRIBUZIONE NORMALE standardizzata

In realtà quando si usa l'errore standard stimato se=s/radice di n

Il numero di se con cui y lineetta si differenzia da μ, ha una distribuzione t

Metodo statistico è robusto SE permette di ottenere risultati validi anche quando il requisito non è rispettato

- l'intervallo di confidenza per una media determinato utilizzando la distribuzione t è robusto rispetto alla violazione dell'ipotesi di normalità della popolazione

ANCHE se la popolazione non è normale, gli intervalli di confidenza basati sulla distribuzione t funzionano bene, specialmente quando n è superiore a 15

- all'AUMENTARE della dimensione campionaria, l'ipotesi di normalità della popolazione diventa meno importante grazie al Teorema del Limite Centrale

La distribuzione campionaria della media campionaria è campanulare anche quando la distribuzione della popolazione non lo è

Il metodo non funziona quando i dati sono estremamente asimmetrici o contengono outlier

LA NORMALE STANDARDIZZATA È UNA SITRIBUZIONE T CON GDL = INFINITO

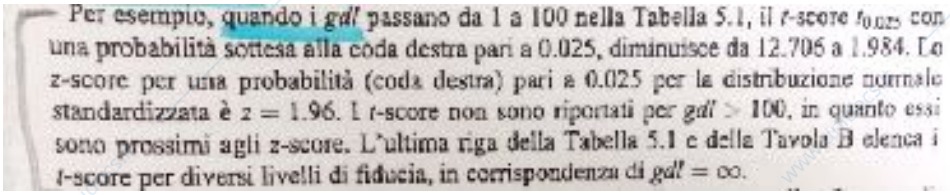
Si osservi la tabella dei t-score (TAVOLA B)

Al crescere dei gdl, ci si sposta verso la parte inferiore della tabella

- il t-score diminuisce e tende sempre più allo z-score di una distribuzione normale standardizzata

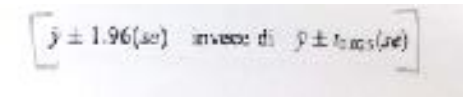
ALLORA

- distribuzione normale standardizzata è una: distribuzione t con gdl = infinito


Per esempio, quando i *gdl* passano da 1 a 100 nella Tabella 5.1, il *t*-score $t_{0.025}$ con una probabilità sottesa alla coda destra pari a 0.025, diminuisce da 12.706 a 1.984. Lo z-score per una probabilità (coda destra) pari a 0.025 per la distribuzione normale standardizzata è $z = 1.96$. I *t*-score non sono riportati per $gdl > 100$, in quanto essi sono prossimi agli z-score. L'ultima riga della Tabella 5.1 e della Tavola B elenca i *t*-score per diversi livelli di fiducia, in corrispondenza di $gdl = \infty$.

Per dei valori dei gdl più grandi di quelli mostrati nella Tavola B, si può usare uno z-score per approssimare il t-score

Per un intervallo di confidenza al 95%, quindi, si userà


$\bar{y} \pm 1.96(se)$ invece di $\bar{y} \pm t_{0.05}(se)$

Non si otterrà lo STESSO risultato

Perché al crescere di n (e quindi dei gdl), la distribuzione t assomiglia sempre più alla distribuzione normale standardizzata?

Perché s è sempre più preciso come stima puntuale di σ nell'approssimare, con la formula se = s/ radice di n, il vero errore standard

σ / radice di n

L'errore campionario aggiuntivo che si ha nel caso di piccoli campioni, viene tenuto in conto dalla distribuzione campionaria t essendo essa più dispesa della normale standardizzata

UN AVVERTIMENTO SULL'UTILIZZO DEL SOFTWARE

Il semplice fatto di ottenere risultati in una schermata di output, non significa che essi siano corretti o che i requisiti alla base dell'applicazione di un certo metodo siano soddisfatti

## 5.4 SCELTA DELLA DIMENSIONE CAMPIONARIA

Il margine di errore per un intervallo di confidenza, dipende dall'ERRORE STANDARD della stima puntuale

Così le basi della adeguatezza del metodo per fare inferenza risiedono nelle formule per il calcolo dell'errore standard

PRIMA di iniziare la raccolta dei dati, nella maggiore parte degli studi si cera di determinare la dimensione campionaria che permetterà di ottenere un determinato grado di precisione nella stima

Una misura importante di cui tenere conto è: VALORE n per il quale un intervallo di confidenza per il parametro di interesse ha un margine di errore corrispondente a un certo valore

Gli elementi chiave per determinare l'AMPIEZZA CAMPIONARIA sono i seguenti:

1. MARGINE DI ERRORE: dipende direttamente dall'errore standard della distribuzione campionaria dello stimatore puntuale

2. ERRORE STANDARD: dipende dalla dimensione campionaria

DIMENSIONE CAMPIONARIA PER LA STIMA DI PROPORZIONI

Determinare la dimensione campionaria:

a- STABILIRE QUALE deve essere il margine di errore

- in alcune indagini, avere una stima altamente precisa NON è così importante come per le altre indagini

b- POI si deve specificare la PROBABILITÀ con la quale si ottiene quel margine di errore

- questa probabilità deve essere stabilita preliminarmente in modo che, con qualsiasi ampiezza campionaria, l'errore non sia maggiore di 0.04 con una certa probabilità

Come determinare l'ampiezza campionaria per stimare una proporzione della popolazione

VEDI ESEMPIO 5.6 PAG 128

FORMULA PER LA DETERMINAZIONE DELLA DIMENSIONE CAMPIONARIA

Formula generale per determinare la dimensione campionaria

- M = margine di errore desiderato

- z-score generico determinato sulla base della probabilità che l'errore non sia più grande di M

**Dimensione campionaria per stimare una proporzione dalla popolazione π**

L'ampiezza $n$ di un campione casuale che ha un margine di errore pari a $M$ nella stima di $\pi$ attraverso la proporzione campionaria $\hat{\pi}$ è pari a

$$n = \hat{\pi}(1-\hat{\pi})\left(\frac{z}{M}\right)^2$$

Lo z-score è lo stesso scelto per definire un intervallo di confidenza con un livello di fiducia prestabilito (ad esempio, $z = 1.96$ per il livello 0.95). È necessario stabilire il valore di $\pi$ o accettare l'approccio prudenziale che fissa $\hat{\pi} = 0.50$.

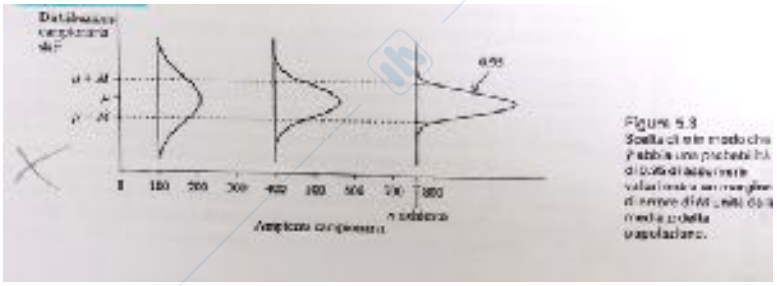
Vedi es. pag 131

DIMENSIONE CAMPIONARIA PER STIMARE LE MEDIE

Un analogo risultato vale per la stima di una media della popolazione μ

- vogliamo determinare quando deve essere grande n in modo che la distribuzione campionaria y lineetta abbia un margine di errore pari a M: la distribuzione campionaria tende a restringersi al crescere di n fino a che, in corrispondenza del valore di n richiesto, il 95% della distribuzione si trova all'interno del margine di errore prescelto

UTILIZZANDO la distribuzione campionaria normale per grandi campionari di y lineetta si ottiene:



**Dimensione campionaria per stimare una media della popolazione μ**

La dimensione $n$ del campione casuale per stimare $M$ tramite la media campionaria $\bar{y}$ con un margine di errore $M$ è:

$$n = \sigma^2\left(\frac{z}{M}\right)^2$$

Lo z-score è quello scelto per un intervallo di confidenza con un certo livello di fiducia prestabilito (ad esempio, $z = 1.96$ per il livello 0.95). È necessario conoscere la deviazione standard $\sigma$.

MAGGIORE variabilità della distribuzione della popolazione (misurata dalla sua deviazione standard σ), MAGGIORE la dimensione campionaria necessaria per raggiungere un certo margine di errore

- POCA VARIABILITÀ SE meno dati rispetto a quelli necessari nel caso di alta eterogeneità delle osservazioni

Pratica: il valore di σ non è noto

QUINDI necessità di sostituirlo con un valore appropriato

COMPLICAZIONE: non conoscendo σ, è necessario usare, per fare inferenza, la distribuzione t piuttosto che la normale standardizzata

MA se non si conosce n, non si sanno neppure i gradi di libertà e i t-score a essi associati

TUTTAVIA il t-score è prossimo allo z-score (tranne quando gdl piccoli): QUINDI non è necessario di preoccuparsi di tale complicazione

L'ERRORE DI APPROSSIMAZIONE nel sostituire t-score con uno z-score è molto piccolo

VEDI ESEMPIO 5.7 PAG 132

ULTERIORI CONSIDERAZIONI SULLA DETERMINAZIONE DELLA DIMENSIONE CAMPIONARIA

La dimensione campionaria dipende dalla PRECISIONE e dal LIVELLO DI FIDUCIA desiderati

- PRECISIONE: margine di errore

- FIDUCIA: probabilità che l'intervallo di confidenza contenga il vero valore del parametro

- dipende anche dalla VARIABILITÀ della POPOLAZIONE

Altre considerazioni in merito alla dimensione campionaria

**COMPLESSITÀ DELL'ANALISI PIANIFICATA**: più complessa è l'analisi (es. nel caso studino più variabili), più grande dovrà essere il campione

- per studiare una singola variabile attraverso una media, un campione relativamente piccolo sarà sufficiente

Prevedere confronti tra diversi gruppi attraverso metodi multivariati richiede un campione più ampio

**TEMPO, DENARO E ALTRE RISORSE:** i campioni grandi sono molto più costosi e richiedono molto tempo per l'indagine e molte più risorse di quelle disponibili

Nessuna formula è in grado di fornire un'appropriata dimensione campionaria!

SE da un lato la dimensione campionaria è un aspetto importante, la sua scelta dipende dalle risorse e dalla pianificazione della ricerca

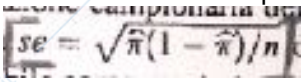
COSA FARE SE SI PUÒ DISPORRE SOLO DI UN PICCOLO CAMPIONE?

Talvolta, non è possibile selezionare un campione della dimensione desiderata: in che modo è compromessa la validità dei metodi di stima per gli intervalli di confidenza, se n è più piccolo di quanto ci si era aspettato?

I METODI per la media basati sulla distribuzione t possono essere usati per qualsiasi valore di n

- quando n PICCOLO: si ha la necessità di essere prudenti e di controllare la presenza di outlier

Formula per la determinazione dell'intervallo di confidenza per una popolazione: richiede che si disponga di almeno 15 osservazioni per ciascuna modalità

SE NON ACCADE, la distribuzione campionaria della proporzione campionaria non tende alla normale e la stima 

del vero errore standard $\sigma_{\tilde{x}} = \sqrt{\pi(1-\pi)/n} = \sqrt{0.50(0.50)/n} = 0.50/\sqrt{n}.$

può risultare poco attendibile come mostrato nel PROSSIMO ESEMPIO

VEDI ESEMPIO 5.8 PAG 135

## 5.5 INTERVALLI DI CONFIDENZA PER LA MEDIANA E ALTRI PARAMETRI

INEFFICIENZA DELLA MEDIANA CAMPIONARIA PER DATI DISTRIBUITI NORMALMENTE

Quando la distribuzione della popolazione è normale e il campione è casuale

ALLORA l'errore standard della mediana campionaria viene calcolato impiegando una formula simile a quella che si usa per la media campionaria

ERRORE STANDARD è pari a 1.25 σ / radice di n

MEDIANA della popolazione per una distribuzione normale, è pari alla MEDIA della popolazione μ

DI CONSEGUENZA la MEDIANA CAMPIONARIA e la MEDIA CAMPIONARIA sono entrambe stime puntuali dello stesso valore

MEDIANA CAMPIONARIA: non è così efficace come la media campionaria, perché il suo errore standard è più grande del 25%

- QUANDO la distribuzione della popolazione è approssimativamente normale: MEDIA CAMPIONARIA è uno dei migliori stimatori del centro di una distribuzione

Questa è una delle ragioni per cui nell'inferenza statistica la media è utilizzata più frequentemente rispetto alla mediana

QUANDO la distribuzione della popolazione è fortemente asimmetrica,

la MEDIANA della popolazione è spesso una misura di sintesi più utile rispetto alla media della popolazione

MEDIA CAMPIONARIA: si usa per stimare la mediana della popolazione

Formula per il calcolo dell'errore standard

1.25 σ / radice di n

È valida SOLO quando: distribuzione della popolazione è approssimativamente normale

INTERVALLO DI CONFIDENZA DELLA MEDIANA PER GRANDI CAMPIONI

L'intervallo di confidenza per la MEDIANA: valido nel caso di grandi campioni (n almeno 20-30 unità)

- non richiede assunzioni a parte quella di essere CONTINUA

0.50 = probabilità pi greco che un'osservazione selezionata casualmente si trovi prima della mediana

Così per un campione casuale di dimensione n, la probabilità campionaria che si trova prima della mediana ha

media 0.50

ed errore standard $\sqrt{\pi(1-\pi)/n}$

Si ha un probabilità di circa 0.95 che la proporzione campionaria di osservazioni che ricadono al di SOTTO della mediana sia entro:

2 errori standard o 1/radice di n da 0.50

NUMERO CAMPIONARIO di osservazioni che si trova PRIMA DELLA MEDIANA è n volte tale proporzione campionaria

Perciò

0.95 = probabilità che il numero di osservazioni che cade PRIMA della mediana sia entro n (1 /radice di n)

0.95 = probabilità che il numero di osservazioni che si trova OLTRE la mediana sia entro n (1/ radice di n) = radice di n dalla metà del campione

Per un campione ordinato di dimensione n, la

MEDIANA = misura centrale

- la posizione è (n+1)/2

- estremo inferiore: l'osservazione con la posizione (n+1)/2 – radice di n = è l'estremo inferiore dell'intervallo di confidenza al 95% per la mediana

- stremo superiore: l'osservazione con la posizione (n+1)/2 + radice di n = rappresenta l'estremo superiore di tale intervallo

VEDI ESEMPIO 5.9 PAG 136

IL BOOTSTRAP

Per alcuni parametri, non è possibile scrivere una formula per l'intervallo di confidenza che funzioni bene indipendentemente dalla distribuzione della popolazione o dalla dimensione campionaria o dalla tecnica di campionamento

BOOTSTRAP = metodo che considera la distribuzione campionaria come se essa fosse la vera distribuzione della popolazione

- vengono campionate n osservazioni DA questa distribuzione, dove $\hat{\pi}$ ciascuno degli n dati originati ha probabilità 1/n di essere selezionato

- per questo "nuovo" campione di dimensione n, si può costruire la stima puntuale del parametro

CAPITOLO 6: **INFERENZA STATISTICA - TEST DI SIGNIFICATIVITA**

*IPOTESI*: in statistica, un'ipotesi è un'affermazione sulla popolazione. Solitamente è la previsione che un parametro impiegato per descrivere alcune caratteristiche di una variabile assuma un particolare valore numerico o ricada in un certo intervallo di valori.

Un test di significatività utilizza i dati campionari per riassumere le evidenze empiriche riferibili a una certa ipotesi (confrontando le stime puntuali dei parametri con valori previsti dalle ipotesi)

Esempio 6.1 -> verifica sulla *discriminazione di genere* nella selezione dei manager

## 6.1 LE CINQUE PARTI DI UN TEST DI SIGNIFICATIVITA

Ogni test (verifica di ipotesi) è formata da 5 parti:

- *Assunzioni*
- *Ipotesi*
- *Test*
- *P-valore*
- *Conclusioni*

### *1. ASSUNZIONI*

Ciascun test si basa su ***determinate assunzioni*** che riguardano i seguenti aspetti:

- *Tipo di dati* -> ciascun test si applica a dati quantitativi o dati categoriali
- *Casualizzazione* (o randomizzazione)-> un test assume che i dati siano ottenuti casualmente
- *Distribuzione della popolazione* -> si assume che la variabile abbia una particolare distribuzione
- *Dimensione campionaria* -> la validità di molti test cresce all'aumentare della dimensione campionaria

### *2. IPOTESI*

IPOTESI NULLA: è un'affermazione su un particolare valore assunto dal parametro

IPOTESI ALTERNATIVA: attesta che un parametro prende valori in qualsiasi altro intervallo di valori. Solitamente il valore posto nell'ipotesi nulla corrisponde, in un certo qual modo, all'ASSENZA di EFFETTO. I valori nell'ipotesi alternativa, rappresentano un effetto di qualunque tipo.

NOTAZIONE PER LE IPOTESI:

**$H_0$** rappresenta l'ipotesi nulla

**$H_a$** rappresenta l'ipotesi alternativa o ipotesi di ricerca

Un test di significatività valuta l'evidenza campionaria dell'ipotesi nulla H0:

- il test investiga se i dati contraddicono H0 in maniera da suggerire che Ha sia vera

- DIMOSTRAZIONE PER CONTRADDIZIONE: presuppone che l'ipotesi nulla sia vera; sulla base di questo presupposto, se si riscontra che i dati osservati sono poco verosimili, si propende per l'ipotesi alternativa

### *3. TEST*

**TEST**: la statistica test permette di valutare quanto questa stima ricada vicino al valore del parametro ipotizzato con $H_0$. Spesso questa distanza è espressa nei termini di QUANTI ERRORI STANDARD separano la stima del valore del parametro ipotizzato sotto $H_0$.

### *4. P-VALORE*

**P-VALORE**: per interpretare il valore della statistica test si determina la probabilità che riassuma le evidenze statistiche che contraddicono $H_0$ 8 probabilità: distribuzione campionaria della statistica test sotto l'ipotesi che $H_0$ sia vera).

➔ Si tratta di valutare la probabilità dei possibili valori della statistica test che forniscono almeno la stessa evidenza contro $H_0$ del valore effettivamente osservato del test. Questa probabilità è definita come P-VALORE.

DEFINIZIONE: il **P-VALORE** è la probabilità che la statistica test sia pari al valore osservato o a uno più grande nella direzione prevista da $H_a$; è determinato presumendo che $H_0$ sia vera. Il P-VALORE è indicato con P.

Un piccolo P-valore significa che i dati osservati dovrebbero essere considerati come insoliti se $H_0$ fosse vera. *Quanto più piccolo è il P-valore, tanto più forte è l'evidenza statistica contro $H_0$.*

### *5. CONCLUSIONI*

Il P-valore sintetizza le evidenze empiriche contro $H_0$. Se il P-valore è sufficientemente piccolo, diciamo che rifiutiamo $H_0$ e accettiamo $H_a$. La maggior parte degli studi che si osservino dei P-valori molto piccoli (es: $P \leq 0.05$), per rifiutare $H_0$. In alcune situazioni, si dice che i risultati sono *SIGNIFICATIVI AL LIVELLO* 0.05. Questo significa che se $H_0$ fosse vera, la possibilità di ottenere risultati cosi estremi come quelli ottenuti con i dati campionari non dovrebbe essere più grande di 0.05.

Prendere una decisione rifiutando o meno un'ipotesi nulla è una parte opzionale di un test di significatività.

## 6.2 TEST DI SIGNIFICATIVITÀ PER UNA MEDIA

LE CINQUE PARTI DI UN TEST DI SIGNIFICATIVITA

I test di significatività solitamente fanno riferimento alla media della popolazione μ. Le cinque parti di un test per una media sono le seguenti:

1. **Assunzioni** -> il test assume che i dati siano ottenuti utilizzando la randomizzazione come nel caso di un campione casuale . si ipotizza che la variabile quantitativa abbia una distribuzione normale nella popolazione.

2. **Ipotesi** -> l'ipotesi nulla su una media della popolazione assume una forma

   $H_0 : \mu = \mu_0$

   dove $\mu_0$ è un valore particolare per la media della popolazione.

   Il valore ipotizzato per μ in $H_0$ è un valore singolo. Questa ipotesi solitamente si riferisce *all'assenza di effetto* o *all'assenza di cambiamento* ( cioè l'ipotesi nulla $H_0 : \mu = 0$) rispetto a una qualche situazione standard.

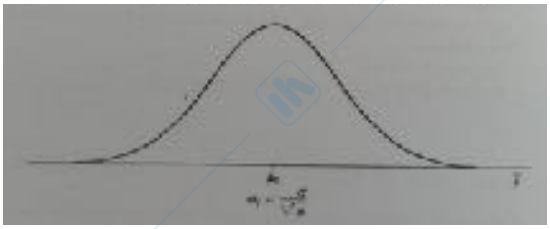
   L'IPOTESI bidirezionale

   $H_a : \mu \neq \mu_0$ cioè $H_a : \mu \neq 0$

   è un'ipotesi alternativa che contiene valori sia inferiori sia superiori al valore presente in $H_0$.

3. **Test** -> la media campionaria *y* stima la media della popolazione μ. Quando la distribuzione della popolazione è normale, la distribuzione campionaria di y è normale centrata su μ. Questo risultato è approssimativamente vero, grazie al TEOREMA DEL LIMITE CENTRALE, anche quando la distribuzione della popolazione non è normale ma la dimensione del campione casuale è abbastanza grande.

Sotto l'ipotesi che $H_0 : \mu = \mu_0$ sia vera, il centro della distribuzione ampionaria di $\bar{y}$ è il valore $\mu_0$. Un valore di $\bar{y}$ che ricade in prossimità delle code della distribuzione fornisce una forte evidenza contro $H_0$ : sarebbe, infatti, piuttosto insolito osservarlo se realmente $\mu = \mu_0$. L'evidenza su $H_0$ è sintetizzata dal numero di errori standard che separa $\bar{y}$ dal valore dell'ipotesi nulla $H_0$.



Il vero ERRORE STANDARD di $\bar{y}$ è $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. sostituisco la deviazione standard campionaria s al posto della ignota deviazione standard della popolazione σ per ottenere l'errore standard stimato se = s/ √n. La statistica test è il *t-score:*

$$t = \frac{\bar{y} - \mu 0}{se} \text{ , dove se} = \frac{s}{\sqrt{n}}$$

Quanto più grande la differenza tra $\bar{y}$ e $\mu_0$ , tanto più grande sarà il valore assoluto della statistica del test. Quindi, tanto più grande è il valore di |t|, tanto più forte sarà l'evidenza contro $H_0$.

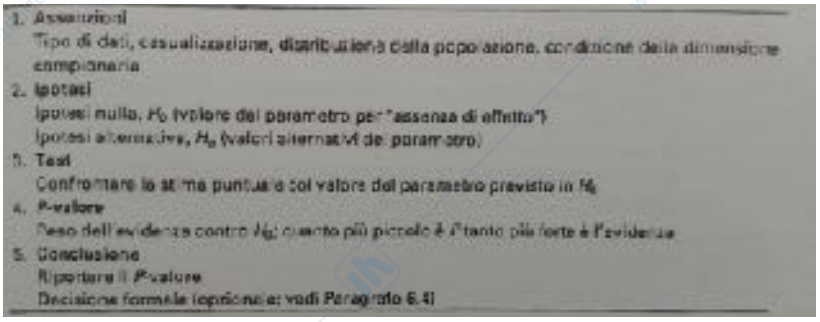
La distribuzione campionaria della statistica test t è la distribuzione t sotto l'ipotesi nulla ( assomiglia alla distribuzione normale standardizzata, poiché ha la media pari a 0. Ma risulta più dispersa, in particolare per valori di n piccoli -> è specificata dai suoi gradi di libertà, gdl = n-1).

4. **P-valore** -> la statistica test riassume quanto i dati osservati sono lontano da ciò che è ipotizzato in $H_0$. Test diversi utilizzano differenti statistiche test (formule) e la loro interpretazione deriva dalla trasformazione dei loro valori in probabilità (cioè nella scala 0-1), attraverso il P-valore. Il P-valore è la probabilità, se è vera $H_0$, che la statistica test sia pari al valore osservato o a uno compreso nell'insieme dei valori più estremi che forniscono evidenze ancor più forti contro $H_0$. Nel caso

   $H_0 : \mu \neq \mu_0$ , i valori più estremi di t sono quelli in corrispondenza delle code della distribuzione t. così il P-valore è la probabilità, a due code, che la statistica test t sia

   almeno pari, in valore assoluto, al valore ottenuto del test. Questa è anche la probabilità che $\bar{y}$ ricada oltre una certa distanza da $\mu_0$ in entrambe le direzioni.

5. **Conclusioni** -> per finire, lo studio dovrebbe interpretare il P-valore rispetto al contesto empirico dell'analisi. Quanto più piccolo risulta P, tanto più forte è l'evidenza contro $H_0$ e a favore di $H_a$




Esempio 6.2 (sul libro) posizioni conservatrici e liberali in politica (spiega questi 5 punti con un esempio IMPORTANTE LEGGI)

Corrispondenza tra test bidirezionali e gli intervalli di confidenza

Le conclusioni cui si perviene utilizzando **test di significatività bidirezionali (o, a due code)** sono coerenti con quelle cui si giunge utilizzando gli intervalli di confidenza. Se un test sostiene che un particolare valore è plausibile per il parametro, lo stesso accade per un intervallo di confidenza.

Esempio 6.3 intervallo di confidenza per la media dell'ideologia politica

Test di significatività a una coda

Una diversa ipotesi alternativa è talvolta utilizzata quando un ricercatore prevede una deviazione da $H_0$ in una particolare direzione. Essa assume la forma

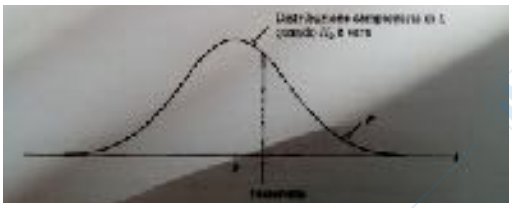
$$\mathbf{H_a : \mu > \mu_0} \qquad \text{o} \qquad \mathbf{H_a : \mu < \mu_0}$$

L'alternativa $H_a : \mu > \mu_0$ è utilizzata per valutare se $\mu$ è più grande del particolare valore $\mu_0$, mentre $H_a : \mu < \mu_0$ è utilizzata per valutare se $\mu$ è più piccolo di quel valore.

Queste ipotesi sono chiamate **UNIDIREZIONALI** o **A UNA CODA**. Al contrario, l'alternativa $H_a$ a DUE CODE (o BIDIREZIONALE) è utilizzata per valutare qualsiasi tipo di allontanamento da $H_0$. La scelta sul tipo di test deve essere fatta prima di analizzare i dati.

Per $H_a : \mu > \mu_0$ il P-valore è la probabilità (se $H_0$ è vera) che ha di osservare per il puro effetto del caso un T-SCORE oltre il T-SCORE EMPIRICO.; cioè, alla sua destra nell'insieme dei numeri reali. Quanti T-SCORE forniscono una maggiore evidenza empirica rispetto al valore osservato in favore di $H_a : \mu > \mu_0$. Così, P è pari alla probabilità sottesa alla coda destra della curva t. Un t-score pari a 0,68 corrisponde a un P= 0,25 rispetto a questa ipotesi alternativa.

Per $H_a : \mu < \mu_0$ il P-valore è la probabilità sottesa alla coda sinistra, (al di sotto) del t-score osservato. Un t-score pari a t= - 0,68 corrisponde a un P= 0,25 per questa ipotesi alternativa. Un t-score pari a 0.68 corrisponde a P= 1 – 0.25 = 0,75



Esempio 6.4 cambiamento medio del peso per ragazze anoressiche

Scelta di test bidirezionali anziché unidirezionali

Nella pratica, i test bidirezionali sono costruiti più frequentemente di quelli unidirezionali. Anche se un ricercatore prevede la direzione in cui agisce un certo effetto, i test bidirezionali possono

rilevare un effetto che agisce nella direzione opposta. Nella maggior parte degli articoli scientifici, i test di significatività utilizzano **P-valori bidirezionali** (APPROCCIO *OGGETTIVO* ALLA RICERCA).

Utilizzando P-valori bidirezionali i ricercatori *evitano di alimentare il sospetto che scelgano $H_a$ avendo visto la direzione verso la quale tendono i dati empirici*. Un simile approccio non sarebbe eticamente corretto.

Gli intervalli di confidenza sono bidirezionali, ottenuti aggiungendo e sottraendo una certa quantità alla stima puntuale. Nella pratica gli intervalli unidirezionali sono utilizzati *RARAMENTE*.

Nella scelta di un $H_a$ uni o bidirezionale si deve considerare il **contesto** in cui si opera (es: "verifica se la media è cambiata" -> alternativa bidirezionale….. " verifica se la media è aumentata" -> alternativa unidirezionale). Le ipotesi si riferiscono SEMPRE ai parametri della popolazione e non a statistiche campionarie. Non si esprime **mai** un'ipotesi utilizzando la notazione usata per le statistiche campionarie.

Il livello α: uso del P-valore per prendere una decisione

Un test di significatività valuta la forza dell'evidenza empirica contraria all'ipotesi nulla, $H_0$. Se $H_0$ è vera e calcolo il P-valore, analizziamo quanto i dati possono essere considerati INSOLITI. Se il P-valore è piccolo, i dati contraddicono $H_0$ e supportano $H_a$.

***Quando $H_0$ è vera, il P-valore ha all'incirca la stessa probabilità di ricadere ovunque tra i valori 0 e 1. Al contrario quando $H_0$ è FALSA è molto più probabile che il P-valore sia vicino a 0 piuttosto che a 1.***

**LIVELLO α** : *il livello α è un valore SOGLIA; rifiutiamo $H_0$ se il P-valore è inferiore o uguale a esso. Il livello α è anche chiamato **livello di significatività**. Nella pratica, i più comuni livelli α sono 0,05 e 0,01.*

La scelta di α riflette quanto l'analista dei dati vuole essere prudente nel ricavare le proprie conclusioni inferenziali. Quanto più piccolo è il livello α, tanto più forte deve essere l'evidenza empirica per rifiutare $H_0$. È necessario scegliere α prima di analizzare i dati.

Esempio 6.5 decisioni inferenziali riferite agli esempi presentati

Robustezza rispetto a violazioni dell'assunzione di normalità

Il test t per una media ipotizza che la distribuzione della popolazione sia normale. Questo assicura che la distribuzione campionaria della media campionaria y sia normale e, dopo che *s* viene utilizzato per stimare σ nel calcolo dello *s,* che il test t abbia una distribuzione t. al crescere dell'ampiezza campionaria, questa assunzione di normalità diventa meno importante.

Un metodo statistico è ROBUSTO se4 i suoi risultati sono attendibili anche quando un'assunzione è violata. La **teoria statistica** ha dimostrato che l'inferenza bilaterale per una media che utilizza la distribuzione t è *robusta* rispetto alle violazioni dell'assunzione di normalità della popolazione. Il test non funziona bene nel caso di verifica di ipotesi unilaterale con piccoli n specie quando la distribuzione della popolazione è fortemente asimmetrica.

## 6.4 TEST DI SIGNIFICATIVITÀ PER UNA POPOLAZIONE

Per una variabile categoriale, il parametro di interesse è la proporzione di una categoria nella popolazione. Il test per una popolazione, come il test per la media, trova un P-valore per un test statistico misurando il numero di errori standard tra una stima puntuale e il valor ipotizzato per il parametro in $H_0$.

Le cinque fasi di un test di significatività per una proporzione

1. **Assunzioni** -> anche in questo test si ipotizza che i dati sono stati ottenuti casualmente, come avviene nel caso della selezione di un campione casuale. La dimensione

campionaria deve essere sufficientemente ampia in modo che la distribuzione campionaria di π sia approssimativamente normale.

2. **Ipotesi** -> ipotesi nulla ssume la forma => **$H_0 : \pi = \pi_0$** cioè **$H_0 : \pi = 0.50$**

   In questo caso $\pi_0$ indica un particolare valore per la proporzione compreso tra 0 e 1. La più comune ipotesi alternativa è : **$H_0 : \pi \neq \pi_0$** cioè **$H_a : \pi \neq 0.50$**

   Questa ***alternativa bidirezionale*** afferma che la proporzione della popolazione differisce dal valore $H_0$. Le alternative *UNIDIREZIONALI*: **$H_a : \pi > \pi_0$** e **$H_a : \pi < \pi_0$** si applicano quando il ricercatore prevede un allontanamento in un adirezione determinata dal valore in $H_0$.

3. **Test** -> la distribuzione campionaria $\bar{\pi}$ ha media π ed errore standard $\sigma_{\pi} = \sqrt{\frac{\pi(1-\pi)}{n}}$.

   Quando $H_0$ è vera, $\pi = \pi_0$ , così che l'errore standard è $se_0 = \sqrt{\pi 0(1-\pi 0)/n}$ . Si utilizza la notazione $se_0$ per indicare che questo è l'errore standard sotto l'ipotesi che $H_0$ sia vera. La statistica test è

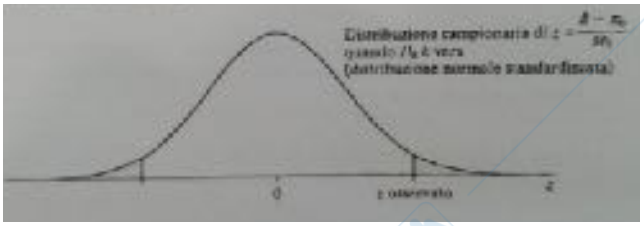
   $$z = \frac{\pi - \pi 0}{se0}, \qquad \text{dove} \qquad se_0 = \sqrt{\frac{\pi 0(1-\pi 0)}{n}}$$

   Questa formula misura il numero di errori standard che separano la proporzione campionaria $\bar{\pi}$ da $\pi_0$. Nel caso di GRANDI CAMPIONAMENTI, se $H_0$ è vera, la distibuzione camionaria del test statistico z è la distribuzione normale standardizzata.

   FORMA DI UN TEST: $z = \dfrac{stima\ del\ parametro - valore\ del\ parametro\ nell'ipotesi\ nulla}{errore\ standard\ dello\ stimatore}$

4. **P-valore** -> è la probabilità a una o due code; si fa riferimento alla distribuzione normale pittosto che alla distribuzione t. per $H_A : \pi \neq \pi_0$, P è la probabilitèà a due code.

   Questa probabilità è il doppio della probabilità sottesa alla singola coda oltre il valore osservato z.



   Nel caso di alternative unidirezionali, il P-valore è la probabilità a una coda. Poiche $H_a : \pi > \pi_0$ prevede che la proporzione della popolazione sia PIU GRANDE rispetto al valore ipotizzato in $H_0$, il suo P-valore è la probabilità **oltre** (a destra del) il valore osservato z. Per $H_a : \pi < \pi_0$, il P-valore è la probabilita **prima** (a sinistra) del valore osservato per z.

5. **Conclusioni** -> quanto più piccolo è il P-valore, tanto più chiaramente i dati contaddistinguono $H_0$ e favoriscono $H_a$ (x prendere una decisone -> rifiutiamo $H_0$ se $P \leq \alpha$ per un prestabilito livello α)

Esempio 6.6 ridurre i servizi o aumentare le tasse (guarda libro, è utile)

MANCA

## 6.4 DECISIONI E TIPI DI ERRORE NELLE VERIFICHE DI IPOTESI

Quando occorre decidere se l'evidenza empirica contro $H_0$ è abbastanza forte per rifiutarla, si è visto che rifiutiamo $H_0$ se $P \leq \alpha$ per un prefissato valore $\alpha = 0.05$. la tabella riassume le due possibili conclusioni per un livello $\alpha = 0.05$

| P-VALORE | CONCLUSIONE | |
|---|---|---|
| | $H_0$ | $H_a$ |
| $P \leq 0.05$<br>$P > 0.05$ | Rifiuto<br>Non rifiuto | Accetto<br>Non accetto |

Errori di I e di II tipo per le decisioni

***Quando $H_0$ è vera, si verifica un ERRORE DI I TIPO se $H_0$ è rifiutata***

***Quando $H_0$ è falsa, si verifica un ERRORE DEL II TIPO se $H_0$ non è rifiutata***

Ci sono 4 possibili risultati:

| | DECISIONI | |
|---|---|---|
| Condizione per $H_0$ $H_0$ vera<br>$H_0$ falsa | ERRORE I TIPO<br>DECISIONE CORRETTA | DECISIONE CORRETTA<br>ERRORE II TIPO |

Regioni di rifiuto

L'insieme dei valori della statistica test per il quale il test porta al rifiuto di $H_0$ è chiamato ***REGIONE DI RIFIUTO***. Ad esempio, la regione di rifiuto per un test al livello $\alpha = 0.05$ è costituito dall'insieme dei valori del test statistico per i quali $P \leq 0.05$

Il livello $\alpha$ è la probabilità dell'errore di I tipo

Troviamo la probabilità dell'errore di I tipo quando $H_0$ è vera. Supponiamo che $\alpha = 0.05$. la probabilità di rifiutare $H_0$ è esattamente 0.05 perchè la probabilità di osservare valori compresi in questa regione di rifiuto di una curva normale standardizzata è 0.05. questo è esattamente il livello $\alpha$.

-> La probabilità di un errore di I tipo è il livello $\alpha$ del test.

Con $\alpha = 0.05$, se $H_0$ è vera, la probabilità di commettere un errore di I tipo e quindi rifiutare l'ipotesi (vera) $H_0$ è pari a 0.05.

Quanto più importanti sono le conseguenze di un errore di I tipo, tanto più piccolo dovrebbe essere $\alpha$. Il valore 0.05 del livello $\alpha$ è il valore scelto più frequente, proprio come la probabilità di errore di 0.05 è il valore più comune nella costruzione degli intervalli di confidenza (cioè, una fiducia del 95%).

Quando P (errore del I tipo) diminuisce, la P ( errore del II tipo) aumenta

Quando stabiliamo un $\alpha$ più piccolo in un test di significatività, abbiamo bisogno di un P-valore più piccolo per rifiutare $H_0$. Diventa più difficile rifiutare $H_0$. Ma ciò significa che questo sarà più difficile anche se $H_0$ è falsa.

➔ ***Quanto più piccola è P(errore del I tipo) che noi fissiamo, tanto più grande diventa la P ( errore del II tipo), cioè la probabilità di non riuscire a rifiutare $H_0$ quando è falsa.***

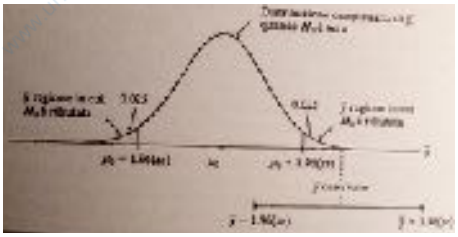
Nel paragrafo 6.6 verrà mostrato come la P(errore del II tipo) dipende solo da quanto lontano è il vero valore del parametro da quello ipotizzato in $H_0$. -> quanto più lontano il parametro dista da quello ipotizzato, tanto meno probabilmente i risultati campionari ci faranno commettere un errore del II tipo.

Per una fissata P (errore del I tipo), la P (errore del II tipo) dipende anche dalla dimensione campionaria. Tanto più grande è la dimensione campionaria con tanta maggiore probabilità rifiuteremo una falsa $H_0$. *Per mantenere a un basso livello sia la P (errore del I tipo) sia P(errore del II tipo) è necessario utilizzare un campione di dimensioni molto elevate.*

La P (errore del I tipo) può essere piuttosto elevata quando la dimensione campionaria è piccola a meno che il parametro non cada molto lontano dal valore ipotizzato sotto $H_0$.

Equivalenza tra intervalli di confidenza e decisioni dei test

Consideriamo un test per grandi campioni:



Quando P < 0.05, $H_0$ viene rifiutata al livello $\alpha = 0.05$. questo accade quando la statistica test $t = (y - \mu_0)/se$ più grande di 1.96 (circa in valore assoluto (quando *n* è grande)-> y dista più di 1.96(*se*), non contiene il valore dell'ipotesi nulla $\mu_0$.

➔ *Nel verificare $H_0 : \mu = \mu_0$ contro $H_a : \mu \neq \mu_0$ al livello a di 0.05: l'intervallo di confidenza al 95% per $\mu$ non contiene $\mu_0$. L'intervallo di confidenza al 95% per $\mu$ consiste di quei valori $\mu_0$ per i quali non rifiutiamo $H_0 : \mu = \mu_0$ al livello a dello 0.05.*

Prendere decisioni oppure riportare il P-valore

APPROCCIO FORMALE ALLA VERIFICA DELLE IPOTESI-> Formulare un ipotesi nulla e alternativa, selezionare un livello α per la P(errore del I tipo), determinare la regione di rifiuto dei valori del test statistico che fornisce sufficiente evidenza per rifiutare $H_0$ e, quindi, prendere la decisione se rifiutare $H_0$. Con questo approccio, non è strettamente necessario trovare un P-valore. La scelta del livello α determina la regione di rifiuto, che insieme al test statistico determina la decisione.

## 6.5 LIMITI DEI TEST DI SIGNIFICATIVITÀ

Un test di significatività viene applicato per fare inferenza sul fatto che un parametro differisca o meno da un valore ipotizzato in $H_0$ nonchè per valutare la direzione dell'eventuale allontanamento.

Significatività statistica e significatività pratica

Un piccolo P-valore, come P=0.001, è altamente significativo da un punto di vista statistico. Fornisce una forte evidenza contro $H_0$. Ciò non implica automaticamente, tuttavia, che esso sia anche un importante risultato sotto l'aspetto pratico. Un piccolo P-valore significa solamente che se $H_0$ fosse vera, i dati osservati dovrebbero essere insoliti. Ciò non vuol dire che il vero valore del parametro sia lontano da $H_0$ in termini pratici.

Esempio 6.7 ideologia politica media della popolazione statunitense

I test di significatività sono meno utili degli intervalli di confidenza

Le ipotesi nulle che fanno riferimento a singoli valori raramente sono vere (raramente il parametro è esattamente uguale al valore indicato in $H_0$).

**PREFERENZA: costruire intervalli di confidenza** per i parametri piuttosto che realizzare solo test di significatività

**Un test** indica solamente se il particolare valore in $H_0$ sia o no verosimile, non ci dice quali altri potenziali valori siano plausibili.

**L'intervallo di confidenza** mostra l'intero insieme di valori credibili. Mostra la misura in cui $H_0$ può essere falsa mostrando se i valori dell'intervallo sono lontani da quello ipotizzato in $H_0$. Esso ci aiuto a determinare se il rifiuto di $H_0$ abbia o meno un'importanza pratica.

Errate interpretazioni dei test di significatività e dei P-valori

Non è appropriato accettare $H_0$. La significatività statistica non implica quella pratica. Altre possibili errate interpretazioni dei test di significatività.

- *È fuorviante riportare i risultati solo se questi sono statisticamente significativi* (alcune riviste scientifiche pubblicano solo i risultati di uno studio solo se il P-valore ≤ 0.05)
- *Al cui test possono essere statisticamente significativi solo per caso* (Bisogna essere diffidenti di studi che riportano solo risultati significativi: potrebbero essere dovuti alla normale variabilità campionaria)
- *Non è corretto interpretare il P-valore come la probabilità che $H_0$ sia vera* (I metodi statistici classici non calcolano probabilità riferite a parametri ma a statistiche che variano casualmente da campione a campione -> Le statistiche hanno distribuzioni campionarie, i parametri no!)
- *Gli effetti veri possono essere più piccoli di quelli stimati* (Anche se un risultato statisticamente significativo rappresenta un effetto reale, il vero effetto può essere più piccolo di quello stimato)

Esempio 6.8 Molte "scoperte" mediche sono in realtà errori del I tipo

## 6.6 CALCOLARE LA P (ERRORE DEL II TIPO)

L'errore del II tipo deriva dal non rifiutare $H_0$ quando è falsa. Questa probabilità può assumere più di un valore perchè $H_a$ comprende un intervallo di possibili valori. Ciascun valore in $H_a$ ha la propria P( errore del II tipo).

Esempio 6.9 verificare se l'astrologia funziona veramente (leggerlo!)

La probabilità dell'errore del II tipo cresce quando il valore del parametro si avvicina a quello ipotizzato in $H_0$. Se il parametro si trova vicino al valore di $H_0$, ci può essere una concreta possibilità di sbagliare nel non rifiutare $H_0$ quando è falsa. Allo stesso modo quanto più lontano si trova il parametro da $H_0$, tanto meno probabile sarà l'errore del II tipo.

Per un fissato livello α e un valore differente del parametro, la P (errore del II tipo) diminuisce quando la dimensione campionaria aumenta. Con più dati sarà meno probabile commettere questo tipo di errore.

I test con α più piccoli hanno P (errore del II tipo) più grandi

Quanto più piccolo è α = P (errore del I tipo) in un test, tanto più grande è la P(errore del II tipo).

La ragione per la quale valori estremante bassi per α non sono solitamente utilizzati è che la P (errore del II tipo) è troppo elevata.

P (errore del II tipo) DIMINUISCE:

- Il valore del parametro è lontano da quello ipotizzato in $H_0$

- La dimensione campionaria cresce
- La P(errore del I tipo) cresce

La potenza del test

Quando $H_0$ è falsa, si desidera che la probabilità di rifiutare $H_0$ sia elevata. La probabilità di rifiutare $H_0$ quando è falsa è chiamata *POTENZA del test*. Per un particolare valore del parametro appartenente all'intervallo definito in $H_a$,

**potenza = 1 – P(errore del II tipo)**

la potenza cresce per valori del parametro che si trovano più lontani da quello ipotizzato in $H_0$.

Le ricerche empiriche idealmente dovrebbero avere un'elevata potenza.

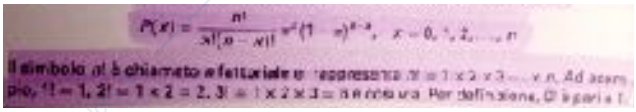
## 6.7 TEST UNA POPOLAZIONE NEL CASO DI PICCOLI CMAPIONI – LA DISTRIBUZIONE BINOMIALE

La distribuzione binomiale

Per dati categoriali, possono verificarsi le seguenti condizioni:

1. Ciascuna osservazione ricade in una delle due categorie
2. Le probabilità per le due categorie sono le stesse per ciascuna osservazione. Indichiamo le probabilità con π per la categoria 1 e (1 – π) per la categoria 2.
3. I risultati di osservazioni successive sono indipendenti. Cioè, il risultato per una osservazione non dipende dal risultato delle altre osservazioni.

La formula per le probabilità binomiali->sia π la probabilità che un'osservazione appartenga alla categoria1. Nel caso di n osservazioni indipendenti, la probabilità di x successi per la categoria 1 è:



Per i prefissati valori di π e n, sostituendo i possibili valori per x nella formula per P(x), si ottengono le probabilità dei possibili risultati. La somma delle probabilità è pari a 1.0.

Esempio 6.10 genere e scelta degli allievi per il corso di management

Proprietà di una distribuzione binomiale

La proporzione campionaria $\bar{\pi}$ è legata a x da: $\bar{\pi} = x/n$

La distribuzione binomiale ha la media e deviazione standard pari a $\mu = n\pi$, $\sigma = \sqrt{n\pi(1-\pi)}$

Esempio 6.11 quanta variabilità può avere un exit poll?

Il test binomiale -> pg 177 leggere tutto importante