Metodi statistici per la ricerca sociale

Richiami di Inferenza

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM–88 SOCIOLOGIA E RICERCA SOCIALE

Statistica e fenomeni collettivi

- Obiettivo: Studio in termini quantitativi di fenomeni collettivi
- Studio quantitativo: Informazioni espresse numericamente (percentuali, medie, ecc.)
- Lo studio di un fenomeno collettivo richiede l'osservazione di un insieme di manifestazioni individuali relative a un insieme di entità omogenee da qualche punto di vista
- Fenomeni collettivi: Reddito di un insieme di individui; Consumo di un determinato bene in un periodo di tempo fissato; Soddisfazione per un servizio; Preferenze degli elettori; Fecondità; Povertà; Performance scolastiche degli studenti; Scelte in materia di tutela della salute
- Raccogliere informazioni e elaborarle attraverso opportune metodologie statistiche al fine di studiare i fattori che esercitano la maggiore influenza sul fenomeno di interesse
- Importante: Saper leggere e interpretare documenti contenenti informazioni statistiche

I dati

Attraverso la raccolta di informazioni sono ottenute le **osservazioni** impiegate per l'analisi statistica

- La raccolta di informazioni è il cuore della scienza
- Le informazioni raccolte per studiare un determinato fenomeno (o rispondere alla domanda di ricerca di interesse) sono chiamate nel loro insieme **dati**
- Data Base: Raccolte dati contenute in archivi

Un piccolo insieme di osservazioni: La matrice dei dati

Individuo (u_i)	Età (Y_1)	Sesso (Y_2)	Titolo di studio (Y_3)	Reddito netto (Y_4)
1	56	Maschio	Diploma	30 000
2	50	Maschio	Licenza Media Inferiore	28 000
3	44	Femmina	Diploma	15 000
4	68	Femmina	Nessun titolo	9 477
5	21	Maschio	Diploma	22 000
6	26	Femmina	Laurea	20 843
7	61	Maschio	Nessun titolo	17 591
8	44	Maschio	Elementare	18 200
9	47	Maschio	Laurea	40 000
10	25	Maschio	Laurea	24 200

- Ciascuna riga contiene le osservazioni riferite a un particolare soggetto
- Ciascuna colonna contiene le osservazioni raccolte per ognuna delle caratteristiche esaminate

Fonti di dati statistici nazionali e internazionali

- ISTAT – Istituto Nazionale di Statistica (http://www.istat.it/it/)
  - ✓ Censimento della popolazione, dell'industria e servizi, dell'agricoltura (ogni 10 anni in Italia)
  - ✓ Indagine sulle Forze di lavoro (dati trimestrali)
  - ✓ FSS - Indagine multi-scopo sulle famiglie e soggetti sociali (2003, 2009)
- Eurostat – Ufficio statistico dell'Unione Europea (http://ec.europa.eu/eurostat)
  - ✓ EU-SILC – European Union Statistics on Income and Living Conditions
- Banca d'Italia, Camere di Commercio, INPS, INAIL, ACI
- Banca centrale Europea (BCE) e Banca modiale (world bank)
- Commissione statistica delle Nazioni Unite
- Organizzazione per la cooperazione e lo sviluppo economico (OECD)
- Organizzazione mondiale della sanità (WHO)
- Organizzazione delle Nazioni Unite per l'alimentazione e l'agricoltura (FAO)
- GSS – General Social Survey
  - ✓ Web site: http://www3.norc.org/GSS+Website/
  - ✓ Statistiche di sintesi: http://sda.berkeley.edu/gss

Unità Statistiche e Collettivo

- Si definisce **unità statistica** l'unità elementare su cui vengono osservate le variabili oggetto di studio
- Un insieme di unità statistiche omogenee rispetto a una o più caratteristiche costituisce un **collettivo** statistico o una **popolazione**
  - ✓ La popolazione è costituita dal totale delle unità statistiche (soggetti) di interesse in uno studio
- **Popolazione reale**: Insieme reale di unità statistiche (Tutte le unità che costituiscono la popolazione sono effettivamente osservabili)
  - ✓ Esempio: Insieme degli studenti immatricolati presso la scuola di Scienze Politiche dell'ateneo di Firenze nell'a.a. 2015/2016
- **Popolazione concettuale**: Non è possibile osservare tutte le unità che costituiscono la popolazione
  - ✓ Esempio: Insieme concettuale di tutti gli studenti che si sono o potranno immatricolarsi presso la scuola di Scienze Politiche dell'ateneo di Firenze
- Popolazione finita (composta da N unità) e popolazione infinita

Caratteri statistici

- Carattere statistico = Variabile (statistica)
- Una variabile è una qualunque caratteristica misurata su ciascuna unità
- Una variabile può assume modalità (valori) differenti in corrispondenza delle diverse unità statistiche della popolazione
- L'insieme dei valori della variabile deve essere esaustivo e i valori che la variabile può assumere devono essere non sovrapposti
- Esaustività: L'insieme dei valori della variabile deve includere tutte i possibili valori della variabile
- Non sovrapposizione: La variabile non può assumere valori diversi su una stessa unità
- Scala di misura: È importante capire come misurare i valori osservati

Classificazione delle variabili statistiche

Variabile qualitativa (categoriale)

- I valori della variabile sono espressi da parole (nomi e/o attributi)
- Opportunità di utilizzare codici
- Esempi: Sesso, Professione, Mezzo impiegato per raggiungere il luogo di studio/lavoro, Titolo di studio, Livello di soddisfazione

Variabile quantitativa

- I valori della variabile sono valori numerici (numeri)
- Si ottengono da operazioni di conteggio o di misurazione
- Esempi: Età, Peso, Altezza, Reddito, Numero di figli, Numero di anni di istruzione

N.B. Diversi metodi statistici si applicano per la sintesi e l'analisi di variabili qualitative e variabili quantitative

Classificazione delle variabili qualitative

Variabili Qualitative Sconnesse (Misurabili su scala nominale)

- Date due modalità è possibile solo dire se sono uguali o diverse
- Esempi: Sesso, Professione, Tipo scuola all'Università

Variabili Qualitative Ordinate (Misurabili su scala ordinale)

- Date due modalità è possibile solo definire un ordine
- Esempi: Titolo di studio, Livello di soddisfazione

Le variabili qualitative (nominali e ordinali) sono sempre discrete perché l'insieme delle modalità è composto da un numero finito di elementi

I valori di variabili qualitative sono anche chiamati *livelli* o *categorie*

Codifica dei valori di una variabile qualitativa

Classificazione delle variabili quantitative

Variabili Quantitative Discrete

- L'insieme delle modalità è un sottoinsieme dei numeri interi
- I possibili valori della variabile formano un insieme di numeri distinti come $0, 1, 2, 3, \ldots$
- Esempi: Età in anni compiuti, Numero di figli, Numero di anni di istruzione

Variabili Quantitative Continue

- L'insieme delle modalità è un sottoinsieme dei numeri reali
- La variabile può assumere come valore ogni possibile numero reale incluso in un continuum infinito
- Esempi: Età esatta, Peso, Altezza, Reddito

N.B. Esistono altre classificazioni

Rilevazione delle informazioni

- Indagine statistica
- Rilevazione sperimentale
- Rilevazione osservazionale

Indagine statistica

- L'indagine statistica è la principale tecnica con cui si possono ottenere informazioni sulle manifestazioni di un fenomeno su una data popolazione
- Un'indagine statistica ha come obiettivo la conoscenza di una popolazione intesa come insieme di unità elementari su cui si manifesta il fenomeno oggetto di studio
- L'indagine è detta **completa** (**censuaria**) se si rilevano i caratteri di interesse su tutte le unità della popolazione
  - ✓ Censimento generale della popolazione e delle abitazioni, Censimento generale dell'industria, del commercio, dei servizi e dell'artigianato; Censimento generale dell'agricoltura (ISTAT)
- L'indagine è detta **parziale** (**campionaria**) se si rilevano i caratteri di interesse su un sottoinsieme di unità statistiche della popolazione, detto **campione**
  - ✓ Indagine sulle forze di lavoro (trimestrale); Indagine sui consumi delle famiglie; Indagine multiscopo (ISTAT)
- Si definisce **campione** un qualsiasi sottoinsieme di n (**dimensione del campione**) unità della popolazione

Indagine statistica completa e parziale

- L'indagine completa è teoricamente semplice, ma vincoli legati
  - ✓ alla numerosità della popolazione (spesso non finita)
  - ✓ ai costi e/o ai tempi dell'indagine

  inducono a optare per un'indagine campionaria
- Campionamento: Modalità di estrazione del campione dalla popolazione
  - ✓ La metodologia statistica permette di misurare e controllare l'attendibilità delle informazioni provenienti da un campione se il campione è casuale o probabilistico (estratto con meccanismi di selezione casuale)
    - ✓ Esempio: Campionamento casuale semplice
  - ✓ Campioni non probabilistici: La motivazione principale per cui si ricorre a campioni non probabilistici è l'assenza di liste e la necessità di limitare i costi
    - ✓ Esempio: Campionamento volontario

Studi sperimentali

- Una situazione di rilevazione sperimentale è caratterizzata dalla possibilità di controllare sia le condizioni in cui l'esperimento si svolge, sia le caratteristiche delle unità statistiche da impiegare
- Negli studi sperimentali si assume che i ricercatori abbiano il pieno controllo delle condizioni sperimentali:
  - ✓ Fattori sperimentali (trattamenti): Variabili su cui l'esperimento è chiamato a fornire una verifica del loro diverso effetto
  - ✓ Fattori di stratificazione: Variabili riguardanti la composizione delle unità sperimentali
- Disegno sperimentale e randomizzazione
  - ✓ Disegno sperimentale: Controllo diretto dei fattori e di stratificazione
  - ✓ Randomizzazione: Controllo indiretto

Studio sperimentale - Esempio

- Obiettivo: Valutare l'effetto della dimensione delle classi in asili sul punteggio a un test cognitivo
- Unità di analisi: Insegnante
- Trattamento: Classe piccola (13-17 bambini) versus classe regolare (22-25 bambini)
- Variabile di controllo (Fattore di stratificazione): Asilo
- Variabile risultato (variabile al livello di classe): media dei punteggi al test cognitivo dei bambini della classe a cui un insegnante è assegnato
- Disegno sperimentale e randomizzazione
  - ✓ Suddividere gli insegnati in gruppi secondo l'asilo in cui insegnano
  - ✓ All'interno di ciascun asilo assegnare casualmente gli insegnanti a classi piccole e classi regolari

Studi osservazionali

- In una situazione di rilevazione osservazionale non si ha la possibilità di controllare le condizioni sotto le quali si svolge l'osservazione e solo in parte si possono controllare le caratteristiche delle unità statistiche.
- Negli studi osservazionali il ricercatore osserva (rileva) le manifestazioni delle variabili di interesse sulle diverse unità statistiche, ma non ha la possibilità di controllo sperimentale sulle stesse
- Le unità statistiche decidono (in un certo senso) il trattamento a cui esporsi

Studi osservazionali - Esempio

- Obiettivo: Valutare gli effetti del fumo sulla salute
- Trattamento: Appartenenza alla categoria di "fumatore" o di "Non fumatore"
- Variabile risposta: Insorgenza di patologie associate al fumo (infarto, cancro ai polmoni ecc)
- Importante avere informazioni e controllare per fattori/caratteristiche (come ad esempio sesso, età, posizione lavorativa ecc) che possono essere associate sia con la decisione di fumare sia con la variabile risposta
- È importante che il gruppo dei fumatori (trattati) e il gruppo dei non fumatori (gruppo di controllo) siano simili rispetto a caratteristiche che possono essere associate sia alla decisione di fumare che alla variabile risposta

Statistica

La **statistica** è l'insieme delle metodologie finalizzate alla raccolta e all'analisi dei dati che permettono lo studio in termini quantitativi di fenomeni collettivi

- **Progettazione**: Disegno dello studio
  - ✓ Pianificazione della raccolta dati (rilevazione)
- **Descrizione** – **Statistica descrittiva**
  - ✓ Metodi per sintetizzare i dati per meglio comprendere le informazioni in essi contenute
  - ✓ **Statistiche descrittive** (Grafici, tabelle e sintesi numeriche)
- **Inferenza** – **Statistica Inferenziale**
  - ✓ Far inferenza significa fare *previsioni* (dedurre informazioni) sul fenomeno di interesse nella popolazione sulla base delle informazioni raccolte su un campione selezionato dalla popolazione
  - ✓ L'inferenza è un processo di generalizzazione attraverso cui i risultati ottenuti su un campione vengono estesi alla popolazione

Inferenza

- Tipicamente l'inferenza riguarda alcuni parametri della popolazione
- Un **parametro** è una caratteristica della popolazione, ossia un indice relativo alla distribuzione del carattere di interesse nella popolazione (come ad esempio la media, la mediana, la deviazione standard, etc)
- I parametri della popolazione non sono usualmente noti
- Informazioni sui parametri della popolazione possono essere ottenute usando il campione: Facciamo inferenza sulla popolazione esaminando i risultati campionari
- Per fare inferenza si usano delle **statistiche**
- Una **statistica** è una sintesi numerica di dati campionari

Un parametro è una caratteristica della popolazione. Una statistica è una sintesi numerica di dati campionari

Esempi

- Reddito annuale delle famiglie italiane

Popolazione	=	Insieme delle famiglie Italiane
Parametro	=	Reddito medio
Campione	=	Un sottoinsieme di famiglie italiane
Statistica	=	Reddito medio calcolato sulle famiglie del campione

- Credere nel paradiso
  GSS = General Social Survey 2004 (http://www.norc.org/GSS+Website/)

Popolazione	=	Popolazione degli adulti USA nel 2004 (oltre 200 milioni di individui)
Parametro	=	Percentuale di coloro che credono nel paradiso
Campione	=	1158 individui adulti con cittadinanza statunitense
Statistica	=	Proporzione di individui del campione che dichiara di credere nel paradiso

Esempio: Punteggio al test di ammissione

- Popolazione: 4 candidati
- Variabile: Punteggio (in centesimi)

Candidato	Punteggio
1	72
2	69
3	85
4	94

- Parametro: Punteggio medio

$$\frac{72 + 69 + 85 + 94}{4} = 80$$

- Campione: Due candidati ($n = 2$)
- Statistica: Punteggio medio calcolato sui candidati del campione

$$\overline{Y} = \frac{Y_1 + Y_2}{2}$$

Esempio: Punteggio al test di ammissione

- Campione composto dai candidati 1 e 4

Candidato	Punteggio
1	72
4	94

- Statistica: Punteggio medio calcolato sui candidati del campione

$$\bar{y} = \frac{72 + 94}{2} = 83$$

- Campione composto dai candidati 1 e 3

Candidato	Punteggio
1	72
3	85

- Statistica: Punteggio medio calcolato sui candidati del campione

$$\bar{y} = \frac{72 + 85}{2} = 78.5$$

Il valore della statistica cambia al variare del campione!

Deduzione versus Induzione

Deduzione

- Un processo di deduzione va dal generale al particolare
- Il processo di deduzione è tipico della logica e della matematica
- Un processo deduttivo porta a conclusioni certe
- Esempio. Per un matematico è noto che la somma degli angoli interni di un triangolo è pari a 180 gradi sessagesimali. Quindi, nota l'ampiezza di due dei tre angoli, il matematico può dedurre l'ampiezza del terzo angolo, senza che abbia mai visto un triangolo del genere.

Induzione

- Un processo di induzione va dal particolare al generale
- Il processo di induzione è tipico delle discipline scientifiche e sperimentali
- Un processo di induzione porta a conclusioni incerte
- Esempio. Il colore dei cigni.
  - ✓ Ho visto un cigno ed era bianco; Ho visto un secondo cigno ed era bianco; Ho visto un terzo cigno ed era bianco; ...
  - ✓ Conclusione: Il prossimo cigno che vedrò sarà probabilmente bianco / Tutti i cigni sono probabilmente bianchi
  - ✓ Il cigno nero (Nassim Nicholas Taleb, 2007)

Inferenza statistica e Induzione

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

- L'incertezza è dovuta principalmente alla **variabilità campionaria**: In linea di principio, tutti i possibili campioni sono diversi e quindi la loro analisi produce risultati diversi. In pratica si dispone di un solo campione
- **Errori di misurazione:** In molti casi ripetendo la misurazione della stessa entità si ottengono valori diversi. Tipicamente si dispone di una sola misurazione per ogni entità
- Un aspetto importante dell'inferenza statistica riguarda l'*accuratezza* delle statistiche campionarie

Popolazione infinita

- La popolazione è infinita tutte le volte che non è esattamente delimitata, cioè non è concettualmente possibile elencare i suoi membri
  - ✓ In un'indagine sulla soddisfazione per i servizi sanitari si seleziona un campione di $n = 200$ cittadini
  - ✓ Chiaramente non si è interessati ai quei 200 cittadini, ma ai cittadini in generale, quindi i dati relativi ai 200 intervistati devono essere generalizzati a una popolazione più ampia
  - ✓ La popolazione è costituita da tutti i cittadini, residenti in varie aree geografiche, di oggi e di domani etc
  - ✓ Si tratta quindi di una popolazione non esattamente definita, quindi di numerosità infinita
- La popolazione è infinita anche nei casi in cui non vi è alcuna estrazione di unità, ma (al fine di generalizzare i risultati) è opportuno considerare il valore della variabile di interesse, Y, nelle unità sotto osservazione come realizzazione di un processo aleatorio:
  - ✓ Una nuova terapia viene applicata a 20 soggetti "omogenei", cioè soggetti con caratteristiche simili, caratterizzati ad esempio dalla stessa probabilità di guarire

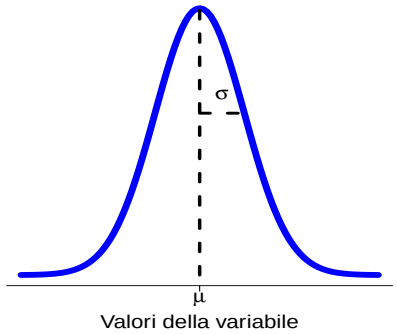
Modelli probabilistici per popolazioni infinite

- Una popolazione infinita è composta da tutte le unità potenzialmente osservabili e non necessariamente già esistenti fisicamente
- Il fenomeno di interesse può essere rappresentato da una variabile aleatoria Y avente una certa distribuzione di probabilità nella popolazione
- Per brevità si fa riferimento a Y come alla "Popolazione"
  - ✓ Il termine popolazione identifica i possibili esiti della ripetizione, teoricamente illimitata, nelle stesse condizioni di un esperimento causale
- Ipotesi: Esiste un modello probabilistico capace di rappresentare adeguatamente la distribuzione del fenomeno che interessa studiare nella popolazione
  - ✓ Esempio 1: Y = Status occupazionale (0 = Disoccupato, 1 = Occupato). Ipotesi: $Y \sim Ber(\pi)$
  - ✓ Esempio 2: Y = Livello di colesterolo. Ipotesi $Y \sim N(\mu, \sigma^2)$

Modelli probabilistici per popolazioni infinite

- Se Y è v.a. discreta, la distribuzione della variabile nella popolazione è usualmente descritta dalla funzione di massa di probabilità: $P(Y = y; \theta)$
  - ✓ Esempio 1: Y = Status occupazionale (0 = Disoccupato, 1 = Occupato). Ipotesi: $Y \sim Ber(\pi)$ con $\theta = \pi$, quindi
$$P(Y = y; \pi) = \pi^y (1 - \pi)^{1-y} \qquad y = 0, 1$$
- Se Y è v.a. continua, la distribuzione della variabile nella popolazione è usualmente descritta dalla funzione di densità di probabilità: $f(y; \theta)$
  - ✓ Esempio 2: Y = Livello di colesterolo. Ipotesi: $Y \sim N(\mu, \sigma^2)$ con $\theta = (\mu, \sigma^2)$, quindi

$$f\left(y; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}(y - \mu)^2\right\}$$

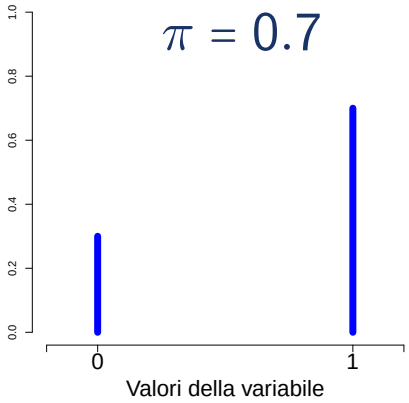


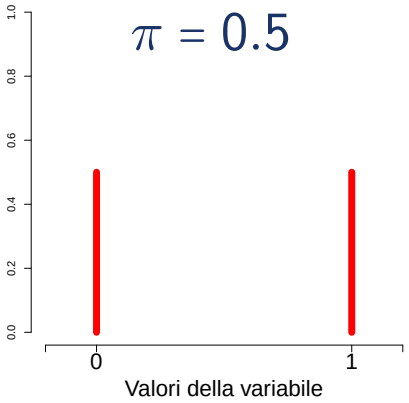
- La distribuzione di probabilità di Y nella popolazione non è completamente nota: La distribuzione di Y dipende da un parametro θ che non è completamente noto
  - ✓ Esempio 1: $Y \sim Ber(\pi)$ ma π non è noto
  - ✓ Esempio 2: $Y \sim N(\mu, \sigma^2)$ ma μ e/o σ^2 non sono noti

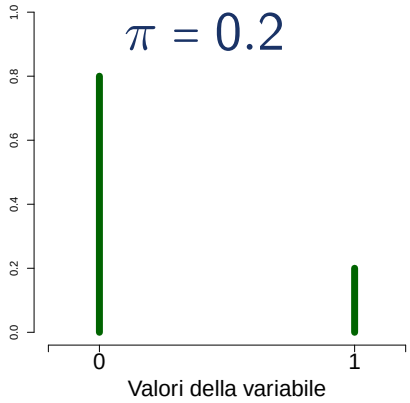
Modelli probabilistici per popolazioni infinite

- Popolazione Bernoulliana: $Y \sim Ber(\pi)$, $\theta = \pi =$ Probabilità di successo

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}, y = 0, 1$$

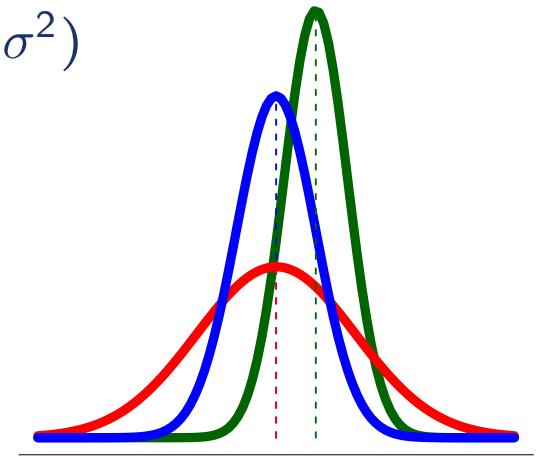






- Popolazione Normale: $Y \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \mu)^2\right\}$$



Due importanti parametri di una popolazione infinita

- Media della popolazione (valore atteso): $\mu_Y = \mathbb{E}[Y]$
- Varianza della popolazione: $\sigma^2_Y = \mathbb{E}[(Y - \mu_Y)^2]$
- Deviazione standard della popolazione: $\sigma_Y = \sqrt{\sigma^2_Y} = \sqrt{\mathbb{E}[(Y - \mu_Y)^2]}$
- Esempio: Se Y è una variabile aleatoria discreta, allora

$$\mu_Y = \sum_y y \cdot P(Y = y) \qquad e \qquad \sigma^2_Y = \sum_y (y - E(Y))^2 \cdot P(Y = y)$$

  Se Y è una variabile aleatoria continua il calcolo della media e della varianza è più complesso
- Esempi: Popolazione Bernoulliana e Popolazione Normale

Popolazione	Media	Varianza	DS
$Y \sim Ber(\pi)$	$\mu_Y = \pi$	$\sigma^2_Y = \pi(1-\pi)$	$\sigma_Y = \sqrt{\pi(1-\pi)}$
$Y \sim N(\mu, \sigma^2)$	$\mu_Y = \mu$	$\sigma^2_Y = \sigma^2$	$\sigma_Y = \sqrt{\sigma^2}$

Logica del campionamento

- La teoria statistica si basa sul concetto di campione casuale
- Si supponga di essere interessati allo studio di un certo fenomeno
- Si supponga che il fenomeno sia rappresentabile da una variabile Y la quale ha una certa distribuzione nella popolazione
  - ✓ Ad esempio la distribuzione del fenomeno nella popolazione può essere data dalla distribuzione delle frequenze relative della variabile Y

y_i	y_1	y_2	$\cdots$	y_K
f_i	f_1	f_2	$\cdots$	f_K

- Si decide di estrarre in modo casuale (con equiprobabilità) un'unità dalla popolazione
- Il procedimento dell'estrazione può essere pensato come un esperimento il cui esito non è noto a priori (non è noto il valore della variabile che si osserverà sull'unità prima di effettuare l'estrazione)
- L'esito dell'esperimento può essere rappresentato da una v.a. Y_1 la cui distribuzione coincide con quella della variabile di interesse Y nella popolazione

Logica del campionamento

- Si supponga di ripetere la prova n volte
- Prima di effettuare l'estrazione/esperimento, ossia prima di osservare il valore della variabile su ogni unità, il valore del carattere Y sulle n unità non è noto, ma può essere pensato come una v.a.: il campione è quindi un insieme di n variabili aleatorie $Y_1, \ldots, Y_i, \ldots, Y_n$
- Dopo l'esperimento si ottiene un insieme di n numeri

$$y_1, y_2, \ldots, y_n$$

  che chiamiamo **campione osservato** di n unità estratto dalla popolazione (v.a.) Y
- I valori osservati $y_1, \ldots, y_i, \ldots, y_n$ sono realizzazioni delle v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$
- Tutto ciò è vero a prescindere dal fatto che la popolazione sia finita o infinita
- La natura finita o infinita cambia il modo di estrarre il campione

Campionamento da una popolazione infinita: Campione casuale

Un insieme di v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$ ottenuto con un procedimento di estrazione dalla popolazione Y è un **campione casuale** di dimensione n se le v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$ sono indipendenti e identicamente distribuite (i.i.d.) con distribuzione di probabilità di ciascuna v.a. Y_i, $i = 1, \ldots, n$ uguale alla distribuzione di probabilità della popolazione Y

- Il procedimento dell'estrazione può essere pensato come un esperimento composto da n prove i cui esiti possono essere rappresentati da n v.a. $Y_1, \ldots, Y_n$ i.i.d. (con distribuzione coincidente con quella della variabile Y nella popolazione)
- Indipendenza: La distribuzione di probabilità di un elemento campionario Y_i non dipende dai valori assunti dagli altri elementi campionari
  - ✓ Questo accade se vi è indipendenza nella popolazione e il metodo di campionamento preserva tale indipendenza (ad esempio estrazione casuale con equiprobabilità e con reinserimento di n unità dalla popolazione)
- Identica distribuzione: Tutti gli elementi campionari hanno la stessa distribuzione (sono dei cloni) e tale distribuzione è la stessa del carattere nella popolazione (ossia, Y_i ha la stessa distribuzione di Y).
  - ✓ Questo accade se le probabilità di estrazione sono identiche per tutte le unità della popolazione e non vi sono problemi dovuti a fattori di distorsione

Campione casuale e campione osservato

Estrazione	Campione		Definizione
Prima	Campione casuale	$Y_1, \ldots, Y_i, \ldots, Y_n$	n v.a. i.i.d.
Dopo	Campione osservato	$y_1, \ldots, y_i, \ldots, y_n$	n numeri reali generati dall'estrazione

Campione casuale da popolazione Normale

$Y_1, \ldots, Y_i, \ldots, Y_n$: Campione casuale di dimensione n da una popolazione Normale

- La distribuzione del carattere Y nella popolazione è Normale con media μ e varianza σ^2 in genere non note
- Le v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$ sono indipendenti
- Le v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$ hanno tutte la stessa distribuzione e tale distribuzione è Normale con media e varianza coincidenti con media e varianza della variabile Y nella popolazione
- Esempio: Campione casuale di dimensione $n = 3$ da una popolazione Normale con media $\mu = 100$ e varianza $\sigma^2 = 16$

$$Y_1 \sim N(100, 16) \quad Y_2 \sim N(100, 16) \quad Y_3 \sim N(100, 16)$$
$$\text{Indipendenti}$$

- Nota: Se la variabile di interesse è continua l'insieme dei possibili campioni è infinito. Ad esempio l'insieme dei campioni di ampiezza 3 da una popolazione Normale è formato da tutte le possibili terne di numeri reali

Campione casuale da popolazione Bernoulliana

$Y_1, \ldots, Y_i, \ldots, Y_n$: Campione casuale di dimensione n da una popolazione Bernoulliana

- La distribuzione del carattere Y nella popolazione è Bernoulliana con probabilità di successo π
- Le v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$ sono indipendenti
- Le v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$ hanno tutte la stessa distribuzione e tale distribuzione è Bernoulliana con probabilità di successo coincidente con probabilità di successo di Y nella popolazione
- Esempio: Campione casuale di dimensione $n = 3$ da una popolazione Bernoulliana con probabilità di successo $\pi = 0.6$

$$Y_1 \sim Ber(0.6) \quad Y_2 \sim Ber(0.6) \quad Y_3 \sim Ber(0.6)$$
$$\text{Indipendenti}$$

- Nota: In questo caso i possibili campioni sono 8:

$$\{(0,0,0); (0,0,1); (0,1,0); (1,0,0); (1,1,0); (1,0,1); (0,1,1); (1,1,1)\}$$

Statistica

Sia $Y_1, \ldots, Y_i, \ldots, Y_n$ un campione casuale di dimensione n. Definiamo **statistica**

$$T(Y_1, \ldots, Y_i, \ldots, Y_n)$$

una qualunque funzione a valori reali del campione casuale $Y_1, \ldots, Y_i, \ldots, Y_n$ che non dipende da altre quantità incognite.

- Una statistica è una v.a. in quanto funzione delle variabili aleatorie $Y_1, \ldots, Y_i, \ldots, Y_n$
- Definiamo **statistica calcolata**

$$t = T(y_1, \ldots, y_i, \ldots, y_n)$$

il valore che la statistica T assume sul particolare campione osservato $y_1, \ldots, y_i, \ldots, y_n$

- La statistica calcolata è un numero reale (valore della statistica sul campione osservato)

Alcune statistiche notevoli

- Media campionaria/Proporzione campionaria

$$\overline{Y} = \frac{Y_1 + \cdots + Y_i + \cdots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Varianza campionaria (corretta)

$$S^2 = \frac{(Y_1 - \overline{Y})^2 + \cdots + (Y_i - \overline{Y})^2 + \cdots + (Y_n - \overline{Y})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Deviazione standard campionaria

$$S = \sqrt{S^2}$$

Esempio: Numero di dipendenti

- Popolazione: popolazione finita composta da 5 unità (imprese)

u_i	y_i
1	18
2	20
3	21
4	23
5	23

- Numero medio di dipendenti nella popolazione (parametro di interesse)

$$\mu_Y = \frac{18 + 20 + 21 + 23 + 23}{5} = \frac{105}{5} = 21$$

- Varianza del numero di dipendenti nella popolazione (si divide per N)

$$\sigma^2_Y = \frac{(18-21)^2 + (20-21)^2 + (21-21)^2 + (23-21)^2 + (23-21)^2}{5} = \frac{18}{5} = 3.6$$

Esempio: Numero di dipendenti (Campioni di dimensione $n = 2$)

- Si supponga di estrarre un campione causale di dimensione $n = 2$ dalla popolazione
- Campionamento con ripetizione: I possibili campioni di dimensione $n = 2$ sono $N^n = 5^2 = 25$ e ogni campione ha probabilità $1/25$ di essere estratto
- In questo esempio i possibili campioni sono pochi quindi è possibile elencarli
- Si noti che nella pratica si seleziona un solo campione: Una volta fatta l'estrazione si osserva un solo campione

Esempio: Numero di dipendenti (Campiononamento con ripetizione)

Campione	Imprese	Osservazione y_1	Osservazione y_2	Media campionaria	Varianza campionaria	Dev.Std campionaria
1	1, 1	18	18	18.0	0.0	0.000
2	1, 2	18	20	19.0	2.0	1.414
3	1, 3	18	21	19.5	4.5	2.121
4	1, 4	18	23	20.5	12.5	3.536
5	1, 5	18	23	20.5	12.5	3.536
6	2, 1	20	18	19.0	2.0	1.414
7	2, 2	20	20	20.0	0.0	0.000
8	2, 3	20	21	20.5	0.5	0.707
9	2, 4	20	23	21.5	4.5	2.121
10	2, 5	20	23	21.5	4.5	2.121
11	3, 1	21	18	19.5	4.5	2.121
12	3, 2	21	20	20.5	0.5	0.707
13	3, 3	21	21	21.0	0.0	0.000
14	3, 4	21	23	22.0	2.0	1.414
15	3, 5	21	23	22.0	2.0	1.414
16	4, 1	23	18	20.5	12.5	3.536
17	4, 2	23	20	21.5	4.5	2.121
18	4, 3	23	21	22.0	2.0	1.414
19	4, 4	23	23	23.0	0.0	0.000
20	4, 5	23	23	23.0	0.0	0.000
21	5, 1	23	18	20.5	12.5	3.536
22	5, 2	23	20	21.5	4.5	2.121
23	5, 3	23	21	22.0	2.0	1.414
24	5, 4	23	23	23.0	0.0	0.000
25	5, 5	23	23	23.0	0.0	0.000

Distribuzione campionaria

- Una statistica $T(Y_1, \ldots, Y_n)$ assume valori diversi a seconda del particolare campione osservato
- La probabilità che una statistica assuma un certo valore $T(y_1, \ldots, y_n)$ dipende da tutti i possibili campioni su cui la statistica assume tale valore
- La distribuzione di probabilità (calcolata sullo spazio di tutti i possibili campioni della stessa dimensione) che fornisce la probabilità per i possibili valori che una statistica può assumere è chiamata distribuzione campionaria della statistica

Distribuzione di probabilità di una statistica T

$\Updownarrow$

Distribuzione campionaria di T

- La distribuzione campionaria di una statistica è la distribuzione dei valori che la statistica assume su tutti i possibili campioni della stessa dimensione estratti dalla stessa popolazione

Esempio: Numero di dipendenti - Distribuzioni campionarie

Media campionaria		
Campione	$\overline{y}$	$P(\overline{Y} = \overline{y})$
1	18.0	$1/25 = 0.04$
2, 6	19.0	$2/25 = 0.08$
3, 11	19.5	$2/25 = 0.08$
7	20.0	$1/25 = 0.04$
4, 5, 8, 12, 16, 21	20.5	$6/25 = 0.24$
13	21.0	$1/25 = 0.04$
9, 10, 17, 22	21.5	$4/25 = 0.16$
14, 15, 18, 23	22.0	$4/25 = 0.16$
19, 20, 24, 25	23.0	$4/25 = 0.16$

	Varianza campionaria (corretta)	
Campione	s^2	$P(S^2 = s^2)$
1, 7, 13, 19, 20, 24, 25	0.0	$7/25$=0.28
8,12	0.5	$2/25$=0.08
2, 6, 14, 15, 18, 23	2.0	$6/25$=0.24
3, 9, 10, 11, 17, 22	4.5	$6/25$=0.24
4, 5, 16, 21	12.5	$4/25$=0.16

Distribuzioni campionarie

- L'estrazione di un campione e il calcolo di una statistica (media campionaria, varianza campionaria, deviazione standard campionaria) sul campione estratto è un esperimento aleatorio: prima di estrarre il campione il valore della statistica che si otterrà è ignoto, ma è possibile determinare quali valori si potranno osservare e con quale probabilità
- Prima di estrarre il campione, una statistica è una variabile aleatoria
- Nell'esempio, media campionaria, varianza campionaria, deviazione standard campionaria sono v.a. discrete
- Ad esempio la media campionaria assume il valore 18 con probabilità 0.04 (e il valore 18 è osservato se viene estratto il campione n. 1), 19 con probabilità 0.08 (il valore 19 è osservato se viene estratto il campione n. 2, oppure n. 6), ect.

Distribuzione campionaria della media campionaria

- Sia Y un fenomeno di interesse con media $\mu_Y = \mu$ e varianza $\sigma^2_Y = \sigma^2$ nella popolazione
- $Y_1, \ldots, Y_i, \ldots, Y_n$: campione casuale di dimensione n dalla popolazione Y
- Il campione casuale $Y_1, \ldots, Y_i, \ldots, Y_n$ è formato da v.a. i.i.d., quindi

$$\mathbb{E}[Y_i] = \mu \qquad \text{e} \qquad Var[Y_i] = \sigma^2 \quad i = 1, \ldots, n$$

- Analizziamo le caratteristiche della statistica media campionaria, $\overline{Y}$, sotto tali condizioni

Distribuzione campionaria della media campionaria

Il valore atteso (media) della media campionaria è uguale alla media della popolazione

$$\mu_{\overline{Y}} = \mu$$

- La distribuzione della media campionaria è centrata sulla media della popolazione
- Esempio: Numero di dipendenti

$$\mu_{\overline{Y}} = \frac{1}{25}[18.0 + 19.0 + 19.5 + 20.5 + 20.5 + 19.0 + 20.0 + 20.5 + 21.5 + 21.5 + 19.5 + 20.5 + 21.0 + 22.0 + 22.0 + 20.5 + 21.5 + 22.0 + 23.0 + 23.0 + 20.5 + 21.5 + 22.0 + 23.0 + 23.0] = \frac{525}{25} = 21$$

ossia (utilizzando la distribuzione campionaria)

$$\mu_{\overline{Y}} = 18 \cdot 0.04 + 19 \cdot 0.08 + 19.5 \cdot 0.08 + 20 \cdot 0.04 + 20.5 \cdot 0.24 + 21 \cdot 0.04 + 21.5 \cdot 0.16 + 22 \cdot 0.16 + 23 \cdot 0.16 = 21$$

Distribuzione campionaria della media campionaria

La varianza della media campionaria è uguale alla varianza della popolazione divisa per la dimensione campionaria:

$$\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$$

La deviazione standard della media campionaria, chiamata anche errore standard della media campionaria, è

$$\sigma_{\overline{Y}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Distribuzione campionaria della media campionaria

- L'errore standard di $\overline{Y}$ indica quanto varia la media campionaria nell'insieme dei possibili campioni di dimensione n estraibili dalla popolazione
- L'errore standard della media campionaria è
  - ✓ Direttamente proporzionale alla deviazione standard del carattere nella popolazione: Quanto più il carattere varia nella popolazione, tanto più la media campionaria varia da campione a campione
  - ✓ Inversamente proporzionale alla radice quadrata della dimensione del campione: Quanto più grande è il campione, tanto meno la media campionaria varia da campione a campione
- La deviazione standard della media campionaria è più piccola della deviazione standard della popolazione
- Le medie campionarie nell'insieme dei possibili campioni variano meno delle osservazioni individuali
- Nel calcolare la media i valori grandi e piccoli si compensano, quindi la media è meno variabile delle singole osservazioni

Esempio: Numero di dipendenti

- Varianza di Y nella popolazione $\sigma^2_Y = 3.6$
- Varianza della media campionaria (campioni casuali di dimensione $n = 2$)

$$\begin{aligned}\sigma^2_{\overline{Y}} &= (18-21)^2 \cdot 0.04 + (19-21)^2 \cdot 0.08 + (19.5-21)^2 \cdot 0.08 + \\ &\quad (20-21)^2 \cdot 0.04 + (20.5-21)^2 \cdot 0.24 + (21-21)^2 \cdot 0.04 + \\ &\quad (21.5-21)^2 \cdot 0.16 + (22-21)^2 \cdot 0.16 + (23-21)^2 \cdot 0.16 \\ &= 1.8\end{aligned}$$

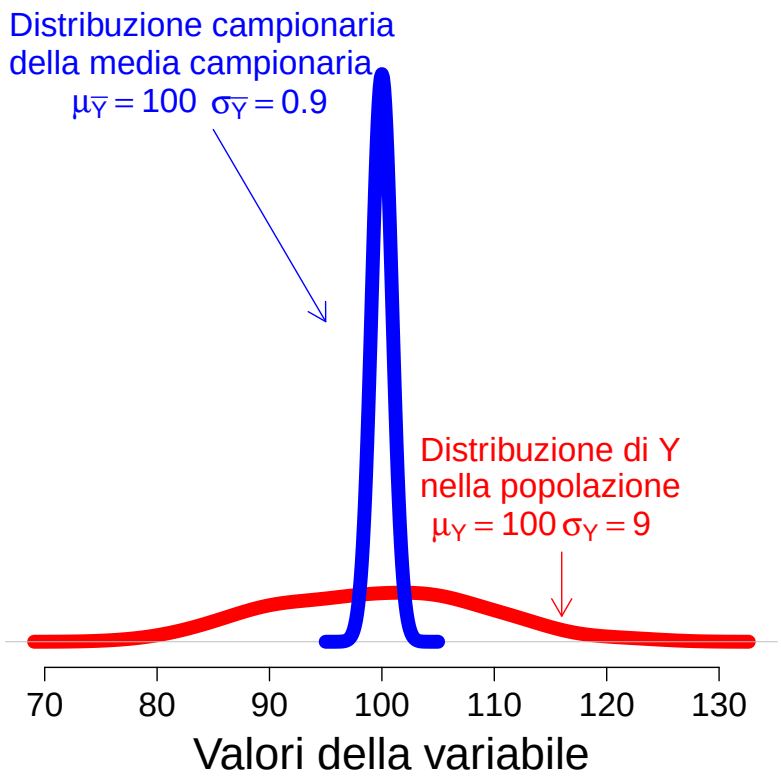
Si nota che

$$\sigma^2_{\overline{Y}} = \frac{3.6}{2} = 1.8$$

Distribuzione campionaria della media campionaria (popolazioni infinite)

- Popolazione: Y v.a. continua con $\mu_Y = 100$ e $\sigma^2_Y = 81$
- Con un campione di dimensione $n = 100$ si ha

$$\mu_{\overline{Y}} = 100 \qquad e \qquad \sigma_{\overline{Y}} = 0.9$$

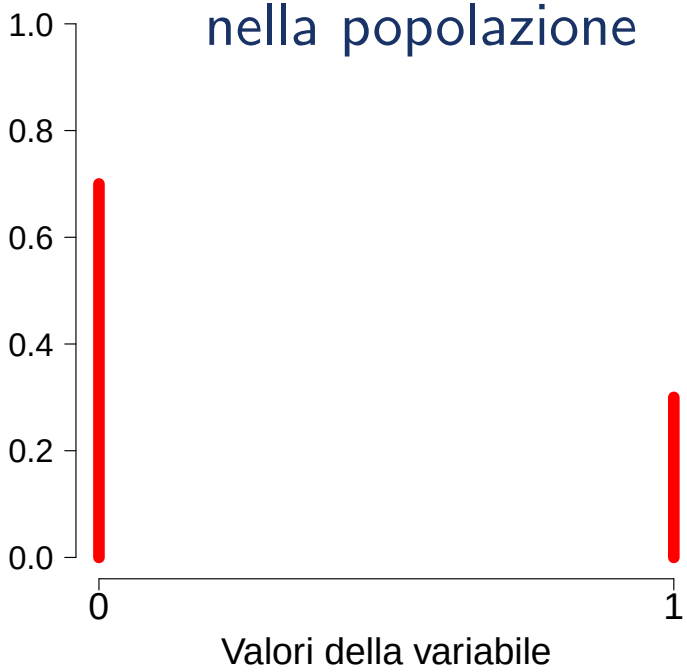


Distribuzione campionaria della media campionaria (popolazioni infinite)

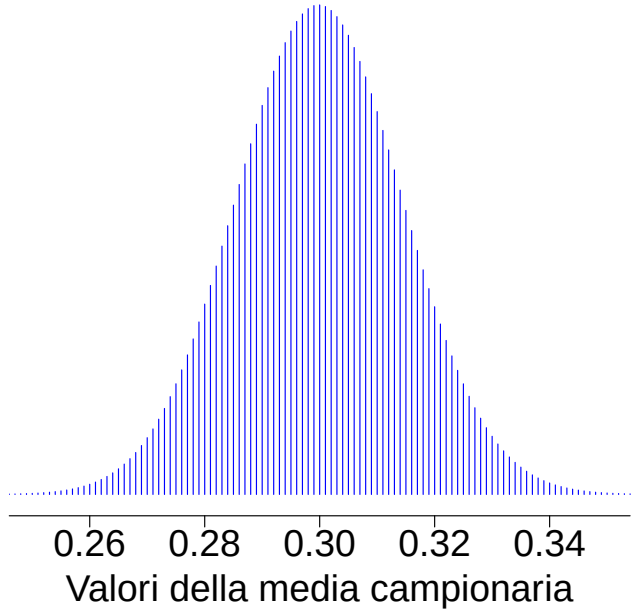
- Popolazione: Y v.a. binaria con $\mu_Y = 0.3$ e $\sigma^2_Y = 0.3 \cdot (1 - 0.3) = 0.21$
- Con un campione di dimensione $n = 1000$ si ha

$$\mu_{\overline{Y}} = 0.3 \qquad e \qquad \sigma^2_{\overline{Y}} = 0.00021$$

Distribuzione di Y nella popolazione



Distribuzione della media campionaria



Distribuzione campionaria della media campionaria

- Qualunque sia la distribuzione del carattere di interesse Y nella popolazione, la media campionaria calcolata su un campione casuale di dimensione n

$$\overline{Y} = \frac{Y_1 + \cdots + Y_i + \cdots + Y_n}{n}$$

ha media uguale alla media di Y nella popolazione

$$\mu_{\overline{Y}} = \mu_Y$$

e errore standard uguale alla deviazione standard di Y nella popolazione diviso la radice quadrata della dimensione del campione

$$\sigma_{\overline{Y}} = \frac{\sigma_Y}{\sqrt{n}}$$

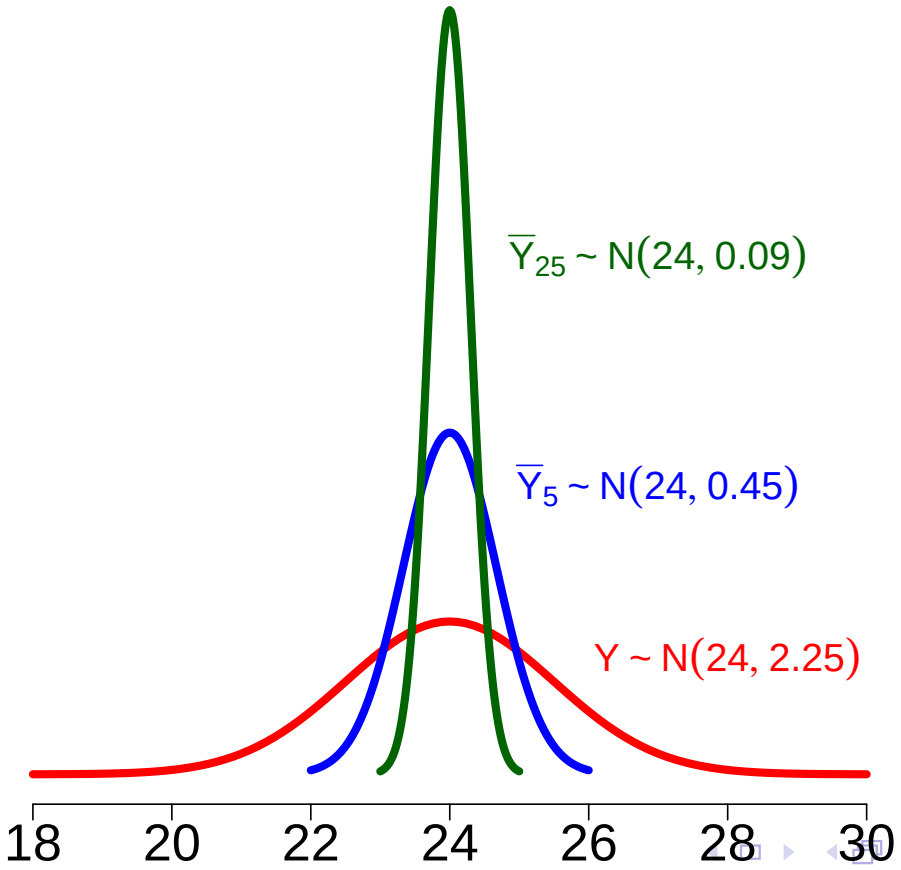
- Se la variabile di interesse Y nella popolazione ha distribuzione Normale allora la distribuzione della media campionaria sarà ancora normale

$$\text{Se} \quad Y \sim N(\mu, \sigma^2) \quad \text{allora} \quad \overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Distribuzione campionaria della media campionaria (popolazione Normale)

- Y = Punteggio a un esame: $Y \sim N(24, 2.25)$
- La media campionaria $\overline{Y}$ ha distribuzione normale con

$$\mu_{\overline{Y}} = \mu_Y = 24 \qquad e \qquad \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n} = \frac{2.25}{n}$$



Distribuzione campionaria della media campionaria (popolazione Bernoulliana)

- In molte applicazioni il carattere di interesse è qualitativo con due modalità (sì/no, soddisfatto/insoddisfatto, occupato/disoccupato, sopravvivenza/morte . . .)
- In tali casi si dice anche che i dati sono binari o dicotomici
- In tali casi la distribuzione del carattere nella popolazione è necessariamente Bernoulli (successo/insuccesso)
  - ✓ Successo = presenza della caratteristica di interesse (sì, soddisfatto, occupato, sopravvivenza . . .)
- L'unico parametro è π = probabilità di successo = "probabilità che un'unità a caso della popolazione presenti la caratteristica di interesse"

Distribuzione campionaria della media campionaria (popolazione Bernoulliana)

Proporzione campionaria

$$P = \frac{\text{Numero di successi}}{\text{Numero di prove}} = \frac{\text{Numero di casi che presentano la caratteristica di interesse}}{\text{Numero di osservazioni}}$$

- Codificando il successo con 1 e l'insuccesso con 0 il campione $Y_1, Y_2, \ldots, Y_n$ è una sequenza di numeri 0 e 1
- In tal caso la proporzione campionaria coincide con la media campionaria calcolata sugli elementi $Y_1, Y_2, \ldots, Y_n$

$$P = \overline{Y} = \frac{Y_1 + \cdots + Y_i + \cdots + Y_n}{n}$$

- La proporzione campionaria è un tipo di media campionaria, quindi valgono tutte le proprietà viste in generale per la media campionaria

Distribuzione campionaria della media campionaria (popolazione Bernoulliana)

Se la popolazione Y possiede una **distribuzione di Bernoulli** con parametro π, la distribuzione della media campionaria $\overline{Y}$ sarà tale che, qualunque sia il valore di π

- La media della proporzione campionaria è uguale alla proporzione di successi nella popolazione

$$\mu_{\overline{Y}} = \pi$$

- La proporzione campionaria ha varianza campionaria uguale alla varianza della popolazione (varianza di una v.a. di Bernoulli) diviso la dimensione del campione

$$\sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n} = \frac{\pi \cdot (1 - \pi)}{n}$$

- Errore standard della proporzione campionaria

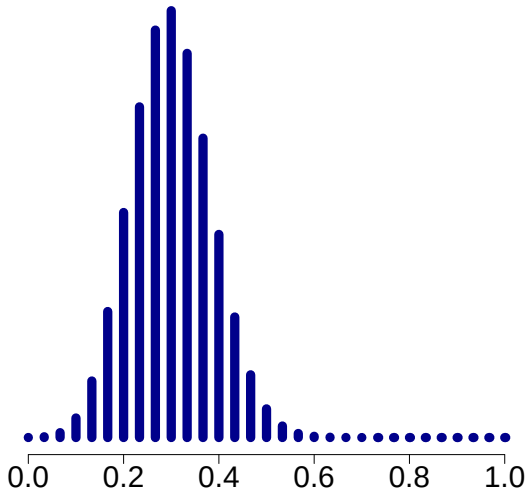
$$\sigma_{\overline{Y}} = \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

Distribuzione campionaria della media campionaria (popolazione Bernoulliana)

- Popolazione: $Y \sim \text{Bernoulli}(\pi = 0.3)$
- Dimensione campionaria: $n = 30$

$$\mu_{\overline{Y}} = 0.3$$

$$\sigma^2_{\overline{Y}} = \frac{0.3 \cdot (1 - 0.3)}{30} = 0.007 \quad e \quad \sigma_{\overline{Y}} = \sqrt{0.007} = 0.08367$$



Distribuzione campionaria della media campionaria (popolazione Bernoulliana)

- Valori della proporzione campionaria: $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$
- La proporzione campionaria ha una distribuzione nota
- Quando l'ampiezza campionaria n è grande il calcolo delle probabilità relative alla distribuzione campionaria della proporzione campionaria è complesso
- Approssimazione Normale: Quando n è grande, per il Teorema del Limite Centrale la distribuzione della proporzione campionaria è ben approssimata dalla distribuzione Normale

Il Teorema del Limite Centrale

- Sia $Y_1, \ldots, Y_n$ un campione casuale (composto da v.a. i.i.d) da una popolazione qualsiasi (non Normale)
  - ✓ La variabile Y di interesse può essere binaria oppure continua con forte asimmetria
- Se la dimensione del campione è elevata allora la media campionaria ha distribuzione approssimativamente Normale
- Questa è una conseguenza del Teorema del Limite Centrale

Teorema del Limite Centrale:

Sia Y un carattere di interesse con una certa distribuzione di media μ_Y e varianza σ^2_Y nella popolazione. Sia $Y_1, \ldots, Y_n$ un campione casuale (composto da v.a. i.i.d) da Y. Allora, per n sufficientemente grande,

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n} \approx N\left(\mu_Y, \frac{\sigma^2_Y}{n}\right)$$

- Il Teorema del Limite Centrale è un risultato asintotico, cioè indica quello che accade quando n, la dimensione del campione, tende all'infinito
- Al crescere della dimensione del campione n l'approssimazione migliora

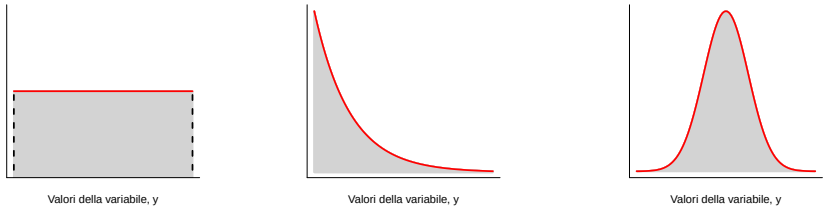
Il Teorema del Limite Centrale: Osservazioni

- Problema pratico: Quanto grande deve essere la dimensione del campione n affinché l'approssimazione sia buona?
- Nelle applicazioni si dispone di un campione di una certa ampiezza n e si deve valutare se l'approssimazione è accettabile
  - ✓ Se si ritiene che n sia sufficientemente grande, allora si può usare l'approssimazione Normale
  - ✓ Se si ritiene che n non sia sufficientemente grande, allora occorre seguire altre strade (alquanto impervie, che noi non vedremo)
- Quanto più la distribuzione del carattere nella popolazione è simmetrica e campanulare tanto più bassa è la dimensione campionaria per la quale l'approssimazione alla Normale è buona (nei casi favorevoli $n = 5$ è sufficiente)
- Regola pratica prudenziale: Un campione di ampiezza $n = 30$ è sufficiente per una buona approssimazione nella maggior parte dei casi

Il Teorema del Limite Centrale

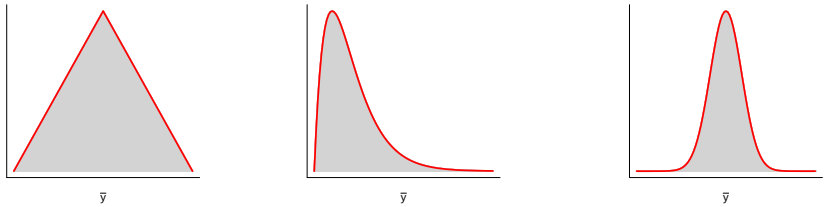
Distribuzione della media campionaria per campioni di diversa ampiezza ($n = 2, 5, 30$) estratti da tre popolazioni con diversa distribuzione

Distribuzione di Y nella popolazione

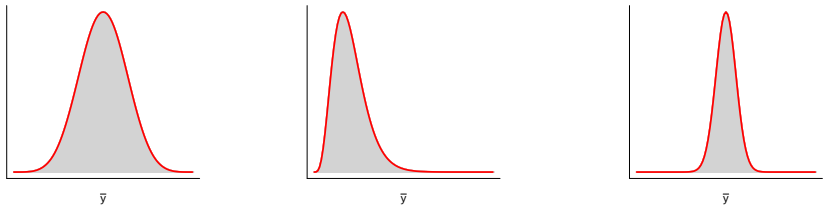


Distribuzione della media campionaria, $\overline{Y}$

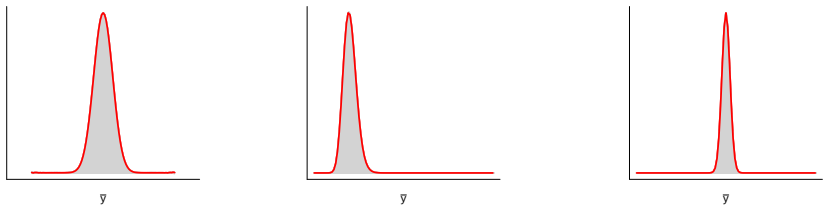
$n = 2$



$n = 5$



$n = 30$



Proporzione campionaria: Approssimazione Normale

- Quando la dimensione del campione n è grande, per il Teorema del Limite Centrale la distribuzione della proporzione campionaria è ben approssimata dalla Normale
- Un criterio generale per giudicare se l'approssimazione Normale per la distribuzione della proporzione campionaria può essere considerata ragionevole stabilisce che si dovrebbe disporre di almeno 15 osservazione per categoria (ossia almeno 15 successi e almeno 15 insuccessi)
- Esistono altri criteri

Distribuzione campionaria della media campionaria: Riepilogo

- Sia Y il carattere di interesse con di media μ_Y e varianza σ^2_Y e distribuzione qualsiasi nella popolazione
- Dato un campione casuale $Y_1, \ldots, Y_n$ da Y, allora la media campionaria $\overline{Y}$ ha media $\mu_{\overline{Y}} = \mu_Y$ e varianza $\sigma^2_{\overline{Y}} = \sigma^2_Y/n$
- Se Y ha distribuzione Normale, allora $\overline{Y}$ ha distribuzione Normale
- Se Y non ha distribuzione Normale, per il teorema del limite centrale, per n sufficientemente grande $\overline{Y}$ ha distribuzione approssimativamente Normale
- L'approssimazione alla normale della distribuzione campionaria della media campionaria si applica qualunque sia la forma della distribuzione del carattere di interesse nella popolazione
- Sapere che la distribuzione campionaria è esattamente o approssimativamente Normale (Teorema del Limite Centrale), permette di determinare le probabilità dei possibili valori di $\overline{Y}$
  - ✓ La media campionaria dovrebbe ricadere con elevata probabilità (approssimativamente 0.997) entro 3 deviazioni standard dalla media μ della popolazione

Inferenza Statistica: Stima

- Obiettivo: Utilizzare i dati campionari per stimare i parametri della popolazione
- Stimare i parametri della popolazione significa ottenere informazioni sul valore dei parametri utilizzando i dati campionari
- Se la variabile di interesse Y è una variabile binaria, l'obiettivo sarà stimare la proporzione dei successi nella popolazione (proporzione di unità statistiche che presentano una certa caratteristica)
  - ✓ Esempio: Il problema del fumo tra gli adolescenti
    - Y = Adolescente fumatore versus non fumatore
    - Obiettivo = Stimare la proporzione di adolescenti che fumano
- Se la variabile di interesse Y è una variabile continua, l'obiettivo sarà stimare la media di Y nella popolazione
  - ✓ Esempio: Prove INVALSI di matematica
    - Y = Punteggio totale alla prova INVALSI di matematica (percentuale di risposte corrette sul totale delle singole domande)
    - Obiettivo = Stimare la media del punteggio totale nella popolazione di studenti

Stima puntuale

Quando un parametro θ della popolazione è stimato attraverso un singolo valore, tale valore è chiamato stima puntuale del parametro θ

Esempio: Il problema del fumo tra gli adolescenti

- Per analizzare il fenomeno del fumo tra gli adolescenti, si estrae un campione casuale semplice di 1000 soggetti.
- In tale campione il numero di adolescenti che fumano è pari a 200.
- Obiettivo: Stimare la proporzione di adolescenti che fumano nell'intera popolazione degli adolescenti
- Popolazione: $Y \sim \text{Bernoulli}(\pi)$
- Parametro da stimare: probabilità di successo (Fumare) = π
- Stima = Valore della proporzione campionaria sul campione osservato

$$\widehat{\pi} = \frac{200}{1000} = 0.2$$

Esempio: Prove INVALSI di matematica (dati fittizi)

- Risultati alla prova INVALSI di matematica di un campione di 5 studenti di terza media

Soggetto	1	2	3	4	5
Punteggio	40	60	80	50	70

- Obiettivo: Stimare il punteggio medio alla prova INVALSI di matematica per l'intera popolazione degli studenti di terza media
- Popolazione: Y = Punteggio totale alla prova INVALSI di matematica con media μ e varianza σ^2
- Stima = Valore della media campionaria sul campione osservato

$$\widehat{\mu} = \overline{y} = \frac{40 + 60 + 80 + 50 + 70}{5} = \frac{300}{5} = 60$$

Stimatore

Uno **stimatore** è una statistica

$$T = T(Y_1, \ldots, Y_i, \ldots, Y_n)$$

utilizzata per stimare una determinata caratteristica (parametro), θ, della popolazione

Il valore assunto da uno stimatore in corrispondenza di un particolare campione osservato $y_1, \ldots, y_i, \ldots, y_n$ è chiamato **stima** del parametro θ e indicato con

$$\widehat{\theta} = T(y_1, \ldots, y_i, \ldots, y_n)$$

- Uno stimatore è una funzione a valori reali del campione casuale $Y_1, \ldots, Y_i, \ldots, Y_n$ che non dipende da altre quantità incognite

Stimatore e stima (puntuale)

Campione casuale
$Y_1, \ldots, Y_i, \ldots, Y_n$
Variabili aleatorie

$\Downarrow$

Campione osservato
$y_1, \ldots, y_i, \ldots, y_n$
Realizzazioni

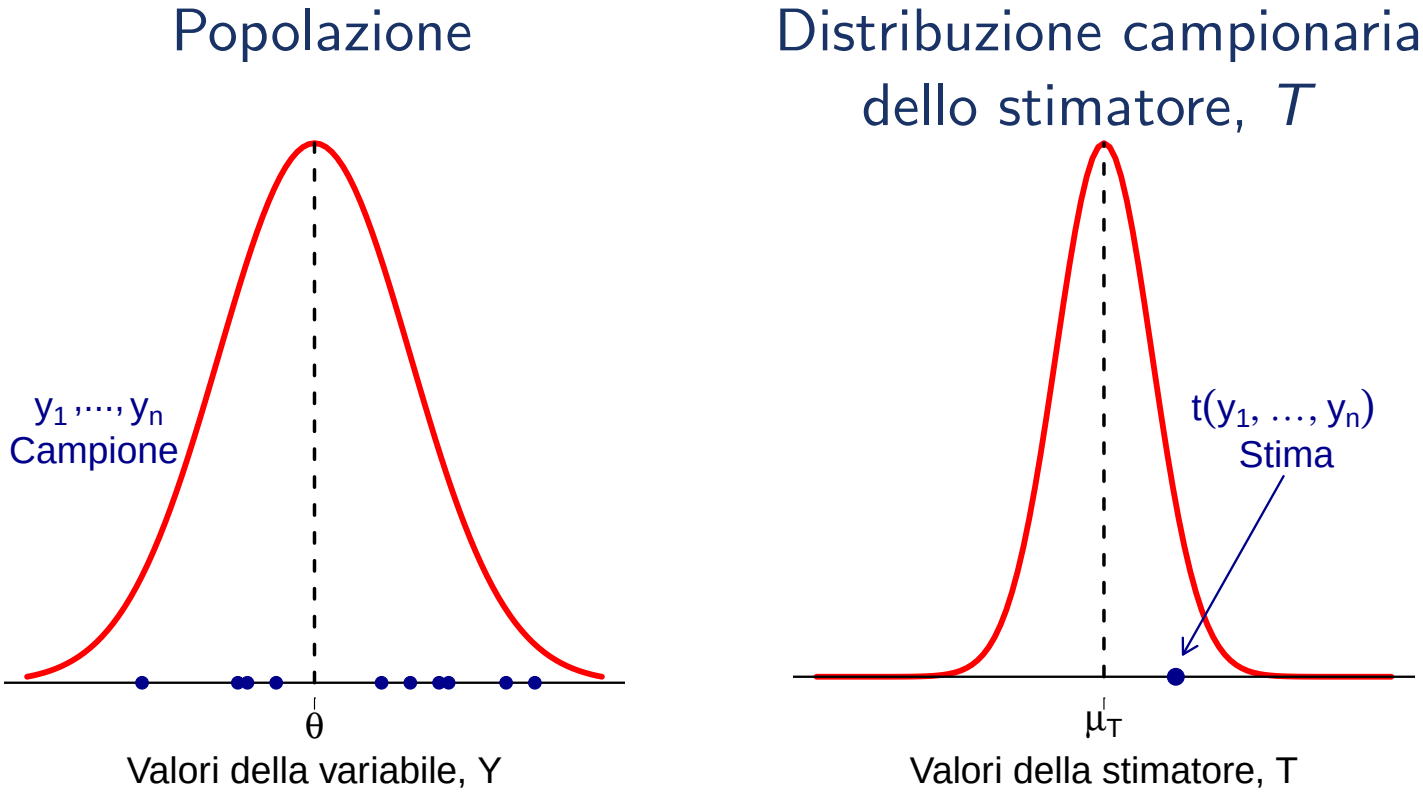
Stimatore
$T = T(Y_1, \ldots, Y_i, \ldots, Y_n)$
Variabile aleatoria

$\Downarrow$

Stima
$\widehat{\theta} = T(y_1, \ldots, y_i, \ldots, y_n)$
Realizzazione

Stimatore e stima (puntuale)

Lo stimatore, dipendendo dal campione casuale, è una variabile aleatoria e quindi possiede una distribuzione campionaria la cui conoscenza permette di capire se lo stimatore scelto produrrà con elevata probabilità stime "vicine" al valore vero del parametro



Proprietà degli stimatori

- Ciascun parametro può avere diversi possibili stimatori
  - ✓ Esempio: $Y \sim N(\mu, \sigma^2)$. Il parametro μ è la media, la mediana e la moda di Y, quindi la media campionaria, la mediana campionaria e la moda campionaria potrebbero essere tre stimatori possibili di μ
- La scelta di uno stimatore si basa sulla sua "bontà"
- Per valutare la "bontà" di uno stimatore T si può guardare alle sue proprietà:
  - ✓ Correttezza
  - ✓ Efficienza
  - ✓ Consistenza

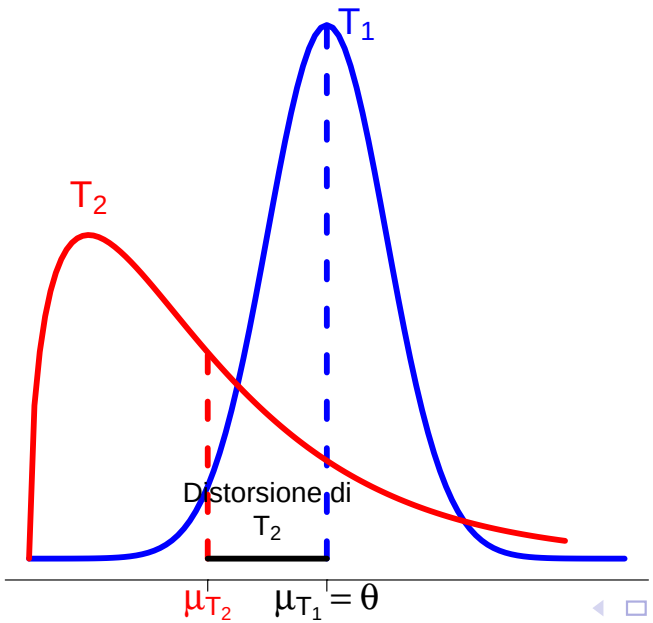
Stimatori corretti (non distorti)

Uno stimatore T è uno stimatore corretto (non distorto) per un parametro θ se

$$\mu_T = \mathbb{E}[T(Y_1, \ldots, Y_n)] = \theta$$

per tutti i possibili valori di θ

- Uno stimatore T è corretto se la sia distribuzione campionaria è centrata sul vero valore del parametro
- La distorsione di uno stimatore è uguale a: $B(T) = \mu_T - \theta$
- Uno stimatore distorto tende, in media, a sottostimare o sovrastimare il parametro



Non distorsione della media campionaria

La media campionaria è uno stimatore non distorto per la media della popolazione:

- Popolazione: Y con media $\mu_Y = \mu$
- Parametro: Media della popolazione μ
- Campione casuale di n osservazioni: $Y_1, \ldots, Y_n$
- La media campionaria

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

è uno stimatore non distorto per μ:

$$\mu_{\overline{Y}} = \mathbb{E}\left[\frac{Y_1 + \cdots + Y_n}{n}\right] = \frac{\mathbb{E}[Y_1] + \cdots + \mathbb{E}[Y_n]}{n} = \frac{n \cdot \mu}{n} = \mu$$

Esempio: Numero di dipendenti

- Popolazione: popolazione finita composta da 5 unità (imprese)

u_i	y_i
1	18
2	20
3	21
4	23
5	23

- Numero medio di dipendenti nella popolazione (parametro di interesse)

$$\mu_Y = \frac{18 + 20 + 21 + 23 + 23}{5} = \frac{105}{5} = 21$$

- Varianza del numero di dipendenti nella popolazione (si divide per N)

$$\sigma^2_Y = \frac{(18-21)^2 + (20-21)^2 + (21-21)^2 + (23-21)^2 + (23-21)^2}{5} = \frac{18}{5} = 3.6$$

Esempio: Numero di dipendenti (Campiononamento con ripetizione, $n = 2$)

Campione	Imprese	Osservazione y_1	Osservazione y_2	Media campionaria
1	1, 1	18	18	18.0
2	1, 2	18	20	19.0
3	1, 3	18	21	19.5
4	1, 4	18	23	20.5
5	1, 5	18	23	20.5
6	2, 1	20	18	19.0
7	2, 2	20	20	20.0
8	2, 3	20	21	20.5
9	2, 4	20	23	21.5
10	2, 5	20	23	21.5
11	3, 1	21	18	19.5
12	3, 2	21	20	20.5
13	3, 3	21	21	21.0
14	3, 4	21	23	22.0
15	3, 5	21	23	22.0
16	4, 1	23	18	20.5
17	4, 2	23	20	21.5
18	4, 3	23	21	22.0
19	4, 4	23	23	23.0
20	4, 5	23	23	23.0
21	5, 1	23	18	20.5
22	5, 2	23	20	21.5
23	5, 3	23	21	22.0
24	5, 4	23	23	23.0
25	5, 5	23	23	23.0

$$\mu_{\overline{Y}} = \frac{1}{25}[18.0 + 19.0 + 19.5 + 20.5 + 20.5 + 19.0 + 20.0 + 20.5 + 21.5 + 21.5 + 19.5 + 20.5 + 21.0 + 22.0 + 22.0 + 20.5 + 21.5 + 22.0 + 23.0 + 23.0 + 20.5 + 21.5 + 22.0 + 23.0 + 23.0]$$

$$= \frac{525}{25} = 21$$

Errore quadratico medio (Mean Square Error, MSE)

Sia T uno stimatore di un parametro θ. L'errore quadratico medio è il valore atteso della differenza al quadrato tra lo stimatore e il parametro:

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

- L'errore quadratico medio permette di valutare la capacità di uno stimatore T di fornire valori prossimi al vero valore del parametro, θ.
- Si noti che
$$MSE(T) = \mathbb{E}[(T - \theta)^2] = \sigma_T^2 + B(T)^2$$
- Se T è uno stimatore non distorto, $\mu_T = \theta$, allora $MSE_T = \sigma_T^2$

Stimatori efficienti

Siano T_1 e T_2 due stimatori di un parametro θ. Si dice che T_1 è più efficiente di T_2 se

$$MSE(T_1) < MSE(T_2)$$

per tutti i possibili valori di θ

- Siano T_1 e T_2 due stimatori non distorti di un parametro θ:

  $$\mu_{T_1} = \theta \qquad \text{e} \qquad \mu_{T_2} = \theta$$

  Si dice che T_1 è più efficiente di T_2 se la varianza di T_1 è più piccola della varianza di T_2

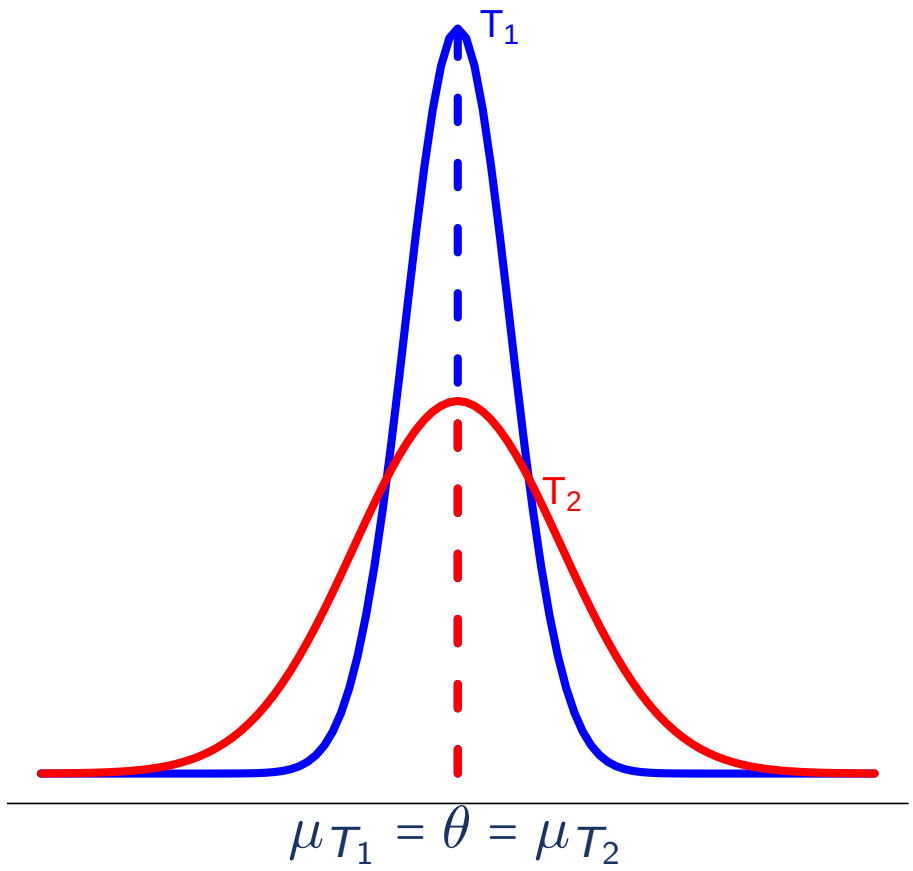
  $$\sigma^2_{T_1} < \sigma^2_{T_2}$$

  per tutti i possibili valori di θ

- T_1 è più efficiente di T_2 se l'errore standard di T_1 è più piccolo dell'errore standard di T_2: $\sigma_{T_1} < \sigma_{T_2}$
- Uno stimatore corretto per θ che ha un errore standard più piccolo di quello di tutti gli altri stimatori corretti di θ è definito stimatore (corretto) efficiente

Stimatori efficienti

Distribuzioni campionarie di due stimatori corretti per un parametro θ: T_1 possiede un errore quadratico medio (ossia una varianza e quindi un errore standard) più piccolo di T_2



Consistenza

Sia T_n uno stimatore di un parametro θ. Si dice che T_n è consistente se al crescere della dimensione del campione n l'errore quadratico medio di T_n tende a zero:

$$\lim_{n \to +\infty} MSE(T_n) = 0$$

- Si noti che

$$\lim_{n \to +\infty} MSE(T_n) = 0 \quad \Longleftrightarrow \quad \lim_{n \to +\infty} \sigma^2_{T_n} = 0 \quad \text{e} \quad \lim_{n \to +\infty} B(T_n) = 0$$

- Se T_n è uno stimatore non distorto per θ, $\mu_{T_n} = \theta$ per ogni θ, allora T_n è consistente se e solo se la varianza di T_n tende a zero al crescere della dimensione del campione: $\lim_{n \to +\infty} \sigma^2_{T_n} = 0$

Stima puntuale della media e della varianza di una popolazione

- Sia Y un carattere quantitativo con media μ e varianza σ^2 nella popolazione
- Sia $Y_1, Y_2, \ldots, Y_n$ un campione casuale di dimensione n
- Obiettivo 1: Trovare uno stimatore per la media di Y nella popolazione μ
- Obiettivo 2: Trovare uno stimatore per la varianza di Y della popolazione σ^2

Stima puntuale della media di una popolazione

- Lo stimatore media campionaria

$$T(Y_1, Y_2, \ldots, Y_n) = \overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

ha le seguenti proprietà

✓ È uno stimatore corretto per la media della popolazione μ:

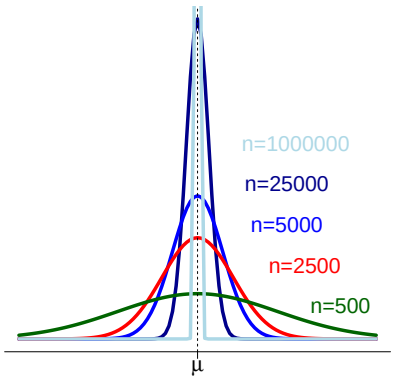
$$\mu_{\overline{Y}} = \mu \qquad \text{per ogni } \mu \in \mathbb{R}$$

✓ Ha errore quadratico medio, uguale alla sua varianza (per la correttezza):

$$MSE(\overline{Y}) = \sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$$

✓ È uno stimatore consistente

$$\lim_{n \to +\infty} MSE(\overline{Y}) = \lim_{n \to +\infty} \sigma^2_{\overline{Y}} = \lim_{n \to +\infty} \frac{\sigma^2}{n} = 0$$



✓ Per n sufficientemente grande, $\overline{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ e

se $Y \sim N(\mu, \sigma^2)$, allora $\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Stima puntuale della varianza di una popolazione

- Lo stimatore varianza campionaria corretta

$$T(Y_1, Y_2, \ldots, Y_n) = S^2 = \frac{(Y_1 - \overline{Y})^2 + \cdots + (Y_n - \overline{Y})^2}{n-1}$$

ha le seguenti proprietà

- ✓ È uno stimatore corretto per la varianza di Y della popolazione σ^2:

$$\mu_{S^2} = \sigma^2 \qquad \text{per ogni } \sigma^2 > 0$$

- ✓ Ha errore quadratico medio che coincide con la sua varianza (per la correttezza): $MSE(S^2) = \sigma^2_{S^2}$
- ✓ È uno stimatore consistente

$$\lim_{n \to +\infty} MSE(S_n^2) = \lim_{n \to +\infty} \sigma^2_{S_n^2} = 0$$

Stima puntuale della media di una popolazione Bernoulliana

- Sia Y un carattere binario con distribuzione di Bernoulli di parametro π nella popolazione: $Y \sim Ber(\pi)$
- Sia $Y_1, Y_2, \ldots, Y_n$ un campione casuale di dimensione n
- Obiettivo: Trovare uno stimatore per la probabilità di successo della popolazione π
- Lo stimatore proporzione campionaria

$$T(Y_1, Y_2, \ldots, Y_n) = \overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

ha le seguenti proprietà

✓ È uno stimatore corretto per la proporzione di successi nella popolazione π:

$$\mu_{\overline{Y}} = \pi \qquad \text{per ogni } \pi \in (0,1)$$

✓ Ha errore quadratico medio, uguale alla sua varianza (per la correttezza):

$$MSE(\overline{Y}) = \sigma^2_{\overline{Y}} = \frac{\pi \cdot (1-\pi)}{n}$$

✓ È uno stimatore consistente

$$\lim_{n \to +\infty} MSE(\overline{Y}) = \lim_{n \to +\infty} \sigma^2_{\overline{Y}} = \lim_{n \to +\infty} \frac{\pi \cdot (1-\pi)}{n} = 0$$

✓ Per n sufficientemente grande, $\overline{Y} \approx N\left(\pi, \frac{\pi \cdot (1-\pi)}{n}\right)$

Intervalli di confidenza

- La stima puntuale utilizza le osservazioni di un campione casuale per ottenere una stima del parametro tramite un singolo valore numerico
- Tale approccio possiede un punto di debolezza: La stima ottenuta sul campione osservato potrebbe differire molto dal valore del parametro nella popolazione
- È dunque opportuno che l'inferenza su un certo parametro si basi non solo sulla stima puntuale ma dia informazioni anche su quanto precisa sia la stima, ossia su quanto è probabile che la stima sia vicina al vero valore del parametro
- A tal fine si considera oltre alla stima puntuale, un *intervallo* di stime plausibili al quale sia associato un fissato *livello di fiducia/confidenza*
- Obiettivo: Determinare un intervallo di numeri intorno alla stima puntuale che ci aspettiamo contenga, con un certo livello di fiducia, il valore del parametro

Intervalli di confidenza

- Un intervallo di confidenza o stima intervallare è un intervallo di numeri intorno alla stima puntuale che con un fissato livello di confidenza contiene il vero valore del parametro
- La probabilità che il **metodo** utilizzato per costruire l'intervallo di confidenza produca un intervallo di confidenza che contiene il vero valore del parametro è detto livello di confidenza o livello di fiducia e indicato con $1 - \alpha$
- Il livello di confidenza per un intervallo di confidenza descrive come si comporta il metodo utilizzato per la costruzione dell'intervallo di confidenza quando viene utilizzato più e più volte con differenti campioni casuali
- Se venissero selezionati più campioni casuali di una certa dimensione e ogni volta venisse costruito un intervallo di confidenza al livello di confidenza $1 - \alpha$ allora circa il $(1 - \alpha)\%$ degli intervalli conterrebbe il vero valore del parametro
  - ✓ Se venissero selezionati più campioni casuali di una certa dimensione e ogni volta venisse costruito un intervallo di confidenza al livello di confidenza $1 - \alpha = 0.90$ allora circa il $(1 - \alpha)\% = 90\%$ degli intervalli conterrebbe il vero valore del parametro

Interpretazione del livello di confidenza

- Maggiore è il livello di confidenza, maggiore sarà la possibilità che l'intervallo di confidenza contenga il vero valore del parametro
- Maggiore è il livello di confidenza, maggiore è l'ampiezza dell'intervallo di confidenza (ossia maggiore è il margine di errore) e quindi minore la precisione (accuratezza) della stima
- La scelta del livello di confidenza è il risultato di un compromesso tra desiderio che l'inferenza sia corretta e precisione della stima: al migliorare di un aspetto l'altro peggiora e viceversa
- Ad esempio con un livello di confidenza $1 - \alpha = 1$, l'intervallo di confidenza per la proporzione di una popolazione Bernoulliana sarebbe $[0, 1]$, che non è di alcun aiuto perché include tutti i possibili valori per π
- In genere si scelgono livelli di confidenza prossimi a 1: $0.9, 0.95, 0.99$

Intervallo di confidenza per un parametro

Obiettivo: Determinare due statistiche campionarie:

$$L_I = L_I(Y_1, \ldots, Y_n) \qquad \text{e} \qquad L_S = L_S(Y_1, \ldots, Y_n)$$

tali che

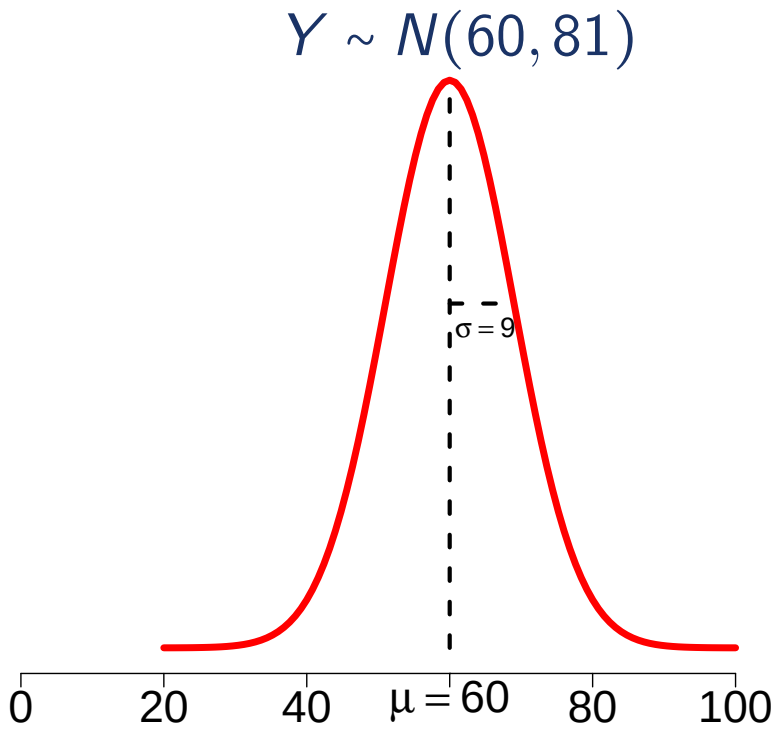
- $L_I \leq L_S$ per ogni possibile campione; e
- L'intervallo $[L_I, L_S]$ contiene il parametro θ con probabilità $1 - \alpha$

$$P(L_I \leq \theta \leq L_S) = P[L_I(Y_1, \ldots, Y_n) \leq \theta \leq L_S(Y_1, \ldots, Y_n)] = 1 - \alpha$$

- $1 - \alpha$ è detto livello di fiducia o livello di confidenza
- Una volta estratto il campione si ottiene l'intervallo di confidenza stimato: $[\ell_I; \ell_S]$
- Non è possibile sapere se l'intervallo stimato contenga o meno il valore vero del parametro
- La chiave per costruire un intervallo di confidenza è la distribuzione campionaria dello stimatore utilizzato per ottenere la stima puntuale
- La distribuzione campionaria dello stimatore permette di determinare la probabilità che lo stimatore produca una stima che cade entro una certa distanza dal parametro

Esempio: Punteggio alla prova INVALSI di matematica

- Si supponga che nella popolazione il punteggio alla prova INVALSI di matematica abbia distribuzione Normale con media $\mu = 60$ e varianza $\sigma^2 = 81$:



- Si supponga di conoscere la varianza della popolazione: $\sigma^2 = 81$ ma non la media μ

Esempio – Punteggio alla prova INVALSI di matematica: $Y \sim N(60, 81)$

- Si estraggono dalla popolazione 10 campioni di dimensione $n = 25$

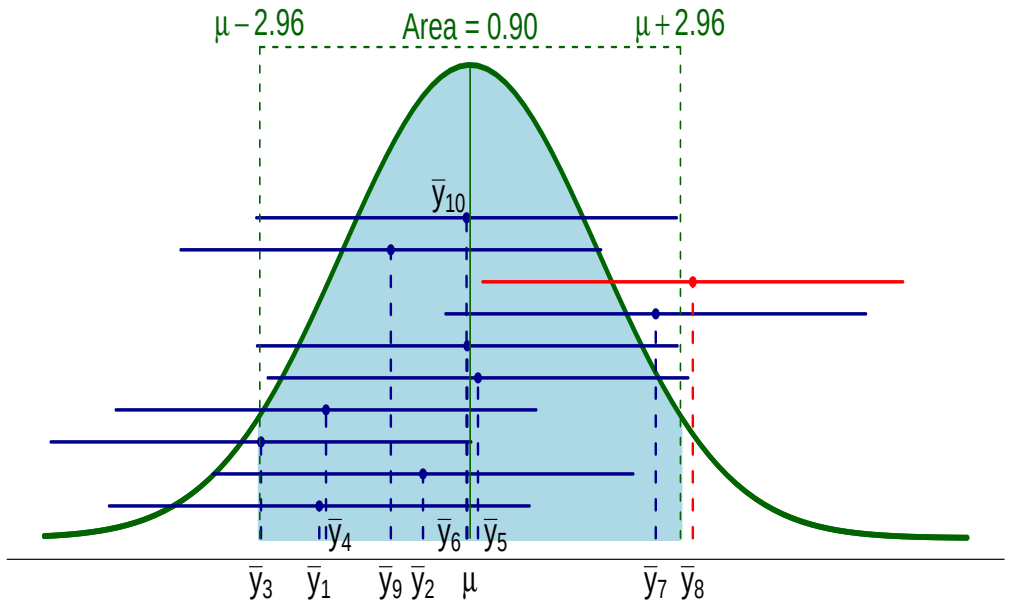
Campione	Media Campionaria
1	57.88
2	59.34
3	57.06
4	57.97
5	60.11
6	59.96
7	62.61
8	63.13
9	58.88
10	59.95

- L'errore standard della media campionaria è

$$e.s.(\overline{Y}) = \frac{\sigma}{\sqrt{n}} = \frac{9}{\sqrt{25}} = 1.8$$

Esempio – Punteggio alla prova INVALSI di matematica: $Y \sim N(60, 81)$
Intervalli di confidenza con livello di confidenza del 90%

Campione ($n = 25$)	$\overline{y}$	ℓ_I	ℓ_S
1	57.88	54.92	60.84
2	59.34	56.38	62.30
3	57.06	54.10	60.02
4	57.97	55.01	60.93
5	60.11	57.15	63.07
6	59.96	57.00	62.92
7	62.61	59.65	65.57
8	**63.13**	**60.17**	**66.10**
9	58.88	55.92	61.84
10	59.95	56.99	62.91



Il 10% degli intervalli di confidenza a livello di confidenza del 90% non includono il vero valore del parametro

Stima puntuale vs stima intervallare

	Stima puntuale	Stima intervallare
Campione casuale	$Y_1, \ldots, Y_n$	$Y_1, \ldots, Y_n$
Obiettivo	Stima puntuale per θ	Stima per intervallo per θ
Strumento	Stimatore puntuale $T = T(Y_1, \ldots, Y_n)$	Intervallo di confidenza $[L_I, L_S] =$ $[L_I(Y_1, \ldots, Y_n); L_S(Y_1, \ldots, Y_n)]$
Accuratezza	Errore quadratico Medio $MSE(T) = \mathbb{E}[(T-\theta)^2]$	Livello di confidenza $P(L_I \leq \theta \leq L_S) = 1-\alpha$
Campione osservato	$y_1, \ldots, y_n$	$y_1, \ldots, y_n$
Risultato	$t = T(y_1, \ldots, y_n)$	$[\ell_I, \ell_S] =$ $[L_I(y_1, \ldots, y_n); L_S(y_1, \ldots, y_n)]$

Costruzione di un intervalli di confidenza

- Se la distribuzione campionaria dello stimatore è Normale centrata sul vero valore del parametro (come la media campionaria nel caso di popolazioni Normali), allora
  - ✓ con probabilità di circa il 95% lo stimatore produrrà una stima del parametro che ricade a 2 errori standard dal parametro
  - ✓ con probabilità di circa il 99.7% lo stimatore produrrà una stima del parametro che ricade a 3 errori standard dal parametro
  - ✓ minore è l'errore standard, maggiore è la precisione dello stimatore
- In pratica la distribuzione campionaria dello stimatore è solo approssimativamente Normale
- Un intervallo di confidenza si può dunque costruire aggiungendo e sottraendo dalla stima puntuale un multiplo dell'errore standard dello stimatore
- Il multiplo dell'errore standard dello stimatore è detto margine di errore e dipende dall'ampiezza (errore standard) della distribuzione campionaria dello stimatore
- Forma tipica degli intervalli di confidenza:

$$\text{Stima puntuale} \pm \text{Margine di Errore}$$

Intervallo di confidenza per la media

- Supponiamo che il carattere di interesse Y sia quantitativo con media μ nella popolazione
- L'intervallo di di confidenza per la media μ ha la forma

$$\text{Stima puntuale} \pm \text{Margine di errore}$$

dove il margine di errore è un multiplo dell'errore standard dello stimatore

- Si distinguono tre casi
  - ✓ Intervallo di confidenza per la media di una *popolazione Normale con varianza nota*
  - ✓ Intervallo di confidenza per la media di una *popolazione Normale con varianza non nota*
  - ✓ Intervallo di confidenza per la media di una *popolazione non Normale per campioni di dimensione elevata*

Intervallo di confidenza per la media di una popolazione Normale con varianza nota

- Distribuzione del carattere nella popolazione: $Y \sim N(\mu, \sigma^2)$ con σ^2 nota
- Campione casuale di dimensione n (qualsiasi): $Y_1, \ldots, Y_n$
- Stimatore della media = Media campionaria

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- È noto che se $Y \sim N(\mu, \sigma^2)$ allora $\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ quindi $\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- Fissato $(1-\alpha)$ con probabilità $(1-\alpha)\%$ lo stimatore media campionaria produce valori della media campionaria entro $z_{\alpha/2}$ errori standard dalla media (non nota) della popolazione

Intervallo di confidenza per la media di una popolazione Normale con varianza nota

- Una volta osservato il campione, $y_1, \ldots, y_n$, si ha un solo valore dello stimatore

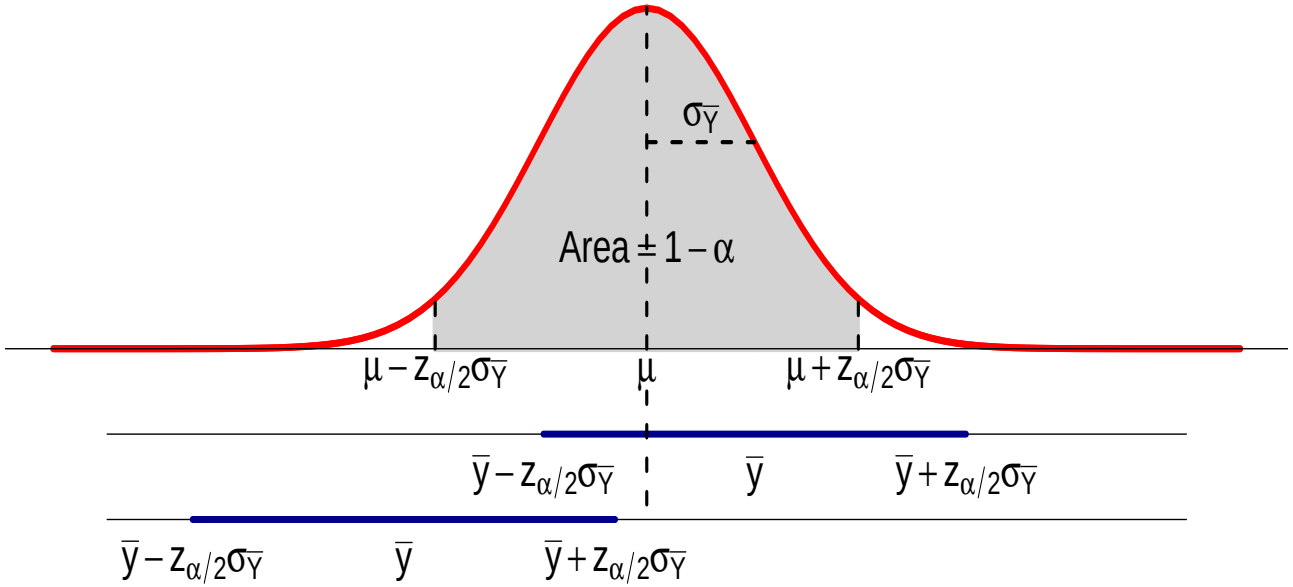
$$\overline{y} = \frac{y_1 + \ldots + y_n}{n}$$

e non è noto se tale valore si trova entro $z_{\alpha/2} \cdot \sigma_{\overline{Y}}$ unità da μ

- Se $\overline{y}$ si trova entro $z_{\alpha/2} \cdot \sigma_{\overline{Y}}$ unità da μ allora l'intervallo di estremi

$$\overline{y} \pm z_{\alpha/2} \cdot \sigma_{\overline{Y}}$$

contiene μ, altrimenti tale intervallo non contiene μ



Intervallo di confidenza per la media di una popolazione Normale con varianza nota

- Quindi se $Y \sim N(\mu, \sigma^2)$ con σ^2 nota, il margine di errore può essere ottenuto moltiplicando l'errore standard della media campionaria, $\sigma_{\overline{Y}} = \sigma/\sqrt{n}$, per un opportuno valore z della distribuzione Normale

$$\text{Margine di errore } = z \cdot \frac{\sigma}{\sqrt{n}}$$

dove il valore z dipende dal livello di confidenza

Intervallo di confidenza al livello di confidenza $1-\alpha$ per la media μ di una popolazione Normale con varianza σ^2 nota

$$IC_{1-\alpha}(\mu) = \left[\overline{Y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{Y} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

dove $z_{\alpha/2}$ è il valore che nella Normale standard lascia alla sua destra un'area pari a $\alpha/2$: $P(Z > z_{\alpha/2}) = \alpha/2$

Esempio: Punteggio alla prova INVALSI di matematica

- Distribuzione di Y nella popolazione di studenti di terza media: $Y \sim N(\mu, \sigma^2 = 81)$
- Si estrae dalla popolazione 1 campione casuale di dimensione $n = 25$

u_i	y_i	u_i	y_i
1	47.68	14	51.55
2	59.48	15	51.36
3	62.63	16	72.85
4	52.65	17	57.91
5	57.17	18	58.67
6	61.52	19	50.98
7	50.72	20	62.91
8	60.98	21	65.67
9	44.11	22	56.00
10	60.31	23	67.28
11	59.82	24	53.97
12	73.17	25	50.63
13	56.89		

- Media campionaria

$$\overline{y} = \frac{1446.912}{25} = 57.88$$

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$

  Dalle tavole della normale standard $z_{\alpha/2} = z_{0.025} = 1.96$

  Quindi

$$\left[57.88 - 1.96\frac{9}{\sqrt{25}}; 57.88 + 1.96\frac{9}{\sqrt{25}}\right] = [54.35; 61.40]$$

- Il punteggio medio alla prova INVALSI di matematica per studenti della terza media è compreso tra 54.35 e 61.40 al livello di confidenza del 95%

Intervallo di confidenza per la media di una popolazione Normale con varianza non nota

- Distribuzione del carattere nella popolazione: $Y \sim N(\mu, \sigma^2)$ con μ e σ^2 non note
- Campione casuale di dimensione n (qualsiasi): $Y_1, \ldots, Y_n$
- Stimatore della media = Media campionaria

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- La varianza non è nota, quindi deve essere stimata
- Per stimare la varianza della popolazione si utilizza lo stimatore varianza campionaria corretta:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \frac{\sum_{i=1}^{n} Y_i^2 - n \cdot \overline{Y}^2}{n-1}$$

Intervallo di confidenza per la media di una popolazione Normale con varianza non nota

- Se $Y \sim N(\mu, \sigma^2)$, allora $\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ e quindi $\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- Stimatore della varianza e dell'errore standard della media campionaria

$$\hat{\sigma}^2_{\overline{Y}} = \frac{S^2}{n} \qquad e \qquad se_{\overline{Y}} = \hat{\sigma}_{\overline{Y}} = \frac{S}{\sqrt{n}} \qquad \text{dove } S = \sqrt{S^2}$$

- Usando la notazione del libro, il simbolo *se* (standard error) rappresenta la stima dell'errore standard
- Se si sostituisce la deviazione standard di Y, σ, con la deviazione standard campionaria s per ottenere l'errore standard stimato $s/\sqrt{n}$, si introduce un ulteriore fonte di errore
- Sostituendo la varianza incognita con un suo stimatore si ottiene

$$\frac{\overline{Y} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$$

dove t_{n-1} è la distribuzione t di Student con $n-1$ gradi di libertà (gdl)

La distribuzione t–Student

William Sealy Gosset

- William Sealy Gosset (Canterbury, 13 giugno 1876 – Beaconsfield, 16 ottobre 1937) è stato uno statistico inglese, meglio noto in statistica come il signor Student della distribuzione t di Student
- Nel 1908 pubblica con lo pseudonimo Student l'articolo nel quale introduce la distribuzione t di Student
- Gosset dovette usare uno pseudonimo poiché la fabbrica Guinness presso la quale lavorava vietava la pubblicazione di articoli per evitare la divulgazione dei segreti di produzione della birra

La distribuzione t–Student

- La distribuzione t di Student dipende da un parametro a valori interi positivi detto *gradi di libertà* (gdl)
- La distribuzione t di Student ha una forma campanulare, simmetrica intorno alla media uguale a zero
- La deviazione standard della distribuzione t di Student è leggermente più grande di 1 (il valore esatto dipende da quelli che vengono chiamati gradi di libertà indicati con gdl)
- La distribuzione t di Student presenta un'ampiezza leggermente diversa per ciascun differente valore dei gdl
- La distribuzione t di Student presenta aree sulle code più grandi (più pesanti) ed è più dispersa rispetto alla distribuzione normale standard
- Quanto più elevato è il valore dei gdl tanto più la distribuzione tenderà a assomigliare a una distribuzione normale standard

La distribuzione t–Student

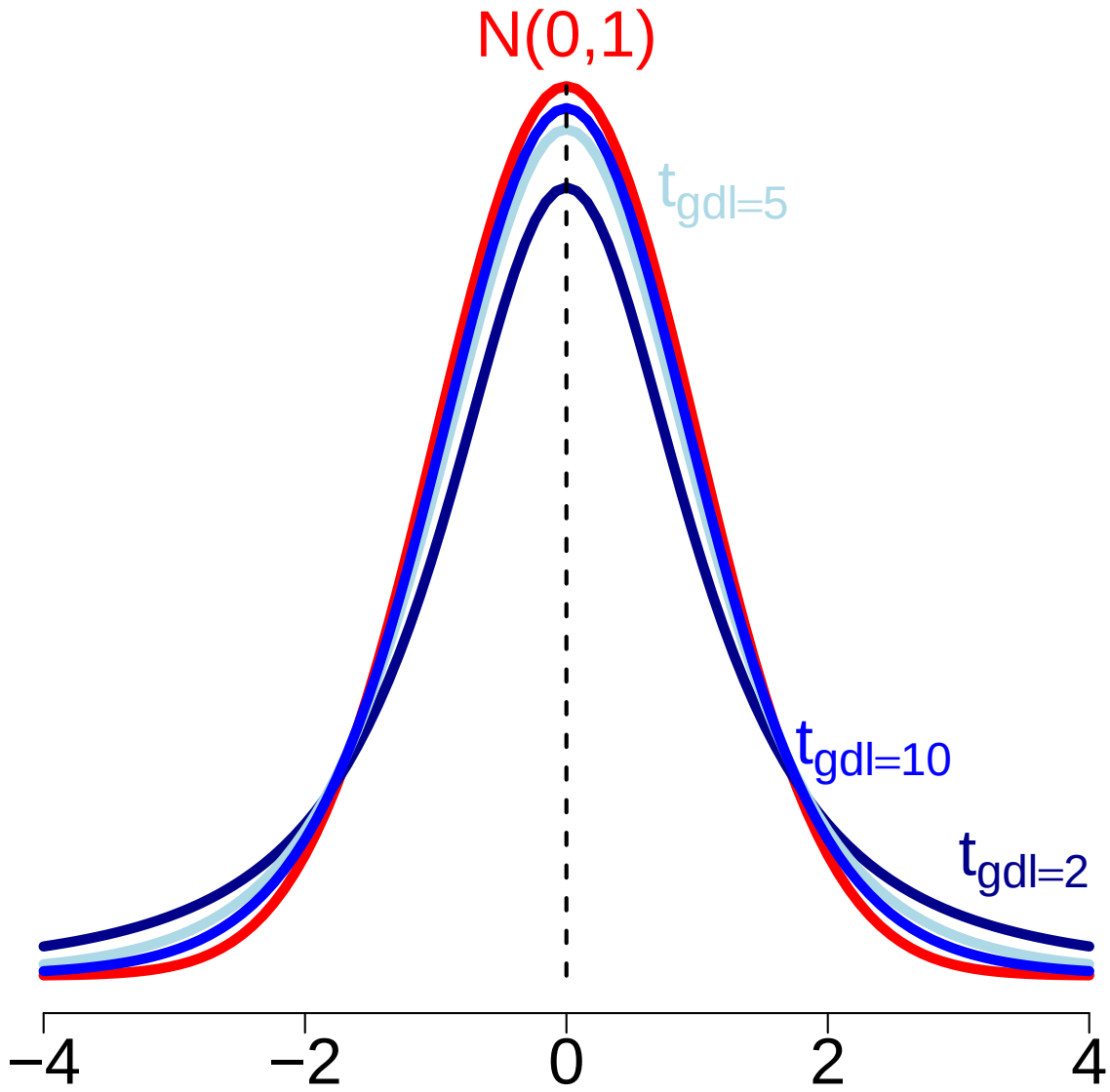
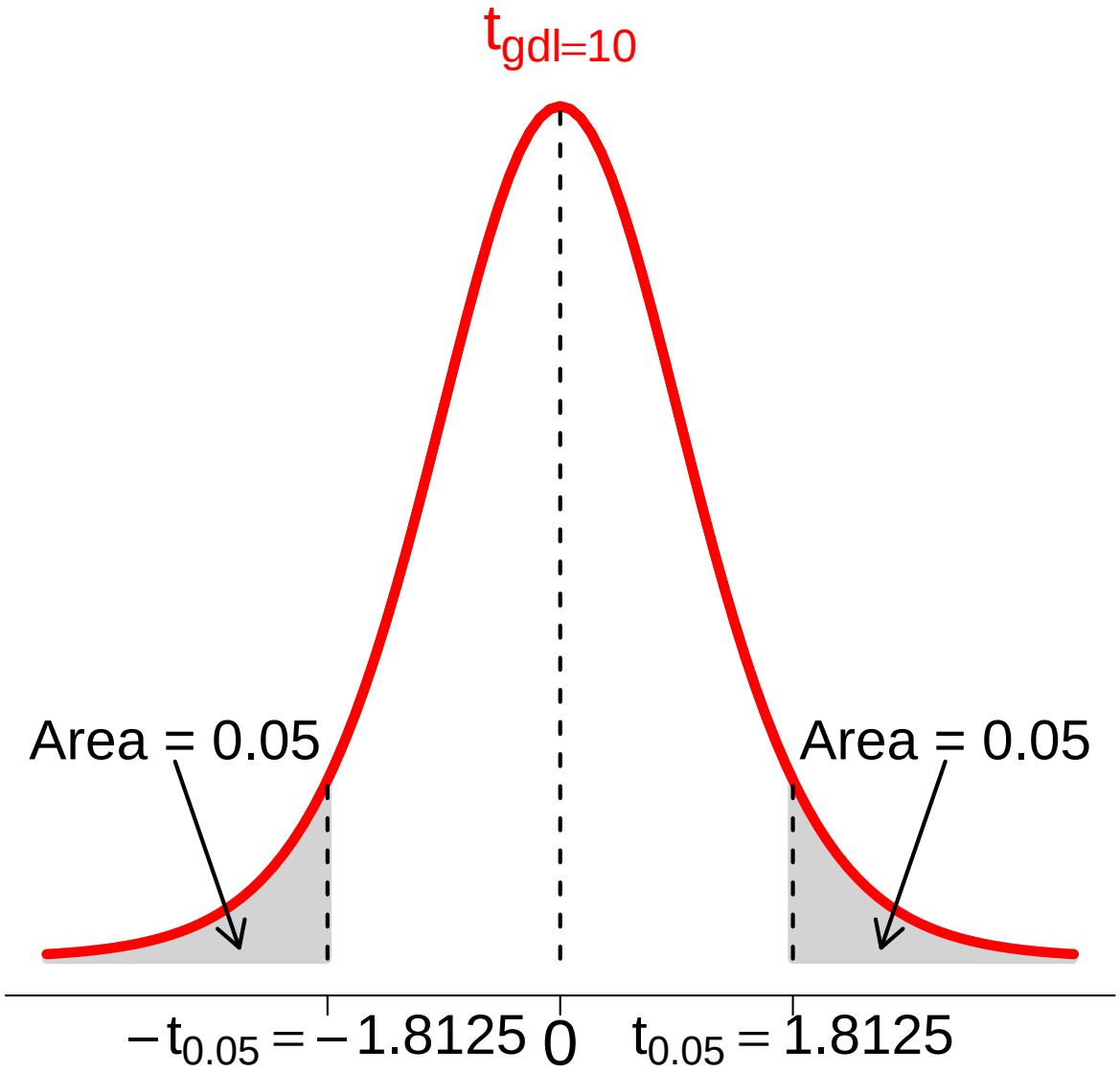


Tavola della distribuzione t– Student

Gradi di Libertà	Area della coda destra della distribuzione *t di Student*						
	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614	3.4960
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317	3.4602
70	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108	3.4350
80	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953	3.4163
90	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019
100	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
Infinito	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

Leggere la tavola della distribuzione t–Student



Intervallo di confidenza per la media di una popolazione Normale con varianza non nota

Dato un campione casuale di dimensione n estratto da una popolazione Normale con media e varianza entrambe ignote, l'intervallo di confidenza per la media a livello di confidenza $1-\alpha$ è dato da:

$$\left[\overline{Y} - t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}; \overline{Y} + t_{(n-1),\alpha/2} \cdot \frac{S}{\sqrt{n}}\right]$$

dove $t_{(n-1),\alpha/2}$ è il valore che nella distribuzione t–Student con $n-1$ gdl lascia alla sua destra un'area pari a $\alpha/2$: $P(T_{n-1} > t_{(n-1),\alpha/2}) = \alpha/2$

Esempio – La perdita di calcio

- Il latte materno contiene una certa quantità di calcio, una parte del quale deriva direttamente dal calcio contenuto nelle loro ossa
- Alcune donne, quindi, durante l'allattamento possono andare incontro a demineralizzazione ossea
- I ricercatori hanno misurato la variazione percentuale di calcio nelle vertebre di 16 mamme nel corso di tre mesi d'allattamento:

Mamma_i	1	2	3	4	5	6	7	8
y_i	−4.7	0.4	−1.0	−0.8	−6.5	−6.5	−4.0	0.2
Mamma_i	9	10	11	12	13	14	15	16
y_i	0.3	−5.2	−3.1	−2.0	−3.0	−0.3	−4.7	1.7

- Obiettivo: Costruire un intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$ per la media della variazione percentuale di calcio nelle vertebre nella popolazione delle mamme
- Condizioni
  - ✓ Le osservazioni sono realizzazioni di un campione casuale di dimensione $n = 16$, $Y_1, \ldots, Y_{16}$, estratto dalla popolazione delle mamme
  - ✓ Nella popolazione delle mamme la variazione percentuale di calcio nelle vertebre, Y, ha distribuzione Normale: $Y \sim N(\mu, \sigma^2)$

Esempio – La perdita di calcio

Mamma_i	y_i	$(y_i - \bar{y})^2$	y_i^2
1	−4.7	5.0625	22.09
2	−1.0	2.1025	1.00
3	−6.5	16.4025	42.25
4	−4.0	2.4025	16.00
5	0.3	7.5625	0.09
6	−3.1	0.4225	9.61
7	−3.0	0.3025	9.00
8	−4.7	5.0625	22.09
9	0.4	8.1225	0.16
10	−0.8	2.7225	0.64
11	−6.5	16.4025	42.25
12	0.2	7.0225	0.04
13	−5.2	7.5625	27.04
14	−2.0	0.2025	4.00
15	−0.3	4.6225	0.09
16	1.7	17.2225	2.89
Totale	−39.2	103.2000	199.24

- Stima della media di Y

$$\bar{y} = \frac{-39.2}{16} = -2.45$$

- Stima della varianza di Y

$$s^2 = \frac{103.2000}{16-1} = \frac{199.24 - 16 \cdot (-2.45)^2}{16-1} = 6.88$$

- Stima della deviazione standard di Y

$$s = \sqrt{6.88} = 2.623$$

Esempio – La perdita di calcio

- Stima dell'errore standard della media campionaria

$$se_{\overline{Y}} = \frac{s}{\sqrt{n}} = \frac{2.623}{\sqrt{16}} = 0.656$$

- Valore $t_{n-1,\alpha/2}$ per $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Longrightarrow \alpha = 0.05 \Longrightarrow \alpha/2 = 0.025$$

  Quindi $t_{(n-1),\alpha/2} = t_{15,0.025} = 2.131$

- Intervallo di confidenza per la media μ della variazione percentuale di calcio nelle vertebre delle mamme al livello di confidenza $1 - \alpha = 0.95$

$$IC_{0.95}(\mu) = -2.45 \pm 2.131 \cdot \frac{2.623}{\sqrt{16}} =$$

$$\left[-2.45 - 2.131 \cdot \frac{2.623}{\sqrt{16}}; -2.45 + 2.131 \cdot \frac{2.623}{\sqrt{16}}\right] = [-3.85; -1.05]$$

La distribuzione t–Student e la distribuzione Normale

- Al crescere dei gradi di libertà la distribuzione t–Student tende (assomiglia sempre di più) alla distribuzione Normale standard: Se i *gdl* sono sufficientemente grandi, $t_{gdl} \approx N(0,1)$
- Al crescere della dimensione del campione, la distribuzione t–Student diventa sempre meno dispersa e assomiglia sempre di più alla distribuzione Normale
- Se i gradi di libertà sono sufficientemente grandi, $t_{gdl,\alpha/2} \approx z_{\alpha/2}$: Si confrontino i valori sulle tavole

La distribuzione t–Student e la distribuzione Normale

	Area della coda destra della distribuzione *t di Student*						
gdl	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019
100	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
Infinito	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

Valore	Area della coda destra della distribuzione *Normale*						
z	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
z	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

- Per $gdl > 100$ le differenze tra distribuzione Normale e distribuzione t–Student con gdl gradi di libertà sono molto piccole, quindi non si riportano nella tavola i valori della t

La distribuzione t–Student e la distribuzione Normale

- Si noti che se $Y \sim N(\mu, \sigma^2)$ allora $\overline{Y} \sim N(\mu, \sigma^2/n)$
- La distribuzione t–Student viene introdotta per tener conto dell'incertezza sulla varianza σ^2, che se non nota deve essere stimata con S^2
- La distribuzione t–Student è infatti più dispersa della distribuzione Normale standard
- Nella costruzione dell'intervallo di confidenza per la media di una variabile Y che ha distribuzione nella popolazione Normale di media μ e varianza σ^2 non noti, i *gdl* della distribuzione t–Student dipendono dalla dimensione del campione
- Se la dimensione del campione n è sufficientemente grande la distribuzione t–Student tende (assomiglia sempre di più) alla distribuzione Normale
- Quindi, se la dimensione del campione n è sufficientemente grande,

$$IC_{1-\alpha}(\mu) = \overline{y} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} \approx \overline{y} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- Lo stimatore S^2 è uno stimatore consistente: Al crescere della dimensione del campione S^2 si avvicina sempre più al vero valore della varianza σ^2
- Al crescere della dimensione del campione l'errore standard stimato della media campionaria, $\frac{s}{\sqrt{n}}$, approssima sempre meglio il vero errore standard, $\frac{\sigma}{\sqrt{n}}$

Esempio – Qualità delle acque

- Un'associazione ambientalista raccoglie un litro d'acqua in 101 punti scelti a caso lungo un fiume e misura, in ogni punto, la quantità di ossigeno presente
- La media campionaria è $\bar{y} = 4.62$ milligrammi (mg) e la deviazione standard (campionaria) è $s = 0.92$ mg
- Si ipotizza che la quantità di ossigeno presente nel fiume abbia una distribuzione Normale di media μ e varianza σ^2, non note
- Intervallo di confidenza (esatto) al livello di confidenza del 95%: Il valore $t_{101-1,0.025} = 1.984$, quindi

$$IC_{0.95}(\mu) = \left[4.62 - 1.984 \cdot \frac{0.92}{\sqrt{101}}; 4.62 + 1.984 \cdot \frac{0.92}{\sqrt{101}}\right] = [4.438; 4.802]$$

- Intervallo di confidenza (approssimato) al livello di confidenza del 95%: Il valore $z_{0.025} = 1.96$, quindi

$$IC_{0.95}(\mu) = \left[4.62 - 1.96 \cdot \frac{0.92}{\sqrt{101}}; 4.62 + 1.96 \cdot \frac{0.92}{\sqrt{101}}\right] = [4.441; 4.799]$$

Intervallo di confidenza per la media di una popolazione non Normale

- Sia Y una variabile continua che rappresenta il fenomeno di interesse nella popolazione
- La distribuzione di Y nella popolazione non è nota (non Normale)
- Obiettivo: Trovare un intervallo di confidenza per la media di Y nella popolazione
  - ✓ Campioni di dimensione elevata
  - ✓ Campioni di dimensione piccola

Intervallo di confidenza per la media di una popolazione non Normale: Campioni di dimensione elevata

- Si supponga di essere interessati a un carattere Y (continuo) con distribuzione nella popolazione non nota
- Obiettivo: Costruire un intervallo di confidenza per la media, μ, di Y nella popolazione utilizzando un campione casuale $Y_1, \ldots, Y_n$ di dimensione n
- Si ricorda che per il teorema del limite centrale, se la dimensione del campione, n è sufficiente grande, la distribuzione campionaria della media campionaria può essere approssimata dalla distribuzione Normale:

$$\overline{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ per } n \text{ sufficientemente grande}$$

Intervallo di confidenza per la media di una popolazione non Normale: Campioni di dimensione elevata

Quindi, per n sufficientemente grande l'intervallo di confidenza per la media μ di Y al livello di confidenza $1-\alpha$ può essere approssimato come segue

- Se la varianza di Y nella popolazione, σ^2, è nota:

$$IC_{1-\alpha}(\mu) \approx \left[\overline{Y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{Y} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

- Se la varianza di Y nella popolazione, σ^2 è non nota

$$IC_{1-\alpha}(\mu) \approx \left[\overline{Y} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}; \overline{Y} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}\right]$$

dove $S = \sqrt{S^2}$

Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

- Si supponga che la variabile di interesse Y sia binaria
- La distribuzione di Y nella popolazione è Bernoulliana con probabilità di successo π: $Y \sim \text{Bernoulli}(\pi)$ e π è il parametro di interesse
- Stimatore puntuale di π:

Proporzione di successi campionaria = Media campionaria

$$\hat{\pi} = \overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- Si ricorda che

$$\mu_{\hat{\pi}} = \pi \qquad \text{e} \qquad \sigma^2_{\hat{\pi}} = \frac{\pi \cdot (1 - \pi)}{n}$$

Quindi l'errore standard della proporzione campionaria è

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

- La proporzione campionaria $\hat{\pi}$ è una media campionaria, quindi per campioni di dimensioni sufficientemente elevate la sua distribuzione campionaria si può approssimare con una distribuzione Normale per il teorema del limite centrale
- Formalmente, per il il teorema del limite centrale, se n è sufficientemente grande

$$\hat{\pi} \approx N\left(\pi, \frac{\pi \cdot (1-\pi)}{n}\right) \quad \text{e quindi} \quad \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}} \approx N(0,1)$$

- In pratica, il valore dell'errore standard della proporzione campionaria,

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

non è noto perché dipende dal parametro π ignoto che interessa stimare

- L'errore standard della proporzione campionaria viene stimato sostituendo a π la proporzione campionaria

$$se_{\hat{\pi}} = \hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}$$

Intervallo di confidenza per una proporzione: Campioni di dimensione elevata

Intervallo di confidenza al livello di confidenza $1-\alpha$ per la proporzione

$$IC_{1-\alpha}(\pi) = \left[\hat{\pi} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}; \hat{\pi} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}}\right]$$

dove $z_{\alpha/2}$ è il valore che nella Normale standard lascia alla sua destra un'area pari a $\alpha/2$: $P(Z > z_{\alpha/2}) = \alpha/2$

- Con un livello di confidenza pari a $1-\alpha$ si ha una probabilità pari a α che il metodo produca un intervallo di confidenza che *non* contiene il vero valore del parametro

Esempio: Il problema del fumo negli adolescenti

- Per analizzare il fenomeno del fumo tra gli adolescenti, si estrae un campione casuale semplice di $n = 900$ adolescenti.
- In tale campione il numero di adolescenti che fumano è pari a 180
- Nella popolazione la variabile di interesse $Y \sim \text{Bernoulli}(\pi)$
- Obiettivo: Stimare π, la proporzione di adolescenti che fumano nell'intera popolazione degli adolescenti
- Stima puntuale e stima dell'errore standard

$$\hat{\pi} = \overline{y} = \frac{180}{900} = 0.2 \quad e \quad se_{\hat{\pi}} = \sqrt{\frac{\hat{\pi} \cdot (1 - \hat{\pi})}{n}} = \sqrt{\frac{0.2 \cdot 0.8}{900}} = \sqrt{0.00018} = 0.0133$$

- Intervallo di confidenza al livello di confidenza del 95%

$$IC_{1-0.05}(\pi) = [0.2 - 1.96 \cdot 0.0133; 0.2 + 1.96 \cdot 0.0133] = [0.174; 0.226]$$

- La percentuale dei adolescenti che fumano è (al livello di confidenza del 95%) non meno di 17.4% e non più del 22.6%
- Tutti i valori contenuti nell'intervallo di confidenza sono inferiori a 0.25, quindi i dati osservati suggeriscono che meno di 1/4 della popolazione degli adolescenti fuma

Riepilogo

Parametro	Stima puntuale	Errore Standard	Intervallo di confidenza
Media μ $Y \sim N(\mu, \sigma^2)$ (σ^2 nota)	$\bar{y}$	$\frac{\sigma}{\sqrt{n}}$	$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Media μ $Y \sim N(\mu, \sigma^2)$ (σ^2 non nota)	$\bar{y}$	$\frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{n}}$	$\bar{y} \pm t_{(n-1),\alpha/2}\frac{s}{\sqrt{n}}$
Media μ $Y \sim ?$ (σ^2 non nota)	$\bar{y}$	$\frac{s}{\sqrt{n}}$	$\bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
Proporzione π	$\hat{\pi} = \bar{y}$	$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

Inferenza Statistica: Test delle ipotesi

- Un'ipotesi statistica è un'affermazione sulla popolazione
- In molti studi, un'ipotesi statistica è un'affermazione su un parametro che descrive una caratteristica di una variabile di interesse nella popolazione
- In particolare un'ipotesi statistica è un'affermazione relativa ai valori del parametro: Un'ipotesi statistica specifica che il parametro assume un particolare valore (o valori in un particolare intervallo)
- Esempi
  - ✓ Un giocatore di basket afferma di realizzare l'80% dei tiri liberi
  - ✓ Metà o più cittadini di un certo paese è soddisfatto del servizio sanitario pubblico
  - ✓ Secondo il produttore di un certo tipo di batterie per autovetture, la durata media delle batterie è di 3 400 ore
  - ✓ Il reddito medio delle famiglie italiane è inferiore alla media europea (pari a 2 500 Euro lordi mensili)
- Un test di significatività quantifica l'evidenza fornita dai dati nei confronti di una certa ipotesi riguardante la popolazione

Condizioni di base

- Definizione della variabile di interesse
  - ✓ Importante tener conto della natura della variabile
  - ✓ Procedure test diverse per caratteri quantitativi e caratteri categoriali
- Campione casuale
- Distribuzione della variabile di interesse nella popolazione
  - ✓ In alcuni casi si assume che la variabile di interesse abbia una particolare distribuzione nella popolazione
  - ✓ Esempi

$$\begin{aligned} &\text{Variabile continua:} && Y \sim N(\mu, \sigma^2) \\ &\text{Variabile binaria:} && Y \sim Ber(\pi) \end{aligned}$$

- Dimensione campionaria: All'aumentare della dimensione campionaria migliora la performance (attendibilità) di un test

Ipotesi

- Ipotesi nulla, H_0: Ipotesi preesistente all'osservazione dei dati campionari, ritenuta vera fino a prova contraria
- Ipotesi alternativa, H_a: Ipotesi che si contrappone all'ipotesi nulla
- Spesso l'ipotesi nulla specifica un particolare valore del parametro di interesse nella popolazione
- L'ipotesi alternativa può specificare un altro particolare valore del parametro di interesse nella popolazione oppure affermare che il parametro prende valore in qualsiasi altro intervallo (che non comprende il valore ipotizzato sotto H_0)
- Ipotesi unilaterali (unidirezionali o a una coda) e ipotesi bilaterali (bidirezionali o a due code)
- Esempi

$$
\begin{array}{lcll}
H_0 : \theta = \theta_0 & \textit{versus} & H_a : \theta = \theta_a & \\
H_0 : \theta = \theta_0 & \textit{versus} & H_a : \theta \neq \theta_0 & (H_a \text{ bilaterale}) \\
H_0 : \theta = \theta_0 & \textit{versus} & H_a : \theta < \theta_0 & (H_a \text{ unilaterale sinistra}) \\
H_0 : \theta = \theta_0 & \textit{versus} & H_a : \theta > \theta_0 & (H_a \text{ unilaterale destra})
\end{array}
$$

Si noti che sotto H_0 si specifica un *solo* valore mentre sotto H_a si considera anche un intervallo di valori

Ipotesi: Esempi

Altezza media degli abitanti di un paese: $Y \sim N(\mu, \sigma^2)$

- $H_0 : \mu = \mu_0 = 175$ *versus* $H_a : \mu = 178$
- $H_0 : \mu = \mu_0 = 175$ *versus* $H_a : \mu \neq 175$
- $H_0 : \mu = \mu_0 = 175$ *versus* $H_a : \mu > 175$
- $H_0 : \mu = \mu_0 = 175$ *versus* $H_a : \mu < 175$

Proporzione di dipendenti femmine al corso di formazione: $Y \sim Ber(\pi)$

- $H_0 : \pi = \pi_0 = 0.5$ *versus* $H_a : \pi = 0.4$
- $H_0 : \pi = \pi_0 = 0.5$ *versus* $H_a : \pi \neq 0.5$
- $H_0 : \pi = \pi_0 = 0.5$ *versus* $H_a : \pi < 0.5$
- $H_0 : \pi = \pi_0 = 0.5$ *versus* $H_a : \pi > 0.5$

Dimostrazione per contraddizione

- Un test di significatività è una regola che permette di valutare l'evidenza campionaria contro l'ipotesi nulla
- Un test si pone l'obiettivo di investigare se i dati contraddicono l'ipotesi nulla (a favore dell'ipotesi alternativa): *Dimostrazione per contraddizione*
- Procedura pratica
  - ✓ Supporre che l'ipotesi nulla, H_0, è vera
  - ✓ Nell'ipotesi che H_0 sia vera, valutare se i dati osservati sono verosimili
  - ✓ Se nell'ipotesi che H_0 sia vera i dati osservati risultano poco verosimili, allora si rifiuta l'ipotesi nulla a favore dell'ipotesi alternativa
- L'ipotesi nulla è *rifiutata* o *non rifiutata*
  - ✓ Per variabilità campionaria, esiste una gamma di valori possibili per il parametro: il valore ipotizzato in H_0 è uno dei tanti possibili
  - ✓ L'intervallo di confidenza fornisce un intervallo di valori possibili per un parametro
  - ✓ Si dice dunque "Non si rifiuta l'ipotesi nulla H_0" piuttosto che "Si accetta l'ipotesi nulla H_0" per mettere in evidenza che il valore ipotizzato sotto H_0 è solamente uno dei tanti possibili

Esempio – Un grande realizzatore di tiri liberi

- Un giocatore di basket afferma di realizzare l'80% dei tiri liberi
- L'ipotesi che interessa sottoporre a verifica è che la probabilità che il giocatore realizzi un tiro libero è 0.8. L'ipotesi 'alternativa' è che il giocatore di basket realizzi meno dell'80% dei tiri liberi:

$$H_0 : \pi = 0.8 \qquad \textit{versus} \qquad H_a : \pi < 0.8$$

- Al fine di verificare tale ipotesi si seleziona un campione casuale di $n = 20$ tiri liberi del giocatore
- Se H_0 fosse vera dovremmo osservare che il giocatore realizza circa l'80% dei tiri liberi, sebbene per variabilità campionaria non ci aspettiamo che il giocatore realizzi esattamente l'80% (ossia 16) dei tiri liberi selezionati
- Si supponga che su 20 liberi il giocatore ne realizzi solo 8, quindi $\hat{\pi} = 8/20 = 0.4$
- Quesito: Quanto è improbabile osservare un risultato di questo tipo?
- Quanto deve essere inferiore a 0.8 la proporzione campionaria dei tiri liberi realizzati per non credere all'affermazione del giocatore?
  - ✓ Se si ritiene che la differenza tra proporzione campionaria, $\hat{\pi} = 0.4$, e il valore $\pi = 0.8$ ipotizzato sotto H_0 è eccessiva, si tenderà a rifiutare l'ipotesi nulla, ossia si tenderà a non credere all'affermazione del giocatore
  - ✓ Se si ritiene che la differenza tra proporzione campionaria, $\hat{\pi} = 0.4$, e il valore $\pi = 0.8$ ipotizzato sotto H_0 sia semplicemente dovuta a variabilità campionaria, allora si conclude che non ci sono sufficienti evidenze empiriche per rifiutare H_0

Errore di primo e secondo tipo

- A causa della variabilità campionaria, le decisioni che si prendono utilizzando i test statistici hanno sempre un certo grado di incertezza
- Una decisione quindi può essere sbagliata
- Esistono due tipi di errore: Errore di I tipo e errore di II tipo
- Errore del I tipo: Si rifiuta l'ipotesi nulla mentre questa è vera.
- Errore del II tipo: Non si rifiuta l'ipotesi nulla mentre questa è falsa
- Quattro possibili risultati:

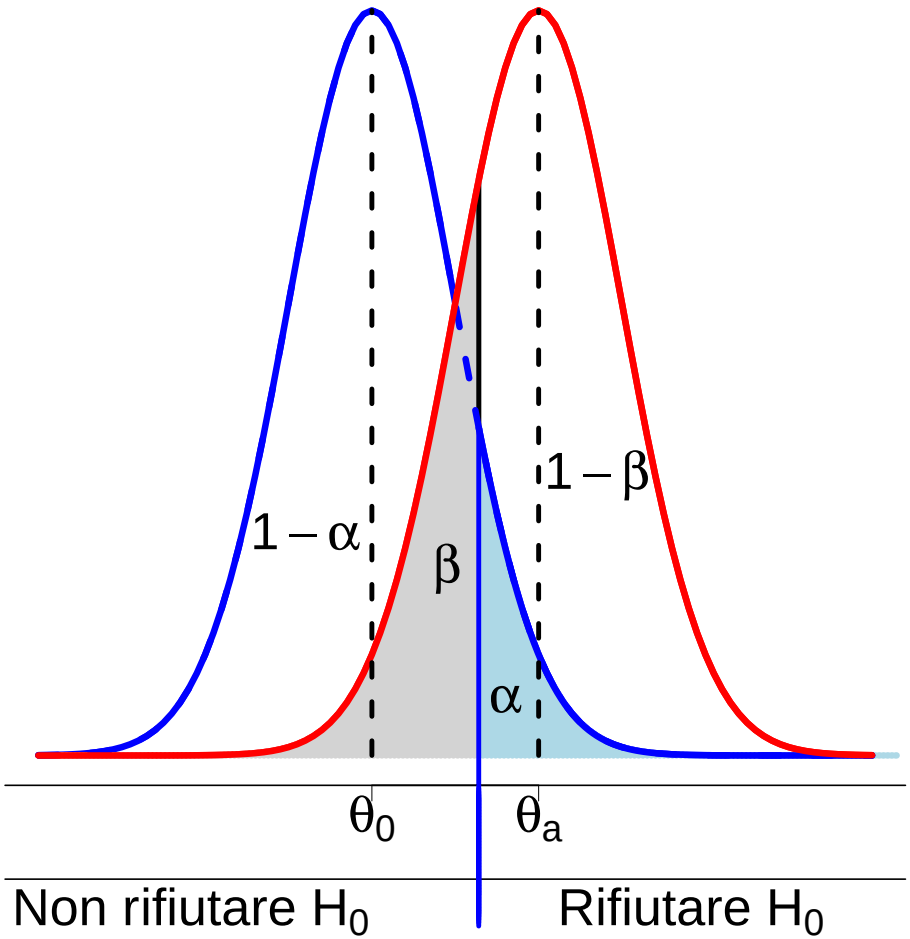
	Decisione	
	Non si rifiuta H_0	Si rifiuta H_0
H_0 è vera	Decisione corretta	Errore I tipo
H_0 è falsa	Errore II tipo	Decisione corretta

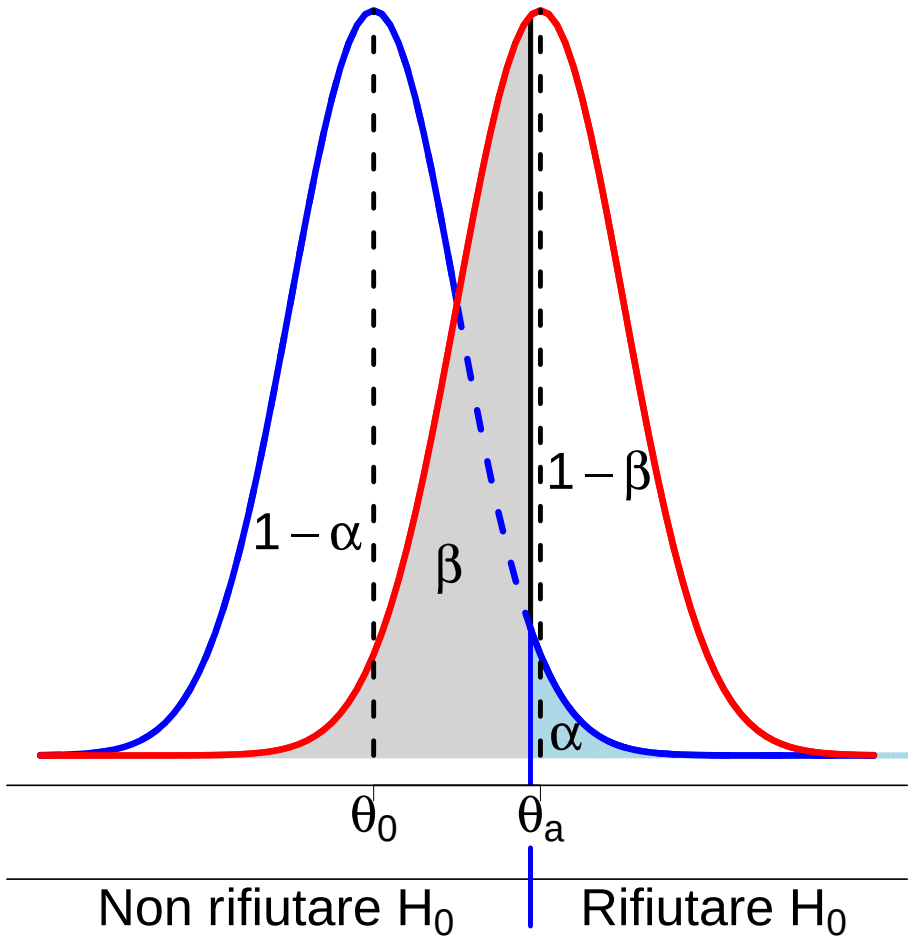
Probabilità di commettere l'errore di primo e secondo tipo

- Quando si prende una decisione, non sappiamo se tale decisione è corretta oppure se abbiamo commesso un errore di I o II tipo
- α = Probabilità di commettere l'errore del I tipo
- β = Probabilità di commettere l'errore del II tipo
- $\gamma = 1 - \beta$ = Potenza del test: Probabilità di rifiutare l'ipotesi nulla quando questa è falsa
- La probabilità di commettere l'errore del I tipo è anche chiamata livello di significatività del test
- Situazione ideale: Prendere *certamente* la decisione corretta, ossia prendere decisioni con $\alpha = 0$ e $\beta = 0$
- Nella realtà la situazione ideale è utopistica
- Tra α e β esiste una relazione inversa: Tanto più piccola è la probabilità di commettere l'errore del I tipo, tanto più grande è probabilità di commettere l'errore del II tipo

Probabilità di commettere l'errore di primo e secondo tipo

$H_0 : \theta = \theta_0$ *versus* $H_a : \theta = \theta_a$





Probabilità di commettere l'errore di primo e secondo tipo

- Tradizionalmente, nella teoria dei test di ipotesi, si ritiene più grave commettere un errore di primo tipo che di secondo tipo
- Esempio – Nuovo farmaco
  - ✓ Un'industria farmaceutica deve decidere se immettere sul mercato un nuovo farmaco
  - ✓ Si vuole immettere sul mercato il nuovo farmaco solo se è più efficace di quello già esistente in commercio
  - ✓ Quando si verifica l'efficacia del farmaco si possono commette due errori:
    - Errore di tipo I: Si mette in commercio il nuovo farmaco pur essendo equivalente al farmaco precedente
    - Errore di tipo II: Si mantiene in commercio il vecchio farmaco che però non è superiore al nuovo farmaco
  - ✓ Un approccio prudenziale porta a mantenere sul mercato il vecchio farmaco

Probabilità di commettere l'errore di primo e secondo tipo

- La probabilità di commettere l'errore del II tipo diminuisce se
  - ✓ aumenta la probabilità di commettere l'errore del I tipo (tra α e β esiste una relazione inversa)
  - ✓ aumenta la distanza tra il valore del parametro e il valore ipotizzato sotto l'ipotesi nulla, H_0
  - ✓ aumenta la dimensione campionaria
- Se diminuisce β aumenta $\gamma = 1 - \beta$, la potenza del test
- Procedura pratica: Fissare il livello di significatività del test, α, e cercare una regola che permetta di rendere più piccola possibile (minima, se possibile) la probabilità di commettere l'errore del II tipo
- La scelta di α riflette quanto l'analista vuole essere prudente nel prendere decisioni
- Tanto più è piccolo il livello di significatività α, tanto più forte deve essere l'evidenza empirica per rifiutare l'ipotesi nulla
- Valori tipici di α: $0.10, 0.05, 0.01$

Regione di accettazione e regione di rifiuto (critica)

- In genere le ipotesi riguardano parametri della popolazione che possono essere stimati con opportuni stimatori
- Il test statistico (test delle ipotesi) è una regola che permette di discriminare i campioni che portano a non rifiutare l'ipotesi nulla da quelli che portano a rifiutare l'ipotesi nulla
- Il test si basa sul valore assunto da una statistica, detta **statistica test**
  - ✓ La statistica test permette di valutare quanto il valore stimato del parametro sulla base del campione osservato si discosta dal valore ipotizzato con H_0
  - ✓ In genere la distanza tra stima e valore del parametro sotto H_0 è espressa in termini di errori standard
- L'insieme dei valori della statistica test che portano a non rifiutare l'ipotesi nulla è chiamata **regione di accettazione**
- L'insieme dei valori della statistica test che portano a rifiutare l'ipotesi nulla è chiamata **regione di rifiuto**
- La statistica test è una statistica la cui distribuzione campionaria deve essere completamente nota sotto l'ipotesi nulla
- Il livello di significatività del test determina i valori della statistica test che appartengono alla regione di accettazione e alla regione di rifiuto

Regione di accettazione e regione di rifiuto – Esempio: Statura e alimentazione

- È noto che la statura media degli abitanti di un certo paese è di 175 cm
- Un ricercatore, sulla base dei cambiamenti nell'alimentazione avvenuti negli ultimi anni, ipotizza che la statura media dei giovani sia cambiata rispetto a quella dei loro genitori
- Ipotesi nulla: $H_0 : \mu = 175$
- Ipotesi alternative: $H_a : \mu \neq 175$ oppure $H_a : \mu > 175$ oppure $H_a : \mu < 175$
- Una possibile statistica test è la media campionaria
- Si ipotizzi che nella popolazione la statura abbia una distribuzione Normale: $Y \sim N(\mu, \sigma^2 = 324)$
- Distribuzione campionaria della statistica test: $\overline{Y} \sim N(\mu, \sigma^2_{\overline{Y}} = 324/n)$
- Si consideri un campione di $n = 81$ giovani e si ipotizzi di osservare su tale campione $\overline{y} = 180$
- Se l'ipotesi nulla fosse vera, ossia se $\mu = 175$, allora

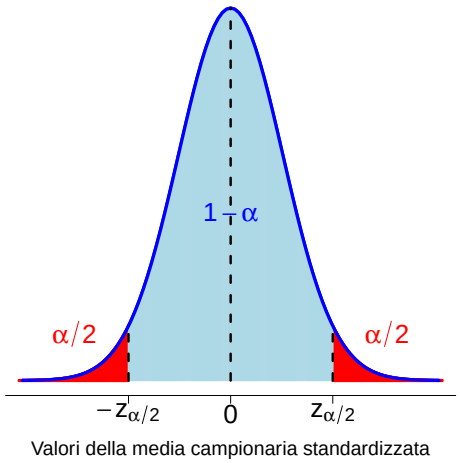
$$\overline{Y} \sim N(175, \sigma^2_{\overline{Y}} = 324/81) \qquad e \qquad \frac{\overline{Y} - 175}{\sqrt{324/81}} \sim N(0, 1)$$

- Obiettivo: Valutare se il valore stimato $\overline{y} = 180$ è abbastanza distante dal valore ipotizzato sotto $H_0 : \mu = 175$ (ossia se il valore stimato standardizzato $(180 - 175)/\sqrt{4} = 2.5$ è abbastanza distante da $0 = (175 - 175)/2$) per rifiutare H_0

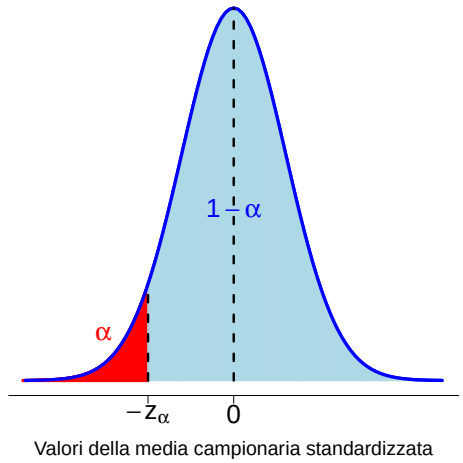
Regione di accettazione e regione di rifiuto – Esempio: Statura e alimentazione

- La distribuzione campionaria della statistica test (media campionaria) standardizzata è Normale standard
- Distribuzione campionaria della statistica test e regione di rifiuto

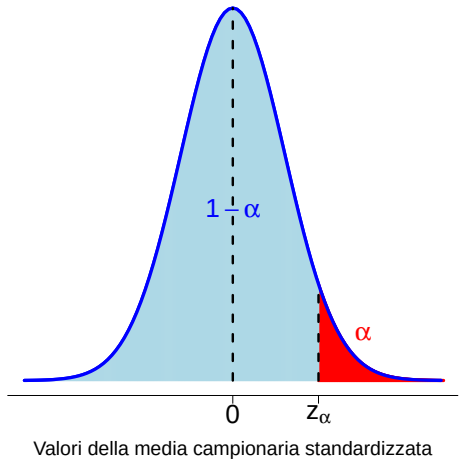
$H_a : \mu \neq \mu_0$



$H_a : \mu < \mu_0$



$H_a : \mu > \mu_0$



- Regione di rifiuto:
  - ✓ $H_a : \mu \neq \mu_0$: Regione di rifiuto = Insieme dei valori della statistica test minori di $z_{\alpha/2}$ o maggiori di $z_{\alpha/2}$ ($z_{0.025} = 1.96$ per $\alpha = 0.05$)
  - ✓ $H_a : \mu > \mu_0$: Regione di rifiuto = Insieme dei valori della statistica test maggiori di z_α ($z_{0.05} = 1.645$ per $\alpha = 0.05$)
  - ✓ $H_a : \mu < \mu_0$: Regione di rifiuto = Insieme dei valori della statistica test minori di $-z_\alpha$ ($-z_{0.05} = -1.645$ per $\alpha = 0.05$)

Passi da seguire nella verifica di ipotesi

- Definizione delle ipotesi
- Scelta della statistica test
- Scelta del livello di significatività, α
- Definizione della regione di rifiuto (che dipende da α)
- Estrazione del campione
- Calcolo del valore osservato della statistica test
- Decisione: Se il valore osservato della statistica test cade nella regione di rifiuto, allora si rifiuta H_0, altrimenti non si rifiuta H_0 al livello di significatività fissato

Tale approccio è stato sviluppato nel periodo tra il 1928 e il 1938 da Jerzy Neyman e Egon Pearson

p–valore (p–value)

- Un approccio alternativo per prendere decisioni si basa sul considerare la probabilità che la statistica test produca un valore inverosimile se H_0 è vera
- La decisione si basa sul determinare (in termini probabilistici) quanto insolito è il valore osservato della statistica test rispetto al valore ipotizzato sotto H_0
- Tale approccio utilizza il concetto di p–valore
- Il p–valore da informazioni su quanto probabili sono i possibili valori della statistica test che fornirebbero almeno la stessa evidenza del valore effettivamente osservato contro l'ipotesi nulla H_0 (se H_0 è vera)

Il p–valore (p-value o livello di significatività osservato), che indicheremo con P, è la probabilità che la statistica test assuma un valore uguale o più estremo (nella direzione prevista dall'ipotesi alternativa) di quello osservato quando è vera l'ipotesi nulla H_0

- L'approccio del p–valore è stato proposto da R. A. Fisher che suggerì di riportare solo il p–valore piuttosto che prendere una decisione formale
- Tale approccio si è diffuso ampiamente anche grazie al fatto che oggi i software permettono di calcolare facilmente il p–valore per molti test di ipotesi
- In generale può essere sempre utile calcolare e riportare il p–valore in quanto permette di valutare la forza dell'evidenza empirica contro l'ipotesi nulla

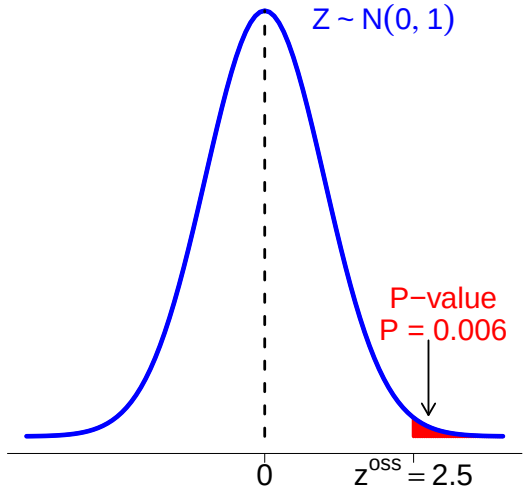
Interpretare il p–valore

- Il p–valore misura l'evidenza fornita dai dati contro l'ipotesi nulla: Un valore del p–valore piccolo suggerisce che il risultato della statistica test è insolito se H_0 è vera
- Quanto più è piccolo il p–valore, tanto più forte è l'evidenza statistica contro l'ipotesi nulla H_0: Quando il p–valore è piccolo il campione osservato dovrebbe essere insolito se H_0 fosse vera
- Quando l'ipotesi nulla è vera, il p–valore può assumere un qualunque valore tra 0 e 1 con la stessa probabilità
- Quando l'ipotesi nulla è falsa, è più probabile che il p–valore assuma un valore vicino a 0 piuttosto che un valore vicino a 1
- Generalmente l'evidenza contro l'ipotesi nulla si considera forte se il p–valore è molto piccolo, inferiore, ad esempio, a 0.05 oppure a 0.01
- Nella pratica l'interpretazione del p–valore, ossia la decisione se l'evidenza contro l'ipotesi nulla sia sufficiente forte da poter rifiutare tale ipotesi, si basa sul confronto tra il p–valore, P, e il valore fissato α del livello di significatività:
  - ✓ $P < \alpha$: I dati mostrano evidenza contraria ad H_0 al livello α (Rifiutare H_0)
  - ✓ $P > \alpha$: I dati *non* mostrano evidenza contraria ad H_0 al livello α (Non rifiutare H_0)
- Il p–valore è minore di α se il valore osservato della statistica test cade nella regione critica

Interpretare il p–valore – Esempio: Statura e alimentazione

- Ipotesi: $H_0 : \mu = 175$ versus $H_a : \mu > 175$
- Popolazione: $Y \sim N(\mu, \sigma^2 = 324)$
- Statistica test = Media campionaria: $\overline{Y} \sim N(\mu, \sigma^2_{\overline{Y}} = 324/n)$
- Osservazioni campionarie: $n = 81$ e $\overline{y} = 180$
- Il p–valore è la probabilità che la media campionaria assuma valori maggiori di $\overline{y} = 180$ sotto H_0
- Se H_0 fosse vera, ossia se $\mu = 175$, allora $\overline{Y} \sim N(175, \sigma^2_{\overline{Y}} = 324/81)$, quindi

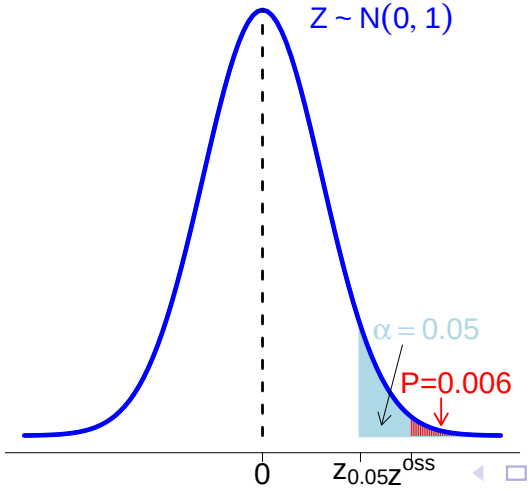
$$
\begin{aligned}
P &= P\left(\overline{Y} > 180\right) = P\left(\frac{\overline{Y} - 175}{2} > \frac{180 - 175}{2}\right) \\
&= P\left(Z > 2.5\right) = 1 - P\left(Z \leq 2.5\right) = 0.00621
\end{aligned}
$$



- Il p–valore è piccolo: se la statura media dei giovani fosse di 175 cm la probabilità di ottenere un valore campionario così estremo, ossia una statura media campionaria di 180 cm, sarebbe inferiore a 0.01

Interpretare il p–valore – Esempio: Statura e alimentazione

- Un valore del p–valore così piccolo fornisce una forte evidenza contro l'ipotesi nulla a favore dell'alternativa: l'ipotesi del ricercatore che la statura media dei giovani sia aumentata rispetto a quella dei loro genitori è verosimile
- Fissato il livello di significatività $\alpha = 0.05$, Il p–valore è minore di α, quindi si rifiuta l'ipotesi nulla $H_0 : \mu = 175$ al livello di significatività α
- Il p–valore è minore di α se il valore osservato della statistica test cade nella regione critica
- Regione critica = Insieme dei valori della statistica test standardizzata (sotto H_0), ossia di $Z = (\overline{Y} - 175)/2$, maggiori di $z_{0.05} = 1.645$
- Valore osservato della statistica test Z (sotto H_0): $z^{oss} = (180 - 175)/2 = 2.5$
- Il valore $z^{oss} = 2.5$ è maggiore di $z_{0.05} = 1.645$, quindi appartiene alla regione critica



Test per la media di una popolazione Normale con varianza nota

- Si supponga che la variabile di interesse Y sia una variabile continua con distribuzione Normale nella popolazione: $Y \sim N(\mu, \sigma^2)$ con σ^2 nota
- Ipotesi

$$\begin{array}{lll} H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu \neq \mu_0 \\ H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu > \mu_0 \\ H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu < \mu_0 \end{array}$$

- Campione casuale: $Y_1, \ldots, Y_n$
- Scelta della statistica test = Media campionaria: $\overline{Y} = \dfrac{Y_1 + \cdots + Y_n}{n}$
- Poiché Y ha distribuzione Normale nella popolazione, la distribuzione campionaria della statistica test è Normale

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Sotto l'ipotesi nulla, ossia sotto l'ipotesi che $H_0 : \mu = \mu_0$ sia vera,

$$\overline{Y} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right) \qquad \text{e quindi} \qquad Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

- Scelta del livello di significatività: Tipicamente $\alpha = 0.1, 0.05, 0.01$

Test per la media di una popolazione Normale con varianza nota

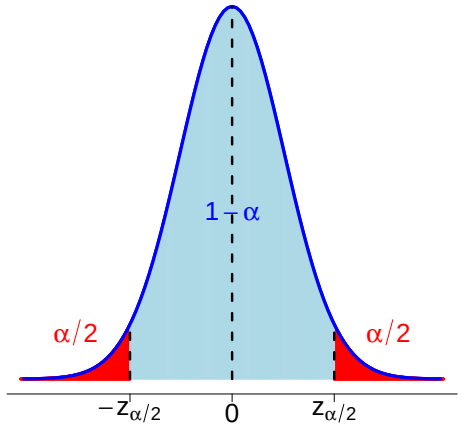
- Definizione della regione di rifiuto (al livello di significatività α): Si rifiuta l'ipotesi nulla per valori della statistica test *lontani* (nella direzione determinata dall'ipotesi alternativa) dal valore ipotizzato sotto H_0
- La regione di rifiuto dipende dal livello di significatività α e dall'ipotesi alternativa

Ipotesi Alternativa	Regione di Rifiuto
$H_a : \mu \neq \mu_0$	$Z \leq -z_{\alpha/2}$ o $Z \geq z_{\alpha/2}$
$H_a : \mu > \mu_0$	$Z \geq z_{\alpha}$
$H_a : \mu < \mu_0$	$Z \leq -z_{\alpha}$

  dove z_α è il valore che nella distribuzione Normale standard lascia alla sua destra un'area pari a α e $z_{\alpha/2}$ è il valore che nella distribuzione Normale standard lascia alla sua destra un'area pari a $\alpha/2$
- I valori $-z_\alpha$ e z_α nel caso di ipotesi alternative unilaterali e i valori $-z_{\alpha/2}$ e $z_{\alpha/2}$ nel caso di ipotesi alternative bilaterali sono chiamati usualmente valori critici

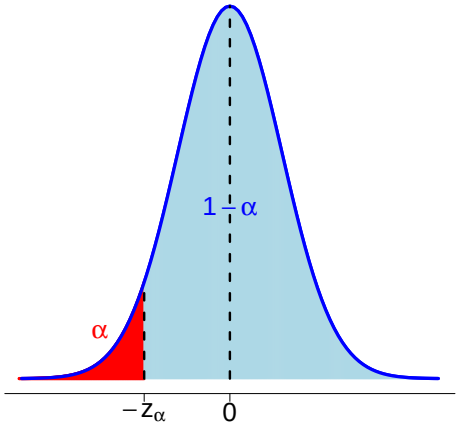
Test per la media di una popolazione Normale con varianza nota
Regione di accettazione e regione di rifiuto

$H_0 : \mu = \mu_0 \quad vs \quad H_a : \mu \neq \mu_0$



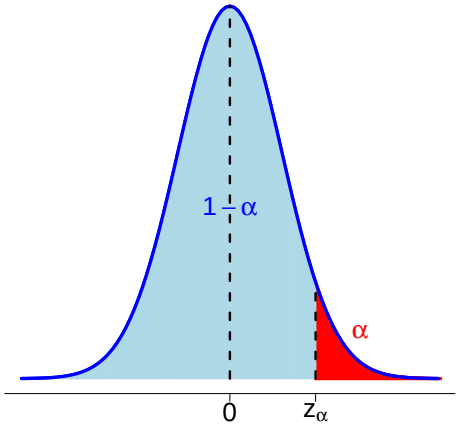
Valori della media campionaria standardizzata

$H_0 : \mu = \mu_0 \quad vs \quad H_a : \mu < \mu_0$



Valori della media campionaria standardizzata

$H_0 : \mu = \mu_0 \quad vs \quad H_a : \mu > \mu_0$



Valori della media campionaria standardizzata

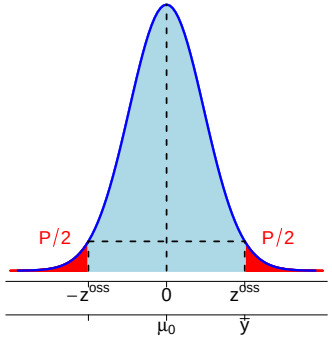
Test per la media di una popolazione Normale con varianza nota

- Estrazione del campione: Si estrae un campione casuale di n unità
- Calcolo della statistica test: Si calcola la media campionaria $\overline{y}$ e il valore della media campionaria standardizzata z^{oss}
  - ✓ Tanto maggiore è la differenza tra $\overline{y}$ e μ_0, tanto maggiore sarà z^{oss} in valore assoluto
  - ✓ Maggiore è il valore di assoluto di z^{oss}, $|z^{oss}|$, maggiore è l'evidenza contro H_0
- Conclusioni formali (approccio di Neyman-Pearson):
  - ✓ Se il valore osservato della media campionaria appartiene alla regione di rifiuto si conclude che i dati mostrano evidenza contraria al livello di significatività fissato: Si rifiuta H_0
  - ✓ Se il valore osservato della media campionaria non appartiene alla regione di rifiuto si conclude che i dati non mostrano evidenza contraria al livello di significatività fissato: Non si rifiuta H_0
- In alternativa il test può essere svolto usando l'approccio del $p-$valore il cui calcolo è comunque utile anche quando è usato l'approccio di Neyman-Pearson
- Ricordando che $Z = \dfrac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$, si ha ...

Test per la media di una popolazione Normale con varianza nota

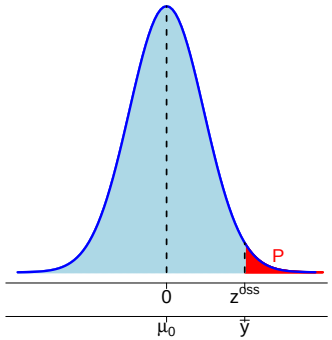
- Ipotesi alternativa bilaterale: $H_a : \mu \neq \mu_0$

$$\begin{aligned} P &= P(Z \leq -|z^{oss}| \text{ o } Z \geq |z^{oss}|) \\ &= 2 \cdot (1 - P(Z \leq |z^{oss}|)) \\ &= 2 \cdot (1 - \Phi(|z^{oss}|)) \end{aligned}$$



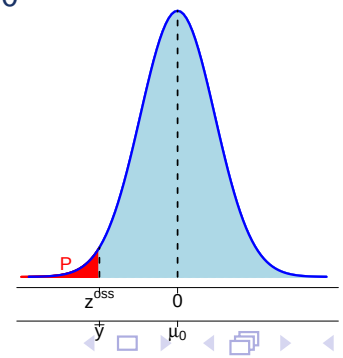
- Ipotesi alternativa unilaterale destra: $H_a : \mu > \mu_0$

$$\begin{aligned} P &= P(\overline{Y} \geq \overline{y}) = P(Z \geq z^{oss}) \\ &= 1 - Pr(Z \leq z^{oss}) = 1 - \Phi(z^{oss}) \end{aligned}$$



- Ipotesi alternativa unilaterale sinistra: $H_a : \mu < \mu_0$

$$P = P(\overline{Y} \leq \overline{y}) = P(Z \leq z^{oss}) = \Phi(z^{oss})$$



Test per la media di una popolazione Normale con varianza nota: Esempio

- Anemia. L'emoglobina è una proteina che trasporta l'ossigeno dai polmoni a tutti i tessuti corporei
- Gli individui con meno di 12 grammi di emoglobina per decilitro di sangue (g/dl) sono considerati anemici
- Si supponga che il livello di emoglobina, Y, segua una distribuzione Normale con deviazione standard $\sigma = 1.6$ g/dl: $Y \sim N(\mu, 1.6^2)$
- Un funzionario della sanità pubblica ipotizza che la media μ per i bambini della sua città sia inferiore a 12:

$$H_0 : \mu = \mu_0 = 12 \quad \textit{versus} \quad H_a : \mu < \mu_0 = 12$$

- Per verificare questa ipotesi il funzionario decide di considerare un campione casuale di $n = 16$ bambini e un livello di significatività $\alpha = 0.01$
- Statistica test = Media campionaria
- Regione di rifiuto: $\alpha = 0.01$ e $z_{0.01} = 2.33$, quindi

$$RC_{0.01} = Z \leq -2.33$$

Test per la media di una popolazione Normale con varianza nota: Esempio

- Campione osservato

Bambino	Livello di emoglobina	Bambino	Livello di emoglobina
1	8.5	9	10.0
2	10.2	10	12.4
3	9.3	11	11.8
4	11.0	12	9.7
5	11.6	13	12.0
6	11.9	14	12.7
7	8.1	15	11.1
8	12.0	16	12.1

- Calcolo della statistica test:

$$\overline{y} = \frac{174.4}{16} = 10.9 \quad \Longrightarrow \quad z^{oss} = \frac{10.9 - 12}{1.6/\sqrt{16}} = -2.75$$

- Decisione (approccio di Neyman-Pearson): il valore osservato della statistica test cade nella regione di rifiuto quindi si conclude che i dati mostrano evidenza contro l'ipotesi nulla al livello di significatività del 1%: L'ipotesi funzionario della sanità pubblica sembra essere verosimile
- P–valore: $P = P(Z \leq -2.75) = 1 - \Phi(2.75) = 0.003$ (Il p– valore è piccolo!)

Test per la media di una popolazione Normale con varianza non nota

- Si supponga che la variabile di interesse Y sia una variabile continua con distribuzione Normale nella popolazione: $Y \sim N(\mu, \sigma^2)$ con σ^2 non nota
- Ipotesi

$$\begin{array}{lll} H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu \neq \mu_0 \\ H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu > \mu_0 \\ H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu < \mu_0 \end{array}$$

- Campione casuale: $Y_1, \ldots, Y_n$
- Scelta della statistica test = Media campionaria: $\overline{Y} = \dfrac{Y_1 + \cdots + Y_n}{n}$
- Poiché Y ha distribuzione Normale nella popolazione, la distribuzione campionaria della statistica test è Normale

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Sotto l'ipotesi nulla, ossia sotto l'ipotesi che $H_0 : \mu = \mu_0$ sia vera,

$$\overline{Y} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right) \qquad \text{e quindi} \qquad Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

- Problema: La varianza di Y nella popolazione, σ^2 non è nota

Test per la media di una popolazione Normale con varianza non nota

- Soluzione: Stimare la varianza usando la varianza campionaria corretta

$$S^2 = \frac{(Y_1 - \overline{Y})^2 + \cdots + (Y_n - \overline{Y})^2}{n-1}$$

e ottenere una stima dell'errore standard della media campionaria

$$se_{\overline{Y}} = \frac{S}{\sqrt{n}}$$

- La media campionaria standardizzata usando una stima dell'errore standard ha distribuzione t di Student con $n-1$ gradi di libertà

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

Quindi nell'ipotesi che l'ipotesi nulla sia vera, ossia che $\mu = \mu_0$

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

- Scelta del livello di significatività: Tipicamente $\alpha = 0.1, 0.05, 0.01$

Test per la media di una popolazione Normale con varianza non nota

- Definizione della regione di rifiuto (al livello di significatività α): Si rifiuta l'ipotesi nulla per valori della statistica test *lontani* (nella direzione determinata dall'ipotesi alternativa) dal valore ipotizzato sotto H_0
- La regione di rifiuto dipende dal livello di significatività α e dall'ipotesi alternativa

Ipotesi Alternativa	Regione di Rifiuto
$H_a : \mu \neq \mu_0$	$T \leq -t_{n-1,\alpha/2}$ o $T \geq t_{n-1,\alpha/2}$
$H_a : \mu > \mu_0$	$T \geq t_{n-1,\alpha}$
$H_a : \mu < \mu_0$	$T \leq -t_{n-1,\alpha}$

dove $t_{n-1,\alpha}$ è il valore che nella distribuzione t di Student con $n-1$ gdl lascia alla sua destra un'area pari a α e $t_{n-1,\alpha/2}$ è il valore che nella distribuzione t di Student con $n-1$ gdl lascia alla sua destra un'area pari a $\alpha/2$

- I valori $-t_{n-1,\alpha}$ e $t_{n-1,\alpha}$ nel caso di ipotesi alternative unilaterali e i valori $-t_{n-1,\alpha}$ e $t_{n-1,\alpha}$ nel caso di ipotesi alternative bilaterali sono chiamati usualmente valori critici

Test per la media di una popolazione Normale con varianza non nota

- Estrazione del campione: Si estrae un campione casuale di n unità
- Calcolo della statistica test: Si calcola la media campionaria $\overline{y}$ e il valore della media campionaria standardizzata t^{oss}
  - ✓ Tanto maggiore è la differenza tra $\overline{y}$ e μ_0, tanto maggiore sarà t^{oss} in valore assoluto
  - ✓ Maggiore è il valore di assoluto di t^{oss}, $|t^{oss}|$, maggiore è l'evidenza contro H_0
- Conclusioni (approccio di Neyman-Pearson):
  - ✓ Se il valore osservato della media campionaria appartiene alla regione di rifiuto si conclude che i dati mostrano evidenza contraria al livello di significatività fissato: Si rifiuta H_0
  - ✓ Se il valore osservato della media campionaria non appartiene alla regione di rifiuto si conclude che i dati non mostrano evidenza contraria al livello di significatività fissato: Non si rifiuta H_0
- È utile calcolare anche il p–valore: Ricordando che $T = \dfrac{\overline{Y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$, il p–valore richiede di calcolare delle probabilità relative alla distribuzione t di Student
- La tavola della distribuzione t permette di calcolare solo un valore approssimato del p–valore
- Se la dimensione del campione è sufficientemente grande, il p–valore può essere approssimato utilizzando la distribuzione Normale

Test per la media di una popolazione Normale con varianza non nota: Esempio

- Un'azienda operante nel e-commerce vuole stabilire se il tempo medio di consegna di 40 ore indicato dal corriere è ragionevole
- Si assume che il tempo di consegna abbia una distribuzione Normale: $Y \sim N(\mu, \sigma^2)$
- Ipotesi

$$H_0 : \mu = \mu_0 = 40 \quad \textit{versus} \quad H_a : \mu \neq \mu_0 = 40$$

- Al fine di valutare l'ipotesi di interesse, l'azienda decide di considerare un livello di significatività $\alpha = 0.01$ e di effettuare un'indagine campionaria fra i suoi clienti circa i tempi di consegna, selezionando un campione casuale di $n = 6$ clienti
- Statistica test = Media campionaria
- Regione di rifiuto: Gdl $= n - 1 = 5$ e $\alpha = 0.01$. Quindi, $\alpha/2 = 0.005$ e $t_{5,0.005} = 4.032$. Quindi

$$RC_{0.01} = T \leq -4.032 \quad \text{o} \quad T \geq 4.032$$

Test per la media di una popolazione Normale con varianza non nota: Esempio

- Campione osservato

Cliente	y_i	y_i^2	$(y_i - \bar{y})^2$
1	37	1369	30.25
2	41	1681	2.25
3	45	2025	6.25
4	38	1444	20.25
5	46	2116	12.25
6	48	2304	30.25
Tot	255	10939	101.50

$$\bar{y} = \frac{255}{6} = 42.5$$

$$s = \sqrt{\frac{10939 - 6 \cdot 255^2}{6-1}} = \sqrt{\frac{101.50}{6-1}} = \sqrt{20.3} = 4.51$$

- Valore osservato della statistica test: $t^{oss} = \dfrac{42.5 - 40}{4.51/\sqrt{6}} = 1.36$
- Decisione (approccio di Neyman-Pearson): il valore osservato della statistica test NON cade nella regione di rifiuto ($-4.032 < T^{oss} = 1.36 < 4.032$) quindi si conclude che i dati non mostrano evidenza contro l'ipotesi nulla per $\alpha = 0.01$
- $P-$valore: $P = 0.2321$
  - ✓ Se il tempo medio di consegna fosse 40 minuti, la probabilità che la media campionaria cada lontano da 40 almeno quanto il valore osservato $\bar{y} = 42.5$ è pari a 0.2321
  - ✓ Il $p-$valore non è piccolo: i dati non mostrano evidenza contraria a H_0

Test per la media di una popolazione non Normale (campioni di grandi dimensioni)

- Si supponga che la variabile di interesse Y sia una variabile continua media μ e varianza σ^2 nella popolazione
- Ipotesi

$$\begin{array}{lll} H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu \neq \mu_0 \\ H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu > \mu_0 \\ H_0 : \mu = \mu_0 & \textit{versus} & H_a : \mu < \mu_0 \end{array}$$

- Campione casuale: $Y_1, \ldots, Y_n$
- Scelta della statistica test: Media campionaria

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- Se la dimensione del campione è sufficientemente grande per il teorema del limite centrale la media campionaria ha distribuzione approssimativamente Normale

$$\overline{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ per } n \text{ sufficientemente grande}$$

Test per la media di una popolazione non Normale (campioni di grandi dimensioni)

- Sotto l'ipotesi nulla, ossia sotto l'ipotesi che $H_0 : \mu = \mu_0$ sia vera, per n sufficientemente grande,

$$\overline{Y} \approx N\left(\mu_0, \frac{\sigma^2}{n}\right) \qquad \text{e quindi} \qquad Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} \approx N(0,1)$$

- Se la varianza σ^2 di Y nella popolazione non è nota si sostituisce con una sua stima, s^2, e sotto l'ipotesi che $H_0 : \mu = \mu_0$ sia vera, per n sufficientemente grande,

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{s^2}/\sqrt{n}} \approx N(0,1)$$

- Regione di rifiuto (al livello di significatività α) approssimata

Ipotesi Alternativa	Regione di Rifiuto
$H_a : \mu \neq \mu_0$	$Z \leq -z_{\alpha/2}$ o $Z \geq z_{\alpha/2}$
$H_a : \mu > \mu_0$	$Z \geq z_\alpha$
$H_a : \mu < \mu_0$	$Z \leq -z_\alpha$

dove $z_\alpha : P(Z \geq z_\alpha) = \alpha$ e $z_{\alpha/2} : P(Z \geq z_{\alpha/2}) = \alpha/2$

Test per la media di una popolazione non Normale (campioni di grandi dimensioni)

- Estrazione del campione: Si estrae un campione casuale di n unità
- Calcolo della statistica test: Si calcola la media campionaria $\overline{y}$ e/o il valore della media campionaria standardizzata z^{oss}
- Conclusioni formali (approccio di Neyman-Pearson):
  - ✓ Se il valore osservato della media campionaria appartiene alla regione di rifiuto si conclude che i dati mostrano evidenza contraria al livello di significatività fissato: Si rifiuta H_0
  - ✓ Se il valore osservato della media campionaria non appartiene alla regione di rifiuto si conclude che i dati non mostrano evidenza contraria al livello di significatività fissato: Non si rifiuta H_0
- P–valore
  - ✓ Ipotesi alternativa bilaterale, $H_a : \mu \neq \mu_0$: $P \approx 2 \cdot (1 - \Phi(|z^{oss}|))$
  - ✓ Ipotesi alternativa unilaterale destra, $H_a : \mu > \mu_0$: $P \approx P(Z \geq z^{oss})$
  - ✓ Ipotesi alternativa unilaterale sinistra, $H_a : \mu < \mu_0$: $P \approx P(Z \leq z^{oss})$

Robustezza rispetto a violazioni dell'assunzione di normalità

- Se il carattere di interesse Y ha distribuzione Normale nella popolazione come media μ e varianza σ^2 entrambe non note, il test sulla media μ è costruito utilizzando la distribuzione t di Student con $n-1$ gradi di libertà
- Al crescere della dimensione del campione l'ipotesi di Normalità diventa sempre meno importante: Per il teorema del limite centrale la distribuzione campionaria della media campionaria è approssimativamente Normale per n sufficientemente grande qualunque sia la distribuzione di Y nella popolazione
  - ✓ Per n sufficientemente grande, la distribuzione t di Student con $n-1$ gdl è ben approssimata dalla distribuzione Normale standard
- Se Y ha una distribuzione non Normale nella popolazione e la dimensione del campione è piccola (inferiore a 30 unità) allora
  - ✓ Il test t sulla media μ con ipotesi alternativa bilaterale funziona bene: Il test t bilaterale è robusto rispetto a violazioni dell'assunzione di Normalità
  - ✓ Il test t sulla media μ con ipotesi alternativa unilaterale in genere non funziona bene, soprattutto se la distribuzione di Y nella popolazione è fortemente asimmetrica

Corrispondenza tra intervalli di confidenza e test bilaterali

- Le conclusioni a cui si perviene utilizzando un test di ipotesi sulla media di un carattere continuo con ipotesi alternativa bilaterale (test a due code) sono coerenti con quelle a cui si giunge utilizzando un intervallo di confidenza
- Se il test (bilaterale) suggerisce che un particolare valore è plausibile per il la media della variabile di interesse, allora lo stesso accade per l'intervallo di confidenza

Formalmente

- Sia RC_α la regione di rifiuto al livello di significatività α per il test bilaterale

$$H_0 : \mu = \mu_0 \qquad \textit{versus} \qquad H_a : \mu \neq \mu_0$$

- Sia $IC_{1-\alpha}(\mu)$ l'intervallo di confidenza per la media μ al livello di confidenza $1 - \alpha$
- I dati non mostrano evidenza contro H_0 al livello di significatività α (p–valore maggiore di α) $\Longleftrightarrow$ L'intervallo di confidenza al livello di confidenza $1 - \alpha$ contiene il valore μ_0
- I dati mostrano evidenza contro H_0 al livello di significatività α (p–valore minore di α) $\Longleftrightarrow$ L'intervallo di confidenza al livello di confidenza $1 - \alpha$ non contiene il valore μ_0

Corrispondenza tra intervalli di confidenza e test bilaterali: Esempio

- Inferenza per la media di una popolazione Normale con varianza non nota: Tempo di consegna
  - ✓ Distribuzione del tempo di consegna nella popolazione: $Y \sim N(\mu, \sigma^2)$
  - ✓ Campione di $n = 6$ clienti per cui $\overline{y} = 42.5$ e $s^2 = 20.3$ quindi $t^{oss} = 1.36$
  - ✓ La regione di rifiuto al livello di significatività $\alpha = 0.01$ per il test bilaterale

$$H_0 : \mu = \mu_0 = 40 \quad \textit{versus} \quad H_a : \mu \neq \mu_0 = 40$$

è

$$RC_{0.01} = T \leq -4.032 \quad \text{o} \quad T \geq 4.032$$

  - ✓ Intervallo di confidenza per il tempo medio di consegna μ al livello di confidenza $1 - \alpha = 0.99$

$$IC_{0.99}(\mu) = 42.5 \pm 4.032\sqrt{\frac{20.3}{6}} = [35.08; 68.84]$$

  - ✓ I dati non mostrano evidenza contro H_0 al livello di significatività $\alpha = 0.01$: $t^{oss} = 1.36$ non cade nella regione di rifiuto ($P = 0.2321 > 0.01 = \alpha$)
  - ✓ L'intervallo di confidenza al livello di confidenza $1 - \alpha = 0.01$ contiene il valore $\mu_0 = 40$

Test per la proporzione (campioni di grandi dimensioni)

- Si supponga che la variabile di interesse Y sia una variabile binaria con distribuzione di Bernoulli nella popolazione con probabilità di successo π: $Y \sim Ber(\pi)$
- Ipotesi

$$\begin{array}{lll} H_0 : \pi = \pi_0 & \textit{versus} & H_a : \pi \neq \pi_0 \\ H_0 : \pi = \pi_0 & \textit{versus} & H_a : \pi > \pi_0 \\ H_0 : \pi = \pi_0 & \textit{versus} & H_a : \pi < \pi_0 \end{array}$$

- Campione casuale: $Y_1, \ldots, Y_n$
- Scelta della statistica test: Proporzione campionaria

$$\hat{\pi} = \overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

- Se la dimensione del campione è sufficientemente grande tale per cui si hanno almeno 15 osservazioni per categoria (almeno 15 successi e almeno 15 insuccessi) per il teorema del limite centrale la proporzione campionaria ha distribuzione approssimativamente Normale

$$\hat{\pi} = \overline{Y} \approx N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \text{ per } n \text{ sufficientemente grande}$$

Test per la proporzione (campioni di grandi dimensioni)

- Sotto l'ipotesi nulla, ossia sotto l'ipotesi che $H_0 : \pi = \pi_0$ sia vera, per n sufficientemente grande,

$$\hat{\pi} \approx N\left(\pi_0, \frac{\pi_0(1-\pi_0)}{n}\right) \qquad \text{e quindi} \qquad Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \approx N(0,1)$$

- Regione di rifiuto (al livello di significatività α) approssimata

Ipotesi Alternativa	Regione di Rifiuto
$H_a : \pi \neq \pi_0$	$Z \leq -z_{\alpha/2}$ o $Z \geq z_{\alpha/2}$
$H_a : \pi > \pi_0$	$Z \geq z_{\alpha}$
$H_a : \pi < \pi_0$	$Z \leq -z_{\alpha}$

Test per la proporzione (campioni di grandi dimensioni)

- Estrazione del campione: Si estrae un campione casuale di n unità
- Calcolo della statistica test: Si calcola la proporzione campionaria $\hat{\pi}$ e/o il valore della proporzione campionaria standardizzata (sotto H_0) z^{oss}
- Conclusioni formali (approccio di Neyman-Pearson):
  - ✓ Se il valore osservato della media campionaria appartiene alla regione di rifiuto si conclude che i dati mostrano evidenza contraria al livello di significatività fissato: Si rifiuta H_0
  - ✓ Se il valore osservato della media campionaria non appartiene alla regione di rifiuto si conclude che i dati non mostrano evidenza contraria al livello di significatività fissato: Non si rifiuta H_0
- P–valore
  - ✓ Ipotesi alternativa bilaterale, $H_a : \pi \neq \pi_0$: $P \approx 2 \cdot (1 - \Phi(|z^{oss}|))$
  - ✓ Ipotesi alternativa unilaterale destra, $H_a : \pi > \pi_0$: $P \approx P(Z \geq z^{oss})$
  - ✓ Ipotesi alternativa unilaterale sinistra, $H_a : \pi < \pi_0$: $P \approx P(Z \leq z^{oss})$

Test per la proporzione (campioni di grandi dimensioni): Esempio

- Si supponga che il ministero delle politiche agricole, alimentari e forestali sia interessato a verificare se nel 2002 la percentuale degli occupati in Italia nel settore agricolo è la stessa del 1991 pari a 8.4% oppure è diminuita:

$$H_0 : \pi = \pi_0 = 0.084 \qquad vs \qquad H_1 : \pi < \pi_0 = 0.084$$

- La variabile di interesse Y (1=occupato agricolo, 0=altrimenti) ha distribuzione di Bernoulli di parametro π nella popolazione
- Si estrae un campione di $n = 1000$ occupati
- Statistica test = Proporzione campionaria
- Il campione è sufficientemente grande da poter considerare l'approssimazione Normale della distribuzione campionaria della proporzione campionaria adeguata
- Regione di rifiuto (approssimata): Per $\alpha = 0.01$, $z_{0.01} = 2.33$. Quindi la regione di rifiuto è $Z \leq -2.33$

Test per la proporzione (campioni di grandi dimensioni): Esempio

- Dei mille estratti, 53 sono occupati nel settore agricolo, pertanto:

$$\hat{\pi} = \overline{y} = \frac{53}{1000} = 0.053 \quad e \quad z = \frac{0.053 - 0.084}{\sqrt{0.084 \cdot (1 - 0.084)/1000}} = -3.534$$

- Decisione (approccio di Neyman-Pearson): il valore osservato della statistica test cade nella regione di rifiuto ($z^{oss} = -3.534 < -2.33$) quindi si conclude che i dati mostrano evidenza contro l'ipotesi nulla per $\alpha = 0.01$
- P–valore: $P \approx P(Z \leq -3.534) \approx 0.0002$ (Evidenza contro l'ipotesi nulla molto forte)