



Dipartimento di Economia Agroforestale e dell'Ambiente Rurale (DEAR)

# **LEZIONI DI STATISTICA**

**Appunti del corso di**  
**STATISTICA E INFORMATICA**

*a cura di*

*Silvio Franco e Barbara Pancino*

*Anno accademico 2004-05*

# INDICE

PRESENTAZIONE

**CONCETTI INTRODUTTIVI** *Pag. 3*

**I. STATISTICA DESCRITTIVA** *Pag. 5*

I.1 L'indagine statistica

I.2 Riepilogo dei dati

I.3 Distribuzioni di frequenze

**II. ELEMENTI DI PROBABILITA'** *Pag. 23*

II.1 Definizioni

II.2 Distribuzioni discrete

II.3 Distribuzione normale

**III. TEORIA DEI CAMPIONI** *Pag. 43*

III.1 Distribuzioni campionarie

III.2 Teoria della stima

**IV. DECISIONI STATISTICHE** *Pag. 63*

IV.1 Concetti generali

IV.2 Uguaglianza di medie

IV.3 Uguaglianza di frequenze

**V. RELAZIONI FRA VARIABILI** *Pag. 83*

V.1 Modello di regressione

V.2 Analisi di correlazione

## **PRESENTAZIONE**

La scelta dell'impostazione con cui affrontare l'insegnamento della Statistica nei corsi di I livello attivati all'interno di una Facoltà di Agraria non è un compito facile.

Le motivazioni di tale affermazione sono riconducibili a due questioni di ordine generale.

La prima è determinata dalla eterogeneità nella formazione di base che contraddistingue i ragazzi che si immatricolano in questa Facoltà. Le indagini condotte a questo proposito hanno evidenziato come la provenienza degli studenti che si iscrivono ai corsi offerti nelle Facoltà di Agraria sia distribuita in modo sostanzialmente uniforme fra licei, istituti tecnici ed istituti professionali. Come è noto, gli obiettivi formativi di tali scuole medie superiori sono molto diversi e, in particolare per le materie scientifiche, differenziano profondamente sia l'approccio concettuale che il metodo di studio. Si trovano così a frequentare il corso di Statistica studenti che hanno livelli di familiarità molto diversi rispetto alla comprensione di riferimenti teorici a carattere generale e di formalizzazioni simboliche degli strumenti di analisi.

La seconda motivazione è legata ai differenti campi di applicazione che la statistica trova all'interno delle discipline che vengono insegnate in una Facoltà di Agraria. Il loro numero e la loro diversificazione sono tali da rendere impossibile l'individuazione di un programma che possa contenere l'illustrazione di tutti gli strumenti statistici con cui potrebbero trovarsi ad operare nei corsi successivi e, soprattutto, nello svolgimento della loro tesi di laurea. Una implicita conferma in questo senso è rappresentata dalla presenza di ulteriori insegnamenti di statistica nella maggior parte dei corsi di laurea specialistica offerti dalle Facoltà di Agraria.

Entrambe le questioni cui si è fatto cenno assumono un rilievo ancora maggiore se si considera che la statistica viene ritenuta una materia di base e, pertanto, il suo insegnamento viene generalmente impartito nel I anno di corso.

Alla luce di questa situazione l'individuazione di un programma ottimale, e, di conseguenza, di un testo di riferimento che lo segua in modo soddisfacente, risulta assai difficoltosa. Si è deciso allora, pur muovendosi nell'ambito degli argomenti normalmente inseriti in un corso di statistica di base, di costruire un programma "su misura" e di procedere alla stesura di un materiale didattico aderente a tale programma, sia nei contenuti, sia nel livello di approfondimento.

Queste "lezioni" costituiscono il prodotto, certamente non definitivo, di tale operazione.

L'obiettivo che abbiamo avuto come punto di riferimento, e che speriamo di aver almeno in parte raggiunto, è stato quello di far comprendere allo studente alcune tipologie di problemi che costituiscono oggetto della statistica, il contesto teorico in cui vanno inquadrati e gli strumenti operativi con cui dovrebbero essere affrontati.

Questo obiettivo ha determinato sia la sequenza cui sono stati trattati i singoli argomenti (prima un cenno alle definizioni ed ai concetti teorici di base, poi

l'illustrazione dei relativi strumenti di analisi) sia la configurazione del testo. Per quanto riguarda quest'ultimo aspetto, si è scelto di dividere ciascuna pagina delle "lezioni" in due riquadri: nel riquadro superiore vengono riportati i concetti generali e/o la notazione formalizzata degli strumenti, in quello inferiore sono inseriti commenti e chiarimenti (nel caso nel riquadro superiore siano riportati dei concetti o delle definizioni) ed esemplificazioni quantitative (quando il riquadro superiore contiene formule o, più in generale, l'illustrazione di metodi statistici).

E' comunque necessario precisare che questo materiale didattico non può essere considerato sufficiente per la piena comprensione degli argomenti affrontati. Esso è nato per fornire un ausilio agli studenti che frequentano il corso e che seguono con regolarità le lezioni e le esercitazioni. Per coloro che non hanno questa possibilità, il testo delle "lezioni" può rappresentare un utile punto di riferimento che, però, deve essere integrato con letture di approfondimento e con lo svolgimento di ulteriori esercizi rispetto a quelli che vengono proposti.

E' d'obbligo, infine, un ringraziamento agli studenti che, avendo studiato su precedenti versioni degli appunti, ci hanno segnalato alcuni errori nei calcoli e, soprattutto, la insufficiente chiarezza di alcune parti; con questa edizione delle "lezioni" speriamo di aver rimediato, almeno in parte, ad entrambi gli inconvenienti.

## CONCETTI INTRODUTTIVI

**Scopo della statistica:** raccogliere, ordinare, riassumere, presentare ed analizzare insiemi di dati per migliorare la conoscenza dei fenomeni e per facilitare i processi decisionali.

I dati oggetto della statistica rappresentano risultati osservati di **variabili casuali**.

Le variabili casuali **qualitative** danno origine a dati categoriali che possono essere misurati su scale nominali o scale ordinali.

Le variabili casuali **quantitative** danno origine a dati numerici che possono essere discreti (frutto di un conteggio) o continui (frutto di una misurazione).

**Popolazione:** totalità degli elementi presi in esame.

- La popolazione è infinita se il numero dei suoi elementi è illimitato;
- La popolazione è finita se il numero dei suoi elementi è limitato.

**Campione:** insieme di elementi estratti dalla popolazione.

- Estrazione senza ripetizione: un elemento può essere selezionato una sola volta;
- Estrazione con ripetizione: un elemento può essere selezionato più di una volta.

Le categorie rappresentano i possibili valori assunti dalle singole osservazioni di una variabile casuale qualitativa.

Quando le categorie non possiedono alcun tipo di ordinamento le osservazioni vengono raccolte in scale nominali. Categorie nominali sono quelle identificate dalle possibili risposte a domande del tipo:

Indicare il colore preferito (Bianco, Blu, Verde, Rosso, ecc.)

Identificare la specie di un albero (Quercia, Olivo, Castagno, Noce, ecc.)

Quando le categorie possono essere ordinate in base ad un criterio le osservazioni vengono raccolte in scale ordinali. Categorie ordinali sono quelle identificate dalle possibili risposte a domande del tipo:

Il giudizio ottenuto in una prova (Insufficiente, Sufficiente, Buono, Distinto, Ottimo)

Il livello di interesse per una materia (Nessuno, Basso, Medio, Elevato)

-----

Una popolazione viene considerata infinita anche quando, pur essendo finita, è costituita da un gran numero di elementi. Ad esempio, gli elettori di una consultazione nazionale o gli alberi in 1.000 ha di bosco possono essere considerate popolazioni infinite.

-----

Un campionamento senza ripetizione viene eseguito escludendo da ogni selezione gli elementi che sono stati già estratti; un campionamento con ripetizione viene eseguito selezionando ogni elemento dall'intera popolazione. Ne consegue che nel campionamento senza ripetizione ogni elemento sarà presente nel campione una sola volta, mentre nel campionamento con ripetizione lo stesso elemento potrà essere presente nel campione più di una volta.

Si consideri l'estrazione di un campione di 5 carte da un mazzo; se ogni carta, dopo essere stata estratta, non viene reinserita si ha un campionamento senza ripetizione, se viene reinserita si ha un campionamento con ripetizione.

## CONCETTI INTRODUTTIVI

Le caratteristiche di un insieme di dati vengono descritte attraverso:

- **parametri**: misure riassuntive della popolazione;
- **statistiche**: misure riassuntive del campione.

### **STATISTICA DESCRITTIVA:**

parte della statistica che studia i metodi per raccogliere e presentare un insieme di dati (popolazione o campione) e per descriverne le caratteristiche attraverso delle misure riassuntive (parametri o statistiche).

### **STATISTICA INFERENZIALE:**

parte della statistica che studia i metodi che permettono di stabilire le caratteristiche o prendere decisioni su una popolazione partendo dalle osservazioni di un campione estratto da essa.

## I.1 STATISTICA DESCRITTIVA - Indagine Statistica

Un'indagine statistica si basa su una raccolta di dati eseguita partendo da:

- fonti di documentazione (fascicoli, banche dati, ecc.);
- risultati di prove sperimentali basate su un piano degli esperimenti (saggi in campo, test di laboratorio, ecc.);
- indagini dirette (interviste, questionari, ecc.).

Al termine della fase di raccolta, e dopo una necessaria verifica preliminare, i dati si presentano in forma grezza.

Per trarre le informazioni dai dati grezzi è necessario eseguirne l'ordinamento, l'esplorazione e la descrizione per giungere poi al riepilogo e alla presentazione.

Nel caso di dati categoriali, la fase di ordinamento consiste nel “contare” il numero di osservazioni che ricadono in ciascuna categoria.

La presentazione del risultato assoluto o relativo (percentuale) avviene in forma di tabelle o di grafici.

I principali grafici utilizzati per la rappresentazione dei dati qualitativi, sia assoluti che relativi, sono l'ortogramma e il diagramma a torta

**Esempio:** Ad un gruppo di 32 studenti universitari viene richiesto di indicare la scuola superiore presso cui hanno conseguito il diploma. L'elenco delle risposte fornite risulta il seguente:

B-C-B-A-C-C-C-B-A-A-D-B-A-D-C-C-B-B-A-A-B-B-B-C-C-A-D-D-B-C-B-B  
(A = Liceo; B = Istituto tecnico; C = Istituto professionale; D = Altro)

Presentare i risultati dell'indagine.

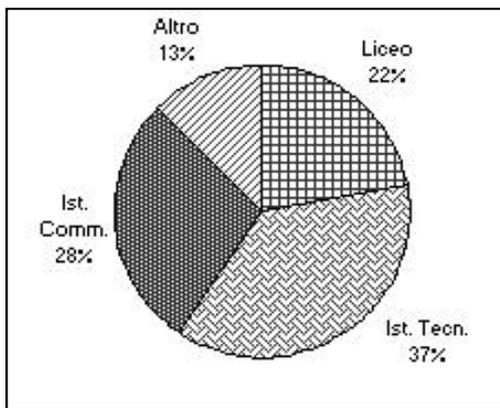
Partendo dai dati in forma grezza si esegue il conteggio dei dati ricadenti in ogni categoria.

Il risultato del conteggio viene presentato in forma di tabella (1) o in forma grafica (2-diagramma a torta; 3-ortogramma).

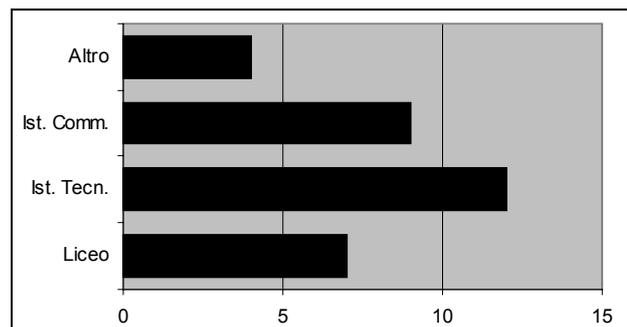
Risposta	Scuola	Numero	Percent.
A	Liceo	7	21,9%
B	Ist. Tecnico	12	37,5%
C	Ist. Professionale	9	28,1%
D	Altro	4	12,5%
Totale		32	100,0%

1

2



3



## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (1)

Il **riepilogo** delle caratteristiche di un insieme di dati quantitativi (popolazione o campione) avviene identificando le sue proprietà principali (parametri o statistiche):

- misure di TENDENZA CENTRALE
- misure di DISPERSIONE
- FORMA

**Misure di Tendenza Centrale o di Posizione:** individuano un valore "centrale" attorno al quale tendono a disporsi i dati di una popolazione o di un campione

Le principali misure di tendenza centrale sono:

- Media aritmetica (o **media**)
- Mediana
- Moda
- Intervallo medio

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (1)

### LA MEDIA

Indicati con:  $n$  = numero di osservazioni del campione;  
 $N$  = numero di osservazioni della popolazione;  
 $x_i$  = generica osservazione del campione o della popolazione.

**Media della popolazione** ( $\mu_x$ ): 
$$\mu_x = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

**Media di un campione** ( $\bar{x}$ ): 
$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

La media possiede due importanti proprietà (ad esempio, per un campione):

- i) La somma degli scostamenti dalla media è 0:  $\sum_{i=1}^n (x_i - \bar{x}) = 0$
- ii) Se  $y_i = x_i + k$ , con  $k$  costante, allora  $\bar{y} = \bar{x} + k$

### **Esempio:**

*Durante un'indagine condotta su un campione di sei aziende agricole sono state rilevate le seguenti produzioni unitarie (tonnellate per ettaro) di grano:*

Az. 1	Az. 2	Az. 3	Az. 4	Az. 5	Az. 6
3,2 t/ha	3,7 t/ha	3,5 t/ha	3,6 t/ha	3,1 t/ha	3,6 t/ha

*Determinare la produzione unitaria media delle aziende.*

Poiché le osservazioni si riferiscono ad un campione, dovrà essere calcolata la statistica  $\bar{x}$ . Considerando il numero di osservazioni ( $n=6$ ) si ha:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{3,2 + 3,7 + 3,5 + 3,6 + 3,1 + 3,6}{6} = \frac{20,7}{6} = 3,45$$

Applicando la prima proprietà della media si ottiene:

$$\sum_{i=1}^n (x_i - \bar{x}) = -0,25 + 0,25 + 0,05 + 0,15 - 0,35 + 0,15 = 0$$

Applicando la seconda proprietà della media, ad esempio ponendo  $k=0,5$ , si ottiene:

$$\bar{y} = \bar{x} + k = 3,45 + 0,5 = 3,95 \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{3,7 + 4,2 + 4,0 + 4,1 + 3,6 + 4,1}{6} = \frac{23,7}{6} = 3,95$$

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (1)

### MEDIANA

La **mediana** è il valore centrale di un insieme ordinato di dati (popolazione o campione).

Per un campione costituito da **n** osservazioni la mediana è data:

- se **n** è dispari, dalla osservazione centrale  $x_{(n+1)/2}$ ;
- se **n** è pari, dalla media delle 2 osservazioni centrali  $\frac{(x_{n/2} + x_{n/2+1})}{2}$

Importanti caratteristiche della mediana sono:

- che non è influenzata dai valori estremi dell'insieme dei dati;
- che ogni osservazione dell'insieme di dati ha il 50% di probabilità di cadere sopra o sotto la mediana.

### **Esempio:**

*Con riferimento all'esercizio precedente, determinare la mediana dell'insieme dei dati.*

*Al campione viene aggiunto il dato relativo ad un'altra azienda agricola la cui produzione unitaria di grano è pari a 4,5 t/ha. Calcolare la media e la mediana del nuovo campione.*

Per calcolare la mediana di un insieme di dati è necessario procedere al loro ordinamento.

Nel caso dell'esempio l'ordinamento produce il seguente risultato: 3,1 - 3,2 - 3,5 - 3,6 - 3,6 - 3,7.

Il campione considerato è costituito da **n=6** osservazioni; essendo **n** pari, la mediana è determinata dalla media delle due osservazioni centrali  $x_{n/2}=x_3=3,5$  e  $x_{n/2+1}=x_4=3,6$ .

$$\text{mediana} = \frac{3,5 + 3,6}{2} = 3,55$$

L'aggiunta del nuovo dato porta la dimensione del campione a **n=7**.

Per quanto riguarda la mediana, considerando il nuovo insieme dei dati ordinati e che **n** è dispari, si ha:

$$\text{mediana} = x_{(n+1)/2} = x_4 = 3,6$$

Per la media si avrà, invece:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{3,2 + 3,7 + 3,5 + 3,6 + 3,1 + 3,6 + 4,5}{7} = \frac{25,2}{7} = 3,6$$

L'osservazione aggiuntiva, che si presenta abbastanza estrema rispetto all'insieme di dati, non ha un grande effetto sulla mediana (se invece di 4,5 t/ha fosse stata 3,6 t/ha il risultato sarebbe stato lo stesso) mentre modifica il valore della media (anche a causa della ridotta dimensione del campione).

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (1)

### MODA

La **moda** è il valore che compare più spesso in un insieme di dati (popolazione o campione).

Se tutti i valori compaiono una sola volta l'insieme di dati non ha moda.

Se diversi valori compaiono lo stesso numero di volte l'insieme dei dati si dice multimodale (si parla, ad esempio, di "campione bimodale").

### INTERVALLO MEDIO

L'**intervallo medio** è definito come la media fra l'osservazione più piccola e la più grande:

$$\text{intervallo medio} = \frac{x_{\min} + x_{\max}}{2} = \frac{x_1 + x_n}{2} \quad (\text{per dati ordinati})$$

L'intervallo medio dipende solo dai valori estremi e quindi è una misura molto sensibile alla presenza di osservazioni estreme o errate.

#### **Esempio:**

*Con riferimento ai due esempi precedenti, calcolare la moda e l'intervallo medio dei due insiemi di dati ( $n=6$  e  $n=7$ ) relativi alla produzione unitaria di grano.*

Per calcolare la moda di un insieme di dati è necessario individuare il valore che compare più spesso. Per entrambi gli insiemi considerati l'unico valore che compare più di una volta è 3,6, ne consegue che

$$\text{moda} = 3,6$$

Per determinare l'intervallo medio è utile ordinare i dati in modo da determinarne i valori estremi.

Nel primo caso ( $n=6$ ) si ottiene  $x_1=3,1$  e  $x_n=3,7$  da cui

$$\text{intervallo medio} = (x_1+x_n)/2 = 3,4$$

Nel secondo caso ( $n=7$ ), invece,  $x_1=3,1$  e  $x_n=4,5$  da cui

$$\text{intervallo medio} = (x_1+x_n)/2 = 3,8$$

Come si osserva, il nuovo dato non ha alcun effetto sulla moda. Viceversa, andando a modificare in misura apprezzabile uno dei valori estremi dell'insieme, altera in modo consistente il valore dell'intervallo medio.

A questo proposito va notata la scarsa capacità descrittiva dell'insieme dei dati offerta dall'intervallo medio; infatti, nel secondo caso, soltanto una osservazione presenta un valore superiore a questa statistica mentre tutte le altre hanno un valore inferiore.

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (1)

### I QUANTILI

Rappresentano i valori che dividono un insieme di dati ordinato in un predefinito numero di parti.

La mediana è un particolare quantile poiché divide l'insieme dei dati in due parti in ognuna delle quali ricade esattamente la metà delle osservazioni

Altri valori comunemente utilizzati per dividere insiemi di dati sono i seguenti:

- QUARTILI ( $Q_1, Q_2, Q_3$ ): divisione in quattro parti
- DECILI ( $D_1, D_2, \dots, D_9$ ): divisione in dieci parti
- PERCENTILI ( $P_1, P_2, \dots, P_{99}$ ): divisione in cento parti

I percentili sono molto utilizzati nell'ambito delle statistiche mediche.

Se, ad esempio, l'altezza di un bambino di 10 anni è al 20° percentile, vuol dire che il 20% di tutti i bambini di 10 anni sono più bassi di lui e l'80% più alti.

Per la definizione stessa di quantili si ha: **mediana** =  $Q_2 = D_5 = P_{50}$ .

**Esempio:** Considerando i seguenti dati relativi alle temperature registrate nei giorni del mese di Giugno, determinare i valori dei quartili e dei decili dell'insieme dei dati.

Data	°C										
1/6	24,0	6/6	27,5	11/6	23,5	16/6	27,0	21/6	23,0	26/6	24,0
2/6	25,5	7/6	24,0	12/6	25,0	17/6	26,0	22/6	25,0	27/6	28,0
3/6	26,5	8/6	28,0	13/6	27,0	18/6	26,5	23/6	25,5	28/6	31,0
4/6	28,5	9/6	23,5	14/6	25,0	19/6	25,0	24/6	28,5	29/6	30,5
5/6	28,5	10/6	24,0	15/6	24,0	20/6	22,0	25/6	24,5	30/6	29,0

Per quanto riguarda i quartili si osserva che la mediana, corrispondente al II quartile, divide l'insieme dei dati ordinati in due gruppi da 15 osservazioni ciascuno.

Il I quartile ed il III quartile rappresentano proprio la mediana di questi due gruppi. Essendo tali gruppi costituiti da un numero dispari di osservazioni, la loro mediana sarà rappresentata dal valore centrale:

I Quartile (Oss. 8) = 24,0°C; II Quartile (Media oss. 15 e 16) = 25,5°C; III Quartile (Oss. 23) = 28,0°C

Per determinare i decili, invece, si osserva che  $30/10=3$  e, quindi, che i valori dei decili sono quelli dividono l'insieme dei dati ordinati a gruppi di tre.

Così il I decile è il valore che separa le prime tre dalle altre osservazioni e, perciò, la media fra la terza e la quarta osservazione; analogamente il V decile, che separa le osservazioni in due gruppi di quindici, sarà dato dalla media delle osservazioni n.15 e 16 e corrisponderà, come detto, alla mediana.

Procedendo in questo modo si ottengono i valori di tutti i decili:

I = 23,5; II = 24; III = 24,25; IV = 25; V = 25,5; VI = 26,5; VII = 27,25; VIII = 28,25; XI = 28,75

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (2)

**Misure di Dispersione**: esprimono il grado o l'intervallo di variabilità complessivo delle osservazioni di un insieme di dati (popolazione o campione).

Alcune misure di dispersione misurano la variabilità dell'insieme di dati rispetto ad un suo valore centrale (solitamente la media), alcune ne considerano particolari quantili, alcune solo i valori estremi.

Le principali misure di dispersione sono:

- **campo di variazione**
- **semi-differenza interquartile**
- **intervallo fra 10° e 90° percentile**
- **scostamento medio dalla media**
- **varianza**
- **scarto quadratico medio (o deviazione standard)**
- **coefficiente di variazione**

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (2)

**Campo di variazione**: differenza fra valore massimo e valore minimo dell'insieme:

$$\text{campo di variazione} = x_{\max} - x_{\min} = x_n - x_1 \text{ (per dati ordinati)}$$

**Semi differenza interquartile**: metà della differenza fra terzo e primo quartile:

$$\text{semi-differenza interquartile} = (Q_3 - Q_1)/2$$

**Intervallo fra 10° e 90° percentile**: differenza fra novantesimo e decimo percentile:

$$\text{intervallo fra } 10^\circ \text{ e } 90^\circ \text{ percentile} = P_{90} - P_{10}$$

**Scostamento medio dalla media**: media degli scostamenti fra le osservazioni e la loro media considerati in valore assoluto:

$$\text{scostamento medio dalla media} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Il campo di variazione è una misura molto semplice che prende in esame solo i valori estremi dell'insieme dei dati e che, quindi, non considera il modo in cui si distribuiscono le osservazioni.

Le due misure basate sui quantili non prendono in considerazione il valore delle singole osservazioni, ma il modo in cui queste si distribuiscono nel loro insieme.

Così la semi-differenza interquartile rappresenta la metà del campo di variazione del 50% delle osservazioni centrali, mentre l'intervallo fra il 10° e il 90° percentile rappresenta il campo di variazione dell'80% delle osservazioni centrali.

Lo scostamento medio dalla media è una misura di dispersione che non presenta i problemi delle precedenti in quanto tiene conto di tutte le osservazioni dell'insieme; tuttavia, è una misura scarsamente utilizzata in quanto, pur avendo una capacità descrittiva molto simile allo scarto quadratico medio, a differenza di questo non possiede le proprietà matematiche che ne consentono l'utilizzazione nell'ambito della statistica inferenziale.

**Esempio:** Con riferimento alle produzioni unitarie di grano indicate nel precedente esercizio determinare il campo di variazione e lo scostamento medio dalla media.

Il campo di variazione è dato dalla differenza fra l'osservazione più grande (3,7) e la più piccola (3,1):

$$\text{campo di variazione} = x_{\max} - x_{\min} = 3,7 - 3,1 = 0,6$$

Per calcolare lo scostamento medio dalla media va considerata la media del campione che è pari a 3,45, per cui si ha:

$$\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} = \frac{0,25 + 0,25 + 0,05 + 0,15 + 0,35 + 0,15}{6} = 0,2$$

In questo caso la dimensione del campione è troppo ridotta per poter calcolare le misure di dispersione basate sui quantili.

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (2)

### VARIANZA

La varianza considera la media degli scarti fra le osservazioni e la loro media ma, a differenza dello scostamento medio dalla media, li elevata al quadrato.

La **varianza di una popolazione**  $\sigma_x^2$  è data dalla media delle differenze fra le osservazioni  $x_i$  e la media della popolazione  $\mu_x$  elevate al quadrato:

$$\sigma_x^2 = \frac{(x_1 - \mu_x)^2 + (x_2 - \mu_x)^2 + \dots + (x_N - \mu_x)^2}{N} = \frac{\sum_{i=1}^N (x_i - \mu_x)^2}{N}$$

La **varianza di un campione**  $s^2$  è "quasi" la media delle differenze fra le osservazioni  $x_i$  e la media del campione  $\bar{x}$  elevate al quadrato:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Rapporto fra media e varianza di popolazione e campione:  $\mu_x = \bar{x}$   $\sigma_x^2 = s^2 \frac{n-1}{N}$

Le motivazioni della presenza al denominatore del termine (n-1), detta anche correzione di Student, sono complesse; la ragione principale è che la statistica  $s^2$  calcolata in questo modo possiede particolari proprietà matematiche che ne consentono l'utilizzo nell'ambito dell'inferenza.

Tali proprietà derivano dal fatto che una volta determinata la media il numero di osservazioni indipendenti, detto anche gradi di libertà, è pari a (n-1) in quanto l'ultimo valore della serie non è libero di assumere qualsiasi valore.

In generale il numero di gradi di libertà è uguale al numero n dei dati meno il numero di costanti che sono già state calcolate o di informazioni che siano già state estratte dai dati stessi. Nel caso della varianza, la costante utilizzata per calcolare gli scarti è la media: quindi i gradi di libertà sono (n-1).

**Esempio:** Con riferimento alle produzioni unitarie di grano indicate nel precedente esercizio determinare la varianza considerando l'insieme dei dati sia come campione che come popolazione.

Nel caso in cui l'insieme delle n=6 aziende venga considerato come campione, ricordando che la media è pari a 3,45, la varianza assume il seguente valore:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{0,25^2 + 0,25^2 + 0,05^2 + 0,15^2 + 0,35^2 + 0,15^2}{5} = 0,059$$

Se, invece, l'insieme delle N=6 aziende viene considerato come popolazione, la varianza è pari a:

$$\sigma_x^2 = \frac{\sum_{i=1}^N (x_i - \mu_x)^2}{N} = \frac{0,25^2 + 0,25^2 + 0,05^2 + 0,15^2 + 0,35^2 + 0,15^2}{6} = 0,049$$

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (2)

### SCARTO QUADRATICO MEDIO

Lo scarto quadratico medio, sia della popolazione ( $\sigma_x$ ) che del campione ( $s$ ), è dato dalla radice quadrata della relativa varianza:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu_x)^2}{N}} \qquad s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

E' espresso nella stessa unità di misura delle osservazioni ed è sempre  $\geq 0$  (=0 solo quando le osservazioni sono tutte uguali).

Lo scarto quadratico medio può essere interpretato come una misura della variabilità media dei dati intorno alla media o, analogamente, l'errore generico che si commette approssimando un qualunque dato di un insieme con la media dell'insieme stesso (per questa ragione viene anche detto deviazione standard).

Lo scarto quadratico medio, per le sue proprietà matematiche, è la misura di dispersione più usata nell'ambito della statistica inferenziale.

Qualunque siano i valori delle osservazioni è possibile stabilire un limite minimo al numero di osservazioni comprese entro un numero  $k$  di scarti quadratici medi intorno alla media.

Questa regola (detta di Bienaymè-Chebyshev) stabilisce che nell'intervallo  $[\mu_x - k\sigma_x; \mu_x + k\sigma_x]$  sono comprese almeno il  $(1-1/k^2) \cdot 100\%$  del totale delle osservazioni.

Ponendo  $k=2$ , si vede che almeno il 75% delle osservazioni di un insieme sono comprese entro 2 scarti quadratici medi intorno alla media dell'insieme stesso.

Il fatto che lo scarto quadratico medio (così come la varianza) del campione venga calcolato ponendo al denominatore  $(n-1)$  conferisce a questa statistica delle importanti proprietà matematiche, che come si vedrà, vengono largamente utilizzate nell'ambito della statistica inferenziale.

**Esempio:** Con riferimento alle produzioni unitarie di grano indicate nel precedente esercizio determinare lo scarto quadratico medio considerando l'insieme dei dati sia come campione che come popolazione.

Essendo lo scarto quadratico medio la radice quadrata della varianza si ottiene rispettivamente per il campione e per la popolazione :

$$s = \sqrt{s^2} = \sqrt{0,059} = 0,243 \qquad \sigma_x = \sqrt{\sigma_x^2} = \sqrt{0,049} = 0,222$$

La regola di Bienaymè-Chebyshev risulta ampiamente verificata in quanto, posto  $k=2$ , nel caso in cui l'insieme dei dati venga considerato come popolazione, si ha che nell'intervallo

$$[\mu_x - k\sigma_x; \mu_x + k\sigma_x] = [3,45 - 2 \cdot 0,222; 3,45 + 2 \cdot 0,222] = [3,01; 3,89]$$

sono comprese il 100% delle osservazioni.

## I.2 STATISTICA DESCRITTIVA - Riepilogo dei dati (2)

### COEFFICIENTE DI VARIAZIONE

Il coefficiente di variazione esprime una misura relativa di dispersione; è un numero puro e viene espresso generalmente in termini percentuali.

I coefficienti di variazione della popolazione e del campione sono rispettivamente:

$$CV_{\text{pop}} = \frac{\sigma_x}{\mu_x} \cdot 100\% \qquad CV = \frac{s}{\bar{x}} \cdot 100\%$$

Il coefficiente di variazione è utile nel confronto della variabilità relativa fra diversi gruppi di osservazioni.

Ciò accade, in particolare, quando i gruppi hanno medie molto differenti o le osservazioni dei gruppi sono misurate in unità diverse.

Il coefficiente di variazione non può essere utilizzato quando la media è prossima a 0, circostanza che tende a verificarsi quando i valori delle osservazioni sono di segno sia positivo che negativo.

#### **Esempio:**

*Un agricoltore nel corso di cinque anni ha rilevato i dati relativi alla produzione (espressa in quintali per ettaro) di grano e girasole nella sua azienda. Sulla base di questi dati, riportati nel prospetto seguente, determinare quale fra le due colture presenta una maggiore variabilità nella produzione.*

Anno	Grano (q/ha)	Girasole (q/ha)
1	34	22
2	32	20
3	42	27
4	37	25
5	35	21

In questo caso va operato un confronto fra la variabilità relativa della produzione delle due colture.

A questo scopo bisogna calcolare i due coefficienti di variazione per la cui determinazione è necessario valutare la media e lo scarto quadratico medio delle due serie di dati.

Considerando i dati raccolti come una popolazione (di dimensione  $N=5$ ), per il grano si ottiene  $\mu_x = 36$  q/ha e  $\sigma_x = 3,41$  q/ha e per il girasole  $\mu_x = 23$  q/ha e  $\sigma_x = 2,61$  q/ha.

Di conseguenza i coefficienti di variazione risultano rispettivamente 9,5% e 11,3%.

Si osserva quindi che la produzione del girasole, nonostante abbia un valore inferiore dello scarto quadratico medio, presenta una maggiore variabilità rispetto al grano.

## I.2 STATISTICA DESCRITTIVA - Standardizzazione

Per confrontare osservazioni che appartengono a diversi insiemi di dati viene usata la tecnica della standardizzazione.

La standardizzazione consente di esprimere il valore di una osservazione in termini relativi rispetto all'insieme di dati cui appartiene.

Il valore standardizzato di una osservazione esprime la distanza (positiva o negativa) dalla media dell'insieme misurata in termini di scarti quadratici medi, che divengono quindi delle unità standard.

Con riferimento ad un insieme di dati, la variabile standardizzata ( $z$ ) ha la seguente espressione:

$$z = \frac{x - \bar{x}}{s} \quad \text{nel caso di un campione}$$

$$z = \frac{x - \mu_x}{\sigma_x} \quad \text{nel caso di una popolazione}$$

### Esempio:

*Con riferimento all'esercizio relativo alla produzione di grano e girasole, determinare se, per l'azienda considerata, è da considerare migliore una produzione di 42 q/ha di grano o di 27 q/ha di girasole.*

Per poter operare il confronto è necessario eseguire la standardizzazione dei due valori rispetto agli insiemi delle osservazioni cui appartengono.

Ricordando che la media e lo scarto quadratico medio delle produzioni di grano e girasole erano rispettivamente:  $\mu_{x_1} = 36$  e  $\sigma_{x_1} = 3,81$  e  $\mu_{x_2} = 23$  e  $\sigma_{x_2} = 2,92$ , si ottiene:

$$z_1 = \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} = \frac{42 - 36}{3,81} = 1,57 \qquad z_2 = \frac{x_2 - \mu_{x_2}}{\sigma_{x_2}} = \frac{27 - 23}{2,92} = 1,37$$

Si osserva, quindi, che la produzione di grano di 42 q/ha è 1,57 unità standard superiore alla media mentre una produzione di 27 q/ha di girasole è 1,37 unità standard superiore alla media; pertanto è da ritenere più elevata in termini relativi una produzione di 42 q/ha di grano.

## I.2 STATISTICA DESCRITTIVA – Forma dei dati

I dati relativi a una popolazione o a un campione possono avere una forma (distribuzione) **simmetrica** o **obliqua**.

- se la media è uguale alla mediana la distribuzione è simmetrica;
- se la media è maggiore della mediana la distribuzione è positivamente asimmetrica o obliqua destra;
- se la media è minore della mediana la distribuzione è negativamente asimmetrica o obliqua sinistra.

L'asimmetria di una distribuzione viene misurata da due coefficienti (detti di Pearson) che rapportano la differenza rispettivamente fra media e moda e fra media e mediana allo scarto quadratico medio. Tali coefficienti, essendo a-dimensionali, consentono di confrontare la forma di diverse distribuzioni: tanto maggiore è il valore dei coefficienti, tanto più asimmetrica sarà la distribuzione

$$\text{I}^\circ \text{ coefficiente di asimmetria di Pearson} = \frac{\bar{x} - \text{moda}}{s}$$

$$\text{II}^\circ \text{ coefficiente di asimmetria di Pearson} = \frac{3(\bar{x} - \text{mediana})}{s}$$

### Esempio:

Con riferimento alle temperature registrate nel mese di Giugno, di seguito elencate, determinare i principali indicatori relativi alla forma dei dati.

Data	°C										
1/6	24,0	6/6	27,5	11/6	23,5	16/6	27,0	21/6	23,0	26/6	24,0
2/6	25,5	7/6	24,0	12/6	25,0	17/6	26,0	22/6	25,0	27/6	28,0
3/6	26,5	8/6	28,0	13/6	27,0	18/6	26,5	23/6	25,5	28/6	31,0
4/6	28,5	9/6	23,5	14/6	25,0	19/6	25,0	24/6	28,5	29/6	30,5
5/6	28,5	10/6	24,0	15/6	24,0	20/6	22,0	25/6	24,5	30/6	29,0

Per la popolazione considerata la media risulta  $\mu_x=26^\circ\text{C}$ , la mediana  $25,5^\circ\text{C}$  e la moda  $24^\circ\text{C}$ .

Essendo il valore della media maggiore della mediana la distribuzione risulta positivamente asimmetrica o obliqua destra.

Per quanto riguarda i due coefficienti di asimmetria di Pearson, dopo aver calcolato lo scarto quadratico medio che risulta  $\sigma_x=2,24^\circ\text{C}$ , si ottengono i seguenti valori:

$$\text{I}^\circ \text{ coefficiente di asimmetria di Pearson} = \frac{\mu_x - \text{moda}}{\sigma_x} = \frac{26 - 24}{2,24} = 0,894$$

$$\text{II}^\circ \text{ coefficiente di asimmetria di Pearson} = \frac{3(\mu_x - \text{mediana})}{\sigma_x} = \frac{3(26 - 25,5)}{2,24} = 0,671$$

### I.3 STATISTICA DESCRITTIVA – Distribuzioni di frequenze

La **distribuzione di frequenza** di un insieme di dati quantitativi (popolazione o campione) è una tabella in cui:

- la prima colonna contiene il campo di variazione delle osservazioni diviso in **g classi**;
- la seconda colonna contiene le **frequenze delle classi**, vale a dire il numero  $f_j$  di osservazioni che ricadono in ciascuna classe  $j$  ( $j=1, \dots, g$ ).

La divisione dei dati in classi segue alcune regole:

- 1 - tutte le classi devono avere la stessa ampiezza;
- 2 - il numero  $g$  di classi dovrebbe essere compreso fra 5 e 15;
- 3 - in ogni classe il limite inferiore è incluso, il superiore escluso;
- 4 - il valore centrale  $m_j$  di una classe  $j$  è la media dei suoi estremi e viene attribuito a tutte le osservazioni della classe  $j$  nel calcolo delle misure di tendenza centrale e di dispersione.

#### Esempio

*Partendo dalle misure di lunghezza eseguite su un campione di 40 foglie (vedi prospetto seguente) costruire una distribuzione di frequenze.*

n.	mm								
1	138	9	146	17	168	25	146	33	161
2	164	10	158	18	126	26	173	34	145
3	150	11	140	19	138	27	142	35	135
4	132	12	147	20	176	28	147	36	142
5	144	13	136	21	163	29	135	37	150
6	125	14	148	22	119	30	153	38	156
7	149	15	152	23	154	31	140	39	145
8	157	16	144	24	165	32	135	40	128

Dopo aver ordinato i dati, si osserva che il campo di variazione è compreso fra 119 e 176 mm. Una distribuzione di frequenze può essere ottenuta definendo 7 classi, ciascuna dell'ampiezza di 10mm:

Classe $j$	Limiti	Val. centr. ( $m_j$ )	Frequenza ( $f_j$ )
1	110-120	115	1
2	120-130	125	3
3	130-140	135	7
4	140-150	145	14
5	150-160	155	8
6	160-170	165	5
7	170-180	175	2

### I.3 STATISTICA DESCRITTIVA – Distribuzioni di frequenze

Nel caso di distribuzioni di frequenze le principali misure di tendenza centrale e di dispersione vengono calcolate come segue:

- Media del campione: 
$$\bar{x} = \frac{m_1 f_1 + m_2 f_2 + \dots + m_g f_g}{n} = \frac{\sum_{j=1}^g m_j f_j}{n}$$

- Media della popolazione: 
$$\mu_x = \frac{m_1 f_1 + m_2 f_2 + \dots + m_g f_g}{N} = \frac{\sum_{j=1}^g m_j f_j}{N}$$

- Moda: valore centrale della classe con più osservazioni (classe modale)

- Mediana: approssimazione grafica tramite ogiva percentuale (vedi pag.21)

- Varianza del campione: 
$$s^2 = \frac{\sum_{j=1}^g (m_j - \bar{x})^2 f_j}{n - 1}$$

- Varianza della popolazione: 
$$\sigma_x^2 = \frac{\sum_{j=1}^g (m_j - \mu_x)^2 f_j}{N}$$

- Scarto quadratico medio del campione (**s**) e della popolazione ( **$\sigma_x$** ): radice quadrata della rispettiva varianza.

#### Esercizio:

Con riferimento alla distribuzione di frequenze costruita nell'esercizio precedente determinare media, moda, varianza, scarto quadratico medio e coefficiente di variazione.

La media del campione è pari a

$$\bar{x} = \frac{\sum_{j=1}^g m_j f_j}{n} = \frac{115 \times 1 + 125 \times 3 + 135 \times 7 + 145 \times 14 + 155 \times 8 + 165 \times 5 + 175 \times 2}{40} = 147 \text{ mm}$$

La moda è il valore centrale della classe modale che in questo caso è la 4° (140-150); ne consegue che

$$\mathbf{moda} = 145 \text{ mm}$$

La varianza del campione è data da

$$s^2 = \frac{\sum_{j=1}^g (m_j - \bar{x})^2 f_j}{n - 1} = \frac{1024 \times 1 + 484 \times 3 + 144 \times 7 + 4 \times 14 + 64 \times 8 + 324 \times 5 + 784 \times 2}{39} = 185,6 \text{ mm}^2$$

Lo scarto quadratico medio è pari a  $s = \sqrt{s^2} = 13,6 \text{ mm}$

Il coefficiente di variazione, rapporto fra scarto quadratico e media, è  $CV = \frac{s}{\bar{x}} 100\% = \frac{13,6}{147} 100\% = 9,3\%$

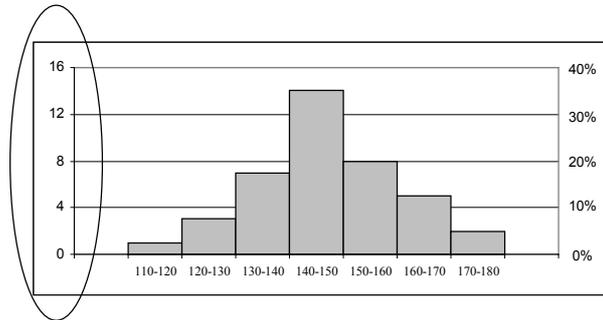
Le statistiche (o i parametri) calcolati dalla distribuzione di frequenze risultano in generale diversi da quelle ottenute partendo dai dati grezzi (nell'esercizio, ad esempio,  $\bar{x} = 146,8$  e  $s = 13,1$ ).

Tale differenza, che dipende dal fatto che nelle distribuzioni di frequenze a tutte le osservazioni di una classe viene attribuito il valore centrale, è tanto minore quanto più aumenta il numero delle classi.

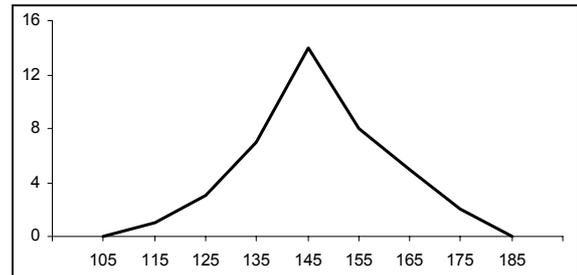
## I.3 STATISTICA DESCRITTIVA – Distribuzioni di frequenze

### Rappresentazione grafica di distribuzioni di frequenze

**Istogramma**: insieme di rettangoli con base sull'asse orizzontale, centrati sul valore centrale della classe, con larghezza pari alla ampiezza ed altezza pari alla frequenza della classe.



**Poligono di frequenze**: grafico lineare passante per i punti identificati dalle frequenze delle classi in corrispondenza dei rispettivi valori centrali.



I grafici di questa pagina e delle pagine seguenti sono riferiti alla distribuzione di frequenze costruita nell'esempio riportato nelle note della pagina precedente.

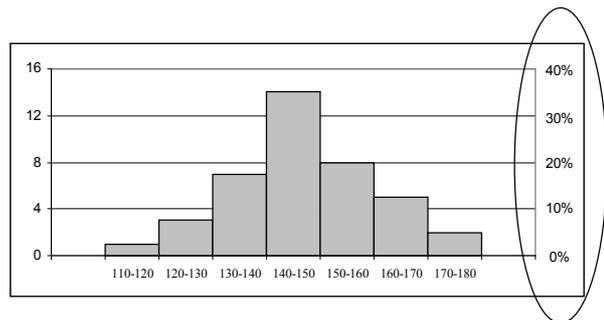
## I.3 STATISTICA DESCRITTIVA – Distribuzioni di frequenze

### Distribuzioni di frequenze relative

Quando, invece delle frequenze assolute, vengono considerate le frequenze relative (percentuali) si costruisce una **distribuzione di frequenze relative**.

Le frequenze espresse in percentuale sono utili per confrontare le distribuzioni assolute di due o più insiemi di dati.

La rappresentazione grafica di una distribuzione di frequenze relative viene eseguita con un istogramma o un poligono di frequenze relative; questi grafici sono identici a quelli costruiti per la distribuzione di frequenze assolute, ma in questo caso sull'asse verticale sono espressi i valori percentuali al posto dei valori assoluti.



#### Esempio:

Con riferimento all'esercizio precedente costruire la distribuzione di frequenze relative.

Per calcolare la frequenza relativa di una classe si divide la frequenza assoluta della classe per il totale delle osservazioni; il risultato viene generalmente espresso in termini percentuali.

Così, ad esempio, per la prima classe si ha che la frequenza relativa è  $(1/40) \times 100\% = 2,5\%$

Procedendo con il calcolo per tutte le classi definite si ottiene la seguente distribuzione:

Classe j	Limiti	Val. centr. (m <sub>i</sub> )	Freq. Ass. (f <sub>j</sub> )	Freq. Rel.
1	110-120	115	1	2,5%
2	120-130	125	3	7,5%
3	130-140	135	7	17,5%
4	140-150	145	14	35,0%
5	150-160	155	8	20,0%
6	160-170	165	5	12,5%
7	170-180	175	2	5,0%

## I.3 STATISTICA DESCRITTIVA – Distribuzioni di frequenze

### Distribuzioni di frequenze cumulate assolute e relative

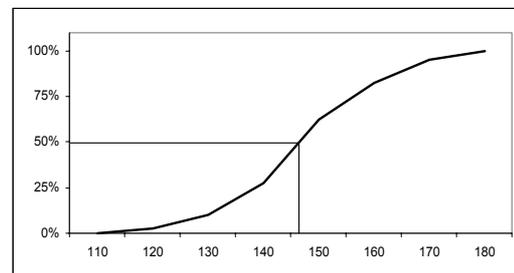
La frequenza assoluta di tutte le osservazioni inferiori al valore superiore di una classe è detta frequenza cumulata e quindi indica il numero totale di osservazioni ricadenti nelle classi precedenti e nella classe stessa.

La tabella che ha nella prima colonna le classi e nella seconda le frequenze cumulate è detta distribuzione cumulata di frequenze.

Analogamente si può costruire una distribuzione cumulata di frequenze relative.

La rappresentazione grafica delle distribuzioni cumulate (assolute o relative) viene chiamata poligono di frequenze cumulate o ogiva.

L'ogiva percentuale viene utilizzata per determinare la mediana (data dal valore sull'asse orizzontale cui corrisponde la frequenza cumulata relativa del 50%) e, più in generale, tutti i quantili.



### Esempio:

Con riferimento all'esercizio precedente costruire la distribuzione cumulata di frequenze assolute e relative.

Per calcolare la distribuzione cumulata di frequenze vengono sommate le osservazioni della classe e di tutte le classi precedenti; tale procedimento si applica sia alle frequenze assolute che alle frequenze relative ottenendo le relative distribuzioni cumulate.

Nel caso dell'esempio la distribuzione completa delle frequenze (assolute, relative e cumulative) è la seguente:

Classe	Limiti	Freq. Ass.	Freq. Rel.	Cum. Ass.	Cum. Rel.
1	110-120	1	2,5%	1	2,5%
2	120-130	3	7,5%	4	10,0%
3	130-140	7	17,5%	11	27,5%
4	140-150	14	35,0%	25	62,5%
5	150-160	8	20,0%	33	82,5%
6	160-170	5	12,5%	38	95,0%
7	170-180	2	5,0%	40	100,0%
Totale		40	100,0%		

## II.1 ELEMENTI DI PROBABILITA' – Definizioni

Detto **A** un certo evento la probabilità del verificarsi di tale evento si indica **P(A)**.

La **P(A)** può essere assegnata in base a:

- un valore teorico (probabilità oggettiva);
- la risultanze di rilevazioni empiriche (probabilità oggettiva);
- un livello di fiducia del decisore (probabilità soggettiva).

Il valore teorico della probabilità di un evento prevede una definizione classica (*rapporto tra il numero dei casi favorevoli e il numero totale dei casi possibili*) e una definizione statistica (*limite della frequenza relativa dell'evento quando il numero di osservazioni tende all'infinito*).

- La probabilità che non si verifichi **A** è indicata con **P( $\bar{A}$ )** e **P( $\bar{A}$ ) = 1 – P(A)**
- Se **A** è un evento certo **P(A) = 1**
- Se **A** è un evento impossibile **P(A) = 0**

Una probabilità assegnata in base ad un valore teorico è, ad esempio, quella dell'uscita di una faccia di un dado. In questo caso, infatti, si fa riferimento alla probabilità (teorica) di 1/6 associata all'uscita di ogni faccia di un dado.

Una probabilità assegnata sulla base delle risultanze di una valutazione empirica del verificarsi di un evento è, ad esempio, quella del numero di giorni piovosi in un mese. In questo caso le osservazioni eseguite negli anni precedenti, che fungono da dati empirici, consentono di formulare una previsione sulla probabilità associata a un certo numero di giorni piovosi.

Una probabilità assegnata in base al livello di fiducia del decisore è, ad esempio, quella che in un incontro calcistico una squadra vinca, perda o pareggi. In questo caso, infatti, l'evento non si è mai ripetuto prima nelle medesime condizioni e quindi la probabilità di ogni evento può essere assegnata solo in modo soggettivo (competenza, capacità previsionale, ecc.).

### **Esempio**

*Una moneta viene lanciata 500 volte ottenendo 262 teste. Qual'è la probabilità che nel successivo lancio esca croce?*

Se è ragionevole ipotizzare che la moneta non sia truccata, la probabilità teorica dell'evento **A**=”uscita croce” è **P(A)=1/2=0,5** indipendentemente dagli esiti precedenti.

Gli esiti osservati mostrano che la probabilità empirica dell'evento **B**=“uscita testa” è **P(B)=262/500=0,524** e, di conseguenza la probabilità empirica dell'evento **A**=”uscita croce” è **P(A)=1-P(B)=1-0,524=0,476**

Un giocatore potrebbe ritenere che l'evento **A**=”uscita croce” essendosi verificato meno volte nei lanci precedenti ha maggiori probabilità di verificarsi dell'evento **B**=”uscita testa”; sulla base di questa considerazione può assegnare una probabilità soggettiva pari, ad esempio, a **P(A)=0,55**.

## II.1 ELEMENTI DI PROBABILITA' – Definizioni

Nel caso in cui si considerino più eventi non si ha più a che fare con delle probabilità semplici, ma vengono considerate le seguenti probabilità:

- Probabilità del verificarsi contemporaneo di due o più eventi, detta **probabilità complessa** (o **composta**)

Detti **A** e **B** due eventi semplici, la probabilità complessa dei due eventi si indica con **P(AB)** ed esprime la probabilità che si verifichino entrambi gli eventi **A** e **B**.

- Probabilità del verificarsi di uno di due o più eventi

Detti **A** e **B** due eventi semplici, la probabilità del verificarsi di uno dei due eventi si indica con **P(AoB)** ed esprime la probabilità che si verifichi o l'evento **A** o l'evento **B**.

- Probabilità del verificarsi di un evento condizionata al verificarsi di un altro evento, detta **probabilità condizionata**

Detti **A** e **B** due eventi semplici, la probabilità condizionata di **A** dato **B** si indica con **P(A|B)** ed esprime la probabilità che si verifichi **A** dato il verificarsi di **B**.

### **Esempio**

*In un'urna ci sono 6 palline rosse, 5 palline blu e 4 palline verdi. Determinare la probabilità:*

- 1) che la prima pallina estratta sia rossa;*
- 2) che la prima pallina estratta non sia blu.*

Detti **R**, **B** e **V** gli eventi “estrazione pallina rossa”, “estrazione pallina blu” e “estrazione pallina verde” le relative probabilità teoriche sono:

$$P(R) = 6/15 = 0,400$$

$$P(B) = 5/15 = 0,333$$

$$P(V) = 4/15 = 0,267$$

- 1) La probabilità che la prima pallina estratta sia rossa è quindi **P(R) = 0,400**
- 2) La probabilità che la prima pallina estratta non sia blu è **P( $\bar{B}$ ) = 1 - P(B) = 1 - 0,333 = 0,667**

## II.1 ELEMENTI DI PROBABILITA' – Definizioni

La probabilità complessa  $P(AB)$  è data dalla probabilità che si verifichi  $A$  una volta verificato  $B$ ,  $P(A|B)$ , per la probabilità che si verifichi  $B$ ,  $P(B)$ :

$$P(AB) = P(A|B) P(B) \quad (\text{regola della moltiplicazione})$$

Se il verificarsi di  $B$  non influenza il verificarsi di  $A$  i due eventi si dicono indipendenti e  $P(A|B) = P(A)$ .

In questo caso la regola della moltiplicazione diviene:

$$P(AB) = P(A) P(B)$$

La probabilità che si verifichi  $A$  o  $B$ ,  $P(A \cup B)$ , è pari a:

$$P(A \cup B) = P(A) + P(B) - P(AB) \quad (\text{regola dell'addizione})$$

Se  $A$  e  $B$  sono eventi che si escludono a vicenda la loro probabilità complessa è nulla,  $P(AB)=0$ , e di conseguenza la regola dell'addizione diviene:

$$P(A \cup B) = P(A) + P(B)$$

**Esempio:** Con riferimento all'esercizio precedente determinare la probabilità che:

- 1) reinserendo la prima pallina, le prime due palline estratte siano verdi;
- 2) non reinserendo la prima pallina, le prime due palline estratte siano verdi;
- 3) tre palline estratte senza reinserimento siano nell'ordine rossa, verde e blu;
- 4) tre palline estratte senza reinserimento siano di tre colori diversi;
- 5) di due palline estratte senza reinserimento almeno una sia rossa.

1) Se ogni pallina estratta viene reinserita prima della successiva estrazione gli eventi complessi sono fra di loro indipendenti, per cui  $P(VV) = P(V) P(V) = 4/15 \cdot 4/15 = 16/225 = 0,071$

2) Se le palline estratte non vengono reinserite ogni estrazione successiva è condizionata dall'esito delle estrazioni precedenti e gli eventi complessi non sono indipendenti. In questo caso, per la regola della moltiplicazione, si ha  $P(VV) = P(V|V) P(V) = 3/14 \cdot 4/15 = 2/35 = 0,057$

3) Applicando per due volte la regola della moltiplicazione la probabilità richiesta è  $P(BVR) = P(B|VR) P(VR) = P(B|VR) P(V|R) P(R) = 5/13 \cdot 4/14 \cdot 6/15 = 4/91 = 0,044$

4) Poiché le 6 combinazioni che danno origine all'estrazione di tre palline di diverso colore, detto evento  $T$ , sono reciprocamente escludentisi si ottiene, dato che  $P(BVR)=4/91$ , la seguente:

$$P(T) = P(BVR) + P(BRV) + P(RBV) + P(RVB) + P(VRB) + P(VBR) = 6 P(BVR) = 24/91 = 0,264$$

5) In questo caso si applica la regola dell'addizione per cui la probabilità che venga estratta almeno una pallina rossa su due palline estratte senza reinserimento è data da

$$P(RoR) = P(R) + P(R) - P(RR) = 2 P(R) - P(R) P(R|R) = 2 \cdot 6/15 - 6/15 \cdot 5/14 = 0,657$$

Si sarebbe giunti allo stesso risultato considerando che l'estrazione di almeno una pallina rossa si ha quando sono rosse solo la prima, solo la seconda o entrambe:

$$P(RoR) = P(R\bar{R}) + P(\bar{R}R) + P(RR) = 9/15 \cdot 6/14 + 6/15 \cdot 9/14 + 6/15 \cdot 5/14 = 0,657$$

## II.2 ELEMENTI DI PROBABILITA' – Distribuzioni discrete

Le distribuzioni di probabilità possono riferirsi a variabili casuali discrete o continue.

Le variabili discrete possono assumere solo alcuni dei valori compresi in un intervallo mentre le variabili continue possono assumere un qualunque valore all'interno di un intervallo.

La distribuzione di probabilità di una variabile casuale discreta  $X$  è un elenco di  $k$  possibili risultati numerici  $x_i$  reciprocamente incompatibili ad ognuno dei quali è associata la probabilità  $P(x_i)$ .

I  $k$  risultati devono esaurire tutti i possibili esiti della variabile casuale, ciò implica che la somma delle  $P(x_i)$  deve essere pari ad 1.

Una distribuzione di probabilità discreta, presentando una configurazione analoga ad una distribuzione di frequenze relative, può essere rappresentata graficamente con un istogramma di frequenze relative le cui barre rappresentano le probabilità  $P(x_i)$  associate ai singoli eventi  $x_i$ .

### Esempio:

*Costruire la distribuzione di probabilità relativa al numero di figlie femmine nelle famiglie con 4 figli.*

Indicando con F una femmina e con M un maschio le possibili sequenze di 4 figli sono le seguenti:

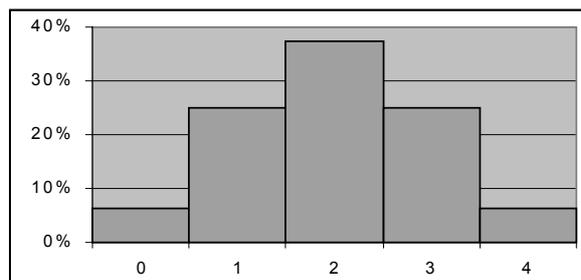
MMMM – MMMF – MMFM – MMFF – MFMM – MFMF – MFFM – MFFF

FMMM – FMMF – FMFM – FMFF – FFMM – FFMF – FFFM – FFFF

Contando il numero di combinazioni in cui ci sono rispettivamente 0, 1, 2, 3 e 4 figlie femmine si determina la relativa distribuzione di probabilità:

$x_i$	0	1	2	3	4
$P(x_i)$	1/16	4/16	6/16	4/16	1/16

Con riferimento a tale distribuzione di probabilità è possibile costruire una rappresentazione grafica; questa si configura come un istogramma le cui barre risultano adiacenti e in cui la probabilità viene riportata sull'asse verticale in valori percentuali.



## II.2 ELEMENTI DI PROBABILITA' – Distribuzioni discrete

Data la distribuzione di probabilità di una variabile casuale discreta  $X$ , si definisce valore atteso (o speranza matematica):

$$E(X) = \mu_X = \sum_{i=1}^k x_i P(x_i)$$

Il valore atteso  $E(X)$ , che corrisponde alla media della distribuzione, può non coincidere con uno dei possibili esiti (ad esempio per il lancio del dado  $E(X)=3,5$ ).

Varianza e scarto quadratico medio di una variabile casuale discreta sono date da:

$$\sigma_X^2 = \sum_{i=1}^k (x_i - \mu_X)^2 P(x_i) \qquad \sigma_X = \sqrt{\sigma_X^2}$$

Le tre principali distribuzioni di probabilità discrete sono le seguenti:

- la **distribuzione binomiale**
- la **distribuzione ipergeometrica**
- la **distribuzione di Poisson**

### Esempio

*Determinare la distribuzione di probabilità della variabile discreta  $X$  relativa all'esito del lancio di due dadi non truccati e calcolarne il valore atteso e lo scarto quadratico medio.*

*Si consideri un gioco in cui si scommette sul risultato del lancio di due dadi e in cui si perde 1€ se esce 5, 6, 7, 8 o 9, si vince 1€ se esce 3, 4, 10 o 11, e si vincono 5€ se esce 2 o 12. Determinare la distribuzione di probabilità dell'esito del gioco e la convenienza a partecipare.*

La distribuzione di probabilità è ottenuta considerando il numero di combinazioni, fra le 36 complessivamente possibili, che origina ognuno dei  $k=11$  possibili risultati:

$x_i$	2	3	4	5	6	7	8	9	10	11	12
$P(x_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Il valore atteso e lo scarto quadratico medio della distribuzione sono:

$$E(X) = \sum_{i=1}^k x_i P(x_i) = 7 \qquad \sigma_X = \sqrt{\sigma_X^2} = \sqrt{\sum_{i=1}^k (x_i - \mu_X)^2 P(x_i)} = \sqrt{5,83} = 2,42$$

Nel gioco ipotizzato i possibili esiti sono 3: perdita di 1€ (-1€), vincita di 1€ (+1€) e vincita di 5€ (+5€). La distribuzione di probabilità del lancio dei dadi e le regole del gioco consentono di attribuire a ciascuno dei 3 eventi la relativa probabilità:  $P(-1€) = 24/36$ ;  $P(+1€) = 10/36$ ;  $P(+5€) = 2/36$

da cui si ricava la seguente distribuzione di probabilità:

$x_i$	-1€	+1€	+5€
$P(x_i)$	24/36	10/36	2/36

Il valore atteso della distribuzione risulta  $-24/36+10/36+10/36=-0,11€$ ; ciò significa che il gioco non è vantaggioso in quanto si perdono, in media, 11 centesimi di euro ad ogni lancio.

## II.2 ELEMENTI DI PROBABILITA' – Distribuzione binomiale

La **distribuzione binomiale** possiede le seguenti proprietà:

- (1) le osservazioni della variabile casuale sono estratte da una popolazione infinita o finita con ripetizione;
- (2) ogni osservazione è classificata in una delle due categorie reciprocamente incompatibili: successo (esito positivo) o insuccesso (esito negativo);
- (3) le probabilità di successo **p** e di insuccesso (**1-p**) rimangono costanti in tutte le osservazioni;
- (4) il risultato di una osservazione è indipendente dal risultato delle precedenti.

Il valore della distribuzione binomiale esprime la probabilità che il numero **X** di successi sia pari ad un prefissato valore **x** noto il numero di osservazioni **n** e la probabilità attesa di successo **p**, indicata con **P(X=x | n, p)**.

Per fenomeni che seguono la distribuzione binomiale la probabilità di ottenere **x** esiti favorevoli in **n** osservazioni è data dalla seguente espressione:

$$P(X = x | n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Il numero di uscite di una particolare faccia di un dado in un certo numero di lanci è un esempio di variabile casuale discreta che segue la distribuzione binomiale in quanto:

- (1) le osservazioni sono estratte da una popolazione finita (6 possibili esiti) con ripetizione (la faccia che esce non viene eliminata prima del lancio successivo e quindi viene sempre “reinserita”);
- (2) il lancio produce un esito favorevole o sfavorevole (ad esempio la faccia con il 5, oppure la faccia con un numero superiore a 4, oppure una faccia con un numero pari, ecc.)
- (3) la probabilità di esito favorevole e sfavorevole rimane la stessa in tutti i lanci;
- (4) l'esito di un lancio è indipendente dall'esito dei precedenti.

Altri fenomeni che seguono la distribuzione binomiale sono:

- il numero di uscite di testa o croce in un certo numero di lanci di una moneta;
- il numero di estrazioni di una carta (o di un tipo di carte) da un mazzo nel caso in cui dopo ogni estrazione la carta estratta venga reinserita nel mazzo;
- il numero di piante infestate da un parassita note la probabilità di infestazione ed il numero di piante osservate (nell'ipotesi in cui la popolazione possa essere considerata infinita e la presenza di infestazione su una pianta non causi infestazione sulle piante vicine).

## II.2 ELEMENTI DI PROBABILITA' – Distribuzione binomiale

Nell'espressione  $\frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$

- il termine  $p^x(1-p)^{n-x}$  esprime la probabilità di ottenere  $x$  successi e  $(n-x)$  insuccessi in  $n$  osservazioni secondo una particolare sequenza;

- il termine  $n!/x!(n-x)!$  esprime il numero di possibili sequenze (combinazioni) con cui si ottengono  $x$  successi su  $n$  osservazioni. Il termine deriva dalle regole di conteggio del calcolo combinatorio e viene sinteticamente indicato come:  $\binom{n}{x}$

Noti i valori di  $p$  ed  $n$  si ottiene la relativa distribuzione di probabilità binomiale che possiede le seguenti caratteristiche:

Valore atteso:  $E(X) = \mu_x = np$

Scarto quadratico medio:  $\sigma_x = \sqrt{np(1-p)}$

Forma: simmetrica se  $p=0,5$ ; tanto più obliqua quanto più  $p \neq 0,5$

### Esempio

*Determinare la probabilità che esca due volte il 6 in 4 lanci di dado e la distribuzione di probabilità del numero di uscite del 6 in 4 lanci di dado.*

Il numero di uscite di una faccia di un dado, come si è visto, è una variabile casuale discreta che segue la distribuzione binomiale.

La probabilità che esca 2 volte ( $x=2$ ) la faccia con il 6 in 4 lanci ( $n=4$ ), considerando che la probabilità di successo (cioè dell'uscita del 6) è  $p=1/6$  è data da

$$P(X=2 | 4, \frac{1}{6}) = \frac{4!}{2!(4-2)!} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{4-2} = 6 \frac{1}{36} \frac{25}{36} = 0,116$$

La distribuzione di probabilità è data dall'elenco di tutte le possibili uscite del 6 nei 4 lanci con la relativa probabilità. Applicando la relazione, come nel caso precedente, si ottiene:

$$P(X=0 | 4, 1/6) = 0,482; P(X=1 | 4, 1/6) = 0,386; P(X=2 | 4, 1/6) = 0,116$$

$$P(X=3 | 4, 1/6) = 0,015; P(X=4 | 4, 1/6) = 0,001$$

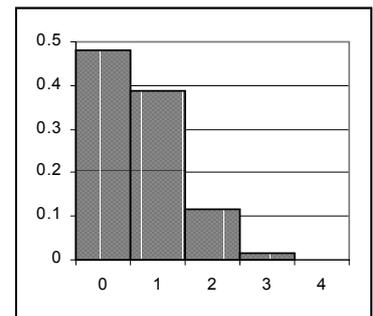
Dalle probabilità così ottenute è possibile costruire la distribuzione di probabilità ed il relativo andamento grafico.

$x_i$	0	1	2	3	4
$P(x_i)$	0,482	0,386	0,116	0,015	0,001

Il valore atteso della distribuzione risulta  $E(X) = np = 0,667$

Lo scarto quadratico medio risulta  $\sigma_x = \sqrt{np(1-p)} = 0,745$

Essendo  $p=1/6$  (distante da  $p=1/2$ ) la distribuzione è molto obliqua.



## II.2 ELEMENTI DI PROBABILITA' – Distrib. ipergeometrica

La **distribuzione ipergeometrica** si riferisce alla probabilità di successo in osservazioni estratte senza ripetizione da una popolazione finita; in questo caso la probabilità di successo varia fra le diverse osservazioni e l'esito di una osservazione è influenzato dal risultato delle precedenti.

La distribuzione ipergeometrica esprime la probabilità che il numero di successi della variabile casuale  $X$  sia pari ad  $x$  dato il numero di osservazioni  $n$ , la dimensione della popolazione  $N$  e il numero atteso di successi  $A$ :

$$P(X = x | n, N, A) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}$$

La distribuzione ipergeometrica ha le seguenti caratteristiche ( $p=A/N$ ):

Media:  $E(X) = \mu_x = np$

Scarto quadratico medio:  $\sigma_x = \sqrt{np(1-p)} \cdot \sqrt{\frac{N-n}{N-1}}$

Il secondo termine di  $\sigma_x$  è il fattore di correzione per popolazioni finite.

### Esempio

Considerando l'estrazione di 5 carte da un mazzo di carte francesi, determinare la probabilità di estrarre 2 carte di fiori e la distribuzione di probabilità dell'estrazione di carte di fiori.

Gli esiti delle estrazioni da un mazzo di carte eseguite senza reinserire le carte già estratte seguono una distribuzione ipergeometrica.

Nel caso dell'estrazione di una carta di un particolare seme da un mazzo di carte francesi si ha  $N=52$ ,  $A=13$ , e, di conseguenza,  $p=13/52=1/4=0,25$ .

La probabilità di estrarre  $x=2$  carte di fiori su  $n=5$  carte estratte è data da:

$$P(X = 2 | 5, 52, 13) = \frac{\binom{13}{2} \binom{39}{3}}{\binom{52}{5}} = \frac{78 \cdot 9.139}{2.598.960} = 0,274$$

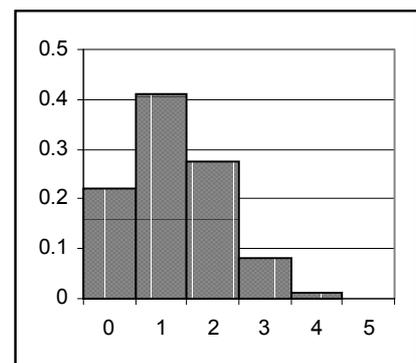
Per determinare la distribuzione di probabilità è necessario calcolare la probabilità associata a tutte le possibili estrazioni di carte di fiori su 5 carte estratte:

$P(X = 0 | 5, 52, 13) = 0,221$ ;  $P(X = 1 | 5, 52, 13) = 0,411$

$P(X = 2 | 5, 52, 13) = 0,274$ ;  $P(X = 3 | 5, 52, 13) = 0,082$

$P(X = 4 | 5, 52, 13) = 0,011$ ;  $P(X = 5 | 5, 52, 13) = 0,001$

L'andamento grafico della distribuzione di probabilità è riportato nella figura seguente.



## II.2 ELEMENTI DI PROBABILITA' – Distrib. di Poisson

La **distribuzione di Poisson** descrive la probabilità di ottenere un numero  $x$  di esiti favorevoli per unità di riferimento noto il numero atteso di esiti favorevoli per unità di riferimento ( $\lambda$ ).

Il numero atteso di successi  $\lambda$  per unità di riferimento rappresenta l'unico parametro della distribuzione i cui valori vengono calcolati come:

$$P(X = x | \lambda) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

dove  $e$  rappresenta il numero di Nepero (2,71828...).

Le caratteristiche della distribuzione di Poisson sono:

Media:  $E(X) = \mu_x = \lambda$

Scarto quadratico medio:  $\sigma_x = \sqrt{\lambda}$

A differenza delle distribuzioni binomiale ed ipergeometrica, la distribuzione di Poisson associa una probabilità a qualunque valore di  $x$ .

**Esempio:** In un negozio entrano in media 2 clienti al minuto, determinare

- la probabilità che in un minuto entri un solo cliente;
- la probabilità che in un minuto entrino più di 3 clienti;
- la distribuzione di probabilità del numero di arrivi di clienti in un minuto.

La probabilità del numero di arrivi di clienti nell'unità di tempo, noto il numero atteso di arrivi per unità di tempo ( $\lambda=2$ ), segue una distribuzione di probabilità di Poisson.

La probabilità che in un minuto entri un cliente ( $x=1$ ) è allora pari a:  $P(X=1|2) = \frac{2^1 \cdot e^{-2}}{1!} = 0,271$

La probabilità che in un minuto entrino più di tre clienti deve essere calcolata per differenza sottraendo da 1 la probabilità che in un minuto entrino tre clienti o meno; è infatti possibile che in un minuto entri un numero qualunque di clienti. Applicando la distribuzione di Poisson si ottiene

$P(X=0|2)=0,135$ ;  $P(X=1|2)=0,271$ ;  $P(X=2|2)=0,271$ ;  $P(X=3|2)=0,180$

Per cui  $P(X>3|2)=1-P(X\leq 3|2) = 1-(0,135+0,271+0,271+0,180)=0,143$

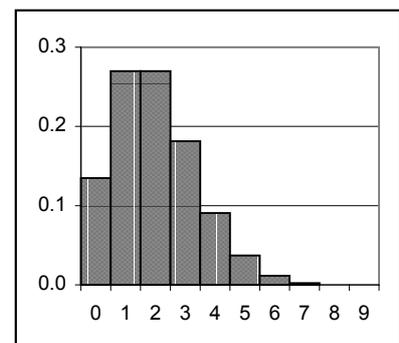
Per determinare la distribuzione di probabilità si calcolano le probabilità associate ad un numero crescente  $x$  di successi, fino a quando queste non assumono valori trascurabili.

Avendo già calcolato le probabilità per  $x=0, 1, 2, 3$ , si ha:

$P(X=4|2)=0,090$ ;  $P(X=5|2)=0,036$ ;  $P(X=6|2)=0,012$ ;

$P(X=7|2)=0,003$ ;  $P(X=8|2)=0,001$ ;  $P(X=9|2)=0,000$

Da cui si ottiene la distribuzione di probabilità e la relativa rappresentazione grafica.



## **II.3 ELEMENTI DI PROBABILITA' – Distribuzioni continue**

Per le variabili casuali continue, che assumono un numero infinito di valori, non è possibile calcolare la probabilità di un particolare risultato.

Per le distribuzioni di probabilità continue, che prendono il nome di funzioni di densità di probabilità, si può invece determinare la probabilità che il valore di una osservazione cada in un determinato intervallo.

La probabilità che una osservazione cada in un intervallo è rappresentata dall'area sottostante la funzione densità di probabilità compresa fra gli estremi dell'intervallo stesso (ne consegue che tutta l'area compresa sotto una funzione densità di probabilità è sempre pari ad 1).

Il calcolo delle probabilità associate ai diversi intervalli (e quindi delle relative aree) richiede la conoscenza dell'espressione analitica della funzione di densità di probabilità ed il ricorso al calcolo integrale.

Per alcune distribuzioni, particolarmente utili nell'analisi statistica, le probabilità associate ai diversi intervalli sono state calcolate e sono state inserite in delle tabelle dette tavole statistiche.

## II.3 ELEMENTI DI PROBABILITA' – Distribuzione normale

La più importante funzione densità di probabilità è la **distribuzione gaussiana** o **distribuzione normale**.

La fondamentale importanza della distribuzione normale è dovuta al fatto che :

- descrive e rappresenta molti fenomeni reali;
- è usata per approssimare diverse distribuzioni di probabilità discrete, evitando calcoli complessi e laboriosi;
- costituisce la base dell'inferenza statistica classica attraverso il teorema del limite centrale.

La distribuzione normale ha le seguenti proprietà:

- forma "a campana" simmetrica;
- variabile casuale  $X$  con un campo di variazione infinito;
- misure di tendenza centrale (media, mediana, moda) coincidenti.

Le probabilità di fenomeni reali che si discostano in misura limitata dalle suddette caratteristiche sono comunque determinabili con buona precisione utilizzando la distribuzione normale.

## II.3 ELEMENTI DI PROBABILITA' – Distribuzione normale

La funzione di densità di probabilità della variabile casuale  $X$  che descrive la distribuzione normale ha la forma seguente:

$$f(X) = \frac{1}{\sigma_X \sqrt{2\pi}} e^{-\frac{(X-\mu_X)^2}{2\sigma_X^2}}$$

La densità di probabilità  $f(X)$  dipende da due parametri della distribuzione:  $\mu_X$  (media) e  $\sigma_X$  (scarto quadratico medio).

Al variare dei due parametri corrispondono diverse distribuzioni normali.

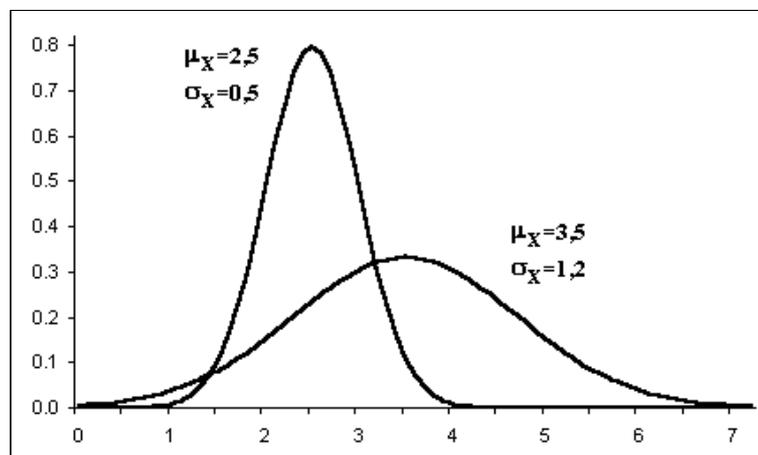
Il calcolo delle aree sotto una curva normale, corrispondenti a diversi intervalli della variabile casuale  $X$ , si presenta molto laborioso.

E' quindi necessario poter disporre di tabelle nelle quali vengono riportate le probabilità associate agli intervalli di una variabile casuale che segue la distribuzione normale, indipendentemente dal valore dei parametri  $\mu_X$  e  $\sigma_X$ .

La posizione della campana relativa alla distribuzione normale rispetto all'asse delle ascisse dipende dalla media della distribuzione ( $\mu_X$ ); tanto maggiore è la media della distribuzione tanto più la campana sarà spostata verso destra.

L'apertura della campana relativa alla distribuzione normale dipende dal valore dello scarto quadratico medio ( $\sigma_X$ ); tanto maggiore è lo scarto quadratico medio della distribuzione tanto più la campana risulterà "larga".

La figura seguente mostra due curve normali diverse per media e per scarto quadratico medio.



## II.3 ELEMENTI DI PROBABILITA' – Distribuzione normale

Per calcolare, indipendentemente dai parametri della distribuzione, le ordinate della distribuzione normale e le aree da essa sottese in corrispondenza degli intervalli della variabile casuale  $X$  è necessario standardizzare la variabile casuale  $X$  definendo la variabile casuale normale  $Z$  i cui valori sono

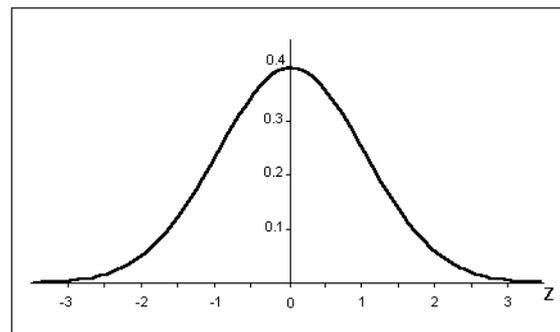
$$z = \frac{x - \mu_x}{\sigma_x}$$

La variabile casuale  $Z$  ha:

- media nulla ( $\mu_z=0$ );
- scarto quad. medio unitario ( $\sigma_z=1$ );
- funzione densità di probabilità:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- andamento grafico della  $f(z)$  come mostrato in figura:



La figura a destra mostra la corrispondenza fra una distribuzione normale generica ( $\mu_x=57$  e  $\sigma_x=0,5$ ) e la distribuzione normale standardizzata.

La corrispondenza dei valori si ottiene attraverso il procedimento della standardizzazione.

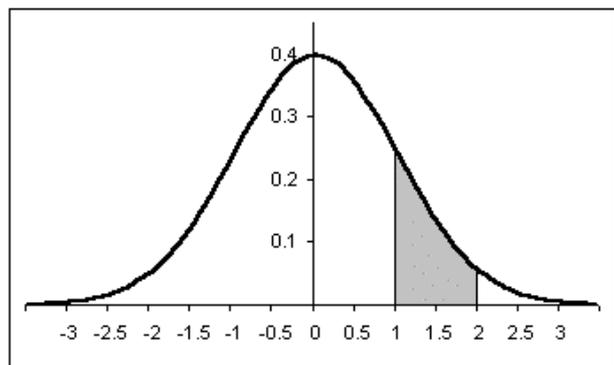
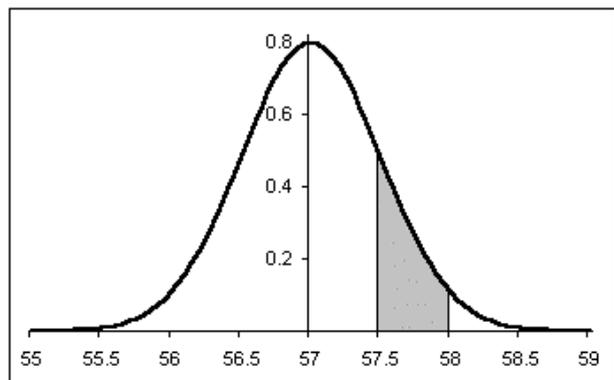
Ad esempio, per i valori riportati, si ha:

$$z_1 = \frac{x_1 - \mu_x}{\sigma_x} = \frac{57,5 - 57}{0,5} = 1$$

$$z_2 = \frac{x_2 - \mu_x}{\sigma_x} = \frac{58 - 57}{0,5} = 2$$

Inoltre si verifica che le due aree evidenziate sotto le curve normali comprese fra i valori della variabile  $X$  ed i corrispondenti valori standardizzati sono uguali.

Ciò, come si vedrà, consente di utilizzare la distribuzione normale standardizzata per calcolare qualunque area compresa sotto una generica distribuzione normale e, quindi, qualunque probabilità relativa a variabili che seguono una tale distribuzione.



## II.3 ELEMENTI DI PROBABILITA' – Distribuzione normale

Le tabelle relative alla distribuzione normale standardizzata riportano

- le ordinate della funzione  $f(\mathbf{Z})$ ;
- l'area  $F(\mathbf{Z})$  compresa sotto la  $f(\mathbf{Z})$  fra 0 e i diversi valori di  $\mathbf{Z}$ , vale a dire:

$$F(\mathbf{Z}) = \int_0^{\mathbf{Z}} f(\mathbf{Z}) d\mathbf{Z}$$

Nelle tabelle vengono considerati solo valori di  $\mathbf{Z} \geq 0$  in quanto, per la simmetria della  $f(\mathbf{Z})$ , da questi possono essere desunti tutti gli altri valori.

Dalla tabella relativa alle aree comprese sotto la distribuzione normale standardizzata si legge, ad esempio, che l'area sotto la distribuzione compresa fra  $z=0$  e  $z=0,32$  è pari 0,1255

$\mathbf{Z}$	0	1	2	...	9
0,0	0,0000	0,0040	0,0080		0,0359
0,1	0,0398	0,0438	0,0478		0,0754
0,2	0,0793	0,0832	0,0871		0,1141
0,3	0,1179	0,1217	0,1255		0,1517
...					
3,9	0,5000	0,5000	0,5000		0,5000

## II.3 ELEMENTI DI PROBABILITA' – Distribuzione normale

La conoscenza dell'area sotto la curva normale compresa fra  $z=0$  ed un qualunque valore  $z$  della variabile casuale  $Z$  consente di determinare:

- la probabilità (area **A1**) che la variabile casuale  $X$  sia maggiore di un valore  $x_a$

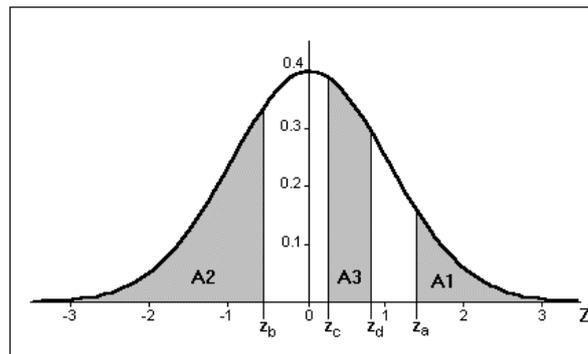
$$X \geq x_a \rightarrow Z \geq z_a \text{ con } z_a = (x_a - \mu_X) / \sigma_X$$

- la probabilità (area **A2**) che la variabile casuale  $X$  sia minore di un valore  $x_b$

$$X \leq x_b \rightarrow Z \leq z_b \text{ con } z_b = (x_b - \mu_X) / \sigma_X$$

- la probabilità (area **A3**) che la variabile casuale  $X$  sia compresa fra i valori  $x_c$  e  $x_d$

$$x_c \leq X \leq x_d \rightarrow z_c \leq Z \leq z_d \text{ con } z_c = (x_c - \mu_X) / \sigma_X \text{ e } z_d = (x_d - \mu_X) / \sigma_X$$



### Esempio

Una segheria produce assi di legno che hanno una lunghezza che segue una distribuzione normale con media di 2,40m e scarto quadratico medio di 2cm. Determinare la probabilità che la lunghezza di un'asse sia maggiore di 2,45m, minore di 2,42m e compresa fra 2,36m e 2,44m;

Per poter determinare le probabilità richieste è necessario determinare le relative aree comprese sotto alla distribuzione normale standardizzata.

A questo scopo bisogna standardizzare le diverse misure e considerare se si è interessati a conoscere l'area sottesa dalla distribuzione normale a destra o a sinistra di queste.

Nel primo caso si ha ( $X \geq x=2,45m$ ):

$$z = \frac{x - \mu_X}{\sigma_X} = \frac{2,45 - 2,40}{0,02} = 2,5$$

L'area che deve essere determinata è quella compresa sotto la distribuzione normale a destra di  $z=2,5$  ed è data da tutta l'area compresa a destra del valore  $z=0$  (pari, per simmetria, a 0,5) meno quella compresa fra 0 e 2,5, il cui valore è riportato in tabella:

$$P(X \geq 2,45) = P(Z \geq 2,5) = 0,5 - F(2,5) = 0,5 - 0,4938 = 0,0062 = 0,62\%$$

Nel secondo caso ( $X \leq x=2,42m$ ) il valore standardizzato è  $z=1$ . L'area che interessa è quella compresa sotto la distribuzione a sinistra di 1 che è pari a 0,5 più l'area compresa fra 0 e 1:

$$P(X \leq 2,42) = P(Z \leq 1) = 0,5 + F(1) = 0,5 + 0,3413 = 0,8413 = 84,13\%$$

Nel terzo caso ( $2,36m \leq X \leq 2,44m$ ) va determinata l'area sotto la distribuzione standardizzata compresa fra  $z_1=-2$  e  $z_2=2$ . Per la simmetria della distribuzione si ha:

$$P(2,36 \leq X \leq 2,44) = P(-2 \leq Z \leq 2) = 2 \cdot F(2) = 2 \cdot 0,4772 = 0,9544 = 95,44\%$$

## II.3 ELEMENTI DI PROBABILITA' – Distribuzione normale

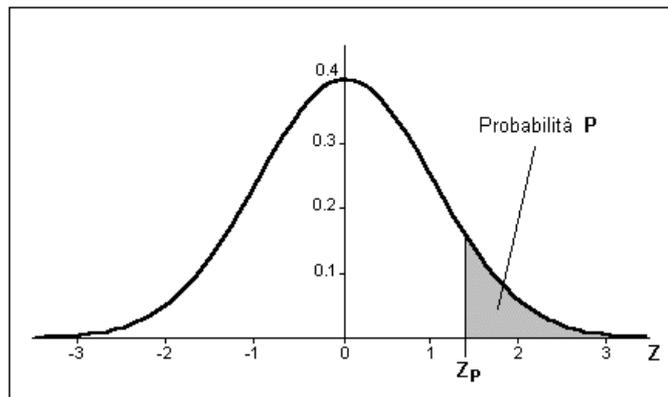
La conoscenza dell'area sotto la curva normale compresa nell'intervallo fra  $z=0$  ed un qualunque valore  $z$  della variabile casuale  $Z$  consente anche di calcolare i valori corrispondenti a probabilità note.

In questo caso si procede determinando per mezzo delle tabelle il valore  $z_p$  della variabile casuale  $Z$  cui corrisponde la probabilità  $P$  (cioè l'area) nota.

Il corrispondente valore  $x_p$  della variabile casuale  $X$  verrà determinato eseguendo la de-standardizzazione del valore  $z_p$  ottenuto:

$$x_p = \mu_X + z_p \sigma_X$$

Il valore di  $z_p$  sarà positivo o negativo a seconda del tipo di probabilità assegnata.



### Esercizio

Con riferimento alla situazione illustrata nel precedente esercizio determinare:

- la misura al di sotto della quale si trova il 3% delle assi;
- la misura al di sopra della quale si trova l'1% delle assi;
- le misure entro cui saranno comprese il 50%, il 90%, il 95%, il 99% delle assi.

Per rispondere ai quesiti è necessario trovare i valori della distribuzione normale corrispondenti alle probabilità indicate e quindi de-standardizzarli.

Nel primo caso il valore  $z_p$  cercato è quello che determina un'area sotto la coda sinistra pari a 0,03. Tale valore si legge sulle tabelle vedendo a quale valore di  $Z$  corrisponde un'area di 0,47 (0,5-0,03).

$$z_p = -F^{-1}(0,47) = -1,88 \quad \text{e} \quad x_p = \mu_X + z_p \sigma_X = 2,40 - 1,88 \cdot 0,02 = 2,362$$

Nel secondo caso il valore  $z_p$  cercato è quello che determina un'area sotto la coda destra pari a 0,01. Con modalità analoghe a quelle seguite nel caso precedente si ha:

$$z_p = F^{-1}(0,49) = 2,33 \quad \text{e} \quad x_p = \mu_X + z_p \sigma_X = 2,40 + 2,33 \cdot 0,02 = 2,447$$

Per i quattro quesiti relativi all'ultimo punto, vista la simmetria della distribuzione, va calcolato il valore corrispondente alla metà della probabilità richiesta. Si ha così per un probabilità del 50%:

$$z_p = F^{-1}(0,25) = \pm 0,675; \quad x_p = \mu_X + z_p \sigma_X = 2,40 \pm 0,675 \cdot 0,02 \quad \text{da cui} \quad 2,387 \leq X \leq 2,413$$

In modo analogo per le altre probabilità indicate si ottiene:

$$z_{90\%} = F^{-1}(0,45) = \pm 1,645 \quad x_p = 2,40 \pm 0,02 \cdot 1,645 \quad 2,367 \leq X \leq 2,433$$

$$z_{95\%} = F^{-1}(0,475) = \pm 1,96 \quad x_p = 2,40 \pm 0,02 \cdot 1,96 \quad 2,361 \leq X \leq 2,439$$

$$z_{99\%} = F^{-1}(0,495) = \pm 2,58 \quad x_p = 2,40 \pm 0,02 \cdot 2,58 \quad 2,348 \leq X \leq 2,452$$

## II.3 ELEMENTI DI PROBABILITA' – Approssimazione normale

La distribuzione normale si presta molto bene per approssimare distribuzioni discrete come la binomiale, l'ipergeometrica e di Poisson. Tale approssimazione risulta utile quando i conteggi da eseguire sono particolarmente laboriosi.

Per le tre distribuzioni discrete considerate l'approssimazione ottenuta con la distribuzione normale è tanto migliore quanto maggiore è il valore atteso della distribuzione; quando  $E(X) < 5$  l'approssimazione non è più accettabile.

Il riferimento alla distribuzione normale viene ottenuto definendo in corrispondenza al numero stabilito di successi  $x$  il valore standardizzato  $z$ :

$$z = \frac{(x \pm 0,5) - \mu_x}{\sigma_x}$$

I termini  $\mu_x$  e  $\sigma_x$  rappresentano la media (valore atteso) e lo scarto quadratico medio della distribuzione di probabilità discreta seguita dalla variabile casuale  $X$ .

Il termine  $\pm 0,5$  costituisce una correzione per continuità che viene inserita proprio per tenere conto della continuità della distribuzione normale.

Il segno  $+$  o  $-$  dipende dal tipo di probabilità che si intende determinare.

Per comprendere la funzione del fattore di correzione per continuità si considerino i casi seguenti relativi al calcolo di probabilità legate ad esiti del lancio di una moneta:

- che esca testa più di 40 volte in 100 lanci;
- che esca testa meno di 55 volte in 100 lanci;
- che esca testa 45 volte in 100 lanci.

Nel primo caso la probabilità richiesta guarda “a destra” del valore  $x=40$  in quanto deve essere considerata la probabilità di 41 uscite ma esclusa quella di 40 uscite. Si sceglierà allora un fattore di correzione per continuità pari a  $+0,5$  che rende il valore da standardizzare pari a  $40,5 \rightarrow P(X > 40,5)$ .

Nel secondo caso la probabilità richiesta guarda “a sinistra” del valore  $x=55$  in quanto deve essere considerata la probabilità di 54 uscite ma esclusa quella di 55 uscite. Si sceglierà allora un fattore di correzione per continuità pari a  $-0,5$  che rende il valore da standardizzare pari a  $54,5 \rightarrow P(X < 54,5)$ .

Nel terzo caso la probabilità richiesta guarda soltanto il valore  $x=45$ , escludendo la probabilità di esiti sia inferiori (da 44 in giù) che superiori (da 46 in su). In questo caso, anche se l'approssimazione normale non viene di solito utilizzata, si sceglie un fattore di correzione per continuità pari sia a  $-0,5$  che a  $+0,5$  rendendo i valori da standardizzare pari a  $44,5$  e  $45,5 \rightarrow P(44,5 < X < 45,5)$ .

## II.3 ELEMENTI DI PROBABILITA' – Approssimazione normale

### Distribuzione binomiale

La distribuzione normale standardizzata approssima la distribuzione binomiale quando  $np \geq 5$  e  $n(1-p) \geq 5$ . L'approssimazione migliora al crescere di  $n$ .

Se sono verificate tali condizioni è possibile valutare la probabilità associata al verificarsi su  $n$  osservazioni di un numero di esiti positivi superiore o inferiore ad  $x$  calcolando il seguente valore standardizzato:

$$z = \frac{(x \pm 0,5) - np}{\sqrt{np(1-p)}}$$

Il termine  $np$  è la media (valore atteso) della distribuzione binomiale e  $\sqrt{np(1-p)}$  ne è lo scarto quadratico medio.

Se, ad esempio, si vuole determinare la probabilità di ottenere più di 20 teste in 50 lanci di una moneta si avrà:  $P(X > x=20 | 50, 1/2) = P(Z > z=-1,273)$

### **Esempio**

*Determinare la probabilità che in 10 lanci di moneta esca testa per meno di 4 volte.*

Applicando la distribuzione binomiale ( $n=10$ ,  $p=1/2$ ) per  $x=0,1,2,3$  si ottiene:

$$P(X=0|10,1/2)=1/1024; \quad P(X=1|10,1/2)=10/1024; \quad P(X=2|10,1/2)=45/1024; \quad P(X=3|10,1/2)=120/1024$$

Sommando queste probabilità si ottiene la probabilità che esca testa per meno di quattro volte:

$$P(X < x=4 | 10, 1/2) = 176/1024 = 17,2\%$$

Considerando che  $np=5$  e  $n(1-p)=5$ , l'approssimazione con la distribuzione normale fornisce un risultato accettabile.

Poiché si vuole determinare la probabilità che  $X$  sia minore di  $x=4$  nella correzione per continuità si sceglie il segno – considerando l'area a sinistra di 3,5.

$$z = \frac{(x - 0,5) - np}{\sqrt{np(1-p)}} = \frac{3,5 - 5}{\sqrt{5(1-0,5)}} = \frac{-1,5}{1,58} = -0,95$$

La probabilità cercata è quella sotto la distribuzione normale standardizzata a sinistra di  $z=-0,95$ :

$$P(Z < z=-0,95) = 0,5 - 0,329 = 0,171 = 17,1\%$$

Come si nota, nonostante ci si trovi nelle condizioni limite di applicazione della approssimazione con la distribuzione normale, i due valori (esatto e approssimato) sono molto simili.

## II.3 ELEMENTI DI PROBABILITA' – Approssimazione normale

### Distribuzione ipergeometrica

La distribuzione normale standardizzata approssima la distribuzione ipergeometrica quando  $np \geq 5$  e  $n(1-p) \geq 5$ . L'approssimazione migliora al crescere di  $n$ .

Se sono verificate tali condizioni, data una popolazione di dimensione  $N$ , è possibile valutare la probabilità associata al verificarsi su  $n$  osservazioni di un numero di esiti positivi superiore o inferiore ad  $x$  calcolando il seguente valore standardizzato:

$$z = \frac{(x \pm 0,5) - np}{\sqrt{np(1-p)} \sqrt{\frac{N-n}{N-1}}}$$

Il termine  $np$  è la media (valore atteso) della distribuzione ipergeometrica mentre per lo scarto quadratico medio, rispetto alla distribuzione binomiale, è presente il fattore di correzione per popolazioni finite.

Nel caso in cui  $n$  risulti molto piccolo rispetto ad  $N$  il fattore di correzione può essere ignorato e l'approssimazione diviene identica al caso della binomiale.

### Esempio

*Per un concorso sono state presentate 1000 domande di partecipazione. La commissione decide di controllare la veridicità delle dichiarazioni dei candidati selezionando un campione di 50 domande. Se le domande con dichiarazioni mendaci sono complessivamente 120, qual è la probabilità che nel campione selezionato ne risultino presenti meno di 3?*

Detta  $N=1000$  la dimensione della popolazione (numero totale di domande),  $n=50$  il numero di esiti considerati (numero di domande selezionate) e  $A=120$  il numero di successi nella popolazione (domande con dichiarazioni mendaci), la probabilità ricercata è la seguente:

$$P(X \leq x=3 | 1000, 120, 50)$$

La probabilità attesa di successi è pari a  $p=A/N=120/1000=0,12$

Ne consegue che il valore riportato sulla distribuzione normale standardizzata corrispondente al valore  $x$  è il seguente:

$$z = \frac{(x \pm 0,5) - np}{\sqrt{np(1-p)} \sqrt{\frac{N-n}{N-1}}} = \frac{(3-0,5) - 6}{\sqrt{5,28} \sqrt{\frac{950}{999}}} = \frac{-3,5}{2,241} = -1,56$$

E' quindi possibile calcolare la probabilità richiesta utilizzando l'approssimazione normale:

$$P(X \leq x=3 | 1000, 120, 50) = P(Z \leq z=-1,56) = 0,5 - 0,441 = 0,059 = 5,9\%$$

## II.3 ELEMENTI DI PROBABILITA' – Approssimazione normale

### Distribuzione di Poisson

La distribuzione normale standardizzata approssima la distribuzione di Poisson quando  $\lambda \geq 5$ .

Se è verificata la precedente condizione, dato il numero di successi attesi nell'unità di riferimento è possibile valutare la probabilità associata al verificarsi di un numero di successi nell'unità di riferimento superiore o inferiore ad  $x$  calcolando il seguente valore standardizzato:

$$z = \frac{(x \pm 0,5) - \lambda}{\sqrt{\lambda}}$$

Il termine  $\lambda$  è la media (valore atteso) della distribuzione di Poisson mentre lo scarto quadratico medio è la radice di  $\lambda$ .

### **Esempio**

*In un parco si vuole realizzare un'area attrezzata con dei tavoli per consentire il pic-nic ai visitatori. Considerando che il parco viene visitato in media da 25 famiglie al giorno, quanti tavoli bisogna prevedere per avere il 99% di probabilità che nessuna famiglia rimanga senza posto?*

Per risolvere il problema utilizzando la distribuzione di Poisson bisognerebbe calcolare le probabilità associate ad un numero di famiglie pari a 0,1,2,...e andare avanti fino a quando la somma di tali probabilità raggiunge il 99%.

Per evitare questo procedimento, che risulta assai laborioso, si può procedere utilizzando l'approssimazione normale (essendo  $\lambda=25 > 5$ ).

Operando con questo metodo si deve determinare il valore di  $z$  per cui la probabilità che  $Z > z$  è 0,01.

Leggendo sulle tavole si ottiene che  $P(Z > z) = 0,01$  quando  $z = 2,33$ .

Dalla relazione inversa che lega il valore standardizzato  $z$  con  $x$  si ricava il valore  $x$  cercato:

$$x = \lambda + z\sqrt{\lambda} = 25 + 2,33 \cdot 5 = 36,65$$

Essendo  $x$  una variabile intera (il numero di tavoli), andrà considerato il numero intero superiore a quello ottenuto per la variabile de-standardizzata.

Si conclude quindi che è necessario prevedere almeno 37 tavoli per avere la probabilità del 99% che tutte le famiglie che visitano il parco possano sedersi per fare il pic-nic.

### III.1 TEORIA DEI CAMPIONI – Concetti generali

La teoria dei campioni studia le relazioni esistenti tra un campione e la popolazione da cui è stato estratto allo scopo di:

- stimare i parametri della popolazione sulla base delle statistiche del campione (**teoria della stima**);
- decidere se le differenze osservate fra due campioni sono da ritenere casuali o da attribuire ad una differenza fra le relative popolazioni (**teoria delle decisioni**).

Queste analisi (**inferenze**) forniscono risultati corretti soltanto se i campioni sono rappresentativi delle popolazioni. Ciò si verifica quando il campione contiene un numero sufficiente di osservazioni estratte in modo casuale.

Sia la modalità di estrazione casuale dei campioni che la determinazione della sua dimensione sono oggetto dello studio del piano degli esperimenti.

La casualità del campione può essere ottenuta con diverse tecniche, la più semplice delle quali consiste nel selezionare le osservazioni con l'aiuto di una tavola dei numeri casuali.

La dimensione del campione deve essere scelta in relazione al livello di affidabilità desiderato sulla stima dei parametri.

### III.1 TEORIA DEI CAMPIONI – Distribuzioni campionarie

Per utilizzare la teoria della probabilità nella inferenza statistica è necessario partire dal concetto di distribuzione campionaria.

Per ciascun campione di dimensione **n** estratto da una popolazione è possibile calcolare la media, la mediana, la varianza, lo scarto quadratico medio, ecc.

I valori di queste statistiche in tutti i possibili campioni di dimensione **n** estratti dalla popolazione danno origine alle distribuzioni delle statistiche campionarie.

La dimensione della popolazione e la modalità di estrazione del campione influenzano la distribuzione delle statistiche campionarie ed i loro parametri.

Il numero dei possibili campioni di dimensione **n** estraibili da una popolazione è illustrato nella seguente tabella:

	Popolazione infinita	Popolazione finita (N)
Campionamento con ripetizione	<b>Infiniti</b> ( $\infty$ )	<b>Finiti</b> ( $N^n$ )
Campionamento senza ripetizione	<b>Infiniti</b> ( $\infty$ )	<b>Finiti</b> ( $N!/n!(N-n)!$ )

### III.1 TEORIA DEI CAMPIONI – Distribuzioni campionarie

Fra le possibili distribuzioni delle statistiche campionarie la più utilizzata nelle analisi inferenziali è la distribuzione della media campionaria ( $\bar{X}$ )

Indicata con  $\mu_X$  la media della popolazione e con  $\mu_{\bar{X}}$  la media della distribuzione della media campionaria accade sempre che:

$$\mu_{\bar{X}} = \mu_X$$

Per questa ragione una media campionaria  $\bar{x}$  è uno **stimatore** della media della popolazione  $\mu_X$  che possiede alcune importanti caratteristiche:

- è **non distorto** (la media della distribuzione del parametro campionario è uguale al parametro della popolazione);
- è **efficiente** (fra tutte le misure di tendenza centrale è quella che varia meno da campione a campione);
- è **consistente** (al crescere di  $n$  la statistica del campione tende sempre più ad avvicinarsi al parametro della popolazione).

Per verificare l'uguaglianza fra la media della popolazione e la media della distribuzione della media campionaria si consideri una popolazione  $X$  costituita dalle seguenti  $N=5$  osservazioni: 2, 5, 2, 8, 3.

La media della popolazione risulta  $\mu_X = 20/5 = 4$

Per quanto riguarda la distribuzione campionaria si consideri, ad esempio, una dimensione del campione pari a  $n=3$  con una modalità di estrazione senza ripetizione.

In queste ipotesi tutti i possibili campioni estraibili dalla popolazione sono i seguenti:

2, 5, 2	2, 5, 8	2, 5, 3	2, 2, 8	2, 2, 3
2, 8, 3	5, 2, 8	5, 2, 3	5, 8, 3	2, 8, 3

Le medie dei 10 possibili campioni sono rispettivamente 3, 5, 10/3, 4, 7/3, 13/3, 5, 10/3, 16/3, 13/3 per cui la media della distribuzione campionaria risulta  $\mu_{\bar{X}} = (69/3+17)/10 = 4$

Risulta, come si voleva provare, che  $\mu_{\bar{X}} = \mu_X$

In generale il numero di campioni di dimensione  $n$  estraibili senza ripetizione da una popolazione finita di dimensione  $N$  è pari a:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Nell'esempio proposto si ha, infatti:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)}{(3 \cdot 2 \cdot 1) \cdot (2 \cdot 1)} = \frac{120}{12} = 10$$

Se il campionamento avviene con ripetizione il numero di possibili campioni di dimensione  $n$  estraibili dalla popolazione di dimensione  $N$  è pari invece a  $N^n$ .

### III.1 TEORIA DEI CAMPIONI – Distribuzioni campionarie

Data una variabile casuale  $X$  con scarto quadratico medio  $\sigma_X$  la distribuzione della media campionaria  $\bar{X}$  ha scarto quadratico medio  $\sigma_{\bar{X}}$

Quando il campionamento avviene da una popolazione infinita o con ripetizione da una popolazione finita lo scarto quadratico medio della distribuzione della media campionaria è pari a:

$$\sigma_{\bar{X}} = \sigma_X / \sqrt{n}$$

Quando il campionamento avviene senza ripetizione da una popolazione finita di dimensione  $N$ , allo scarto quadratico medio della distribuzione della media campionaria va applicato il fattore di correzione per popolazioni finite:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Il termine  $\sigma_{\bar{X}}$  viene indicato anche come errore standard della media in quanto rappresenta una misura della variabilità della media da campione a campione.

Con riferimento alla popolazione precedente ( $X=2,5,2,8,3$ ) lo scarto quadratico medio è  $\sigma_X=2,28$ .

Si consideri la distribuzione della media campionaria con riferimento ai campioni di dimensione  $n=2$  estratti con ripetizione da tale popolazione. Il numero dei possibili campioni è  $N^n=5^2=25$ .

Tali campioni con le relative medie sono i seguenti:

2, 2 → 2,0	2, 5 → 3,5	2, 2 → 2,0	2, 8 → 5,0	2, 3 → 2,5
5, 2 → 3,5	5, 5 → 5,0	5, 2 → 3,5	5, 8 → 6,5	5, 3 → 4,0
2, 2 → 2,0	2, 5 → 3,5	2, 2 → 2,0	2, 8 → 5,0	2, 3 → 2,5
8, 2 → 5,0	8, 5 → 6,5	8, 2 → 5,0	8, 8 → 8,0	8, 3 → 5,5
3, 2 → 2,5	3, 5 → 4,0	3, 2 → 2,5	3, 8 → 5,5	3, 3 → 3,0

La media delle medie campionarie è  $\mu_{\bar{X}}=4$  e, come verificato, è uguale alla media della popolazione.

Lo scarto quadratico medio della media campionaria risulta pari a  $\sigma_{\bar{X}} = \sqrt{65/25} = 1,612$

Considerando che il campionamento è con ripetizione si verifica che  $\sigma_{\bar{X}} = \sigma_X / \sqrt{n} = 2,28 / \sqrt{2} = 1,612$

Se il campionamento avviene senza ripetizione i campioni ( $5!/3!2!=10$ ) con le relative medie sono:

2, 5 → 3,5	2, 2 → 2,0	2, 8 → 5,0	2, 3 → 2,5	5, 2 → 3,5
5, 8 → 6,5	5, 3 → 4,0	2, 8 → 5,0	2, 3 → 2,5	8, 3 → 5,5

La media delle medie campionarie è ancora 4 e lo scarto quadratico medio è  $\sigma_{\bar{X}} = \sqrt{19,5/10} = 1,396$

Essendo il campionamento senza ripetizione va applicato il fattore di correzione per popolazioni finite.

Si verifica allora che  $\sigma_{\bar{X}} = (\sigma_X / \sqrt{n}) \sqrt{(N-n)/(N-1)} = (2,28 / \sqrt{2}) \sqrt{3/4} = 1,396$

### III.1 TEORIA DEI CAMPIONI – Distribuzioni campionarie

Data una variabile casuale  $X$  con media  $\mu_X$  e scarto quadratico medio  $\sigma_X$  si è mostrato che la distribuzione della media campionaria  $\bar{X}$  ha

- media:  $\mu_{\bar{X}} = \mu_X$

- scarto quadratico medio:  $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$  o  $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

Si può verificare inoltre che:

- se la distribuzione  $X$  è normale, la distribuzione  $\bar{X}$  è normale;

- se la distribuzione  $X$  non è normale, la distribuzione  $\bar{X}$  tende alla distribuzione normale al crescere della dimensione del campione (teorema del limite centrale).

Per una dimensione del campione sufficientemente “grande” ( $n > 30$ ) la distribuzione della media campionaria non mostra apprezzabili scostamenti dalla distribuzione normale; per questa ragione si dice che la distribuzione della media campionaria è asintoticamente normale.

#### Esempio

Dagli assi di legno la cui lunghezza segue una distribuzione normale con media di 2,40m e scarto quadratico medio di 2cm viene estratto un campione di 10 assi.

- Qual è la probabilità che la media della lunghezza delle assi del campione sia maggiore di 2,41m?

- Entro quale intervallo si trova il 99% delle medie delle lunghezze delle assi dei campioni di dimensione  $n=10$ ?

Essendo la popolazione praticamente infinita, la media e lo scarto quadratico medio della distribuzione della media campionaria sono:

$$\mu_{\bar{X}} = \mu_X = 2,40\text{m} \quad \sigma_{\bar{X}} = \sigma_X / \sqrt{n} = 2 / \sqrt{10} = 0,63 \text{ cm} = 0,0063 \text{ m}$$

La popolazione è distribuita normalmente per cui la media campionaria segue una distribuzione normale, indipendentemente dalla dimensione  $n$  del campione.

E' quindi possibile, dopo aver standardizzato, calcolare la probabilità richiesta:

$$z = \frac{\bar{x} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{2,41 - 2,40}{0,0063} = 1,58 \quad P(\bar{x} \geq 2,41) = P(Z \geq z = 1,58) = 0,5 - F(1,58) = 0,5 - 0,443 = 0,057 = 5,7\%$$

Per determinare l'intervallo entro cui sarà compreso il 99% delle medie campionarie si procede calcolando i valori sulla distribuzione normale cui corrisponde un'area nelle code pari a 0,01 e quindi si de-standardizza tale valore utilizzando i parametri della distribuzione campionaria.

Il valore  $z_c$  cercato è quello per cui l'area sotto la curva normale standardizzata da 0 a  $z_c$  è di 0,495.

Leggendo sulle tavole si trova che  $z_c = F^{-1}(0,495) = 2,58$  per cui si ha

$$\bar{x} = \mu_{\bar{X}} \pm \sigma_{\bar{X}} z_c = 2,40 \pm 0,0063 \cdot 2,58 = 2,40 \pm 0,016$$

Ne consegue che il 99% dei campioni di 10 assi hanno una lunghezza media fra 2,384m e 2,416m.

### III.1 TEORIA DEI CAMPIONI – Distribuzioni campionarie

Nel caso delle variabili qualitative la caratteristica che viene presa in considerazione è la frequenza relativa di successi  $p$  nella popolazione.

Se si considerano le frequenze relative di successi  $p_s$  in tutti i campioni di dimensione  $n$  estratti dalla popolazione si ottiene la distribuzione campionaria della frequenza relativa che avrà media  $\mu_{p_s}$  e scarto quadratico medio  $\sigma_{p_s}$

Per qualunque dimensione della popolazione e tipo di campionamento

$$\mu_{p_s} = p$$

Per quanto riguarda lo scarto quadratico medio:

- per popolazioni infinite o per popolazioni finite con ripetizione:  $\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}}$

- per popolazioni finite senza ripetizione:  $\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$

Anche in questo la distribuzione campionaria è asintoticamente normale; e quindi per  $n > 30$  può essere considerata normale.

Si consideri la seguente popolazione di  $N=5$  elementi (S=successo, I=insuccesso)

S, I, I, S, I

La frequenza relativa di successi risulta pari a  $p=2/5=0,40$ .

I campioni di dimensione  $n=2$  estratti senza ripetizione dalla popolazione sono i seguenti:

S, I – S, I – S, S – S, I – I, I – I, S – I, I – I, S – I, I – S, I

Le frequenze relative di successi nei 10 campioni sono le seguenti:

0,5 – 0,5 – 1 – 0,5 – 0 – 0,5 – 0 – 0,5 – 0 – 0,5

La media della distribuzione della frequenza campionaria è  $\mu_{p_s}=4/10=0,40$ .

Si verifica quindi che  $p = \mu_{p_s}$

Lo scarto quadratico medio della distribuzione della frequenza relativa campionaria è pari a

$$\sigma_{p_s} = \sqrt{0,9/10} = 0,3$$

e coincide con  $\sigma_{p_s} = \sqrt{p(1-p)/n} \sqrt{(N-n)/(N-1)} = 0,3$

### III.2 TEORIA DELLA STIMA – Definizioni

La **teoria della stima** comprende i metodi per inferire (stimare) i parametri della popolazione partendo dalle statistiche del campione.

La stima di un parametro può essere eseguita definendo:

- un solo valore (**stima puntuale**);
- gli estremi di un intervallo (**stima per intervallo**).

Poiché la media della distribuzione della media campionaria è uguale alla media della popolazione ( $\mu_{\bar{x}} = \mu_x$ ), una qualunque media campionaria ( $\bar{x}$ ) rappresenta una **stima puntuale corretta** (non distorta) della media della popolazione ( $\mu_x$ ).

Con la **stima per intervallo** si ottengono indicazioni migliori in quanto viene determinato un intervallo intorno alla statistica campionaria (ad esempio la media) nel quale è compreso, con una certa probabilità (o **confidenza**), il valore del relativo parametro della popolazione.

Tanto più vogliamo essere confidenti che il parametro si trovi nell'intervallo di stima tanto maggiore deve essere l'ampiezza dell'intervallo stesso.

### III.2 TEORIA DELLA STIMA – Intervalli di stima della media

Nel caso di popolazioni che seguono la distribuzione normale o di grandi campioni ( $n > 30$ ) le statistiche campionarie seguono la distribuzione normale, ciò consente di valutare il legame quantitativo fra il livello di confidenza e l'ampiezza dell'intervallo per mezzo dei valori di questa distribuzione.

Per una variabile casuale  $X$  di media  $\mu_X$  e scarto quadratico medio  $\sigma_X$  da cui vengono estratti campioni di dimensione  $n > 30$  la distribuzione della media campionaria  $\bar{X}$  segue una distribuzione normale.

La relativa distribuzione normale standardizzata è:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}$$

Quando il campionamento viene eseguito senza ripetizione da una popolazione finita, al denominatore della precedente espressione deve essere applicato il fattore di correzione per popolazioni finite.

### III.2 TEORIA DELLA STIMA – Intervalli di stima della media

In corrispondenza di ciascun livello di confidenza è possibile determinare il relativo valore (critico)  $z_C$  della distribuzione normale standardizzata.

Ne consegue che la variabile casuale standardizzata  $Z$  sarà compresa fra  $-z_C$  e  $+z_C$  con una probabilità pari a tale livello di confidenza.

I livelli di confidenza più utilizzati sono il **90%**, il **95%** e il **99%** a cui corrispondono, per la simmetria della distribuzione normale, i seguenti valori:

$$z_{C(90\%)}=F^{-1}(0,45)=1,645 \quad z_{C(95\%)}=F^{-1}(0,475)=1,96 \quad z_{C(99\%)}=F^{-1}(0,495)=2,58$$

Ciò comporta che, per una variabile casuale che segue la distribuzione normale standardizzata, esiste un livello di confidenza (cioè una probabilità):

- del 90% che sia compresa fra  $-1,645$  e  $+1,645$ ;
- del 95% che sia compresa fra  $-1,96$  e  $1,96$ ;
- del 99% che sia compresa fra  $-2,58$  e  $2,58$ .

### III.2 TEORIA DELLA STIMA – Intervalli di stima della media

Considerato un campione di dimensione  $n > 30$  di media  $\bar{x}$  estratto da una popolazione (infinita o finita con ripetizione) di media  $\mu_x$  e scarto quadratico medio  $\sigma_x$  e fissato un certo livello di confidenza, la variabile

$$\frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}$$

sarà compresa fra i relativi valori critici  $-z_c$  e  $+z_c$  con tale livello di confidenza.

Ne consegue che la media della popolazione  $\mu_x$  può essere stimata dalla media di un campione di dimensione  $n > 30$  estratto da essa con la seguente relazione:

$$\bar{x} - z_c \frac{\sigma_x}{\sqrt{n}} \leq \mu_x \leq \bar{x} + z_c \frac{\sigma_x}{\sqrt{n}}$$

Dove  $z_c$  è il valore critico della distribuzione normale corrispondente al livello di confidenza stabilito per la stima.

Si può così affermare che l'espressione

$$\bar{x} \pm z_c \frac{\sigma_x}{\sqrt{n}}$$

rappresenta un intervallo (o forchetta) di stima della media della popolazione.

Fissato un livello di confidenza (ad esempio il 95%) si è dimostrato come l'espressione

$$\frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}$$

risulti compresa fra i valori critici  $-z_c$  (-1,96) e  $+z_c$  (+1,96) con una probabilità pari a tale livello di confidenza (95%):

$$-z_c \leq \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} \leq +z_c$$

Da questa relazione è possibile determinare l'intervallo di stima della media della popolazione.

La relazione precedente può essere divisa in  $\frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} \geq -z_c$        $\frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} \leq +z_c$

Risolvendo entrambe rispetto a  $\mu_x$  si ottiene  $\bar{x} - \mu_x \geq -z_c \frac{\sigma_x}{\sqrt{n}}$        $\bar{x} - \mu_x \leq +z_c \frac{\sigma_x}{\sqrt{n}}$

E quindi  $-\mu_x \geq -\bar{x} - z_c \frac{\sigma_x}{\sqrt{n}}$        $-\mu_x \leq -\bar{x} + z_c \frac{\sigma_x}{\sqrt{n}}$

Cambiando di segno  $\mu_x \leq \bar{x} + z_c \frac{\sigma_x}{\sqrt{n}}$        $\mu_x \geq \bar{x} - z_c \frac{\sigma_x}{\sqrt{n}}$

Ricomponendo, si ricava l'intervallo di stima  $\bar{x} - z_c \frac{\sigma_x}{\sqrt{n}} \leq \mu_x \leq \bar{x} + z_c \frac{\sigma_x}{\sqrt{n}}$

### III.2 TEORIA DELLA STIMA – Intervalli di stima della media

La stima per intervallo della media della popolazione  $\mu_x$  prevede i seguenti passi:

- 1) si estrae un campione di dimensione  $n$  dalla popolazione
- 2) si determina la media campionaria  $\bar{x}$
- 3) si considera lo scarto quadratico medio della popolazione  $\sigma_x$
- 4) si fissa il livello di confidenza e il relativo valore critico  $z_c$
- 5) si calcola l'intervallo di stima di  $\mu_x$  con la seguente relazione:

$$\bar{x} - z_c \frac{\sigma_x}{\sqrt{n}} \leq \mu_x \leq \bar{x} + z_c \frac{\sigma_x}{\sqrt{n}}$$

Tale procedimento è valido nelle seguenti ipotesi:

- è noto lo scarto quadratico medio della popolazione  $\sigma_x$
- la popolazione è distribuita normalmente oppure la popolazione ha una qualsiasi distribuzione ma il campione è di “grande” dimensione ( $n > 30$ );
- la popolazione è infinita oppure la popolazione è finita ma il campionamento viene eseguito con ripetizione.

#### Esercizio

*Il Ministero delle Politiche Agricole e Forestali è interessato a conoscere l'età media degli imprenditori agricoli in Italia. A questo scopo esegue un'indagine campionaria su 135 aziende; la media dell'età dei conduttori delle aziende del campione risulta di 43 anni.*

*Eseguire la stima con un livello di confidenza del 95% e del 99% considerando che dai censimenti precedenti è risultato che lo scarto quadratico medio dell'età degli imprenditori agricoli è di 9 anni.*

In questo caso tutte le condizioni sono verificate in quanto è noto lo scarto quadratico medio della popolazione, il campione è di grande dimensione e la popolazione può essere considerata infinita.

Essendo stati già eseguiti i primi tre punti per la esecuzione della stima:

- 1) estrazione del campione ( $n=135$ );
- 2) determinazione della media campionaria ( $\bar{x} = 43$ );
- 3) valutazione dello scarto quadratico medio della popolazione ( $\sigma_x=9$ );

ed essendo noti i valori critici della distribuzione normale in corrispondenza dei livelli di confidenza specificati:

- 4)  $z_c=1,96$  (95%);  $z_c=2,58$  (99%);

si può procedere alla determinazione degli intervalli di stima della media della popolazione.

- 5) Le “forchette” di stima al 95% e al 99% dell'età media degli imprenditori agricoli italiani sono:

$$43 - 1,96 \frac{9}{\sqrt{135}} \leq \mu_x \leq 43 + 1,96 \frac{9}{\sqrt{135}} \qquad 41,5 \leq \mu_x \leq 44,5$$

$$43 - 2,58 \frac{9}{\sqrt{135}} \leq \mu_x \leq 43 + 2,58 \frac{9}{\sqrt{135}} \qquad 41 \leq \mu_x \leq 45$$

### III.2 TEORIA DELLA STIMA – Intervalli di stima della media

Con il metodo illustrato la stima della media della popolazione  $\mu_x$  richiede la conoscenza dello scarto quadratico medio della popolazione  $\sigma_x$ .

E' molto difficile, però, che di una popolazione di cui si desidera stimare la media si conosca lo scarto quadratico medio.

Quando non si dispone di questo parametro, o di una sua valutazione affidabile, si utilizza la sua stima rappresentata dallo scarto quadratico medio del campione  $s$ .

L'approssimazione di  $\sigma_x$  con  $s$  è tanto migliore quanto maggiore è la dimensione del campione.

Nel caso di "piccoli" campioni ( $n < 30$ ) tale approssimazione non è sufficiente; ciò comporta che in questi casi non è più possibile utilizzare la distribuzione normale per determinare l'intervallo di stima della media.

Nel caso in cui il campione sia "piccolo" ( $n < 30$ ) e non sia noto lo scarto quadratico medio della popolazione è possibile eseguire una stima per intervallo di  $\mu_x$  (utilizzando un'altra distribuzione di probabilità) solo nel caso in cui la popolazione sia distribuita normalmente (o quasi).

#### Esempio

*In un studio sperimentale condotto in un bosco è stato misurato il diametro del tronco di un campione di 36 alberi ottenendo una media di 3,2m e uno scarto quadratico medio di 0,74m.*

*Eseguire una stima del diametro medio del tronco degli alberi del bosco.*

Essendo la dimensione del campione  $n=36 > 30$  è possibile eseguire la stima della media della popolazione  $\mu_x$  utilizzando la distribuzione normale.

Lo scarto quadratico medio della popolazione  $\sigma_x$  non è noto ma, essendo la dimensione del campione maggiore di 30, si utilizza la sua stima rappresentata dallo scarto quadratico medio del campione  $s$ .

Inoltre, dato il numero totale degli alberi del bosco, è possibile considerare la popolazione infinita.

Si ha allora,  $\bar{x}=3,2m$  e  $\sigma_x = s = 0,74m$

Se si stabilisce un livello di confidenza del 95%, cui corrisponde  $z_c=1,96$ , l'intervallo di stima del diametro del tronco degli alberi risulta allora:

$$3,2 - 1,96 \frac{0,74}{\sqrt{36}} \leq \mu_x \leq 3,2 + 1,96 \frac{0,74}{\sqrt{36}} \qquad 2,96m \leq \mu_x \leq 3,44m$$

Con un livello di confidenza del 99% ( $z_c=2,58$ ), l'intervallo di stima sarebbe risultato invece:

$$3,2 - 2,58 \frac{0,74}{\sqrt{36}} \leq \mu_x \leq 3,2 + 2,58 \frac{0,74}{\sqrt{36}} \qquad 2,88m \leq \mu_x \leq 3,52m$$

Chiaramente all'aumentare del livello di confidenza l'intervallo di stima diviene più ampio.

### III.2 TEORIA DELLA STIMA – Distribuzione t di Student

Se una variabile casuale  $X$  è distribuita normalmente la variabile casuale campionaria

$$\frac{\bar{x} - \mu_X}{s / \sqrt{n}}$$

segue una distribuzione detta **t di Student** con  $(n-1)$  gradi di libertà ( $t_{n-1}$ ) indipendentemente dalla dimensione  $n$  del campione.

La distribuzione **t di Student** viene utilizzata per stimare la media di una popolazione che segue una distribuzione normale (o approssimativamente normale) partendo da un campione di qualunque dimensione estratto da essa.

Nel caso di piccoli campioni ( $n < 30$ ), non potendo stimare correttamente  $\sigma_X$  con  $s$ , e non potendo quindi utilizzare la distribuzione normale, è necessario ricorrere alla distribuzione **t di Student**, che rappresenta la distribuzione di probabilità di riferimento nell'ambito della **teoria dei piccoli campioni**.

### III.2 TEORIA DELLA STIMA – Distribuzione t di Student

La distribuzione **t** di **Student** ha un andamento molto simile alla distribuzione normale ma presenta un'area minore al centro e maggiore nelle code.

All'aumentare dei gradi di libertà la distribuzione **t** di **Student** si avvicina progressivamente alla distribuzione normale; per  $n \geq 150$  la differenza fra le due distribuzioni è praticamente trascurabile.

Applicando la **t** di **Student** l'intervallo di stima della media della popolazione è:

$$\bar{x} - t_{(n-1)C} \frac{s}{\sqrt{n}} \leq \mu_X \leq \bar{x} + t_{(n-1)C} \frac{s}{\sqrt{n}}$$

Il termine  $t_{(n-1)C}$  indica il valore critico della distribuzione della **t** di **Student** con  $(n-1)$  gradi di libertà in corrispondenza del livello di confidenza stabilito.

Per la **t** di **Student** sono disponibili delle tabelle che riportano l'area compresa sotto la distribuzione per i diversi gradi di libertà e in corrispondenza dei più comuni livelli di confidenza.

Ad esempio, per  $n=16$ , i valori critici letti dalle tabelle risultano:

$$\text{al } 90\%: t_{(15)C}=1,75; \quad \text{al } 95\%: t_{(15)C}=2,13; \quad \text{al } 99\%: t_{(15)C}=2,95$$

#### Esempio

*Una fabbrica di funi intende mettere in commercio un nuovo tipo di fune e conduce dei test per stabilirne la resistenza a trazione. Vengono sottoposte al test 8 funi che si rompono quando il carico applicato (in kg) è rispettivamente pari a 2250, 1980, 2160, 2270, 2140, 2090, 2040, 2190.*

*Determinare gli intervalli di stima della resistenza media a trazione del nuovo tipo di fune con un livello di confidenza del 95% e 99%.*

Le statistiche (media e scarto quadratico medio) del campione selezionato sono pari a:

$$\bar{x} = 2140 \text{kg} \quad s = 100 \text{kg.}$$

La dimensione del campione è "piccola" ( $n=8$ ) per cui è necessario determinare l'intervallo di stima della media della popolazione facendo ricorso alla distribuzione t di Student.

Per un livello di confidenza del 95% con  $(n-1)=7$  gradi di libertà, dalle tabelle risulta  $t_{(7)C}=2,36$  da cui

$$2140 - 2,36 \frac{100}{\sqrt{8}} \leq \mu_X \leq 2140 + 2,36 \frac{100}{\sqrt{8}} \quad 2057 \leq \mu_X \leq 2223$$

Per il livello di confidenza del 99% si ha  $t_{7,(0,995)}=3,50$  da cui

$$2140 - 3,50 \frac{100}{\sqrt{8}} \leq \mu_X \leq 2140 + 3,50 \frac{100}{\sqrt{8}} \quad 2016 \leq \mu_X \leq 2264$$

### III.2 TEORIA DELLA STIMA – Stima della frequenza relativa

Per la stima della frequenza relativa  $p$  della popolazione (numero di successi in rapporto al numero di osservazioni totali) si procede in modo analogo alla stima della media.

Considerato un campione di dimensione  $n$  estratto dalla popolazione (infinita o finita con ripetizione) e detta  $p_s$  la sua frequenza relativa, l'intervallo (forchetta) di stima della frequenza relativa della popolazione  $p$  è dato da:

$$p_s - z_C \sqrt{\frac{p(1-p)}{n}} \leq p \leq p_s + z_C \sqrt{\frac{p(1-p)}{n}}$$

Non essendo noto il valore della frequenza relativa della popolazione  $p$ , lo scarto quadratico medio della popolazione viene calcolato stimando (puntualmente)  $p$  con  $p_s$ ; per cui l'intervallo di stima risulterà:

$$p_s - z_C \sqrt{\frac{p_s(1-p_s)}{n}} \leq p \leq p_s + z_C \sqrt{\frac{p_s(1-p_s)}{n}}$$

#### Esempio

*Durante un sondaggio elettorale sono stati intervistati 160 elettori 88 dei quali hanno espresso la preferenza per un certo candidato. Stimare con il 95% ed il 99% di confidenza la percentuale effettiva di consensi che tale candidato otterrà alle elezioni.*

La dimensione del campione è  $n=160$ , il numero di successi nel campione è 88; la frequenza relativa del campione è pertanto pari a  $p_s = 88/160 = 0,55 = 55\%$ .

Lo scarto quadratico medio della popolazione, considerando la sua dimensione come infinita, può essere calcolato stimando  $p$  con  $p_s$  e quindi:

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p_s(1-p_s)}{n}} = \sqrt{\frac{0,55(1-0,55)}{160}} = 0,0393$$

Per un livello di confidenza del 95%, cui corrisponde un valore critico  $z_C=1,96$ , l'intervallo di stima della frequenza relativa della popolazione risulta:

$$0,55 - 1,96 \cdot 0,0393 \leq p \leq 0,55 + 1,96 \cdot 0,0393 \quad 47,3\% \leq p \leq 62,7\%$$

Per un livello di confidenza del 99%, cui corrisponde  $z_C=2,58$ , l'intervallo diviene:

$$0,55 - 2,58 \cdot 0,0393 \leq p \leq 0,55 + 2,58 \cdot 0,0393 \quad 44,9\% \leq p \leq 65,1\%$$

### III.2 TEORIA DELLA STIMA – Stima per popolazioni finite

Nel caso della stima di parametri (media o frequenza relativa) di popolazioni finite da cui il campionamento viene eseguito senza ripetizione, va tenuto conto della dimensione  $N$  della popolazione applicando il fattore di correzione per popolazioni finite.

Gli estremi dell'intervallo per la stima della media della popolazione  $\mu_x$  risultano:

$$\bar{x} \pm z_c \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Nel caso di "piccoli campioni" ( $n < 30$ ) gli estremi dell'intervallo per la stima della media della popolazione  $\mu_x$  divengono:

$$\bar{x} \pm t_{n-1} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

La stima della frequenza relativa della popolazione  $p$  in questo caso avviene utilizzando il seguente intervallo:

$$p_s \pm z_c \sqrt{\frac{p_s(1-p_s)}{n}} \sqrt{\frac{N-n}{N-1}}$$

#### Esempio

*Ripetere la stima eseguita nell'esercizio precedente considerando che la consultazione elettorale si svolge in un comune in cui risiedono 2.700 elettori.*

In questo caso la popolazione è finita ( $N=2.700$ ) e il campionamento avviene senza ripetizione in quanto l'intenzione di voto non viene chiesta più volte allo stesso elettore.

Ricordando che la frequenza relativa del campione è  $p_s = 0,55$ , si calcola il nuovo valore dello scarto quadratico medio applicando il fattore di correzione per popolazioni finite.

$$\sqrt{\frac{p_s(1-p_s)}{n}} \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0,55(1-0,55)}{160}} \sqrt{\frac{2700-160}{2700-1}} = 0,0393 \cdot 0,97 = 0,0381$$

Per un livello di confidenza del 95%, cui corrisponde un valore critico  $z_c=1,96$ , l'intervallo di stima della frequenza relativa della popolazione risulta in questo caso:

$$0,55 - 1,96 \cdot 0,0381 \leq p \leq 0,55 + 1,96 \cdot 0,0381 \qquad 47,5\% \leq p \leq 62,5\%$$

Per un livello di confidenza del 99%, cui corrisponde  $z_c=2,58$ , l'intervallo diviene:

$$0,55 - 2,58 \cdot 0,0381 \leq p \leq 0,55 + 2,58 \cdot 0,0381 \qquad 45,2\% \leq p \leq 64,8\%$$

### III.2 TEORIA DELLA STIMA – Dimensione del campione

Nell'impostare una indagine campionaria è importante determinare la dimensione minima del campione che consente di eseguire la stima del parametro della popolazione con un errore massimo prefissato per il livello di confidenza stabilito.

Nel caso della stima dalla media della popolazione l'errore di stima  $e$ , dato dalla differenza fra media campionaria  $\bar{x}$  e media della popolazione  $\mu_x$ , è pari a:

$$e = \mu_x - \bar{x} = z_c \frac{\sigma_x}{\sqrt{n}}$$

e, nel caso di "piccoli campioni"

$$e = \mu_x - \bar{x} = t_{(n-1)c} \frac{s}{\sqrt{n}}$$

Risolviendo per  $n$ , si ottiene  $n = \left( \frac{z_c \sigma_x}{e} \right)^2$  e, per "piccoli campioni"  $n = \left( \frac{t_{n-1} s}{e} \right)^2$

La dimensione del campione per la stima della media è determinata dall'errore ( $e$ ), dal livello di confidenza ( $z_c$  o  $t_{(n-1)c}$ ) e dallo scarto quadratico medio ( $\sigma_x$  o  $s$ ).

### III.2 TEORIA DELLA STIMA – Dimensione del campione

Nel caso della stima della frequenza relativa l'errore di stima  $e$  è dato dalla differenza fra il parametro della popolazione ( $p$ ) e la corrispondente statistica campionaria ( $p_s$ ); per cui si ha:

$$e = p - p_s = z_C \sqrt{\frac{p(1-p)}{n}}$$

Risolviendo per  $n$ , si ottiene

$$n = \frac{z_C^2 p(1-p)}{e^2}$$

La dimensione del campione dipende da  $p$  che, essendo il parametro oggetto della stima, non può essere noto.

Va considerato, comunque, che il termine  $p(1-p)$  assume il valore massimo possibile per  $p=0,5$ ; di conseguenza si può usare questo valore per stabilire la dimensione del campione nel modo più prudente possibile.

#### Esempio

*Un istituto di ricerche statistiche intende effettuare un sondaggio per stimare la percentuale di consensi di un partito. Fissato un errore massimo dell'1% e un livello di confidenza del 90%, determinare il numero minimo di elettori da intervistare nel caso in cui*

- a) il partito nelle precedenti elezioni abbia avuto il 5% dei voti;*
- b) il partito possa ottenere la maggioranza assoluta.*

La dimensione del campione per il sondaggio viene ricavata dalla relazione

$$n = \frac{z_C^2 p(1-p)}{e^2} \quad \text{con } e=0,01 \text{ e } z_C = F^{-1}(0,45) = 1,645$$

Per quanto riguarda il valore della frequenza relativa, nel caso a) si può effettuare la stima ponendo  $p=0,1$  (è improbabile che in due elezioni successive un partito passi dal 5% a più del 10%); nel caso b), essendo possibile il conseguimento della maggioranza, è consigliabile mettersi nell'ipotesi più prudente e porre  $p=0,5$ .

Nel caso a) 
$$n = \frac{1,645^2 \cdot 0,1 \cdot 0,9}{0,01^2} = 2.436$$

Nel caso b) 
$$n = \frac{1,645^2 \cdot 0,5 \cdot 0,5}{0,01^2} = 6.765$$

### III.2 TEORIA DELLA STIMA – Dimensione del campione

Nel caso in cui il campionamento debba essere eseguito senza ripetizione da una popolazione finita, l'errore nella stima della media è pari a:

$$e = \mu_X - \bar{x} = z_C \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

e, per "piccoli campioni"

$$e = \mu_X - \bar{x} = t_{n-1} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Mentre l'errore nella stima della frequenza relativa risulterà:

$$e = p - p_s = z_C \sqrt{\frac{p_s(1-p_s)}{n}} \sqrt{\frac{N-n}{N-1}}$$

In tutti i casi per valutare  $n$  si procede determinando dapprima la dimensione del campione ( $n_0$ ) come se la popolazione fosse infinita e quindi si calcola la dimensione effettiva applicando la seguente relazione:

$$n = \frac{n_0 N}{n_0 + (N - 1)}$$

#### Esercizio

Con riferimento all'esercizio precedente, determinare la dimensione del campione considerando che gli elettori sono complessivamente 200.000.

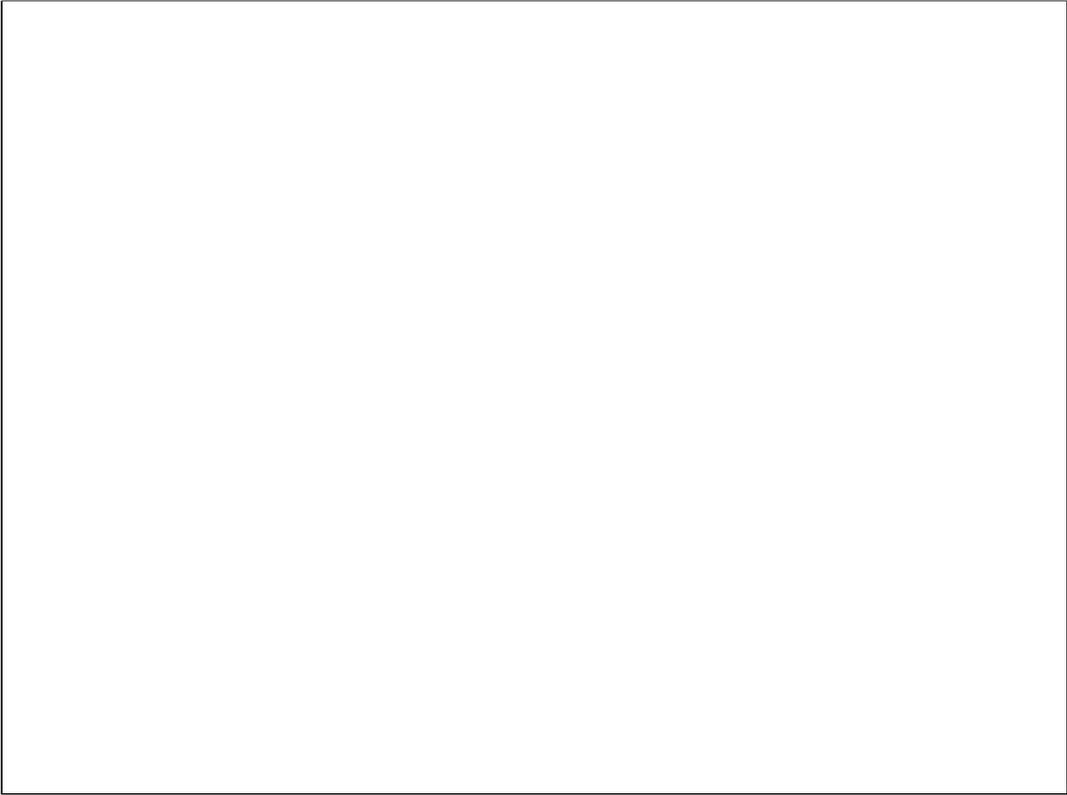
La dimensione del campione, calcolata per un numero infinito di elettori, è risultata nel caso a) pari a  $n_0 = 2.436$  e nel caso b) pari a  $n_0 = 6.765$ .

Applicando la correzione per popolazioni finite si ottiene nei due casi rispettivamente:

$$n = \frac{2.436 \cdot 200.000}{2.436 + (200.000 - 1)} = 2.407$$

$$n = \frac{6.765 \cdot 200.000}{6.765 + (200.000 - 1)} = 6.544$$

Essendo la dimensione della popolazione molto grande rispetto a quella del campione l'effetto del fattore di correzione sulla riduzione della dimensione del campione è molto limitata.



## IV.1 DECISIONI STATISTICHE – Concetti generali

La teoria delle decisioni è la parte della statistica che si occupa di fornire informazioni sulle decisioni da prendere riguardo ad una o più popolazioni in relazione alle informazioni campionarie.

Tali decisioni sono basate sulla verifica di ipotesi statistiche relative ai parametri delle popolazioni.

L'ipotesi statistica relativa alle condizioni "normali", ad esempio che una moneta non sia truccata ( $p=0,5$ ) o che non ci sia differenza fra la media nei risultati di due differenti trattamenti ( $\mu_{x1}=\mu_{x2}$ ), viene detta ipotesi nulla ed indicata con  $H_0$ .

L'ipotesi alternativa, indicata con  $H_1$ , esprime una condizione diversa da quella "normale"; nel caso della moneta sarà  $p \neq 0,5$  e nel caso del confronto fra trattamenti  $\mu_{x1} \neq \mu_{x2}$ .

Se i risultati osservati sui campioni estratti dalle popolazioni non differiscono in misura "significativa" da quelli attesi in relazione all'ipotesi nulla formulata, l'ipotesi nulla  $H_0$  viene "accettata" in caso contrario deve essere "rifiutata".

## IV.1 DECISIONI STATISTICHE – Concetti generali

I procedimenti che consentono di decidere se accettare o rifiutare un'ipotesi sono detti test delle ipotesi o test di significatività e si basano:

- sulla determinazione della distribuzione di una statistica-test;
- sull'individuazione del valore critico della statistica-test che separa la regione di accettazione dalla regione di rifiuto dell'ipotesi;
- sul confronto fra valore osservato e valore critico della statistica-test.

Nel prendere decisioni su una popolazione partendo dai dati campionari è possibile commettere degli errori:

- si verifica un errore di I tipo se l'ipotesi nulla  $H_0$  è vera ma viene rifiutata.
- si verifica un errore di II tipo se l'ipotesi nulla  $H_0$  è falsa ma viene accettata.

La probabilità di commettere un errore di I tipo viene indicata con  $\alpha$ ,  
la probabilità di commettere un errore di II tipo viene indicata con  $\beta$ .

## IV.1 DECISIONI STATISTICHE – Concetti generali

Il valore di  $\alpha$  (detto **livello di significatività**) esprime il livello massimo di rischio di rifiutare una ipotesi vera che si è disposti a tollerare; questo viene generalmente posto pari a 0,05 o a 0,01 in modo da essere confidenti al livello  $(1-\alpha)$ , quindi del 95% o del 99%, di non aver rifiutato una ipotesi vera.

Il valore  $(1-\beta)$  esprime la probabilità di rifiutare l'ipotesi nulla quando questa è falsa e viene chiamata **potenza** di un test statistico.

Le due precedenti definizioni possono essere riepilogate nel seguente prospetto:

	<i>Situazione reale</i>	
<i>Decisione</i>	<b>H<sub>0</sub> vera</b>	<b>H<sub>0</sub> falsa</b>
<b>Accettare H<sub>0</sub></b>	<b>Confidenza (1-<math>\alpha</math>)</b>	<b>Errore di II tipo (<math>\beta</math>)</b>
<b>Rifiutare H<sub>0</sub></b>	<b>Errore di I tipo (<math>\alpha</math>)</b>	<b>Potenza (1-<math>\beta</math>)</b>

## IV.1 DECISIONI STATISTICHE – Concetti generali

Nei procedimenti decisionali condotti attraverso i test delle ipotesi si deve sempre cercare la minimizzazione contemporanea dei due errori  $\alpha$  e  $\beta$ .

Ciò non è facile in quanto, una volta stabilita la dimensione del campione, la riduzione di  $\alpha$  fa aumentare il valore di  $\beta$  e la riduzione di  $\beta$  fa aumentare il valore di  $\alpha$ .

Di conseguenza è necessario valutare le conseguenze pratiche di entrambi i tipi di errore per scegliere quale fra  $\alpha$  e  $\beta$  deve essere maggiormente ridotto.

L'errore di I tipo, essendo proprio uguale ad  $\alpha$ , può essere ridotto diminuendo il livello di significatività.

L'errore di II tipo può essere evitato del tutto non accettando mai l'ipotesi nulla, ma affermando semplicemente che non ci sono elementi sufficienti per respingerla; questo comportamento, tuttavia, non è sempre possibile.

Quando è necessario valutare l'entità dell'errore di II tipo si utilizzano delle "curve di potenza" la cui forma dipende dal tipo di test (a 1 o 2 code), dal livello di significatività  $\alpha$  e dalla dimensione  $n$  del campione.

## IV.1 DECISIONI STATISTICHE – Concetti generali

La formulazione di una decisione statistica basata sul procedimento della verifica delle ipotesi prevede lo svolgimento dei seguenti passi:

- 1) stabilire l'ipotesi nulla  $H_0$  e l'ipotesi alternativa  $H_1$ ;
- 2) specificare il livello di significatività  $\alpha$ ;
- 3) determinare la dimensione  $n$  del campione;
- 4) stabilire la statistica-test;
- 5) determinare in base al valore  $\alpha$  i valori critici della statistica test che separano le regioni di accettazione e di rifiuto dell'ipotesi nulla;
- 6) calcolare il valore campionario (osservato) della statistica test;
- 7) verificare se il valore campionario della statistica test cade nella regione di accettazione o di rifiuto dell'ipotesi nulla;
- 8) formulare la decisione statistica.

### Esempio

*Stabilire un procedimento per verificare che una moneta non sia "truccata".*

- 1) Si stabilisce l'ipotesi nulla  $H_0: p=0,5$  e l'ipotesi alternativa  $H_1: p \neq 0,5$ .
- 2) Si specifica il livello di significatività  $\alpha=0,05$ .
- 3) Si determina la dimensione del campione  $n=64$  lanci.
- 4) Si stabilisce la statistica-test; poiché  $np \geq 5$  e  $n(1-p) \geq 5$  è possibile utilizzare la distribuzione normale per approssimare la distribuzione binomiale:
- 5) Si calcolano i valori critici della distribuzione normale per  $\alpha=0,05$

$$z_c = \pm z_{0,475} = \pm 1,96$$

- 6) Si lancia la moneta **64** volte e si osserva, ad esempio, che è uscita testa per **27** volte. Il valore standardizzato di tale risultato è:

$$z = \frac{(x - np)}{\sqrt{np(1-p)}} = \frac{27 - 32}{\sqrt{16}} = 1,25$$

- 7) Il valore osservato della statistica test ( $z=1,25$ ) cade nella regione di accettazione dell'ipotesi.
- 8) La decisione è di accettare l'ipotesi nulla  $H_0$  vale a dire che la moneta non è truccata.

Il procedimento può essere generalizzato de-standardizzando i valori critici della distribuzione normale; ciò consente di affermare che se in 64 lanci di una moneta esce testa (o croce) meno di 25 volte o più di 39 volte si hanno buone ragioni (più del 95% di probabilità) per ritenere che la moneta non sia buona.

In questo tipo di procedimenti è possibile evitare di incorrere in un errore del II tipo semplicemente affermando in 8) che "non ci sono elementi sufficienti per ritenere che la moneta non sia buona".

## IV.1 DECISIONI STATISTICHE – Test a due code

Quando la verifica delle ipotesi parte da un'ipotesi alternativa  $H_1$  in cui vengono considerate differenze sia positive che negative fra statistiche e parametri, la regione di rifiuto si trova in entrambe le code (destra e sinistra) della distribuzione della statistica-test.

In questi casi vengono eseguiti dei test delle ipotesi bilaterali o **a due code**.

Questa circostanza si verifica, ad esempio, quando sono noti i parametri di una popolazione e si desidera verificare se un determinato campione appartiene a tale popolazione.

Un'applicazione di questa situazione si ha tutte le volte che si vuole controllare la regolarità di un processo di produzione di cui sono note le specifiche (parametri) dei prodotti (media e scarto quadratico medio) attraverso il calcolo delle statistiche di un campione di prodotti prelevati dalla linea di produzione.

### Esempio

*Una macchina per la lavorazione delle arance seleziona i frutti in base al loro diametro. In condizioni normali il diametro delle arance selezionate ha media 95 mm e scarto quadratico medio 4 mm. Durante un controllo di qualità si verifica che 40 arance selezionate dalla macchina hanno un diametro medio di 93,5 mm. Ci sono elementi per ritenere che la macchina non stia funzionando regolarmente?*

L'ipotesi nulla ( $H_0$ ) è che la macchina funzioni regolarmente, cioè che il diametro medio sia di 95 mm, mentre quella alternativa ( $H_1$ ) è che la macchina non funzioni regolarmente e quindi il diametro medio delle arance selezionate sia diverso da 95 mm.

1)  $H_0: \mu_X=95\text{mm}$  e  $H_1: \mu_X \neq 95\text{mm}$

2) Si sceglie un livello di significatività, ad esempio  $\alpha=0,01$

3) Si considera la dimensione del campione  $n=40$

4) Essendo noto  $\sigma_X=4\text{mm}$  ed essendo  $n>30$ , la statistica test è la distribuzione normale:

5) I valori critici della statistica test in corrispondenza di  $\alpha=0,01$  (test a due code) sono:

$$z_c = \pm z_{0,495} = \pm 2,58$$

6) Il valore della statistica test in corrispondenza del valore campionario  $\bar{x}=93,5\text{mm}$  risulta:

$$z = \frac{(\bar{x} - \mu_X)}{\sigma_X / \sqrt{n}} = \frac{93,5 - 95}{4 / \sqrt{40}} = -2,37$$

7) Essendo  $-2,37 > -2,58$  la statistica campionaria cade nella regione di accettazione di  $H_0$ .

8) Non ci sono elementi sufficienti per ritenere che la macchina non funzioni regolarmente.

## IV.1 DECISIONI STATISTICHE – Test a una coda

Quando si è interessati ad eseguire una verifica delle ipotesi rispetto ad una sola parte della distribuzione si è in presenza di un test unilaterale o a una coda.

In questo caso si modifica il valore assunto dalla statistica test per il livello di significatività scelto in quanto l'area corrispondente a tale livello di significatività non sarà ripartita fra le due code ma sarà tutta compresa al di sotto di una sola delle code della distribuzione della statistica test.

Ad esempio, considerando come statistica test la distribuzione normale, in un test a due code con un livello di significatività  $\alpha=0,05$  i valori critici sono  $z_{0,475}=\pm 1,96$ ; se invece il test è a una coda si ha, sempre in corrispondenza di  $\alpha=0,05$ ,  $z_{0,45}=+1,645$  (coda destra) o  $z_{0,45}=-1,645$  (coda sinistra).

Come nel caso del test a due code, un'applicazione di questa situazione si ha tutte le volte che si vuole controllare la regolarità di un processo di produzione; nel test a una coda, però, viene verificata la situazione in cui il dato campionario è o solo minore o solo maggiore del dato relativo alla popolazione.

### Esempio

*Si ripeta la valutazione dell'esercizio precedente considerando che la macchina non funziona regolarmente se vengono selezionate arance di diametro inferiore a quello medio previsto.*

1) In questo caso l'ipotesi nulla e l'ipotesi alternativa divengono:  $H_0: \mu_x \geq 95\text{mm}$  e  $H_1: \mu_x < 95\text{mm}$

2)  $\alpha=0,01$

3)  $n=40$

4) La statistica test è ancora la distribuzione normale

5) Il valore critico della distribuzione (relativo alla sola coda sinistra) è in questo caso:

$$z_c = -z_{0,490} = -2,33$$

6) Il valore campionario della statistica  $\bar{x} = 93,5\text{mm}$  espresso in termini standardizzati è ancora

$$z = \frac{(\bar{x} - \mu_x)}{\sigma_x / \sqrt{n}} = -2,37$$

7) La statistica campionaria cade nella regione di rifiuto di  $H_0$

8) Ci sono elementi per ritenere che la macchina non funzioni regolarmente in quanto, con il 99% di probabilità, seleziona arance di diametro inferiore a quello previsto.

## IV.1 DECISIONI STATISTICHE – Valore-p

Invece di verificare l'appartenenza della statistica campionaria alla regione di accettazione o di rifiuto dell'ipotesi nulla, è possibile determinare il livello di significatività osservato (valore-p) della statistica-test.

Il **valore-p** rappresenta quindi il minimo livello di significatività con cui l'ipotesi nulla  $H_0$  può essere rifiutata.

In un test ad una coda:

- se **valore-p** >  $\alpha$ ,  $H_0$  viene accettata
- se **valore-p** <  $\alpha$ ,  $H_0$  viene rifiutata

Per i test a una coda il livello di confidenza è  $(1 - \alpha)$ , in questo caso la differenza fra la statistica campionaria e il parametro della popolazione viene ritenuta:

- non significativa: **valore-p** > 0,10      confidenza < 90%
- scarsamente significativa:  $0,05 < \text{valore-p} < 0,10$       confidenza fra 90% e 95%
- probabilmente significativa:  $0,01 < \text{valore-p} < 0,05$       confidenza fra 95% e 99%
- altamente significativa: **valore-p** < 0,01      confidenza > 99%

### Esempio

*Con riferimento all'esempio precedente calcolare il valore-p.*

Il **valore-p** corrisponde al valore dell'area della coda della distribuzione normale corrispondente alla standardizzazione del valore osservato.

Nel caso in cui l'ipotesi da verificare sia  $H_0: \mu_x \geq 95\text{mm}$  e  $H_1: \mu_x < 95\text{mm}$  il test è a una coda.

La probabilità osservata, in questo caso pari all'area della coda sinistra, risulta:

$$\text{valore-p} = 0,5 - F(Z) = 0,5 - F\left(\frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}\right) = 0,5 - F\left(\frac{93,5 - 95}{4/\sqrt{40}}\right) = 0,5 - F(-2,37) = 0,5 - 0,4911 = 0,0089$$

Risultando il **valore-p** < 0,01 è possibile rifiutare l'ipotesi nulla con un livello di confidenza superiore al 99% (precisamente del 99,11%) e quindi ritenere la differenza fra media del campione e della popolazione altamente significativa.

Nel caso in cui l'ipotesi da verificare sia  $H_0: \mu_x = 95\text{mm}$  e  $H_1: \mu_x \neq 95\text{mm}$  il test è a due code.

Per avere una confidenza del 99% il livello di significatività deve essere di  $\alpha = 0,005$ , in quanto l'area di 0,01 si divide nelle due code. Analogamente per avere un livello di confidenza del 95% l'area nelle code di 0,05 dà origine ad un livello di significatività  $\alpha = 0,025$ .

Risultando il **valore-p** = 0,0089 compreso fra 0,005 e 0,025 il livello di confidenza è compreso fra il 95% ed il 99% (precisamente è del 98,22%) e quindi la differenza fra media del campione e della popolazione risulta probabilmente significativa.

## IV.1 DECISIONI STATISTICHE – Test di uguaglianza

Le procedure di verifica delle ipotesi vengono utilizzate anche per stabilire se i parametri di diverse popolazioni differiscono in modo significativo in base ai valori assunti dalle statistiche dei relativi campioni.

Le verifiche più importanti riguardano i test di uguaglianza delle medie e delle frequenze relative di due o più popolazioni. Tali verifiche si presentano secondo diverse modalità e, di conseguenza, utilizzano diverse statistiche test:

### **uguaglianza medie** (variabili quantitative)

- due popolazioni, grandi campioni → distribuzione **normale**
- due popolazioni, piccoli campioni → distribuzione **t di Student**
- più di due popolazioni – ANOVA → distribuzione **F di Fisher**

### **uguaglianza frequenze relative** (variabili qualitative)

- due popolazioni, grandi campioni → distribuzione **normale**
- due popolazioni, piccoli campioni o più di due popolazioni → distribuzione  $\chi^2$

## IV.2 DECISIONI STATISTICHE – Uguaglianza medie

Si considerino due campioni di dimensione  $n_1$  e  $n_2$  con medie  $\bar{x}_1$  e  $\bar{x}_2$  estratti da due popolazioni diverse aventi medie  $\mu_{x1}$  e  $\mu_{x2}$  e scarti quadratici medi  $\sigma_{x1}$  e  $\sigma_{x2}$

Quando  $n_1 > 30$  e  $n_2 > 30$  le variabili casuali  $\frac{(\bar{x}_1 - \mu_{x1})}{\sigma_{x1} / \sqrt{n_1}}$  e  $\frac{(\bar{x}_2 - \mu_{x2})}{\sigma_{x2} / \sqrt{n_2}}$

seguono con buona approssimazione la distribuzione normale.

Applicando il teorema del limite centrale si dimostra che la distribuzione della differenza delle medie campionarie  $(\bar{x}_1 - \bar{x}_2)$  segue una distribuzione normale con media  $(\mu_{x1} - \mu_{x2})$  e scarto quadratico medio  $\sqrt{\sigma_{x1}^2 / n_1 + \sigma_{x2}^2 / n_2}$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{x1} - \mu_{x2})}{\sqrt{\sigma_{x1}^2 / n_1 + \sigma_{x2}^2 / n_2}}$$

Se si vuole testare l'ipotesi dell'uguaglianza delle due popolazioni l'ipotesi nulla sarà  $H_0: \mu_{x1} = \mu_{x2}$  e quindi la statistica test diviene:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\sigma_{x1}^2 / n_1 + \sigma_{x2}^2 / n_2}}$$

### Esempio

In due città A e B viene condotta una rilevazione per stabilire la spesa media dei cittadini per l'acquisto di prodotti biologici. Nella città A su un campione di 45 intervistati si è registrata una spesa media di 305€ e scarto quadratico medio 120€, nella città B su un campione di 60 intervistati la spesa media è risultata di 352€ e lo scarto quadratico di 90€. Determinare se esiste una differenza fra i cittadini delle due città nell'acquisto di prodotti biologici.

- 1) L'ipotesi nulla è rappresentata dall'uguaglianza delle due medie  $H_0: \mu_{x1} = \mu_{x2}$  e  $H_1: \mu_{x1} \neq \mu_{x2}$
- 2) Il livello di significatività viene posto  $\alpha = 0,01$
- 3) Le dimensioni dei due campioni sono  $n_1 = 45$  e  $n_2 = 60$
- 4) La statistica test è la distribuzione normale
- 5) I valori critici della distribuzione (test a due code) sono  $z_C = \pm 2,58$
- 6) Il valore campionario della statistica viene calcolato stimando gli scarti quadratici medi delle due popolazioni con i relativi scarti quadratici medi dei campioni ( $\sigma_{x1} = s_1 = 120$  e  $\sigma_{x2} = s_2 = 90$ ):

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\sigma_{x1}^2 / n_1 + \sigma_{x2}^2 / n_2}} = \frac{305 - 352}{\sqrt{(120)^2 / 45 + (90)^2 / 60}} = \frac{-47}{21,33} = -2,2$$

- 7) La statistica campionaria cade nella regione di accettazione di  $H_0$
- 8) Non ci sono elementi per ritenere che esista una differenza nel comportamento dei cittadini delle città A e B rispetto all'acquisto di prodotti biologici.

Da notare che per  $\alpha = 0,05$  i valori critici della distribuzione sono  $z_C = \pm 1,96$  e che, quindi, la statistica campionaria sarebbe caduta nella regione di rifiuto di  $H_0$ .

Se ne conclude che la differenza fra le statistiche campionarie è probabilmente significativa.

## IV.2 DECISIONI STATISTICHE – Uguaglianza medie

Se le dimensioni dei campioni  $n_1$  e  $n_2$  sono inferiori a 30 l'ipotesi di uguaglianza delle medie delle popolazioni deve essere condotta utilizzando come statistica-test la **t di Student** con un numero di gradi di libertà pari a  $(n_1+n_2-2)$ .

L'ipotesi che si intende verificare, cioè che i campioni appartengano alla stessa popolazione ( $H_0: \mu_{X1} = \mu_{X2}$ ), consente di considerare uguali gli scarti quadratici medi ( $s_1 = s_2 = s_p$ ).

La statistica-test per l'uguaglianza delle medie di due popolazioni nel caso di "piccoli campioni" è pertanto:

$$t_{(n_1+n_2-2)} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{1/n_1 + 1/n_2}}$$

Lo scarto quadratico medio dei due campioni combinati  $s_p$  (che rappresenta una stima dello scarto quadratico medio della popolazione) è calcolato come segue:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

### Esempio

*In un campo sperimentale diviso in 24 parcelle si vogliono studiare gli effetti di un fertilizzante sulla produzione di grano. Nelle 14 particelle fertilizzate la produzione media è stata di 51 q/ha e lo scarto quadratico medio di 3,6 q/ha, nelle altre (controllo) la produzione media è stata di 48 q/ha e lo scarto quadratico medio di 4 q/ha. Si può affermare con una confidenza del 99% che il fertilizzante abbia apportato un significativo miglioramento nella produzione di grano?*

- 1) L'ipotesi nulla è che la produzione di grano nelle particelle fertilizzate (popolazione 1) sia uguale o inferiore al controllo (popolazione 2), per cui si pone  $H_0: \mu_{X1} \leq \mu_{X2}$  e  $H_1: \mu_{X1} > \mu_{X2}$
- 2) Il livello di significatività è fissato in  $\alpha=0,01$
- 3) Le dimensioni dei due campioni sono  $n_1=14$  e  $n_2=10$
- 4) Essendo in presenza di piccoli campioni la statistica test è la **t di Student**
- 5) Il valore critico della distribuzione per  $(n_1+n_2-2)=22$  gradi di libertà è (test a una coda)  $t_c=2,51$
- 6) Per calcolare il valore campionario della statistica test è prima necessario stimare lo scarto quadratico medio comune fra i due campioni ( $s_p$ ):

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{13 \cdot 12,96 + 9 \cdot 16}{22}} = 3,78$$
$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{1/n_1 + 1/n_2}} = \frac{51 - 48}{3,78 \cdot \sqrt{1/14 + 1/10}} = 1,92$$

- 7) Essendo  $1,92 < 2,51$  la statistica campionaria cade nella regione di accettazione dell'ipotesi nulla.
- 8) Non ci sono elementi sufficienti per affermare con il 99% di confidenza che il fertilizzante abbia apportato un miglioramento nella produzione di grano.

## IV.2 DECISIONI STATISTICHE – Uguaglianza medie (ANOVA)

Nel caso in cui debba essere verificata l'uguaglianza delle medie di più di due popolazioni viene utilizzato un procedimento noto come **analisi della varianza (ANOVA)** che utilizza come statistica test la distribuzione **F di Fisher**.

In questo caso si è in presenza di un numero  $c$  di popolazioni ( $c > 2$ ) e l'ipotesi nulla che si vuole testare è  $H_0: \mu_{x_1} = \mu_{x_2} = \dots = \mu_{x_c}$  mentre l'ipotesi alternativa è che almeno una media sia diversa dalle altre.

A questo scopo viene definito un livello di significatività  $\alpha$  e vengono estratti dalle  $c$  popolazioni dei campioni di dimensione  $n_1, n_2, \dots, n_c$  per un numero complessivo di osservazioni pari a  $n$ .

$$\sum_{j=1}^c n_j = n$$

Le medie dei  $c$  campioni vengono indicate con  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_c$

La media complessiva, detta  $\bar{x}$ , la generica osservazione  $i$  del campione  $j$ , risulta:

$$\bar{x} = \frac{\sum_{j=1}^c \sum_{i=1}^{n_j} x_{ij}}{n}$$

### Esempio

*In un parco sono stati realizzati tre percorsi: uno sportivo (A), uno botanico (B) e uno per l'avvistamento di animali (C). L'ente gestore del parco è interessato a verificare se esiste una diversa preferenza da parte dei visitatori per i tre percorsi. Considerando le presenze registrate nei tre percorsi in differenti date e riportate nel prospetto seguente, ci sono elementi sufficienti per ritenere con una confidenza del 99% che i tre percorsi abbiano un diverso numero di visitatori?*

A	B	C
18	11	26
25	16	23
31	13	32
20	25	33
	13	

L'ipotesi nulla da cui si parte è che non ci siano differenze nel numero di visitatori dei  $c=3$  percorsi del parco si ha pertanto:

1)  $H_0: \mu_{x_1} = \mu_{x_2} = \mu_{x_3}$

2) Il livello di significatività è  $\alpha=0,01$

3) La dimensione dei campioni risulta  $n_1=4$ ;  $n_2=5$ ;  $n_3=4$  e in totale  $n=13$ ;

4) Considerando che si sta verificando l'indipendenza di tre popolazioni rispetto ad una variabile quantitativa dovrà essere usato un test basato sulla distribuzione **F** di Fisher e, di conseguenza, andrà eseguita un'analisi della varianza (ANOVA).

## IV.2 DECISIONI STATISTICHE – Uguaglianza medie (ANOVA)

La variazione totale delle osservazioni rispetto alla media complessiva  $\bar{x}$  risulta (Sum of Squares Total):

$$SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

Tale variazione viene scomposta secondo due termini: la variazione tra i gruppi (dovuta alla differenza fra le medie dei gruppi e la media complessiva - Sum of Squares Between) e la variazione all'interno dei gruppi (dovuta alla differenza fra le osservazioni di un gruppo e la relativa media – Sum of Squares Within).

La variazione fra i gruppi è data da:  $SSB = \sum_{j=1}^c n_j (\bar{x}_j - \bar{x})^2$

La variazione all'interno dei gruppi è data da:  $SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$

Per quanto detto risulterà che:

$$SST = SSB + SSW$$

Con riferimento all'esempio di pagina precedente andrà condotta un'analisi della varianza per stabilire il valore campionario della statistica test.

5) Le medie campionarie dei  $c=3$  gruppi sono rispettivamente:  $\bar{x}_1 = 23,5$   $\bar{x}_2 = 15,6$   $\bar{x}_3 = 28,5$

Mentre la media totale risulta pari a  $\bar{x} = 22$

La variazione totale delle osservazioni rispetto alla media totale (**SST**), la variazione fra i gruppi (**SSB**) e la variazione all'interno dei gruppi (**SSW**) hanno il seguente valore:

$$SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = (18 - 22)^2 + (25 - 22)^2 + \dots + (33 - 22)^2 = 676,0$$

$$SSB = \sum_{j=1}^c n_j (\bar{x}_j - \bar{x})^2 = 4 \cdot (23,5 - 22)^2 + 5 \cdot (15,6 - 22)^2 + 4 \cdot (28,5 - 22)^2 = 382,8$$

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = (18 - 23,5)^2 + (25 - 23,5)^2 + \dots + (11 - 15,6)^2 + \dots + (33 - 28,5)^2 = 293,2$$

Come previsto risulta  $SSB + SSW = SST$  essendo  $382,8 + 293,2 = 676$

## IV.2 DECISIONI STATISTICHE – Uguaglianza medie (ANOVA)

Dividendo la variazione fra i gruppi (**SSB**) e la variazione all'interno dei gruppi (**SSW**) per i relativi gradi di libertà si ottengono le relative varianze.

I gradi di libertà fra i gruppi sono **c-1** e i gradi di libertà all'interno dei gruppi sono **n-c**, le varianze fra i gruppi ( $s_B^2$ ) e all'interno dei gruppi ( $s_W^2$ ) risultano:

$$s_B^2 = \frac{SSB}{c-1} \qquad s_W^2 = \frac{SSW}{n-c}$$

Il rapporto fra le due varianze segue una distribuzione **F** di Fisher; di conseguenza tale distribuzione diviene la statistica test per questo tipo di verifica delle ipotesi.

$$F_{(c-1),(n-c)} = \frac{s_B^2}{s_W^2}$$

Tale distribuzione dipende dai gradi di libertà del numeratore (**c-1**) e del denominatore (**n-c**). I valori critici della distribuzione **F** in corrispondenza dei principali livelli di significatività possono essere ricavati da apposite tabelle.

Proseguendo l'esemplificazione si osserva che i gradi di libertà fra i gruppi sono  $(c-1)=(3-1)=2$ , mentre i gradi di libertà all'interno dei gruppi sono  $(n-c)=(13-3)=10$ .

Con riferimento a tali gradi di libertà è possibile determinare le varianze fra i gruppi e all'interno dei gruppi che valgono rispettivamente:

$$s_B^2 = \frac{SSB}{c-1} = \frac{382,8}{2} = 191,4 \qquad s_W^2 = \frac{SSW}{n-c} = \frac{293,2}{10} = 29,32$$

Il rapporto fra le due varianze rappresenta il valore campionario (osservato) della statistica-test

$$F_{(c-1),(n-c)} = \frac{s_B^2}{s_W^2} = \frac{191,4}{29,32} = 6,53$$

6) Il valore critico della statistica test **F** di Fisher per 2 gradi di libertà al numeratore e 10 gradi di libertà al denominatore e per un livello di significatività  $\alpha=0,01$  viene letto sulle tabelle e risulta pari a

$$F_{0,01;2,10} = 7,56$$

7) Il valore campionario della statistica test (6,53) risulta minore del valore critico (7,56) e quindi cade nella regione di accettazione dell'ipotesi nulla.

8) Si può concludere che non ci sono elementi per ritenere che esista una differenza significativa (al livello 0,01) fra il numero di visitatori del parco che percorrono i tre diversi sentieri.

## IV.2 DECISIONI STATISTICHE – Uguaglianza medie (ANOVA)

La verifica delle ipotesi viene eseguita leggendo il valore critico della distribuzione **F** per i due gradi di libertà e per il livello di significatività  $\alpha$  scelto.

Se il valore della statistica test risulta maggiore di tale valore critico

$$\frac{s_B^2}{s_W^2} > F_{\alpha; (c-1), (n-c)}$$

significa che la statistica campionaria cade nella regione di rifiuto di  $H_0$  e l'ipotesi nulla deve essere rigettata.

L'analisi della varianza eseguita come illustrato si basa su due ipotesi:

- le **c** popolazioni seguono la distribuzione normale (o quasi);
- le varianze delle **c** popolazioni sono uguali.

La seconda ipotesi è vincolante per l'applicazione del metodo per cui, quando esiste la possibilità che non sia verificata, è necessario un test specifico per essere certi che le varianze dei gruppi non siano significativamente diverse.

### IV.3 DECISIONI STATISTICHE – Uguaglianza di frequenze

L'ipotesi dell'uguaglianza delle frequenze relative di due popolazioni

$$H_0: p_1=p_2 \text{ e } H_1: p_1 \neq p_2$$

può essere verificata nel caso di grandi campioni ( $n > 30$ ) utilizzando come statistica test la distribuzione normale

Detti  $n_{A1}$  il numero di esiti positivi del primo campione di dimensione  $n_1$  e  $n_{A2}$  il numero di esiti positivi del secondo campione di dimensione  $n_2$ , le frequenze relative dei due campioni sono rispettivamente:  $p_{s1} = n_{A1}/n_1$  e  $p_{s2} = n_{A2}/n_2$

La frequenza relativa delle due popolazioni (che per ipotesi è la stessa) viene stimata considerando entrambi i campioni ed è pari al rapporto fra tutti gli esiti positivi ( $n_{A1} + n_{A2}$ ) e la dimensione complessiva dei campioni ( $n_1 + n_2$ ):

$$\bar{p} = \frac{n_{A1} + n_{A2}}{n_1 + n_2}$$

Il valore osservato della statistica test risulta in questo caso:

$$Z = \frac{(p_{s1} - p_{s2})}{\sqrt{\bar{p}(1 - \bar{p})(1/n_1 + 1/n_2)}}$$

#### Esempio

*Durante un sondaggio elettorale per stabilire il gradimento di un candidato vengono intervistati 300 elettori e 200 elettrici; 168 elettori e 96 elettrici esprimono la preferenza per il candidato. Usando un livello di significatività di 0,05 stabilire:*

- se esiste una differenza nella preferenza per il candidato fra elettori ed elettrici;
- se il candidato è preferito dagli elettori maschi.

L'ipotesi nulla è definita dalla uguaglianza delle preferenze per il candidato fra la popolazione degli elettori ( $p_1$ ) e delle elettrici ( $p_2$ ), cioè  $H_0: p_1 = p_2$ .

Dai dati forniti  $n_1=300$ ,  $n_2=200$ ,  $n_{A1}=168$  e  $n_{A2}=96$  si ricavano le frequenze relative dei campioni e la stima della frequenza relativa della popolazione:  $p_{s1}=0,56$ ;  $p_{s2}=0,48$ ;  $\bar{p}=0,528$

Il valore critico della distribuzione normale per  $\alpha=0,05$  (test a due code) è  $z_{0,475}=\pm 1,96$ .

Il valore campionario della statistica-test ha il seguente valore

$$Z = \frac{(p_{s1} - p_{s2})}{\sqrt{\bar{p}(1 - \bar{p})(1/n_1 + 1/n_2)}} = \frac{0,56 - 0,48}{\sqrt{0,528 \cdot 0,472 \cdot 0,0083}} = 1,76$$

Essendo il valore della statistica-test nella regione di accettazione dell'ipotesi nulla non è possibile concludere che ci sono differenze fra elettori maschi e femmine al livello di significatività richiesto.

Il secondo tipo di decisione richiede un test ad una coda il cui valore critico è  $z_{0,45}=1,645$ .

In questo caso la statistica-test (1,76) si trova nella regione di rifiuto dell'ipotesi nulla e pertanto si può concludere, al livello di significatività specificato, che il candidato è preferito dagli elettori maschi.

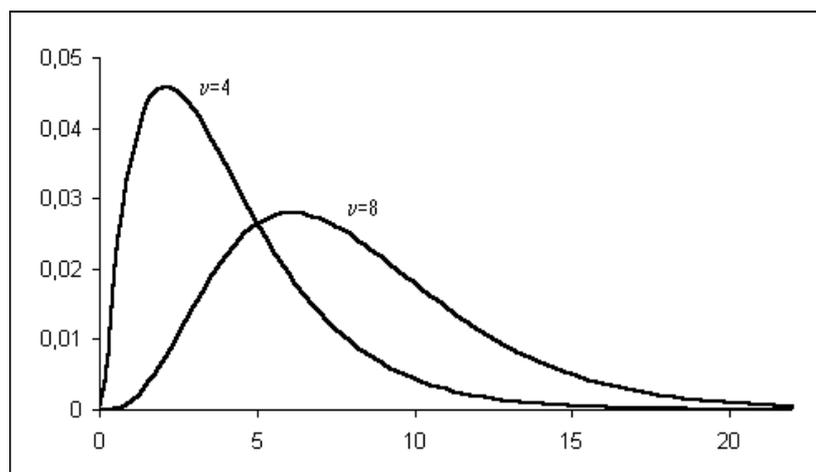
### IV.3 DECISIONI STATISTICHE – Uguaglianza frequenze

Nel caso più generale la verifica dell'uguaglianza delle frequenze relative di due popolazioni basata su rilevazioni campionarie viene condotta con un metodo che utilizza come statistica test la distribuzione  $\chi^2$  (chi-quadrato).

Il valore campionario della statistica test viene determinato per mezzo della costruzione di un prospetto (**tabella di contingenza**) che raccoglie il numero di esiti (positivi e negativi) riscontrati in ciascuno dei due campioni estratti dalle relative popolazioni.

<i>Esiti</i>	<i>Campione</i>		
	<b>Campione 1</b>	<b>Campione 2</b>	<b>Totale (1+2)</b>
<b>Esito A</b>	$n_{A1}$	$n_{A2}$	$n_A$
<b>Esito B</b>	$n_{B1}$	$n_{B2}$	$n_B$
<b>Totale (A+B)</b>	$n_1$	$n_2$	$n$

La distribuzione  $\chi^2$  presenta l'andamento mostrato nella figura seguente il quale, come si osserva, dipende unicamente dal numero dei gradi di libertà  $v$ .



La distribuzione è definita soltanto per valori positivi e, pertanto, i test statistici vengono condotti facendo riferimento alla sola coda (destra) della distribuzione stessa.

I valori del chi-quadrato corrispondenti ad un'area  $\alpha$  sottesa dalla coda superiore per diversi gradi di libertà sono riportati in un'apposita tavola statistica (vedi appendice).

### IV.3 DECISIONI STATISTICHE – Uguaglianza frequenze

Per verificare la significatività della differenza fra le frequenze relative delle due popolazioni viene utilizzata una misura dello scostamento delle frequenze assolute osservate da quelle teoriche, cioè quelle che ci si dovrebbero attendere nel caso in cui fosse verificata l'ipotesi nulla ( $H_0: p_1=p_2$ ) e quindi non ci fosse alcuna differenza fra le popolazioni.

Le frequenze teoriche, calcolate considerando le dimensioni dei campioni ed il numero totale degli esiti, vengono raccolte in un prospetto di struttura analoga alla tabella di contingenza delle frequenze assolute osservate:

Esiti	Campione		
	Campione 1	Campione 2	Totale
<b>Esito A</b>	$\frac{n_A}{n} n_1$	$\frac{n_A}{n} n_2$	$n_A$
<b>Esito B</b>	$\frac{n_B}{n} n_1$	$\frac{n_B}{n} n_2$	$n_B$
<b>Totale</b>	$n_1$	$n_2$	$n$

#### Esempio

Risolvere l'esempio precedente relativo al sondaggio elettorale utilizzando come statistica test la distribuzione chi-quadrato.

La matrice di contingenza delle frequenze osservate costruita in base ai risultati del sondaggio elettorale è la seguente:

	Maschi	Femmine	Totale
Favorevoli	168	96	264
Contrari	132	104	236
Totale	300	200	500

Per calcolare il valore campionario della statistica test è necessario costruire la tabella delle frequenze teoriche, nell'ipotesi in cui non ci sia differenza nella preferenza per il candidato fra le elettrici e gli elettori ( $H_0: p_1=p_2$ ).

Il numero teorico dei maschi favorevoli sarà dato dalla proporzione:

$$\text{maschi favorevoli} : \text{maschi totali } (n_1) = \text{intervistati favorevoli } (n_A) : \text{intervistati totali } (n)$$

Risolvendo la proporzione si ottiene il numero teorico dei maschi favorevoli  $\frac{n_A}{n} n_1 = \frac{264}{500} 300 = 158,4$

Determinando in modo analogo le altre frequenze teoriche si ottiene la seguente tabella:

	Maschi	Femmine	Totale
Favorevoli	158,4	105,6	264
Contrari	141,6	94,4	236
Totale	300	200	500

### IV.3 DECISIONI STATISTICHE – Uguaglianza frequenze

La grandezza che rappresenta lo scostamento fra le frequenze osservate ( $\mathbf{n}_o$ ) e le corrispondenti frequenze teoriche ( $\mathbf{n}_t$ ) nell'insieme delle celle della matrice di contingenza ha la seguente espressione:

$$\sum_{\text{celle}} \frac{(\mathbf{n}_o - \mathbf{n}_t)^2}{\mathbf{n}_t}$$

Questa variabile segue la distribuzione  $\chi^2$  (chi-quadrato), il cui andamento dipende dai gradi di libertà della variabile.

Nel caso delle tabelle di contingenza il numero dei gradi di libertà è rappresentato dal numero di celle che possono essere assegnate liberamente una volta note le dimensioni dei campioni e gli esiti totali. In una tabella 2x2, cioè con 2 righe (**R**) e 2 colonne (**C**), soltanto il valore di una cella può essere attribuito liberamente; di conseguenza la statistica test da usare è  $\chi^2_1$ , cioè il chi-quadrato con un solo grado di libertà.

Il test di verifica dell'ipotesi nulla viene condotto individuando il valore critico della distribuzione  $\chi^2$  corrispondente al livello di significatività e controllando se il valore della statistica test calcolato dalla tabella di contingenza si trova nella regione di accettazione o di rifiuto.

Con riferimento all'esempio precedente la verifica dell'ipotesi nulla prosegue calcolando il valore campionario della statistica test rappresentato dalla misura della differenza fra le frequenze osservate e le frequenze teoriche.

Il valore della statistica-test è:

$$\sum_{\text{celle}} \frac{(\mathbf{n}_o - \mathbf{n}_t)^2}{\mathbf{n}_t} = \frac{(168 - 158,4)^2}{158,4} + \frac{(132 - 141,6)^2}{141,6} + \frac{(96 - 105,6)^2}{105,6} + \frac{(104 - 94,4)^2}{94,4} = 3,08$$

Tale valore deve essere confrontato con il valore critico della distribuzione chi-quadrato per  $\alpha=0,05$  e per un grado di libertà. Dalle tabelle della distribuzione si vede che il valore critico è in questo caso

$$\chi^2_{1,(0,05)} = 3,84$$

Essendo  $3,08 < 3,84$  la statistica campionaria cade nella regione di accettazione dell'ipotesi nulla.

Se ne conclude che non ci sono elementi per ritenere che ci sia una differenza significativa al 95% fra elettori ed elettrici nella preferenza verso il candidato.

### IV.3 DECISIONI STATISTICHE – Indipendenza variabili

Le due modalità per eseguire i test sulla differenza fra frequenze relative, con distribuzione normale e distribuzione chi-quadrato, forniscono sempre lo stesso risultato; ciò potrebbe far preferire come statistica-test la distribuzione normale, considerando anche il vantaggio di poter condurre test unilaterali.

Va considerato, però, che i test eseguiti tramite tabella di contingenza per l'uguaglianza delle frequenze relative possono essere considerati più in generale come test di indipendenza delle variabili presenti nelle righe e nelle colonne della matrice stessa.

Con questo approccio l'ipotesi  $H_0$  dell'esempio precedente diviene "non c'è alcun rapporto fra il sesso dell'elettore e la preferenza per il candidato".

Questa possibilità offerta dai test basati sulla distribuzione chi-quadrato viene utilizzata nel caso di più popolazioni ( $C > 2$ ), di più esiti ( $R > 2$ ) e, in generale, quando sono presenti più popolazioni con diversi possibili esiti ( $C > 2$  e  $R > 2$ ).

Nel caso generale di tabelle di dimensione  $R \times C$  il numero di gradi di libertà è pari a  $(R-1) \times (C-1)$  e la statistica-test di riferimento è  $\chi^2_{(R-1)(C-1)}$

#### Esempio

Ad un campione di 200 studenti viene chiesto di esprimere come "basso", "medio" e "alto" il livello di interesse per la statistica e di abilità in matematica ottenendo i seguenti risultati:

Interesse statistica	Abilità in matematica			Totale
	Basso	Medio	Alto	
Basso	60	15	15	90
Medio	15	45	10	70
Alto	5	10	25	40
Totale	80	70	50	200

E' possibile riscontrare un legame fra abilità in matematica e interesse per la statistica con un livello di significatività di 0,01?

La tabella relativa alle frequenze teoriche è la seguente:

Interesse statistica	Abilità in matematica			Totale
	Basso	Medio	Alto	
Basso	36	31,5	22,5	90
Medio	28	24,5	17,5	70
Alto	16	14	10	40
Totale	80	70	50	200

Lo scostamento fra frequenze osservate e teoriche risulta pari a  $\sum_{\text{celle}} \frac{(n_o - n_t)^2}{n_t} = 84,75$

La statistica-test è la distribuzione  $\chi^2$  con  $(R-1) \times (C-1) = 4$  gradi di libertà. Dalle tabelle della distribuzione chi-quadrato per  $\alpha = 0,01$  si ottiene un valore critico di  $\chi^2_{4,(0,01)} = 13,3$

Poiché il valore osservato della statistica-test cade nella regione di rifiuto dell'ipotesi nulla si può affermare che esiste un legame significativo (99%) fra abilità in matematica e interesse per la statistica.

## V.1 RELAZIONI FRA VARIABILI – Modello di regressione

L'analisi di regressione consiste nello sviluppo di un modello che lega i valori di una variabile dipendente con i valori di una o più variabili indipendenti.

Il prodotto del modello è una curva di regressione la cui forma generale è

$$y = f(x_1, x_2, \dots, x_n)$$

in cui  $y$  è la variabile dipendente e  $x_1, x_2, \dots, x_n$  sono le variabili indipendenti.

Quando la variabile indipendente è soltanto una ( $x$ ) e la relazione che la lega alla variabile dipendente ( $y$ ) è lineare, la curva di regressione è una retta

$$y = b_0 + b_1x$$

In questo caso si parla di regressione lineare semplice.

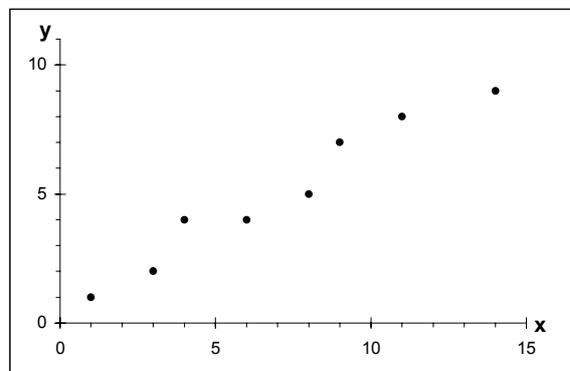
Il primo passo dell'analisi di regressione semplice consiste nel rappresentare in un diagramma a dispersione i valori assunti nelle diverse osservazioni dalle due variabili (indipendente e dipendente).

La semplice osservazione del diagramma a dispersione può consentire di evidenziare il tipo di legame esistente fra le due variabili.

**Esempio:** Costruire il diagramma a dispersione per le osservazioni del prospetto seguente

Osserv.	$x_i$	$y_i$
1	1	1
2	3	2
3	4	4
4	6	4
5	8	5
6	9	7
7	11	8
8	14	9

Il diagramma a dispersione per le osservazioni considerate è il seguente:



Il grafico evidenzia come la variabile dipendente  $y$  cresca al crescere della variabile indipendente  $x$  e che tale crescita segue un andamento di tipo tendenzialmente lineare.

## V.1 RELAZIONI FRA VARIABILI – Modello di regressione

Nel caso generale l'analisi di regressione lineare semplice viene condotta su un campione costituito da un numero finito di osservazioni a partire dalle quali vengono determinati i valori dell'intercetta  $\mathbf{b}_0$  e del coefficiente angolare  $\mathbf{b}_1$  della retta di regressione.

L'analisi di regressione lineare semplice si pone l'obiettivo di determinare la retta di regressione (cioè i coefficienti  $\mathbf{b}_0$  e  $\mathbf{b}_1$ ) in modo da minimizzare il complesso delle differenze fra i valori effettivi della variabile dipendente ( $y_i$ ) e i corrispondenti valori calcolati per mezzo del modello di regressione ( $\hat{y}_i$ )

$$\hat{y}_i = \mathbf{b}_0 + \mathbf{b}_1 x_i$$

Le differenze  $(y_i - \hat{y}_i)$  possono essere positive o negative; per evitare che tali differenze si compensino per ogni osservazione si considera l'errore  $(y_i - \hat{y}_i)^2$

L'errore totale da minimizzare risulta quindi:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (\mathbf{b}_0 + \mathbf{b}_1 x_i)]^2$$

Nel caso in cui la variabile dipendente sia legata alla variabile indipendente da una relazione di tipo lineare è valida la relazione:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

dove:

$\beta_0$  rappresenta la vera y-intercetta della popolazione, vale a dire la costante che rappresenta il valore di  $y$  quando  $x=0$ ;

$\beta_1$  rappresenta il vero coefficiente angolare della popolazione, vale a dire la variazione di  $y$  in corrispondenza di una variazione unitaria di  $x$ .

$\epsilon_i$  rappresenta l'errore casuale relativa alla  $y$  nella generica osservazione  $i$ .

Nel caso generale l'analisi di regressione viene condotta su un campione costituito da un numero finito  $n$  di osservazioni.

Dal campione è possibile determinare i valori campionari dell'intercetta  $\mathbf{b}_0$  e del coefficiente angolare  $\mathbf{b}_1$  che quindi rappresentano una stima dei relativi parametri della popolazione ( $\beta_0$  e  $\beta_1$ ).

## V.1 RELAZIONI FRA VARIABILI – Modello di regressione

I coefficienti incogniti  $b_0$  e  $b_1$  che minimizzano l'errore totale identificano la retta dei minimi quadrati e vengono determinati risolvendo le seguenti equazioni normali:

$$\begin{cases} \sum_{i=1}^n y_i = n \cdot b_0 + b_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 \end{cases}$$

Risolvendo le due equazioni normali nelle due incognite  $b_0$  e  $b_1$  si ottengono i seguenti valori:

$$b_0 = \frac{\sum_{i=1}^n y_i}{n} - b_1 \frac{\sum_{i=1}^n x_i}{n} \quad b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

Per determinare  $b_0$  e  $b_1$  attraverso queste relazioni è utile costruire una tabella in cui, oltre a  $x_i$  e  $y_i$ , vengono inseriti i valori di  $x_i y_i$ , di  $x_i^2$  e la loro somma.

**Esempio:** Determinare la retta di regressione per i valori dell'esempio precedente.

Si costruisce la tabella che raccoglie i dati necessari al calcolo:

Osserv.	$x_i$	$y_i$	$x_i^2$	$x_i y_i$
1	1	1	1	1
2	3	2	9	6
3	4	4	16	16
4	6	4	36	24
5	8	5	64	40
6	9	7	81	63
7	11	8	121	88
8	14	9	196	126
Somme	56	40	524	364

Dalle formule per il calcolo di  $b_0$  e  $b_1$  si ottiene:

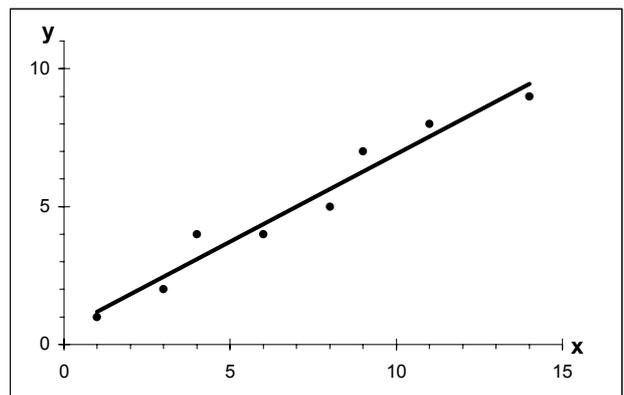
$$b_1 = \frac{8 \cdot 364 - 56 \cdot 40}{8 \cdot 524 - 56^2} = \frac{672}{1056} = 0,636$$

$$b_0 = \frac{40}{8} - 0,636 \frac{56}{8} = 0,545$$

La retta di regressione ha quindi equazione

$$y = 0,545 + 0,636 x$$

e presenta l'andamento mostrato nel grafico.



## V.1 RELAZIONI FRA VARIABILI – Modello di regressione

La retta di regressione esprime il legame "medio" fra due variabili, in quanto i valori  $y_i$  non sono tutti sulla retta; è possibile fare una analogia con la media, che indica il valore centrale di osservazioni che si dispongono attorno ad essa.

La dispersione dei dati attorno alla media si misura con lo scarto quadratico medio, la distanza complessiva delle osservazioni dalla retta di regressione si misura con l'errore standard della stima ( $e_s$ ).

L'errore standard della stima è una statistica che misura la variabilità media delle osservazioni intorno alla retta di regressione e ha la seguente espressione:

$$e_s = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

o, equivalentemente, in termini dei coefficienti della retta di regressione:

$$e_s = \sqrt{\frac{\sum_{i=1}^n y_i^2 - b_0 \sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i y_i}{n - 2}}$$

## V.1 RELAZIONI FRA VARIABILI – Modello di regressione

La retta di regressione, esprimendo il legame medio fra le due variabili, consente di stimare il valore assunto dalla variabile dipendente  $y$  in corrispondenza di un qualunque valore della variabile indipendente  $x$ .

Quando la stima viene eseguita con riferimento ad un valore di  $x$  compreso nel campo di variazione della variabile indipendente si esegue una interpolazione, in caso contrario una estrapolazione.

Il valore stimato (per interpolazione o estrapolazione) è quello che giace sulla retta di regressione; il suo livello di correttezza rispetto al valore reale è determinato dall'errore standard della stima che esprime proprio lo scarto medio fra i valori campionari e i corrispondenti valori sulla retta di regressione.

L'analisi di regressione, e la conseguente possibilità di eseguire estrapolazioni, trova una importante applicazione nel caso in cui la variabile indipendente è il **tempo**; ciò consente lo sviluppo di metodi di previsione nell'ambito delle analisi delle serie temporali, cioè elenchi di dati quantitativi rilevati a scadenze regolari.

### Esempio:

Con riferimento ai dati dell'esercizio precedente, eseguire una stima del valore della variabile  $y$  per  $x=10$  e per  $x=16$ .

L'equazione della retta di regressione relativa ai dati forniti (precedentemente calcolata) è:

$$y = 0,545 + 0,636 x$$

Tramite tale retta deve essere eseguita la stima della  $y$  in corrispondenza dei due valori della  $x$ .

Considerando che il campo di variazione della  $x$  è 1;14, la stima per  $x=10$  è una interpolazione, mentre la stima per  $x=16$  è una estrapolazione.

Nel primo caso ( $x=10$ ) si ottiene  $\hat{y} = b_0 + b_1 x = 0,545 + 0,636 \cdot 10 = 6,905$

e nel secondo ( $x=16$ )  $\hat{y} = b_0 + b_1 x = 0,545 + 0,636 \cdot 16 = 10,721$

L'errore standard della stima per questa analisi di regressione risulta:

$$e_s = \sqrt{\frac{\sum_{i=1}^n y_i^2 - b_0 \sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i y_i}{n - 2}} = \sqrt{\frac{256 - 0,545 \cdot 40 - 0,636 \cdot 364}{8 - 2}} = 0,67$$

E fornisce una indicazione sul margine di affidabilità delle due stime eseguite.

## V.1 RELAZIONI FRA VARIABILI – Metodi di previsione

L'analisi previsionale dei valori futuri di una serie temporale, prevede la scomposizione della serie stessa in diversi elementi:

- tendenza a lungo termine (trend);
- tendenza ciclica a medio termine;
- componente stagionale;
- fluttuazioni irregolari.

L'analisi di regressione viene utilizzata in particolare per determinare il trend lineare di dati annuali.

In questo caso l'inizio della serie (cioè il primo anno considerato) viene posto come  $x_1=0$ , in modo che l'intercetta ( $b_0$ ) identifichi il valore iniziale della serie. Il coefficiente angolare  $b_1$  rappresenta la variazione media annua mostrata dalla variabile dipendente.

La retta di trend viene impiegata per formulare stime del valore atteso della variabile dipendente negli anni successivi al termine della serie (estrapolazione).

L'attendibilità della stima eseguita può essere valutata considerando il valore dell'errore standard della stima o la deviazione assoluta media (DAM).

### Esempio

Il prospetto seguente riporta il numero di iscritti ad un corso universitario nel periodo 1990-2001.

Anno	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Iscritti	230	224	255	301	280	265	295	301	331	291	323	312

Determinare la retta di trend e stimare il numero di iscritti attesi per l'anno 2002 evidenziando il margine di approssimazione della stima.

Per semplificare i calcoli e per ottenere un valore dell'intercetta facilmente interpretabile la variabile indipendente viene modificata ponendo l'inizio della serie (1990) ad un valore 0.

Applicando le formule per il calcolo dei coefficienti della retta di regressione si ottiene  $b_0=240$  e  $b_1=8$

La retta di trend risulta quindi: numero iscritti ( $y$ ) =  $240 + 8$  (anno ( $x$ ) - 1990)

Il valore di  $b_1$  indica che nel periodo considerato si è avuto un incremento medio di 8 iscritti/anno.

Utilizzando l'espressione della retta di trend è possibile stimare il numero di iscritti previsti per l'anno 2002 sostituendo tale valore nell'espressione precedentemente determinata.

$$\text{Iscritti previsti nel 2002} = 240 + 8 (2002-1990) = 336$$

L'affidabilità della stima eseguita può essere valutata attraverso l'errore standard della stima (che in questo vale 19,6) o la deviazione assoluta media. Quest'ultima assume il seguente valore:

$$\text{DAM} = \frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n} = \frac{174}{12} = 14,5$$

## V.2 RELAZIONI FRA VARIABILI – Analisi di correlazione

L'analisi di correlazione viene condotta per determinare il livello del legame fra delle variabili o, da un altro punto di vista, quanto bene la retta di regressione è in grado di spiegare la relazione fra le variabili.

Nel caso della regressione lineare semplice (cioè di una funzione lineare con una sola variabile indipendente) il livello del legame fra le variabili è espresso dalla correlazione lineare semplice.

Per determinare la correlazione fra due variabili ci si basa sulla devianza totale della variabile dipendente espressa come somma delle differenze delle osservazioni dalla loro media elevate al quadrato:

$$\text{devianza totale} = \sum_{i=1}^n (y_i - \bar{y})^2$$

La devianza totale è la somma di due componenti:

- la devianza spiegata dalla retta di regressione;
- la devianza residua (non spiegata) che dipende da fattori diversi dalla relazione fra le variabili stabilita dalla retta di regressione.

## V.2 RELAZIONI FRA VARIABILI – Analisi di correlazione

La devianza spiegata è la somma delle differenze al quadrato fra valori stimati attraverso la retta di regressione e la media dei valori:

$$\text{devianza spiegata} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

La devianza residua è la somma delle differenze al quadrato fra i valori osservati e i corrispondenti valori stimati attraverso la retta di regressione:

$$\text{devianza residua} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Il rapporto fra devianza spiegata e devianza totale, che esprime la proporzione della variazione totale della variabile dipendente spiegata dalla retta di regressione, è detto coefficiente di determinazione e viene indicato con  $r^2$ :

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Partendo dalla relazione

$$\text{devianza totale} = \text{devianza spiegata} + \text{devianza residua}$$

si ottiene dalle precedenti relazioni:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Eseguendo alcuni passaggi algebrici e tenendo conto dei coefficienti della retta di regressione si ottengono le seguenti espressioni:

$$\text{devianza totale} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$$

$$\text{devianza residua} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n y_i^2 - b_0 \sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i y_i$$

dalle quali per differenza si ricava:

$$\text{devianza spiegata} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b_0 \sum_{i=1}^n y_i + b_1 \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$$

## V.2 RELAZIONI FRA VARIABILI – Analisi di correlazione

L'intensità della relazione esistente fra le due variabili è espressa dal coefficiente di correlazione.

Il coefficiente di correlazione fra le variabili viene indicato con  $\rho$ ; la sua stima ottenuta dalle osservazioni campionarie è rappresentata dal coefficiente di correlazione campionario indicato con  $r$  il cui valore è pari alla radice quadrata del coefficiente di determinazione

$$r = \pm\sqrt{r^2}$$

Il segno di  $r$  è determinato dal segno del coefficiente angolare  $b_1$  della retta di regressione. Se la retta di regressione è crescente ( $b_1 > 0$ ) esiste un legame positivo fra le variabili, se è decrescente ( $b_1 < 0$ ) il legame è negativo.

$$\text{Essendo } 0 \leq r^2 \leq 1, \text{ allora } -1 \leq r \leq +1$$

Quando  $r$  è prossimo a  $0$  si dice che le variabili sono incorrelate, quanto più  $r$  si avvicina a  $1$  tanto più le variabili mostrano una correlazione diretta (positiva), quanto più  $r$  si avvicina a  $-1$  tanto più le variabili mostrano una correlazione inversa (negativa).

**Esempio:** Determinare il coefficiente di correlazione fra le variabili dell'esercizio precedente.

Partendo dai dati delle osservazioni è possibile costruire il prospetto che contiene tutti gli elementi necessari per determinare il coefficiente di determinazione, considerando che  $\bar{y} = 5$

Osserv.	$x_i$	$y_i$	$\hat{y}_i$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
1	1	1	1,18	16	14,58
2	3	2	2,45	9	6,49
3	4	4	3,09	1	3,56
4	6	4	4,36	1	0,41
5	8	5	5,63	0	0,40
6	9	7	6,27	4	1,61
7	11	8	7,54	9	6,46
8	14	9	9,45	16	19,79
Somme				56	53,39

$$\text{devianza totale} = \sum_{i=1}^n (y_i - \bar{y})^2 = 56$$

$$\text{devianza spiegata} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 53,39$$

Il valore del coefficiente di determinazione risulta pertanto

$$r^2 = \frac{\text{devianza spiegata}}{\text{devianza totale}} = \frac{53,39}{56} = 0,953$$

e, di conseguenza, il coefficiente di correlazione (essendo  $b_1 > 0$ ):

$$r = +\sqrt{r^2} = +0,977$$

## V.2 RELAZIONI FRA VARIABILI – Analisi di correlazione

Il coefficiente di correlazione determinato partendo dai risultati dell'analisi di regressione, come radice quadrata del rapporto fra devianza spiegata e devianza totale, esprime una misura di quanto bene la retta di regressione è in grado di spiegare il legame fra le due variabili.

Quando si è interessati ad evidenziare il livello di associazione di due variabili, indipendentemente dalla capacità di una di far prevedere l'altra, l'analisi di regressione è superflua e viene eseguita direttamente un'analisi di correlazione.

In questo caso, senza identificare la variabile indipendente e la variabile dipendente, si procede alla determinazione del coefficiente di correlazione  $r$  attraverso la seguente relazione:

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

## V.2 RELAZIONI FRA VARIABILI – Analisi di correlazione

Espresso in questi termini il coefficiente di correlazione rappresenta il rapporto fra la variabilità congiunta (covarianza) delle due variabili rispetto alla loro media ed il prodotto dei loro “quasi” scarti quadratici medi.

Utilizzando la seguente notazione:

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}$$

e indicando la covarianza di x e y come

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

Il coefficiente di correlazione può essere scritto come  $r = \frac{s_{xy}}{s_x s_y}$

### Esempio

Determinare il coefficiente di correlazione fra le variabili dell'esempio precedente utilizzando la formula che utilizza la covarianza fra le variabili.

Per calcolare i termini della relazione che esprime il coefficiente di correlazione viene costruito un prospetto riepilogativo basato sulle differenze dalle medie  $\bar{x} = 7$  e  $\bar{y} = 5$

Osserv.	$x_i$	$y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	1	36	16	24
2	3	2	16	9	12
3	4	4	9	1	3
4	6	4	1	1	1
5	8	5	1	0	0
6	9	7	4	4	4
7	11	8	16	9	12
8	14	9	49	16	28
Somme	56	40	132	56	84

In base ai valori della tabella si ottiene:

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{132}{8}} \quad s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}} = \sqrt{\frac{56}{8}} \quad s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{84}{8}$$

Per cui il coefficiente di correlazione risulta  $r = \frac{s_{xy}}{s_x s_y} = \frac{84}{\sqrt{132} \sqrt{56}} = 0,977$

# **Appendice**

## **TAVOLE STATISTICHE**

**Distribuzione Normale Standardizzata (z)**

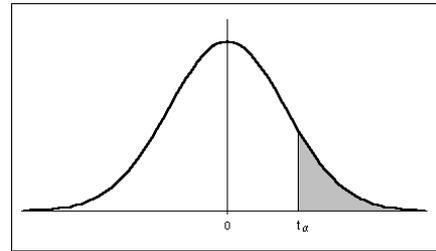
**Distribuzione t di Student (t)**

**Distribuzione Chi-quadrato ( $\chi^2$ )**



## Tavole statistiche – DISTRIBUZIONE t DI STUDENT

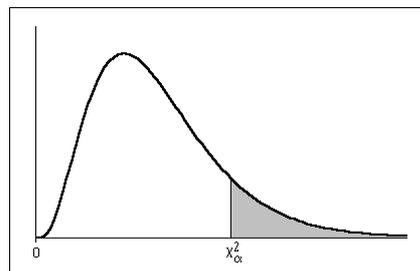
Valori della variabile  $t$  ( $t_\alpha$ ) in  
 corrispondenza di aree  $\alpha$  sotto la coda  
 della distribuzione  $t$  di Student  
 e al variare dei gradi di libertà  $v$



$v$	$\alpha=0,005$	0,01	0,025	0,05	0,100	0,200	0,250	0,300	0,400
1	63,66	31,82	12,71	6,31	3,078	1,376	1,000	0,727	0,325
2	9,92	6,96	4,30	2,92	1,886	1,061	0,816	0,617	0,289
3	5,84	4,54	3,18	2,35	1,638	0,978	0,765	0,584	0,277
4	4,60	3,75	2,78	2,13	1,533	0,941	0,741	0,569	0,271
5	4,03	3,36	2,57	2,02	1,476	0,920	0,727	0,559	0,267
6	3,71	3,14	2,45	1,94	1,440	0,906	0,718	0,553	0,265
7	3,50	3,00	2,36	1,89	1,415	0,896	0,711	0,549	0,263
8	3,36	2,90	2,31	1,86	1,397	0,889	0,706	0,546	0,262
9	3,25	2,82	2,26	1,83	1,383	0,883	0,703	0,543	0,261
10	3,17	2,76	2,23	1,81	1,372	0,879	0,700	0,542	0,260
11	3,11	2,72	2,20	1,80	1,363	0,876	0,697	0,540	0,260
12	3,05	2,68	2,18	1,78	1,356	0,873	0,695	0,539	0,259
13	3,01	2,65	2,16	1,77	1,350	0,870	0,694	0,538	0,259
14	2,98	2,62	2,14	1,76	1,345	0,868	0,692	0,537	0,258
15	2,95	2,60	2,13	1,75	1,341	0,866	0,691	0,536	0,258
16	2,92	2,58	2,12	1,75	1,337	0,865	0,690	0,535	0,258
17	2,90	2,57	2,11	1,74	1,333	0,863	0,689	0,534	0,257
18	2,88	2,55	2,10	1,73	1,330	0,862	0,688	0,534	0,257
19	2,86	2,54	2,09	1,73	1,328	0,861	0,688	0,533	0,257
20	2,85	2,53	2,09	1,72	1,325	0,860	0,687	0,533	0,257
21	2,83	2,52	2,08	1,72	1,323	0,859	0,686	0,532	0,257
22	2,82	2,51	2,07	1,72	1,321	0,858	0,686	0,532	0,256
23	2,81	2,50	2,07	1,71	1,319	0,858	0,685	0,532	0,256
24	2,80	2,49	2,06	1,71	1,318	0,857	0,685	0,531	0,256
25	2,79	2,49	2,06	1,71	1,316	0,856	0,684	0,531	0,256
26	2,78	2,48	2,06	1,71	1,315	0,856	0,684	0,531	0,256
27	2,77	2,47	2,05	1,70	1,314	0,855	0,684	0,531	0,256
28	2,76	2,47	2,05	1,70	1,313	0,855	0,683	0,530	0,256
29	2,76	2,46	2,05	1,70	1,311	0,854	0,683	0,530	0,256
30	2,75	2,46	2,04	1,70	1,310	0,854	0,683	0,530	0,256
40	2,70	2,42	2,02	1,68	1,303	0,851	0,681	0,529	0,255
60	2,66	2,39	2,00	1,67	1,296	0,848	0,679	0,527	0,254
120	2,62	2,36	1,98	1,66	1,289	0,845	0,677	0,526	0,254
$\infty$	2,58	2,33	1,96	1,645	1,282	0,842	0,675	0,524	0,253

## Tavole statistiche – DISTRIBUZIONE CHI-QUADRATO

Valori della variabile  $\chi^2$  ( $\chi^2_\alpha$ ) in  
 corrispondenza di aree  $\alpha$  sotto la coda  
 della distribuzione chi-quadrato  
 e al variare dei gradi di libert   $\nu$



$\nu$	$\alpha=0,005$	0,01	0,025	0,05	0,100	0,200	0,250	0,500
1	7,88	6,63	5,02	3,84	2,71	1,64	1,32	0,45
2	10,60	9,21	7,38	5,99	4,61	3,22	2,77	1,39
3	12,84	11,34	9,35	7,81	6,25	4,64	4,11	2,37
4	14,86	13,28	11,14	9,49	7,78	5,99	5,39	3,36
5	16,75	15,09	12,83	11,07	9,24	7,29	6,63	4,35
6	18,55	16,81	14,45	12,59	10,64	8,56	7,84	5,35
7	20,28	18,48	16,01	14,07	12,02	9,80	9,04	6,35
8	21,95	20,09	17,53	15,51	13,36	11,03	10,22	7,34
9	23,59	21,67	19,02	16,92	14,68	12,24	11,39	8,34
10	25,19	23,21	20,48	18,31	15,99	13,44	12,55	9,34
11	26,76	24,73	21,92	19,68	17,28	14,63	13,70	10,34
12	28,30	26,22	23,34	21,03	18,55	15,81	14,85	11,34
13	29,82	27,69	24,74	22,36	19,81	16,98	15,98	12,34
14	31,32	29,14	26,12	23,68	21,06	18,15	17,12	13,34
15	32,80	30,58	27,49	25,00	22,31	19,31	18,25	14,34
16	34,27	32,00	28,85	26,30	23,54	20,47	19,37	15,34
17	35,72	33,41	30,19	27,59	24,77	21,61	20,49	16,34
18	37,16	34,81	31,53	28,87	25,99	22,76	21,60	17,34
19	38,58	36,19	32,85	30,14	27,20	23,90	22,72	18,34
20	40,00	37,57	34,17	31,41	28,41	25,04	23,83	19,34
21	41,40	38,93	35,48	32,67	29,62	26,17	24,93	20,34
22	42,80	40,29	36,78	33,92	30,81	27,30	26,04	21,34
23	44,18	41,64	38,08	35,17	32,01	28,43	27,14	22,34
24	45,56	42,98	39,36	36,42	33,20	29,55	28,24	23,34
25	46,93	44,31	40,65	37,65	34,38	30,68	29,34	24,34
26	48,29	45,64	41,92	38,89	35,56	31,79	30,43	25,34
27	49,65	46,96	43,19	40,11	36,74	32,91	31,53	26,34
28	50,99	48,28	44,46	41,34	37,92	34,03	32,62	27,34
29	52,34	49,59	45,72	42,56	39,09	35,14	33,71	28,34
30	53,67	50,89	46,98	43,77	40,26	36,25	34,80	29,34
40	66,77	63,69	59,34	55,76	51,81	47,27	45,62	39,34
50	79,49	76,15	71,42	67,50	63,17	58,16	56,33	49,33
60	91,95	88,38	83,30	79,08	74,40	68,97	66,98	59,33
70	104,21	100,43	95,02	90,53	85,53	79,71	77,58	69,33
80	116,32	112,33	106,63	101,88	96,58	90,41	88,13	79,33
90	128,30	124,12	118,14	113,15	107,57	101,05	98,65	89,33
100	140,17	135,81	129,56	124,34	118,50	111,67	109,14	99,33