# PSICOMETRIA 1

# 2000 - 2001

**Psicometria 1 - Lezione 1**
**Lucidi presentati a lezione**

**AA 2000/2001 dott. Corrado Caudek**

# PSICOMETRIA 1

**Corrado Caudek**

**ricevimento: MERCOLEDI' 12:30 - 14:30**

**uff. 515 (V piano) Via S. Anastasio, 12**

**tel. 040 676 2739**

**email: caudek@univ.trieste.it**

**Testi per la preparazione all'esame:**

**- *studenti frequentanti*:**

**(1) Vidotto, Xausa, Pedon. Statistica per psicologi. Ed. Il Mulino.**

**(2) Dispense**

**- *studenti non frequentanti*:**

**da definire con il docente.**

**La frequenza alle *lezioni* ed *esercitazioni* è fortemente consigliata.**

**Le esercitazioni si terranno tutti i giovedì delle settimane in cui si tengono le lezioni e si svolgeranno al piano amezzato di Via S. Anastasio n. 12, aula n. 5.**

**10:00 - 11:30 GRUPPO 1 (lettere A - E)**

**11:30 - 13:00 GRUPPO 2 (lettere F - M)**

**13:00 - 14:30 GRUPPO 3 (lettere N - R)**

**14:30 - 16:00 GRUPPO 4 (lettere S - Z)**

**A lezione verranno assegnati degli *esercizi* da svolgere a casa. Durante le esercitazioni questi esercizi verranno discussi e altri esercizi verranno presentati.**

**Durante le esercitazioni non verranno discussi argomenti teorici non presentati a lezione, ma verranno forniti esempi concreti degli argomenti già discussi.**

**Gli esercizi svolti durante le esercitazioni avranno un grado di complessità simile a quello degli esercizi che saranno contenuti nelle provette.**

**Per essere considerato "frequentante" uno studente dovrà partecipare alle lezioni e alle esercitazioni.**

**Dovrà inoltre consegnare lo svolgimento di almeno 7 degli 8 gruppi di esercizi che verranno proposti durante il corso.**

**Un gruppo di esercizi verrà proposto durante ciascuna delle 8 settimane di svolgimento del corso.**

**Gli esercizi risolti potranno essere consegnati SOLTANTO durante le esercitazioni.**

## Esame studenti frequentanti:

**provetta a metà corso**
**(50% del voto complessivo)**
**+**
**provetta a fine corso**
**(solo sul materiale discusso**
**nella seconda metà del corso)**
**(50% del voto complessivo)**

***Orale opzionale*. Chi vorrà sostenere l'esame orale verrà valutato facendo la media del voto ottenuto nelle provette (50%) e il voto ottenuto all'orale (50%).**

**<u>Le provette si terranno senza l'ausilio del testo e degli appunti</u>. Questo significa che le formule necessarie per risolvere i problemi dovranno essere memorizzate.**

**Durante il corso vi indicherò quali sono le formule che dovrete conoscere per potere superare l'esame.**

**Ciascuna delle due provette avrà la durate di un'ora.**

**Esame studenti NON frequentanti:**

**provetta a fine corso**
**(50% del voto complessivo)**

**+**

**esame orale**
**(50% del voto complessivo)**

# …. come studiare?

# OBIETTIVI DEL CORSO

- introduzione alle tematiche connesse alla misura in psicologia

- trattazione dei concetti di base della statistica applicata alla ricerca e all’elaborazione dei dati di natura psicologica.

indici di tendeza centrale e di variabilità; rappresentazioni grafiche della distribuzione di frequenze; scale di misura; calcolo combinatorio; elementi di teoria della probabiiltà; variabili aleatorie e loro proprietà; principali distribuzioni di probabilità (binomiale, ipergeometrica, normale, $c^{2}$, $F$, $t$); teoria dei campioni, teorema del limite centrale.

parametri, stimatori, stima puntuale e intervallare; verifica di ipotesi, errori di I e II tipo; intervallo di fiducia per la media; intervallo di fiducia per una proporzione; stima per intervalli di una varianza; stima della differenza tra medie; verifica delle ipotesi sulla media; test $t$ di Student; verifica di ipotesi e distribuzione $F$; verifica di ipotesi e distribuzione $c^{2}$; regressione lineare.

- **Variabili**
- **Sommatorie**
- **Scale di misura**
- **Distribuzioni di frequenze**
- **Rappresentazione grafica dei dati**

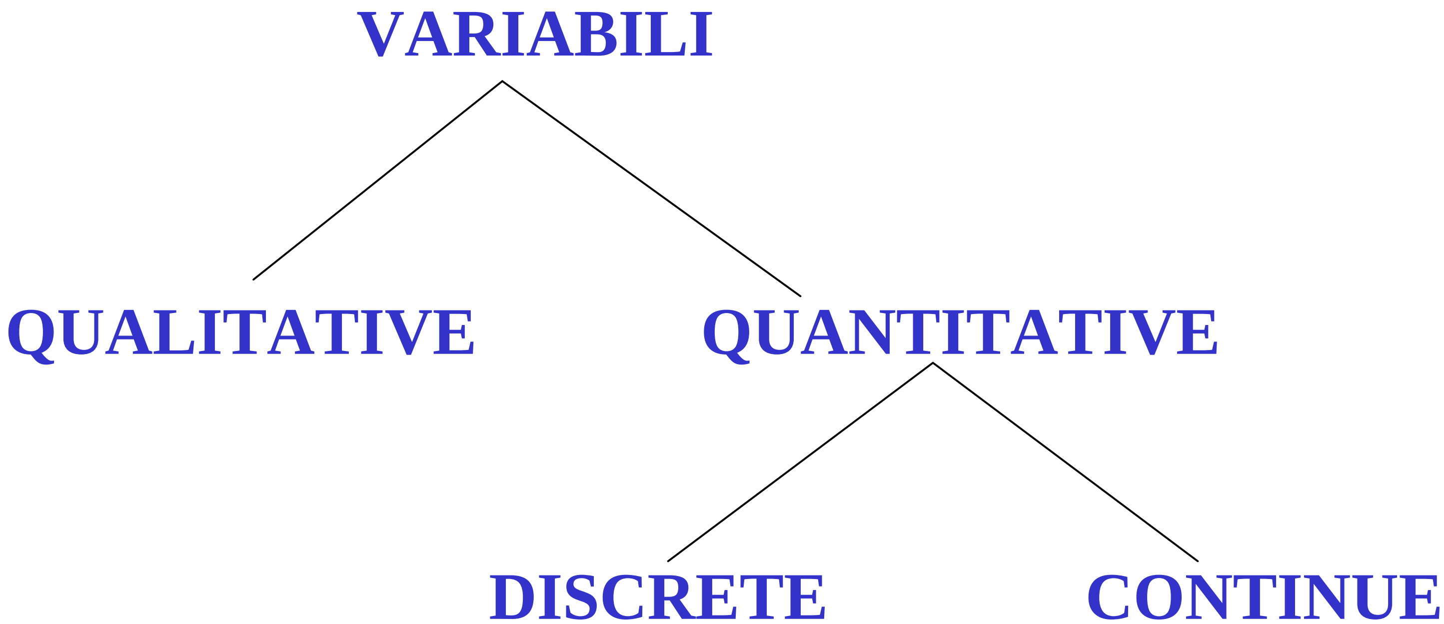
*DEFINIZIONI*

**Unità d’analisi**: gli elementi che costituiscono l’oggetto dell’osservazione

**Popolazione**: l’insieme di tutte le unità d’analisi.

**Variabili**: le proprietà delle unità d’analisi che variano da caso a caso.

**Modalità**: il valore che viene assunto dall’unità d’analisi rispetto alla variabile considerata.


VARIABILI
QUALITATIVE
QUANTITATIVE
DISCRETE
CONTINUE

**VARIABILI QUALITATIVE:**
hanno modalità espresse da “attributi” (sesso, professione, provenienza geografica, …)

**VARIABILI QUANTITATIVE:**
hanno modalità espresse da numeri (età, altezza, QI, ...).

**Le variabili quantitative possono essere *discrete* o *continue*.**

**I valori assunti da una variabile discreta possono differire solo per un ammontare fisso.**

Esempio.

Il *numero di componenti di una famiglia* è una variabile discreta in quanto due famiglie possono differire solo per 0 unità, 1 unità, 2 unità, ecc. Non è ammissibile nessun valore intermedio (due famiglie non possono avere un numero di componenti che differisce per 2.5 unità)

**I valori assunti da una <u>variabile continua</u> possono differire per una qualsiasi quantità arbitraria.**

Esempio.

L'*età* è una variabile continua. La differenza d'età tra due persone, infatti, può essere un qualunque valore (un anno, un giorno, un'ora, …)

## ESERCIZIO

**Classificate le seguenti variabili come qualitative o quantitative e, in quest'ultimo caso, come discrete o continue.**

**a) occupazione**
**b) regione di residenza**
**c) peso**
**d) altezza**
**e) numero di automobili possedute**

# SOMMATORIE

Consideriamo la seguente distribuzione della variabile *X*:

3, 7, 1, 9, 2.

Usiamo l'indice *i* per fare riferimento alle diverse unità di osservazione che costituiscono la distribuzione.

Il valore assunto dalla variabile *X* nel caso di una generica unità di osservazione viene indicato con $X_i$.

Quando l'indice *i* assume il valore *2, 3, o 5*, questo significa che facciamo riferimento al valore assunto da X nel caso della seconda, terza o quinta unita di osservazione.

Esempio: $X_1 = 3$; $X_4 = 9$; $X_2 = 7$

Un simbolo di cui viene fatto un grande uso nella statistica è quello di sommatoria:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$$

Nel caso della distribuzione precedente {3, 7, 1, 9, 2}, la somma sarà:

$$\sum_{i=1}^{5} X_i = 3 + 7 + 1 + 9 + 2 = 22$$

Quando risulta chiaro che vogliamo sommare tutti i valori assunti da *X* nel caso di *n* osservazioni, la notazione precedente può essere semplificata nel modo seguente:

$$\sum_{i} X_i \quad \text{oppure} \quad \sum X_i \,.$$

Un caso più complicato si ha quando le singole osservazioni sono arrangiate in più gruppi.

Per esempio, assumiamo che vi siano 10 studenti in 3 aule. Per distinguere tutte queste osservazioni abbiamo bisogno di 2 indici, uno per gli studenti e uno per le aule. Sia *i* l'indice per gli studenti e *j* l'indice per le aule. Quindi l'indice *i* va da 1 a 10 e l'indice *j* va da 1 a 3.

Ciascuno studente sarà dunque identificato da due indici: $X_{ij}$. Se usiamo la convenzione per cui l'indice *i* precede l'indice *j*, allora $X_{73}$ indica il settimo studente nella terza aula.

Spesso accade che vogliamo sommare i punteggi di tutti gli individui in tutti i gruppi. Per fare questo usiamo la seguente notazione:

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ij}$$

Nel caso dell'esempio precedente (10 studenti per aula) avremo:

$$\sum_{j=1}^{3} \sum_{i=1}^{10} X_{ij} = X_{1,1} + X_{1,2} + ... + X_{10,3}$$

# Regole per la manipolazione delle sommatorie

**Regola 1**. Se $a$ è una costante, allora

$$\sum_{i=1}^{n} a = n \cdot a$$

Esempio. Se $a$ = 10, allora la sommatoria da 1 a $n$ di $a$ (con $n$ = 5) è

$$\sum_{i=1}^{5} 10 = 10 + 10 + 10 + 10 + 10 = 50$$

**Regola 2**. Se ciascuna delle osservazioni che entrano in una sommatoria viene moltiplicata per la costante *a*, allora

$$\sum_{i=1}^{n} a \cdot X_i = a \sum_{i=1}^{n} X_i$$

Esempio. La variabile *X* assume i seguenti valori: {3, 5, 1} e la costante *a* è $a = 2$.

$$\sum_{i=1}^{3} a \cdot X = (2 \times 3) + (2 \times 5) + (2 \times 1)$$

$$= 2(3 + 5 + 1) = 18$$

**Regola 3**. Se dobbiamo eseguire un'operazione algebrica (quadrato, radice, logaritmo, ecc.) sulle singole osservazioni che devono essere sommate, questa operazione deve essere eseguita *prima di sommare* le $n$ osservazioni:

Esempi.

$$\sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

Quindi, $\sum_{i=1}^{n} X_i^2 \neq \left( \sum_{i=1}^{n} X_i \right)^2, \quad \sum_{i=1}^{n} \sqrt{X_i} \neq \sqrt{\sum_{i=1}^{n} X_i}$ .

La regola precedente è specialmente importante quando abbiamo un doppio segno di sommatoria.

L'espressione

$$\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} X_{ij}\right)^2$$

significa che dobbiamo prima fare la sommatoria all'interno dei $J$ gruppi, dopo innalzare al quadrato, e poi sommare per tutti i gruppi.

**Regola 4**. Se la sola operazione che deve essere eseguita sulle osservazioni prima della sommatoria è essa stessa una somma (o sottrazione), allora la sommatoria può essere distribuita:

$$\sum (X_i - Y_i) = \sum X_i - \sum Y_i$$

Esempi.

(1) $$\sum\left(X_i^2-3X_i+10\right)=\sum X_i^2+3\sum X_i+\sum 10\,.$$

(2)

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(X_{ij}-m\right)^2=\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(X_{ij}^2-2mX_{ij}+m^2\right)$$

$$=\sum_{j=1}^{J}\sum_{i=1}^{n_j}X_{ij}^2-\sum_{j=1}^{J}\sum_{i=1}^{n_j}2mX_{ij}+\sum_{j=1}^{J}\sum_{i=1}^{n_j}m^2$$

$$=\sum_{j=1}^{J}\sum_{i=1}^{n_j}X_{ij}^2-2m\sum_{j=1}^{J}\sum_{i=1}^{n_j}X_{ij}+m^2\sum_{j=1}^{J}n_j$$

**Regola 5**. Se ciascuna osservazione ha due punteggi, $X_i$ e $Y_i$, allora

$$\sum_{i=1}^{n} X_i Y_i = X_1 Y_1 + X_2 Y_2 + \ldots + X_n Y_n$$

in altre parole, il prodotto tra i punteggi appaiati deve essere eseguito prima e la sommatoria dopo. Quindi

$$\sum_{i=1}^{n} X_i Y_i \neq \left(\sum Y_i\right)\left(\sum Y_i\right)$$

**Regola 6.**

$$\sum_{i=1}^{n} aX_iY_i = a\sum_{i=1}^{n} X_iY_i$$

**Regola 7**.

$$\sum_{i=1}^{n}(aX_i + bY_i) = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} Y_i$$

**Esempio**. La media della variabile *X* è uguale a:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Provate che la somma di tutti i punteggi dalla media è uguale a zero.

$$\sum_{i=1}^{n} \left( X_i - \overline{X} \right) = 0$$

$$= \sum X_i - \sum \overline{X} = \sum X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$$

## ESERCIZIO

**Scrivete in forma estesa le somme rappresentate dalle seguenti espressioni:**

$$\sum_{i=1}^{4} x_i$$

$$\sum_{i=1}^{3} 4\left(\sum_{j=1}^{2} x_{ij}\right)$$

## ESERCIZIO

**Esprimete le seguenti somme con la notazione appropriata:**

$$\left(x_1 + x_2 + x_3 + x_4\right)/\, n$$

$$\left(3x_1 + 3x_2 + 3x_3 + 3x_4\right)\left(5x_1 + 5x_2 + 5x_3 + 5x_4\right)$$

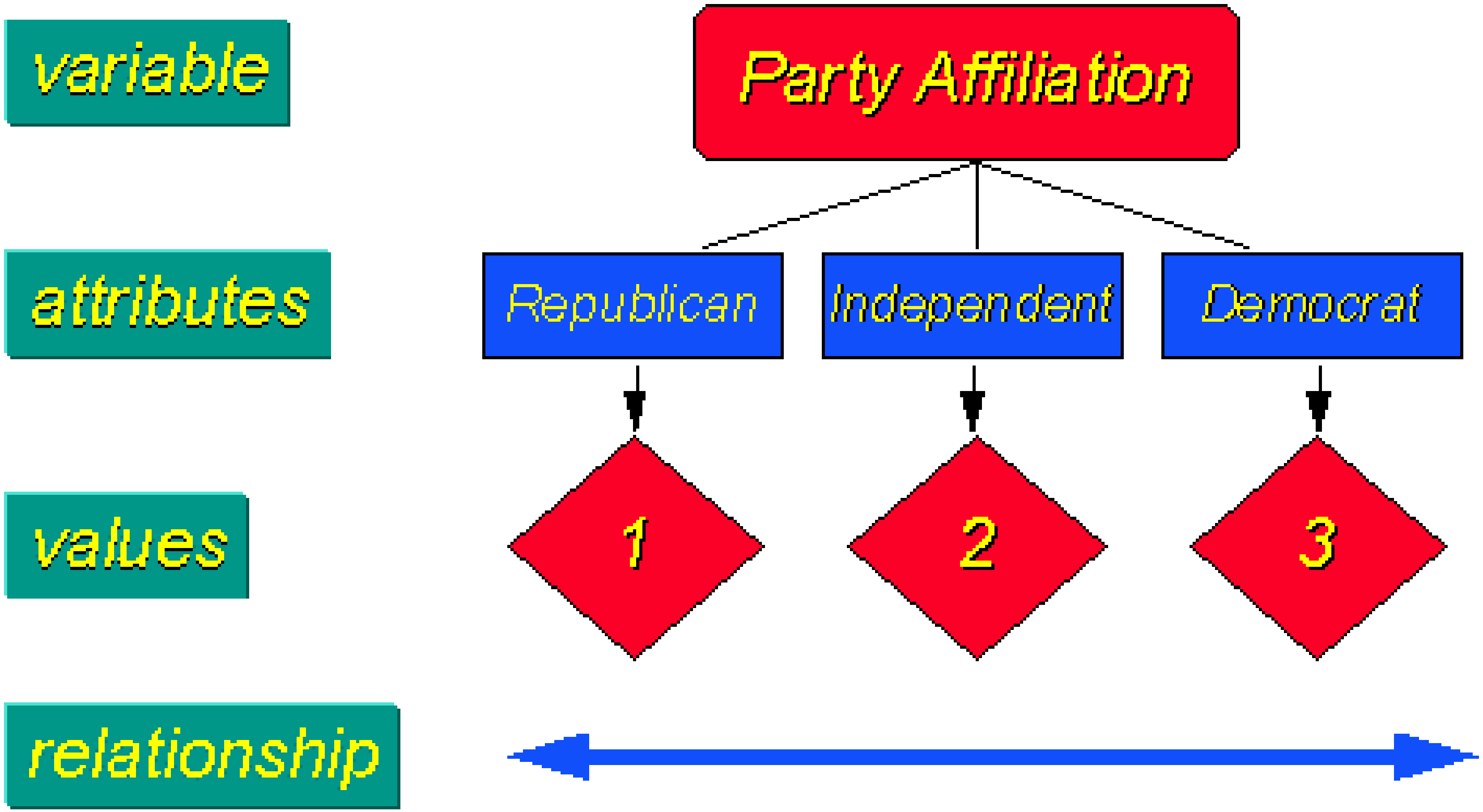
$$\left(x_1 + x_2 + x_3 + x_4\right)^2 \left(y_1^2 + y_2^2 + y_3^2 + y_4^2\right)$$

## ESERCIZIO

**Riducete l'espressione seguente nella sua forma più semplice:**

$$\sum_{i=1}^{n}\left( x_i - \left( \sum_{i=1}^{n} x_i / n \right) \right)$$

# SCALE DI MISURA


variable
Party Affiliation
attributes
Republican
Independent
Democrat
values
1
2
3
relationship

Il livello di misura descrive le relazioni che intercorrono tra i tre valori assunti da una variabile. In questo particolare caso, i numeri vengono usati soltanto come delle etichette che sostituiscono i termini verbali che rappresentano le diverse modalità della variabile considerata.

In questo caso, questi numeri non vogliono trasmettere l’idea che valori più elevati rappresentino una maggiore intensità e numeri più bassi rappresentino una minore intensità di una qualche proprietà.

Il valore 3 non significa che i democratici sono 3 volte quel qualcosa che i repubblicani sono. I numeri sono usati in questo caso solo come delle “etichette” degli attributi.

Questo livello di misurazione si chiama “**nominale**”.

# SCALA NOMINALE

**Classificazione**: le unità d'analisi sono suddivise in *classi di equivalenza*.

A ciascuna classe viene arbitrariamente assegnato un numero o nome.

A ciascuna classe possiamo arbitrariamente cambiare il numero o nome, purché l'appartenenza alla classe sia preservata.

**Esempio.** La relazione "essere coetaneo di" in un insieme $P$ di persone è una relazione di equivalenza.

# SCALA ORDINALE

Consideriamo una certa proprietà (es.,motivazione) posseduta da ciascuna unità d'analisi dell'insieme O.

Supponiamo che a ciascun oggetto $o_i$ sia possibile assegnare il numero $v(o_i)$ che rappresenta l'esatto ammontare della proprietà in esame per l'unità d'analisi $o_i$.

Supponiamo inoltre di non conoscere il numero $v(o_i)$.

Quello che le nostre misurazioni sono in grado di fare è di associare un numero $m(o_i)$ a ciascun $o_i$.

Supponiamo di possedere una procedura di misurazione in grado di associare un numero $m(o_i)$ a ciascun $o_i$ e un diverso numero numero $m(o_j)$ a ciascun $o_j$, per tutte le unità d'analisi dell'insieme O.

Diciamo che le nostre misurazioni sono a livello di scala ordinale se

1. $m(o_i) \neq m(o_j)$ implica che $v(o_i) \neq v(o_j)$

2. $m(o_i) > m(o_j)$ implica che $v(o_i) > v(o_j)$

Esempio: Scala di durezza dei minerali

E’ possibile ordinare le unità d’osservazione rispetto al grado di possesso della proprietà considerata

A ciascuna *classe di equivalenza* viene attributo un numero, dove il valore di questi numeri viene usato per rappresentare ***l’ordinamento*** delle classi, ***non*** la quantità di possesso dell’attributo considerato.

In una scala ordinale, le unità di osservazione possono essere ordinate, ma le distanze tra i valori espressi dalla variabile ordinale non hanno nessun significato.

Esempio.

In un questionario, la seguente codifica potrebbe venire usata:

GRADO DI ISTRUZIONE

0 = scuola elementare;

1 = scuola media inferiore;

2 = scuola media superiore;

3 = università;

4 = specializzazione post-laurea;

In questo esempio, numeri maggiori significano un più elevato grado di istruzione.

Ma la distanza tra 0 e 1 è la stessa della distanza tra 3 e 4?

Ovviamente, per misure a livello ordinale, gli intervalli tra i valori non sono interpretabili

# SCALA AD INTERVALLI

Supponiamo che le nostre misurazioni possiedano,
oltre alle due proprietà precedenti, la seguente proprietà:

1. $m(o_i) \neq m(o_j)$ implica che $v(o_i) \neq v(o_j)$

2. $m(o_i) > m(o_j)$ implica che $v(o_i) > v(o_j)$

3. **Per ciascun unità d'analisi $o_i$, $v(o_i) = x$ se e solo se $m(o_i) = ax + b$, per $a \neq 0$.**

Ovvero, il numero misurato $m(o_i)$ è una funzione lineare della "vera" quantità $x_i$ posseduta dall'unità di osservazione $o_i$

Mentre per una scala ordinale le differenze $m(o_i) - m(o_j)$ non sono rappresentative della differenza $v(o_i) - v(o_j)$, nel caso di una scala ad intervalli la differenza

$$\boldsymbol{m(o_i) - m(o_j) = x}$$

implica

$$\boldsymbol{v(o_i) - v(o_j) = x / a}$$

In altre parole, la misura $m(o_i)$ è una funzione lineare della vera quantità $v(o_i)$.

Esempio. Supponiamo di misurare la temperatura in Fahrenheit di due unità d’osservazione $o_i$ e $o_j$.

La prima misura è $m(o_i) = 180$ e la seconda misura è $m(o_j) = 160$.

Per una scala ad intervalli ha senso dire che la seconda osservazione ha 20 unità di temperatura più della prima.

Le osservazioni misurate su una scala ad intervalli ***possono essere trasformate mediate qualunque trasformazione lineare*** senza per questo distorcere la rappresentazione delle vere proprietà possedute dalle unità d’osservazione considerate.

Esempio. Trasformare la scala di temperature da Fahernheit in Celsius:

$$\boldsymbol{C = 5/9\ (F - 32)}$$

Le differenze tra valori $m(o_i)$ - $m(o_j)$ di una scala ad intervalli rappresentano le differenze $v(o_i)$ - $v(o_j)$ mediante unità di una scala arbitraria.

E' dunque possibile eseguire qualunque operazione aritmetica sulle <u>differenze</u> tra valori di una scala ad intervalli.

# Esempio

$$a = 5*(180-32)/9 = 82.2$$

$$b = 5*(160-32)/9 = 71.1$$

$$c = 5*(140-32)/9 = 60$$

$$d = 5*(100-32)/9 = 37.8$$

$$(c-d)/(a-b) = 2$$

Le differenze tra valori di una scala a intervalli sono interpretabili. Per es., quando misuriamo la temperatura (in Celsius), la differenza tra 10 e 20 ha lo stesso significato della differenza tra 30 e 40. Per questo motivo, ha senso calcolare una media di una variabile a livello a intervalli, mentre non ha senso fare la stessa cosa per una variabile a livello ordinale.

Notate però che **i rapporti tra valori di una scala a intervalli non sono interpretabili**. Per es., nel caso di due misure di temperatura (40° e 20°) non ha senso dire che l'intensità del calore espresso dalla prima misura è doppia rispetto all'intensità del calore espresso dalla seconda misura.

# Esempio

$$(40*9/5)+32 = 104$$

$$(20*9/5)+32 = 68$$

$$104 / 68 = 1.5294$$

# SCALA A RAPPORTI

Supponiamo che le nostre misurazioni possiedano, oltre alle tre proprietà precedenti, la proprietà seguente:

1. $m(o_i) \neq m(o_j)$ implica che $t(o_i) \neq t(o_j)$

2. $m(o_i) > m(o_j)$ implica che $t(o_i) > t(o_j)$

3. *Per ciascun unità d'analisi* $o_i$, $v(o_i) = \mathrm{x}$ se e solo se $m(o_i) = ax + b$, *per* $a \neq 0$.

**4. *Per ciascun unità d'analisi*** $\boldsymbol{o_i}$**,** $\boldsymbol{v(o_i) = x}$ **se e solo se** $\boldsymbol{m(o_i) = ax}$**,** ***per*** $\boldsymbol{a > 0}$**.**

Per una scala a rapporti, i ***rapporti*** tra le misure possono essere considerati uguali ai rapporti tra le le quantità della proprietà in esame che sono “veramente” possedute dalle unità d’analisi considerate.

$$m(o_i) \,/\, m(o_j) = v(o_i) \,/\, v(o_j)$$

Esempio.

La ***lunghezza*** è misurata su una scala a rapporti. Ha dunque senso dire che un primo oggetto ha una lunghezza doppia rispetto ad un secondo oggetto.

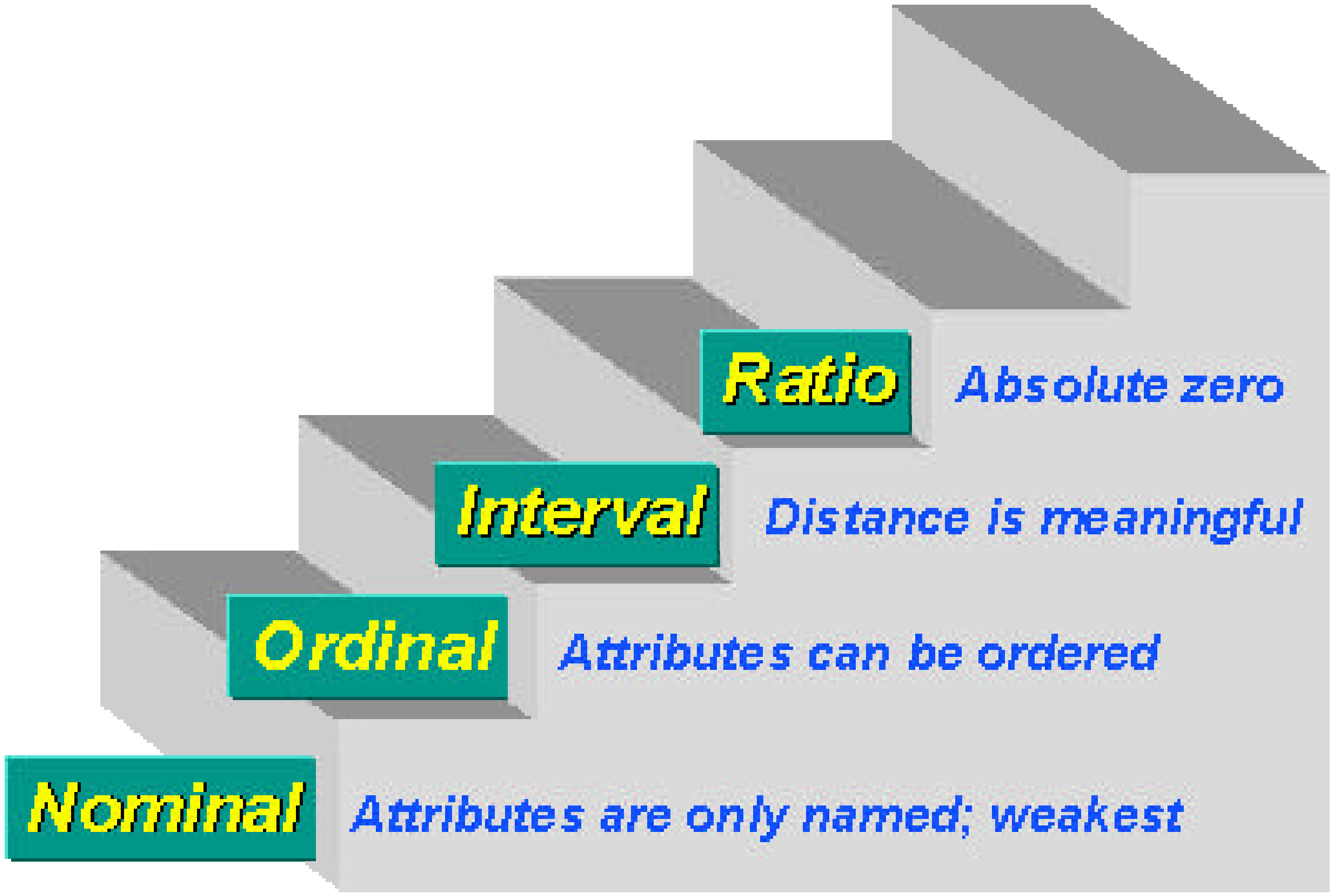
Prima avevamo visto, invece, che non ha senso dire che un oggetto ha una ***temperatura*** doppia rispetto ad un altro oggetto, dato che la temperatura è misurata su una scala ad intervalli.

In una scala di misura a livello di rapporto

***c'è uno zero assoluto non arbitrario***.

Le *frequenze*, per esempio, sono variabili a livello di scala a rapporti.

Perché? In questo caso è sensato dire che un vi sono *zero* casi, e che è stato osservato un numero di casi doppio rispetto ad un'altra occasione.


Ratio
Absolute zero
Interval
Distance is meaningful
Ordinal
Attributes can be ordered
Nominal
Attributes are only named; weakest

# ESERCIZIO

**Individuate la scala di misura appropriata per le seguenti variabili:**

**1) il tipo di musica che un individuo preferisce ascoltare (classica, jazz, …)**
**2) la pressione di una mano misurata da un dinamometro**
**3) la memoria misurata dal numero di parole rievocate a partire da una lista memorizzata in un momento precedente**
**4) la distanza aerea tra Trieste e le altre città europee**
**5) il tempo**
**6) le abilità di lettura dei bambini in età prescolare, misurate con un test standardizzato**
**7) il codice fiscale**

# ESERCIZIO

**Per che tipo di scale di misura sono appropriate le seguenti manipolazioni?**

**1) calcolare la differenza aritmetica tra due valori**
**2) cambiare i numeri o nomi assegnati a classi di variabili**
**3) affermare che, tra due valori assunti da una variabile, uno di essi rappresenta un'intensità maggiore della caratteristica considerata rispetto all'altro.**
**4) calcolare il rapporto tra due valori**
**5) calcolare il rapporto tra le differenze di due valori**
**6) moltiplicare ciascun valore per una costante e poi sommare una costante a ciascun valore**

**STATISTICA DESCRITTIVA**:
riassumere grandi masse di dati

**STATISTICA INFERENZIALE**:
descrivere le proprietà della popolazione sulla base delle informazioni possedute da un campione

# DISTRIBUZIONE DI FREQUENZE

Le misure possono essere a livello nominale, ordinale, a intervalli o a rapporti.

Una distribuzione di frequenze può essere sempre costruita purché ciascuna osservazione venga attribuita ad una ed una sola classe.

## Scala nominale

Supponiamo di suddividere un insieme *O* di unità d’analisi in una serie di classi di equivalenza mutuamente esclusive ed esaustive.

Per ***frequenza*** si intende il numero di unità d’analisi presenti in ciascuna delle classi di equivalenza che sono state definite.

L’elenco delle classi e delle rispettive frequenze è chiamato ***distribuzione di frequenze***.

Una distribuzione di frequenze può essere rappresentata mediante una tabella delle frequenze.

Una ***tabella semplice*** classifica le unità di osservazione secondo le modalità di una sola variabile.

Una ***tabella a doppia entrata*** (o tabella di contingenza) classifica le unità di osservazione secondo le modalità di due variabili.

| | | |
|---|---|---|
| | | |
| Affiliazione politica | Repubblicani | 500 |
| | Indipendenti | 50 |
| | Democratici | 450 |

| | | Regione degli Stati Uniti | | |
|---|---|---|---|---|
| | | Nord Est | Sud Est | Ovest |
| Affiliazione politica | Repubblicani | 200 | 250 | 50 |
| | Indipendenti | 10 | 20 | 30 |
| | Democratici | 150 | 50 | 200 |

# ESERCIZIO

**Sulla base dei dati seguenti, costruite le due possibili distribuzioni di frequenza (tabelle semplici) e una tabella a doppia entrata.**

| Soggetto | Motivazione | Sesso |
|---|---|---|
| 1 | alta | m |
| 2 | media | f |
| 3 | alta | f |
| 4 | bassa | f |
| 5 | alta | m |
| 6 | media | f |
| 7 | alta | f |
| 8 | media | m |
| 9 | bassa | f |
| 10 | alta | f |

## Scala ordinale

In una distribuzione di frequenza per dati a livello nominale, l’ordine delle classi di equivalenza è arbitrario.

Per misure a livello ordinale, invece, le classi di equivalenza possono venire ordinate, solitamente in modo monotòno crescente.

| | | |
|---|---|---|
| Livello di felicità | Molto felice | 467 |
| | Abbastanza felice | 872 |
| | Piuttosto infelice | 165 |

| | | Regione degli Stati Uniti | | |
|---|---|---|---|---|
| | | Nord Est | Sud Est | Ovest |
| Livello di felicità | Molto felice | 185 | 149 | 133 |
| | Abbastanza felice | 412 | 215 | 245 |
| | Piuttosto infelice | 76 | 47 | 42 |

## Scala ad intervalli

Per misure ad intervalli *discrete*, si procede come nel caso di misure a livello ordinale.

Per misure ad intervalli *continue* è necessario dividere il campo di variazione in intervalli più piccoli (non necessariamente di grandezza costante).

Per ciascuna classe si distinguono:
- il ***limite inferiore***
- il ***limite superiore***
- *l'****intervallo di classe***: differenza tra limite superiore e inferiore

## ESEMPIO

Chiamiamo $h$ l’intervallo di classe e poniamo $h = 6$.

Consideriamo l’intervallo 100-105.

Questo sembra suggerire che il numero più piccolo che viene incluso in questa classe è  105 e il numero più grande è 105.

Le cose non stanno così, in quanto i numeri 99.6 o 105.2 sarebbero pure inclusi in questo intervallo.

Si distingue infatti tra *limiti espressi* e *limiti reali*.

**Limite reale inferiore =**

***limite espresso inferiore - .5 (differenza unitaria)***

**Limite reale superiore =**

***limite espresso superiore + .5 (differenza unitaria)***

L’espressione “differenza unitaria” si riferisce al grado di precisione delle nostre misure.  Se le nostre misure sono state arrotondate al numero intero più vicino (es., 103, 101, 104), allora la differenza unitaria è uguale ad un’unità.

**Intervallo espresso: 100  - 105**
**Intervallo reale:       99.5 - 100.5**

Se le nostre misure sono state specificate con una precisione di un decimale (es., 103.2, 101.4, 103,8), allora la differenza unitaria sarà uguale ad un decimo di unità.

**Intervallo espresso: 100    - 105**
**Intervallo reale:       99.95 - 100.05**

# Esempio 1.8 (p. 18), 1.11 (p. 23)

## FREQUENZE ASSOLUTE E RELATIVE

**Frequenze assolute:**
numero di osservazioni all’interno della *i*-esima classe.

**Frequenze relative:**
numero di osservazioni all’interno della *i*-esima classe diviso per il numero totale di osservazioni.

| Punteggio | Frequenza | Proporzione |
|---|---|---|
| 1 | 2 | ,042 |
| 2 | 3 | ,063 |
| 4 | 4 | ,083 |
| 5 | 1 | ,021 |
| 6 | 4 | ,083 |
| 8 | 7 | ,146 |
| 9 | 1 | ,021 |
| 10 | 6 | ,125 |
| 12 | 6 | ,125 |
| 14 | 3 | ,063 |
| 16 | 6 | ,125 |
| 18 | 3 | ,063 |
| 19 | 2 | ,042 |
| Totale | 48 | 1,0 |

# PROPORZIONI E PERCENTUALI

Per confrontare le distribuzioni di due campioni non aventi lo stesso numero di osservazioni è utile ricorrere alle proporzioni (o percentuali) anziché alle frequenze.

Una proporzione non è altro che una frequenza relativa: rapporto tra frequenza di una classe e numero totale di osservazioni del campione.

Una percentuale è una proporzione moltiplicata per 100.

**Punteggio**

| | | Frequenza | Percentuale |
|---|---|---|---|
| Validi | 1 | 2 | 4,2 |
| | 2 | 3 | 6,3 |
| | 4 | 4 | 8,3 |
| | 5 | 1 | 2,1 |
| | 6 | 4 | 8,3 |
| | 8 | 7 | 14,6 |
| | 9 | 1 | 2,1 |
| | 10 | 6 | 12,5 |
| | 12 | 6 | 12,5 |
| | 14 | 3 | 6,3 |
| | 16 | 6 | 12,5 |
| | 18 | 3 | 6,3 |
| | 19 | 2 | 4,2 |
| | Totale | 48 | 100,0 |

# DISTRIBUZIONI CUMULATIVE

La frequenza cumulativa di una data classe è uguale alla somma di tutte le frequenze delle classi precedenti più la frequenza della classe considerata.

$$F_j = \sum_{i=1}^{j} f_i$$

La distribuzione cumulativa delle frequenze è l'elenco che associa agli intervalli di classe, ordinati in senso crescente, le rispettive frequenze cumulative.

**Punteggio**

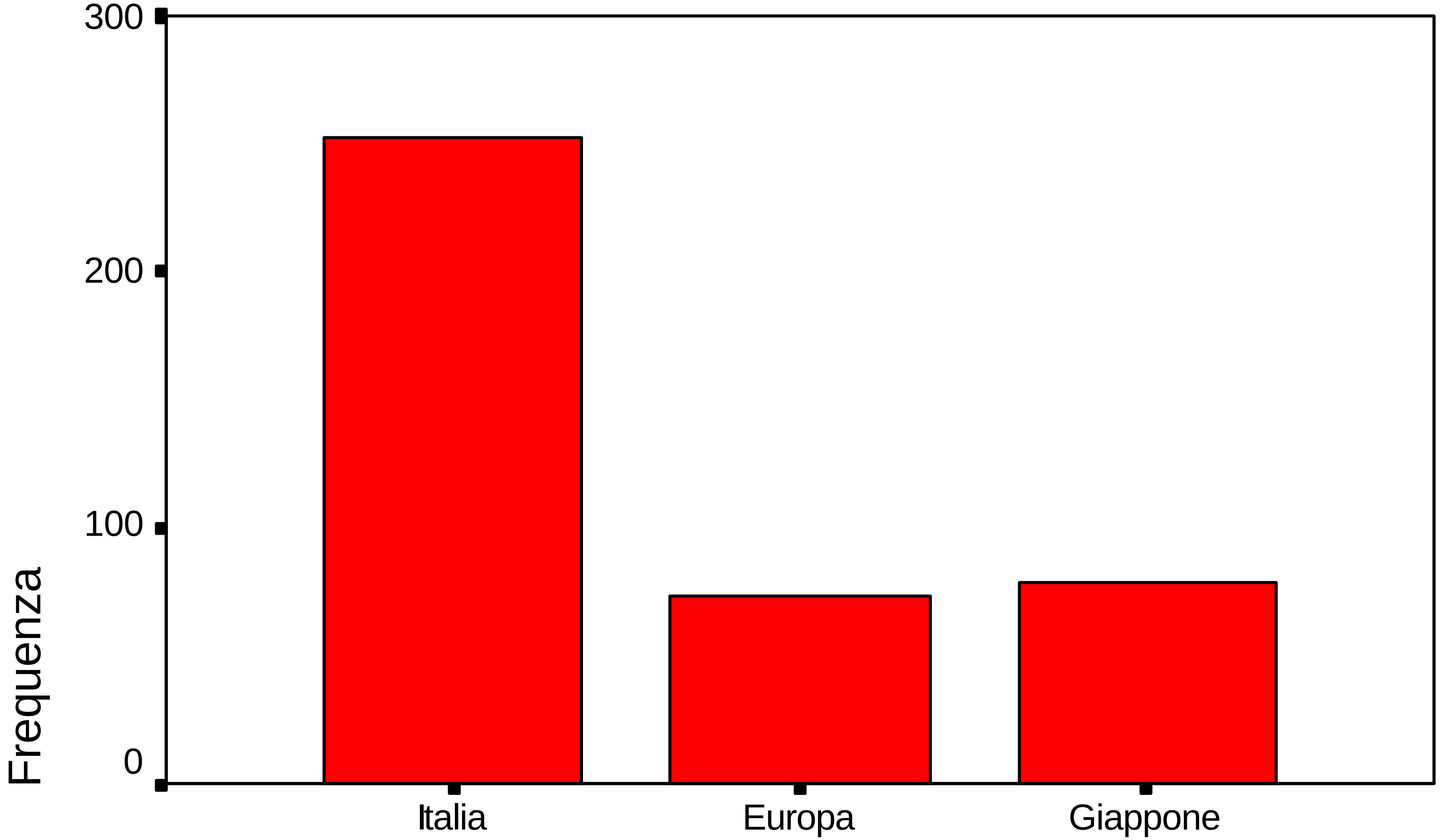
| | | Frequenza | Percentuale | Percentuale cumulata |
|---|---|---|---|---|
| Validi | 1 | 2 | 4,2 | 4,2 |
| | 2 | 3 | 6,3 | 10,4 |
| | 4 | 4 | 8,3 | 18,8 |
| | 5 | 1 | 2,1 | 20,8 |
| | 6 | 4 | 8,3 | 29,2 |
| | 8 | 7 | 14,6 | 43,8 |
| | 9 | 1 | 2,1 | 45,8 |
| | 10 | 6 | 12,5 | 58,3 |
| | 12 | 6 | 12,5 | 70,8 |
| | 14 | 3 | 6,3 | 77,1 |
| | 16 | 6 | 12,5 | 89,6 |
| | 18 | 3 | 6,3 | 95,8 |
| | 19 | 2 | 4,2 | 100,0 |
| | Totale | 48 | 100,0 | |

# RAPPRESENTAZIONE GRAFICA DELLE DISTRIBUZIONI DI FREQUENZE PER DATI QUALITATIVI

**Diagrammi a rettangoli distanziati:** altezze corrispondenti alle frequenze delle diverse classificazioni (Fig. 1.4, p. 29).

Automobili

Nazione di produzione

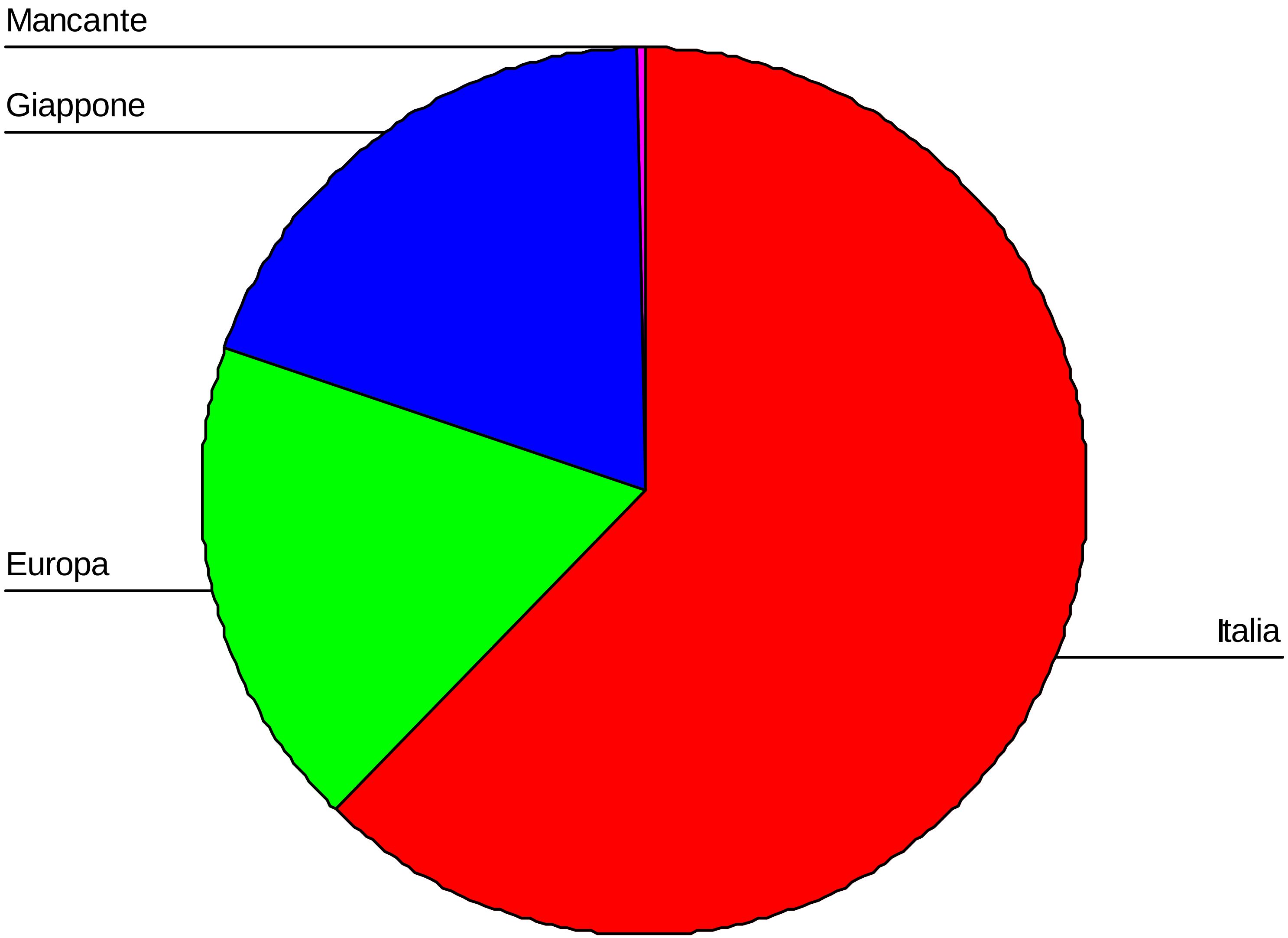



Nazione di produzione

**Diagrammi circolari:** usati per rappresentare frequenze percentuali di dati nominali, in modo da evitare la percezione illusoria di un ordinamento tra le classi che non esiste per dati di questo tipo (Fig. 1.4, p. 30).

$$\mathbf{360 : \alpha = 100 : f\%}$$

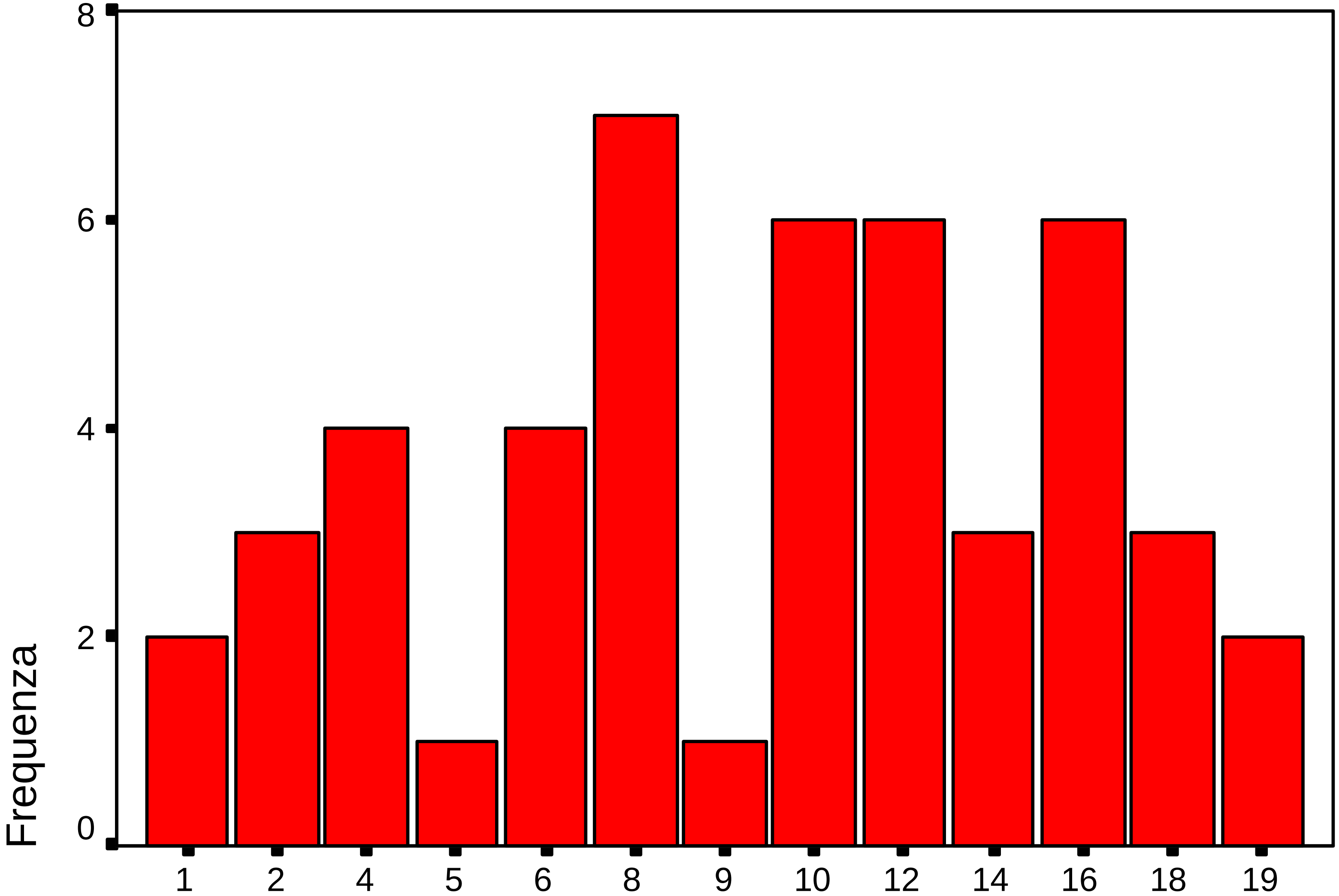
# Nazione di produzione



# RAPPRESENTAZIONE GRAFICA DELLE DISTRIBUZIONI DI FREQUENZE PER DATI QUANTITATIVI DISCRETI

**Diagrammi a barre:** in corrispondenza di ciascun valore $x_i$ (in ascissa) si innalza un segmento proporzionale alla frequenza del valore $x_i$ (Fig. 1.9, p. 33).

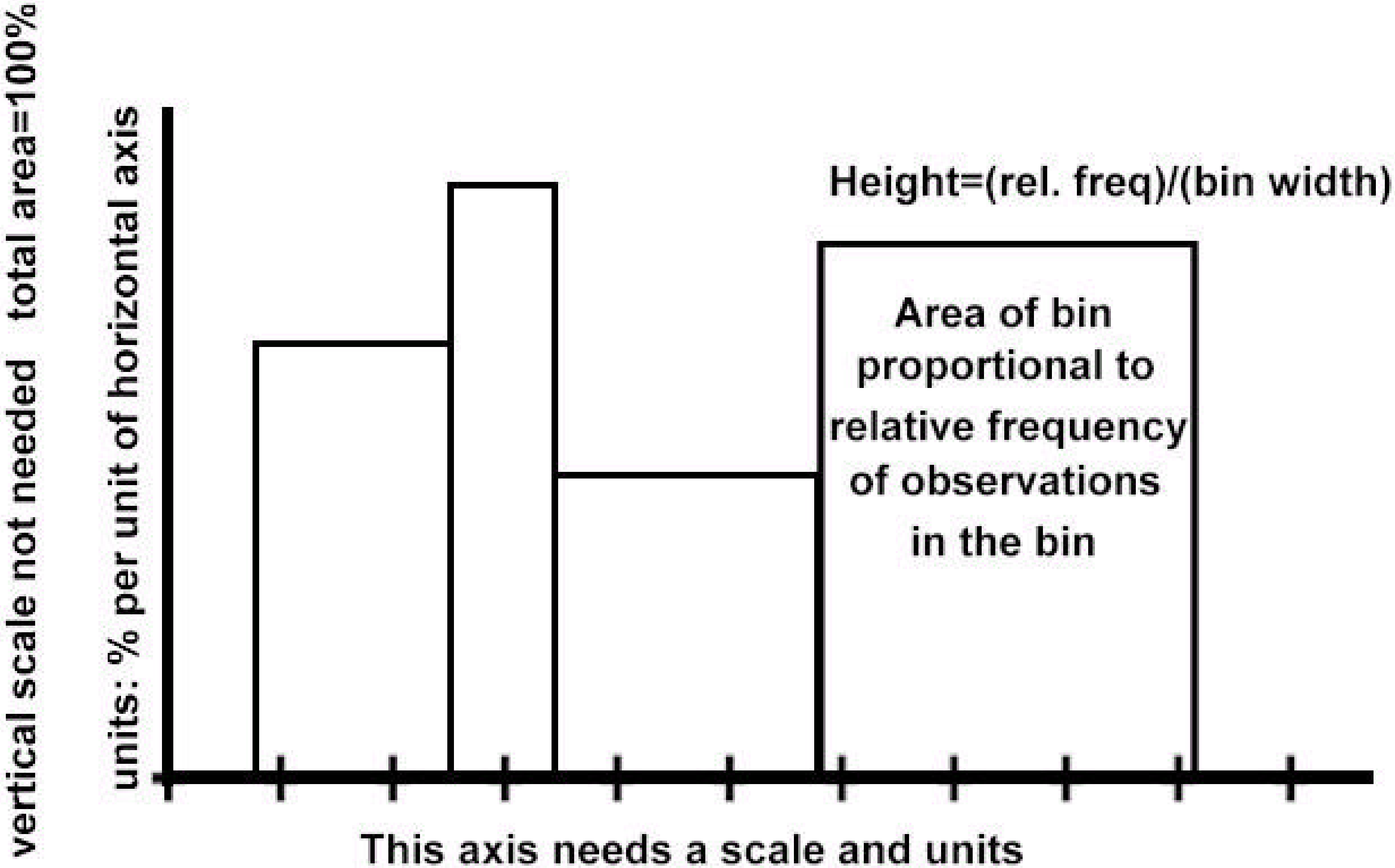
Punteggio



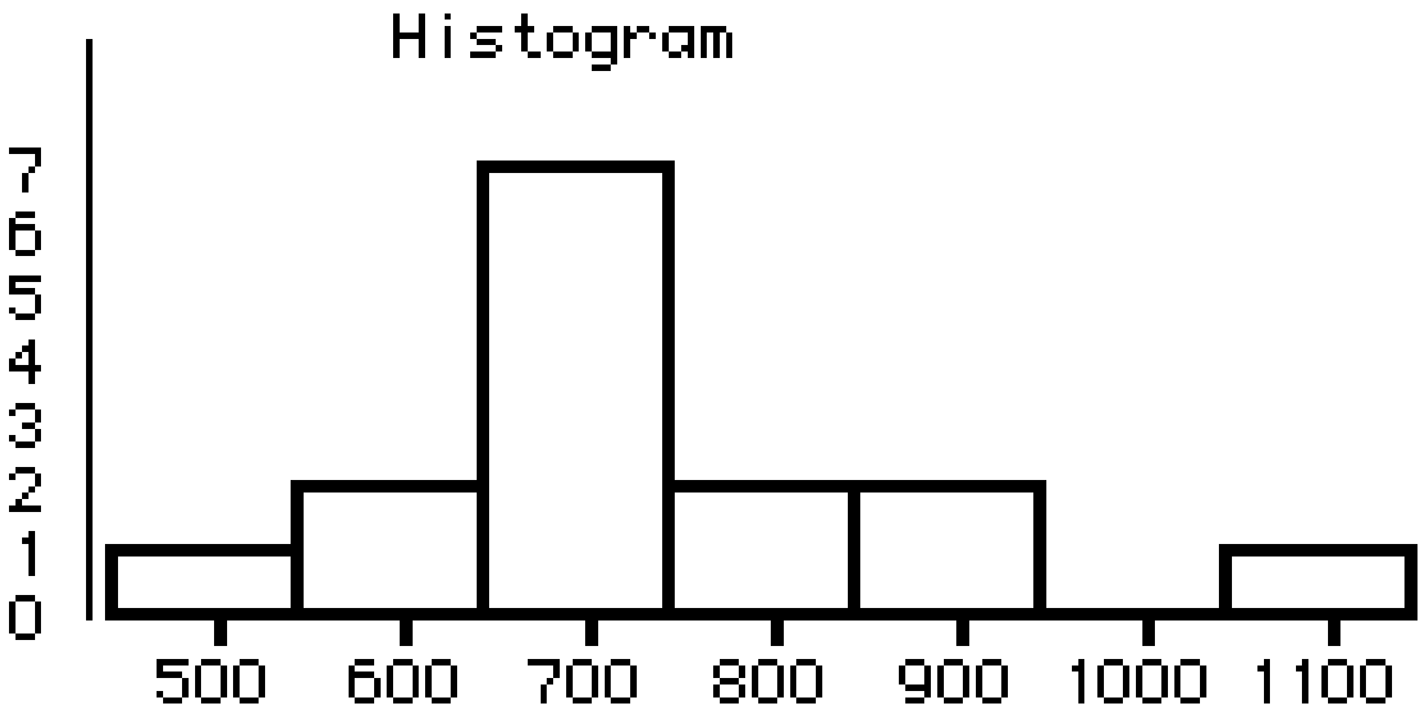

Punteggio

**Istogramma:** è costituito da rettangoli innalzati in corrispondenza dei dati che si considerano.
Le basi dei rettangoli hanno il centro corrispondente al centro dell'intervallo della classe e ampiezza uguale all'ampiezza della classe.

L'altezza del rettangolo è calcolata in modo tale che ***l'area del rettangolo*** sia ***uguale alla frequenza della classe*** (Fig. 1.7, p. 32).

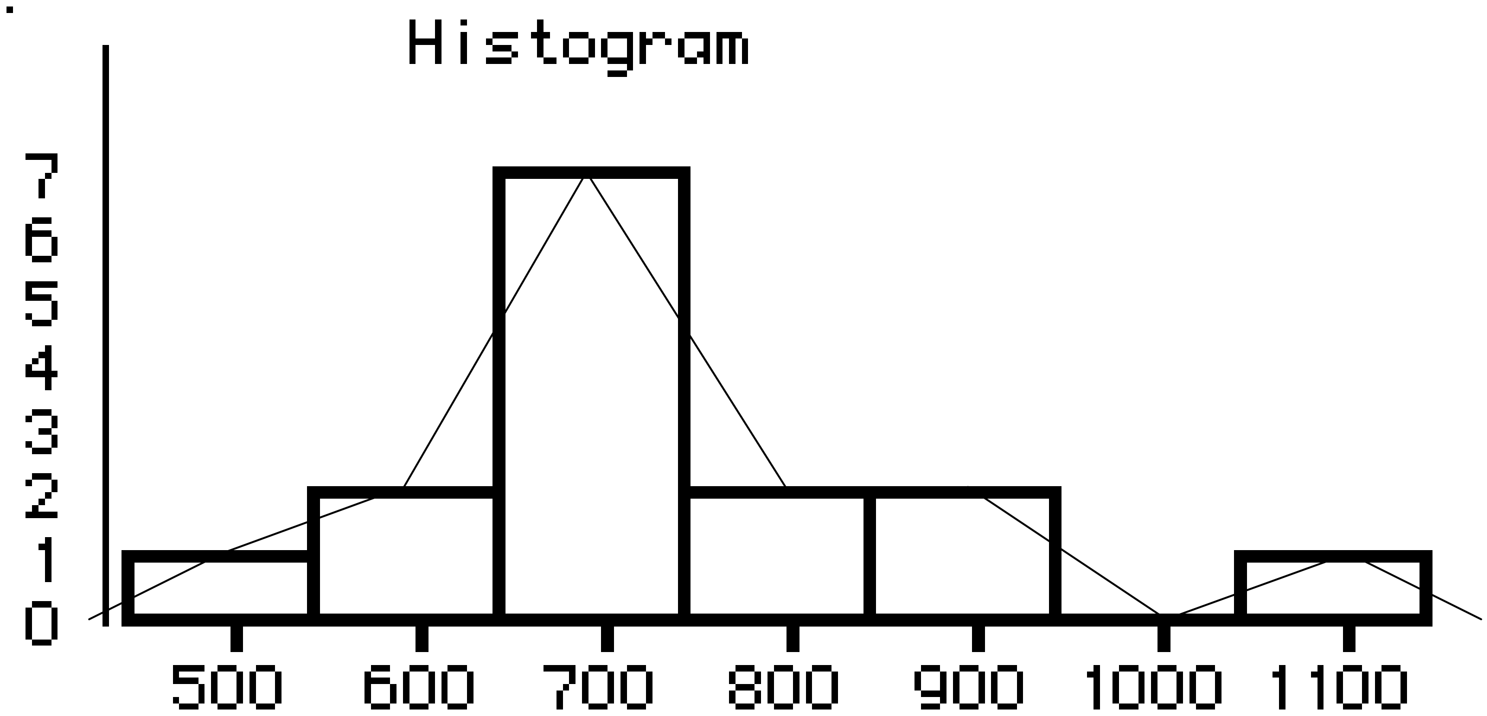

vertical scale not needed
total area=100%
units: % per unit of horizontal axis
Height=(rel. freq)/(bin width)
Area of bin proportional to relative frequency of observations in the bin
This axis needs a scale and units

**Target RT**
700
700
700
600
717
550
867
1067
883
567
833
683
800
750
683

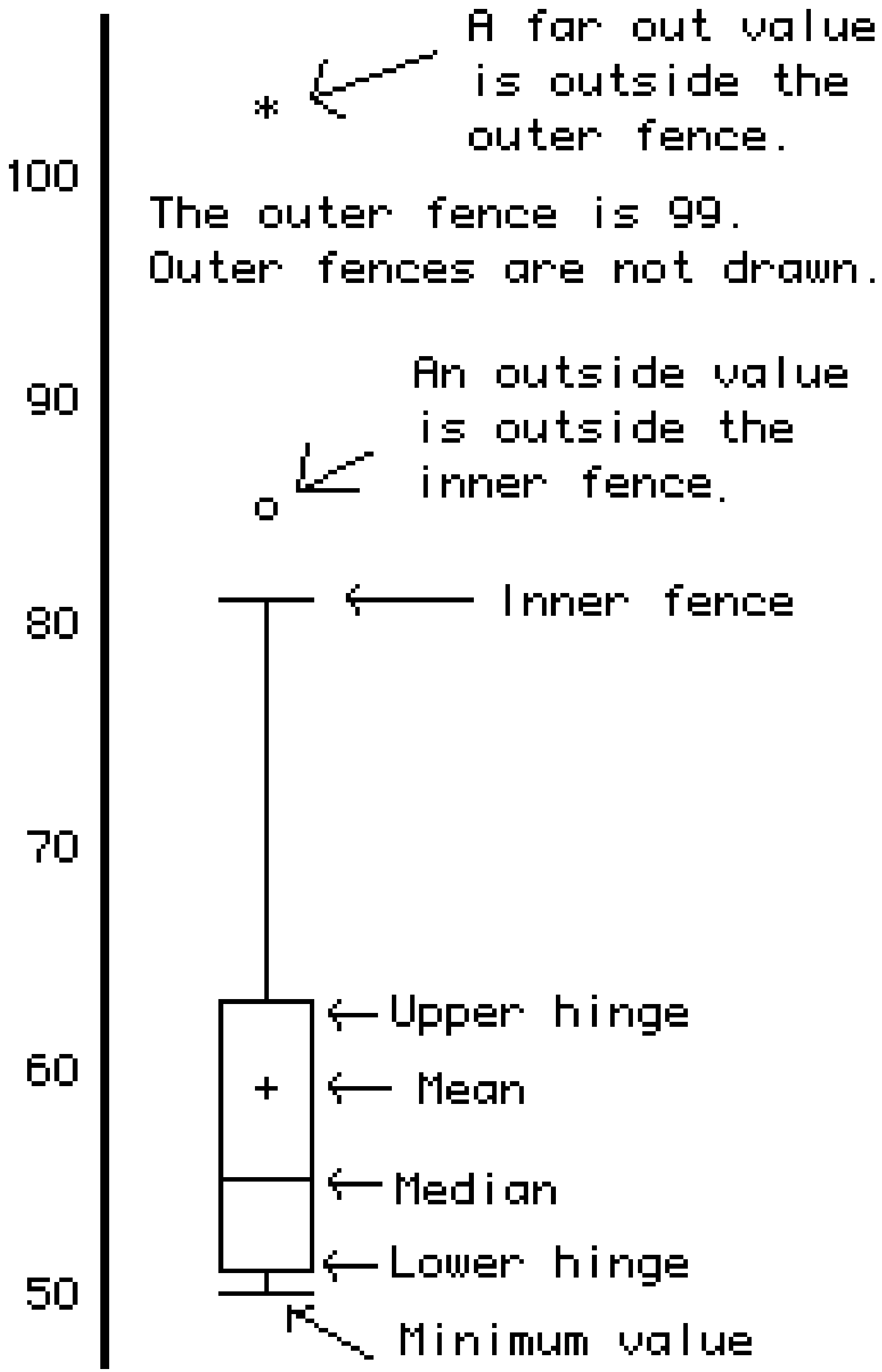


**Poligoni di frequenza:** sono costituiti da una spezzata che unisce i punti centrali delle basi superiori dei successivi rettangoli di un istogramma.

Aggiungendo due rettangoli fittizi di altezza zero all'inizio e alla fine dell'istogramma, si ottiene un poligono chiuso (Fig. 1.10, p. 34).


Histogram
7
6
5
4
3
2
1
0
500
600
700
800
900
1000
1100

# BOX PLOT

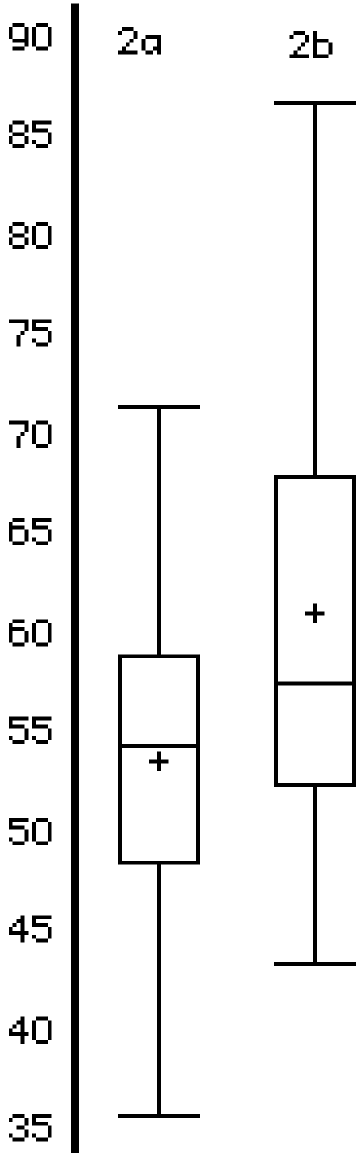



A box is drawn stretching from the lower hinge (defined as the 25th percentile) to the upper hinge (the 75th percentile). This box contains the middle half of the scores in the distribution.

The median is shown as a line across the box. Therefore 1/4 of the distribution is between this line and the top of the box and 1/4 of the distribution is between this line and the bottom of the box. The "H-spread" is defined as the difference between the hinges. A "step" is defined as 1.5 times the H-spread.

Inner fences are 1 step beyond the hinges.

Outer fences are 2 steps beyond the hinges.




Lines (whiskers) are drawn from the box to the smallest and largest values or to the inner fence (whichever is closer). The mean of the distribution is indicated by a plus sign (+). Every score between the inner and outer fences is indicated by an "o"; a score beyond the outer fences is indicated by a "*".

It is often useful to compare data from two or more groups by viewing boxplots from the groups side by side. Plotted are data from Example 2a and Example 2b . The data from 2b are higher, more spread out, and have a positive skew. That the skew is positive can be determined by the fact that the mean is higher than the median and the upper whisker is longer than the lower whisker.

# RAPPRESENTAZIONE GRAFICA DELLE DISTRIBUZIONI DI FREQUENZE PER DATI QUANTITATIVI CONTINUI

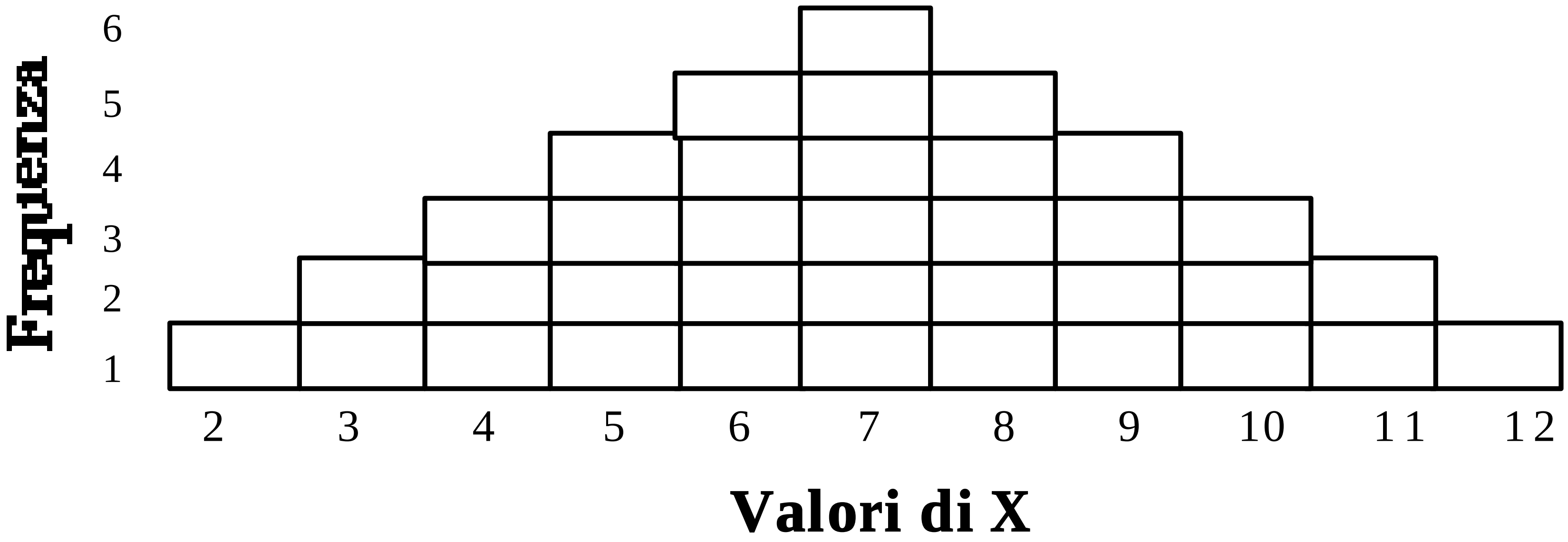
Nel caso di una variabile continua, avremo un numero infinito di osservazioni. Questo significa che, a ciascun intervallo per quanto piccolo della variabile considerata, sarà associata una frequenza diversa da zero.

Per dati di questo tipo, possiamo dunque disegnare un istogramma facendo diminuire sempre di più l'ampiezza degli intervalli di classe. Così facendo otterremo dei rettangoli le cui basi avranno una larghezza sempre minore.

Tanto più l'ampiezza delle classi tenderà a zero, tanto più il poligono di frequenza di un tale istogramma tenderà alla ***curva di frequenza*** di una variabile continua.

Consideriamo la distribuzione di frequenza di una variabile discreta X con $n$ = 36. In corrispondenza di ciascuno dei valori che X può assumere (X=2, X=3, ..., X=12) poniamo un numero di blocchi corrispondenti alla frequenza relativa di quella classe.

Ciascuno dei blocchi posti sopra i valori che X può assumere ha la stessa area. Notiamo dunque che il rapporto tra il numero di blocchi della colonna i-esima e il numero totale di blocchi è uguale alla frequenza relativa della classe X=i.




Abbiamo usato 36 blocchi (dato che $n$ = 36).

Definiamo l'area di ciascun blocco uguale a 1/36.

(1) l'area totale è uguale a 1.0

(2) <u>l'area relativa di ciascuna colonna è uguale alla **frequenza relativa**</u>

Es., la frequenza relativa della classe X=4 è 3/36.

Estendiamo ora questa rappresentazione al caso di una variabile continua.

Consideriamo una variabile $Y$ i cui valori sono ragruppati in intervalli. L'area totale dell'istogramma che rappresenta la distribuzione di $Y$ è uguale a 1.0.

Dunque, l'area di ciascuna barra rappresenterà anche in questo caso la frequenza relativa della variabile $Y$ per l'intervallo considerato.

Supponiamo ora di scegliere degli intervalli più piccoli di Y, e supponiamo inoltre disporre di talmente tante osservazioni per cui la frequenza relativa associata a qualsiasi intervallo non sarà mai uguale a zero, per qualsivoglia piccolo intervallo in Y.

Se procediamo così, succederà che il numero degli intervalli diventerà sempre più grande, e l'area della barra corrispondente a ciascun intervallo diventerà sempre più piccola.

L'area totale dell'istogramma non cambierà se cambiamo le dimensioni degli intervalli (e dunque sarà sempre uguale a 1), ma l'area di ciascun intervallo (ovvero la frequenza relativa) diventerà sempre più piccola.

Immaginiamo ora di racchiudere l'istogramma con una curva continua che si approssima quanto più possibile all'istogramma, <u>in modo tale che l'area della curva sia uguale all'area dell'istogramma</u>.

Consideriamo ora l'area di una barra dell'istogramma corrispondente all'intervallo $\Delta X$.

Se l'intervallo $\Delta X$ è abbastanza piccolo, allora l'area della barra sarà simile all'area sottesa alla curva.

In altre parole, l'area sottesa alla curva in un intervallo sarà molto simile alla <u>frequenza relativa</u> associata a quell'intervallo.

Immaginiamo di continuare a diminuire la grandezza di ciascun intervallo in modo tale da diminuire la differenza tra la frequenza relativa di quell'intervallo e l'area sottesa alla curva in quell'intervallo.

Con il crescere del numero di intervalli, questa differenza diventerà sempre più piccola.

Quando il numero di intervalli diventa infinito, non vi sarà più nessuna differenza tra l'area sottesa alla curva e la frequenza relativa in quell'intervallo.

**Allo stesso tempo, però, l'area di ciascun intervallo sarà diventata uguale a zero.**

Questo è un importante aspetto che differenzia le variabili continue da quelle discrete.

Nel caso di una variabile discreta, infatti, è sempre possibile associare una frequenza relativa maggiore di zero alla classe che rappresenta i casi per cui la variabile $Y$ assume esattamente il valore $y$.

Nel caso di una variabile continua $Y$, invece, non è possibile associare una frequenza relativa diversa da zero alla classe che rappresenta i casi $Y = y$.

Abbiamo visto infatti che, con il diminuire dell'intervallo $DY$, l'area sottesa alla curva in quell'intervallo tende a zero. Questo significa dunque che, <u>per una variable continua, la frequenza relativa di $Y$ = y è uguale a zero.</u>

Per questa ragione, nel caso di una variabile continua *Y* non ha senso chiedersi quale è la frequenza relativa dell'evento $Y = y$ (in quanto questa frequenza relativa è sempre uguale a zero).

Invece ci chiediamo quale è la frequenza relativa con la quale *Y* assume un valore contenuto all'interno di un certo intervallo.

Questi invervalli di valori della variabile corrispondono alle aree sottese da curve come quella che abbiamo descritto in precedenza, per intervalli *DX* diversi da zero.

In conclusione, nel caso di una variabile discreta la frequenza relativa associata all'evento tale per cui *X* assume un valore compreso in un intervallo $a \le X \le b$ è la somma delle frequenze relative associate a ciascuno dei valori di *X* compresi tra *a* e *b*.

In maniera simile, per una variabile continua, la frequenza relativa che *Y* assume un valore compreso in un intervallo *a*, *b* è data dalla somma di tutte le frequenze relative associate agli infinitesimi intervalli *DY* che ci sono tra *a* e *b*.

Dato che ci sono infiniti intervalli tra questi due valori, la somma di questi intervalli si indica con un integrale:

$$f_r(a \leq Y \leq b) = \int_a^b f(y)dy$$

In questa espressione, *f(y)* può essere interpretata come l'altezza e *dy* come la larghezza dell'area creata da ciascuno dei minuscoli intervalli *DY* che sono compresi tra *a* e *b*.

# FORMA DELLE CURVE DI FREQUENZA

(Fig. 1.14, p. 38)

**Curva simmetrica**

**Asimmetria positiva**

**Asimmetria negativa**

**Curva crescente**

**Curva decrescente**

**Curva a U**

**Curva bimodale**

**Curva plurimodale**

- **Variabili**
- **Sommatorie**
- **Scale di misura**
- **Distribuzioni di frequenze**
- **Rappresentazione grafica dei dati**