

Giacomo Bulgarelli
Ufficio Servizi Statistici



SERVIZIO DAF: FONTI STATISTICHE

Mercoledì 3 ottobre 2012

2. La Statistica (I)

La Statistica

*Sai ched'è la statistica? È 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.
Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
secondo le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due.
(Trilussa)*

La Statistica

L'istruttore dice agli allievi paracadutisti: «Oggi impariamo a lanciarsi. Nessuna paura. Le statistiche dicono che avviene un incidente ogni 500 lanci... e siccome voi siete 200 non correte nessun rischio!!!»

Uno statistico può mettere la testa nel forno e i piedi nel ghiaccio e dire che in media si sente bene!

È facile mentire con la statistica, ma è difficile dire la verità senza di essa.

La statistica è come un bikini: ciò che rivela è suggestivo, ma ciò che nasconde è vitale.

La Statistica

Secondo le statistiche l'uomo mangia una prugna ogni venti secondi. Non so chi sia 'sto tizio, ma so dove trovarlo!!!

Lo statistico è colui che viaggia con una bomba nella valigia... Perché due bombe nello stesso aereo è statisticamente impossibile!!!

Le statistiche sostengono che più tempo si passa in automobile sulle strade e più aumenta la probabilità di incidenti; la prudenza consiglia quindi di possedere un'auto molto veloce e di correre a tavoletta!!!

Il fumo è una delle principali cause di statistica.

Etimologia

STATO

Le prime informazioni sui fenomeni reali (per motivi militari, religiosi, economici, sociali, sanitari, ...) sono state raccolte ed organizzate ad opera degli organismi statali, che ne erano anche i principali utilizzatori.

Egitto - I dinastia (3000 a. C.) - Rilevazione ammontare popolazione

II-XX dinastia (2925 a. C. - 1078 a. C.) - Rilevazione di vari beni a fini fiscali, dei componenti delle famiglie dei soldati, dei dipendenti statali, delle merci, delle abitazioni e dei loro abitanti.

Etimologia

STATO

Israele - Libro dei Numeri - Censimento della popolazione nel deserto del Sinai (XII sec. a. C.)

Roma - Servio Tullio, primo censimento - Censimento quinquennale dalla fine del VI sec. a. C., decennale da Augusto

Cina - epoca Ming (XIV sec.) - Censimento decennale

Concilio di Trento (1563): obbligo per i parroci di tenere registri aggiornati di battesimi, matrimoni e, dopo mezzo secolo, l'obbligo si estende alla registrazione dei morti

Etimologia

STATO

La prima apparizione del vocabolo "Statistica" - inteso come raccolta di informazioni organizzate e gestite dallo Stato - si deve a Ghislini (1859), che indica la Statistica come la "descrizione delle qualità che caratterizzano e degli elementi che compongono uno Stato".

Etimologia

STATO

La formalizzazione matematica della statistica è recente:

- XVIII e XIX secolo: calcolo delle probabilità;
- prima metà del XX secolo: inferenza statistica, disegno degli esperimenti, campionamento;
- anni '40 – '70: sviluppi teorici;
- dagli anni '70: sviluppi legati alle capacità di calcolo del computer.

Che cos'è

Scienza delle decisioni in condizioni di incertezza

La Statistica è una disciplina che ha come fine lo studio quantitativo e qualitativo di un particolare fenomeno “collettivo”, in modo da sintetizzarne e comprenderne le caratteristiche. Analizza la realtà (fisica, biologica, economica, sociale, ...) allo scopo di trarne leggi e regole generali per obiettivi predefiniti.

Lo statistico è una sorta di «investigatore» che usa la metodologia statistica e la tecnologia informatica per individuare degli **andamenti**, delle **regolarità** nella complessa realtà che lo circonda.

Come i telescopi, i microscopi, i raggi X e i radar, la Statistica moderna consente di vedere cose invisibili a occhio nudo (David Hand).

Che cos'è

Scienza delle decisioni in condizioni di incertezza

Nel linguaggio comune, con il termine statistica si indica anche il risultato numerico di un processo di sintesi dei dati osservati: tabelle, grafici, indici, ...

Ad es., degli studenti dell'Università di Firenze potremmo essere interessati a conoscerne il numero complessivo, il numero medio di crediti conseguiti in un anno, la variabilità della loro durata degli studi ... Ognuna di queste informazioni è detta anche statistica.

A cosa serve

L'obiettivo specifica la natura delle informazioni da raccogliere e gli strumenti con cui esaminare i dati.



Definito l'**obiettivo**, una corretta indagine inizierà raccogliendo, nel modo più rigoroso possibile, **informazioni** utili connesse con l'obiettivo; successivamente, grazie ai **metodi** propri dell'analisi statistica, il decisore otterrà dei **risultati** che aiuteranno a specificare e rendere efficace il perseguimento degli obiettivi prefissati, concretizzandosi in azioni ben visibili.

A cosa serve



Come in tutte le scienze, nella Statistica si possono ritrovare risultati scientifici notevoli dedotti con modelli rigorosi, ma anche autentiche falsificazioni derivanti da faciloneria, ignoranza o malafede.

I dati e le informazioni non costituiscono verità assolute, ma solo la realizzazione di un'ipotesi che quei dati devono mettere in discussione.

Senza modelli falsificabili dall'esperienza, i dati statistici sono numeri privi di senso. Al di fuori di uno schema concettuale rigoroso, non esiste interesse scientifico, né utilità pratica per qualunque analisi statistica.

Fasi dell'analisi statistica



- 1 - Definizione degli obiettivi della ricerca**
- 2 - Rilevazione dei dati**
- 3 - Elaborazione metodologica**
- 4 - Presentazione ed interpretazione dei risultati**
- 5 - Utilizzazione dei risultati della ricerca**

Fasi dell'analisi statistica



1 - Definizione degli obiettivi della ricerca

Gli **obiettivi** devono essere particolareggiati fino ad individuare con cura le informazioni da ricercare, evitando equivoci nelle definizioni, circoscrivendo con esattezza il territorio e il periodo dell'indagine, fornendo soluzioni alternative in caso di mancate o errate risposte. Per definire correttamente gli obiettivi di una ricerca occorrono studi preliminari per circoscrivere l'ambito dell'indagine e le possibili modalità di risposta.

Fasi dell'analisi statistica



2 - Rilevazione dei dati

La **rilevazione dei dati** può essere **completa** (quando si esaminano tutti gli elementi oggetto di studio) o **parziale** (quando ci si limita a studiare un sottoinsieme, detto **campione**, dell'insieme di riferimento). È inoltre essenziale esplicitare il modo in cui si raccolgono le informazioni, cioè tramite *dichiarazioni* (questionario, intervista, ...) o *misurazioni* (con strumenti), perché la modalità di raccolta di dati determina spesso l'accuratezza delle informazioni su cui si opera, essendo differenti le tipologie di errori, la possibilità di dati mancanti, la presenza di dichiarazioni non veritiere, ...

Fasi dell'analisi statistica



3 - Elaborazione metodologica

In questa fase si applicano schemi formali di tipo logico-deduttivo e inferenziale-induttivo, di natura matematica, cioè gli strumenti propri dell'analisi statistica. È complesso definire univocamente la metodologia più idonea per ciascuna indagine e, molto spesso, una pluralità di approcci consente valutazioni e conferme più accurate; in tal senso, la distinzione tra fenomeni qualitativi e quantitativi gioca un ruolo decisivo per la scelta dei metodi che si possono applicare.

Fasi dell'analisi statistica



4 - Presentazione ed interpretazione dei risultati

Lo statistico deve porre particolare cura nella **presentazione dei risultati** (sotto forma di tabelle, grafici, diagrammi di relazioni, rapporti sintetici) e nell'esame delle implicazioni operative dei medesimi, con particolare riferimento all'interpretazione propria del settore in cui opera. Infatti, gran parte dei decisori economici, finanziari e politici è raramente consapevole delle potenzialità e dei limiti del metodo statistico che, talvolta, viene esaminato con sufficienza o acriticamente a causa di una presentazione poco efficace dei risultati dell'indagine.

Fasi dell'analisi statistica



5 - Utilizzazione dei risultati della ricerca

Lo statistico contribuisce ad una **corretta utilizzazione dei risultati** di un'indagine scientifica, circoscrivendo l'ambito interpretativo e richiamando i vincoli entro cui essa assume validità. L'uso dei risultati di una ricerca è quindi connesso anche alle sue modalità di esecuzione e ai dettagli mediante i quali è stata condotta la misurazione o l'intervista (occasioni, tempi, luoghi, ...), perché solo precisando questi aspetti è lecito ampliare l'interpretazione dei risultati.

Terminologia

“Definire” significa spiegare il significato di vocaboli mediante altri vocaboli di significato noto

- **Rilevazione statistica**
- **Popolazione**
- **Unità statistica**
- **Carattere**
- **Modalità**
- **Frequenza**

Terminologia

Rilevazione statistica

È il complesso di operazioni volte ad acquisire una o più informazioni su un insieme di elementi oggetto di studio.

Età di uno studente	semplice	complessa	Bilancio Università
Questionario VDS	risposte	misure con strumenti	Metro, bilancia, orologio
Censimento	globale	parziale	Campione

Terminologia

Popolazione

(o Universo) è qualsiasi insieme di elementi, reale o ipotetico, che forma l'oggetto di uno studio statistico. L'insieme può essere definito elencando tutti i suoi elementi, oppure indicando le proprietà che ciascun elemento dell'insieme deve possedere.

reale



Effettivamente
esistente e visibile

- *Professori associati Unifi*
- *Lampadine prodotte nell'ultimo mese da un'azienda fiorentina*
- *Spettatori paganti Italia-Germania del 27/06/2012*

virtuale



Definibile con accuratezza, ma
non osservata, né osservabile

- *Possibili cinque estraibili sulla ruota di Milano*
- *Voti conseguibili da uno studente all'esame di Affidabilità e controllo di qualità*
- *Parole italiane di senso compiuto estraendo 6 lettere consecutivamente dall'alfabeto di 21 lettere*

Terminologia

Unità statistica

È l'elemento di base della popolazione sul quale viene effettuata la rilevazione o la misurazione di uno o più fenomeni oggetto dell'indagine.

semplice



Una persona
Un calciatore
Un albero

composta



Una famiglia
Una squadra
Un bosco

Terminologia

Carattere

(o variabile) è il fenomeno oggetto dello studio, rilevato o misurato sulle unità statistiche. È tipico dei fenomeni reali di interesse statistico che i caratteri assumano valori differenti nelle varie unità statistiche.

Sulle popolazioni, e quindi sulle unità statistiche, possono essere rilevati uno o più caratteri di interesse: genere, religione, età, professione, altezza, titolo di studio, reddito mensile, stato civile, ...

Terminologia

Modalità

È l'espressione concreta del carattere nelle unità statistiche, cioè il numero (per caratteri quantitativi) o l'attributo (per caratteri qualitativi) che l'unità statistica manifesta. L'elenco di tutte le possibili modalità di un carattere si dice **esaustivo** se è completo e le modalità si dicono **disgiunte** se una unità statistica può manifestare il carattere in una ed una sola modalità tra quelle indicate.

GENERE - F, M

VOTO AGLI ESAMI - 18, 19, 20, 21, 22, ... , 27, 28, 29, 30, 30 e lode

POSIZIONE ECONOMICA – B1, B2, B3, B4, ... , EP4, EP5, EP6, EP7

Esempio 1

In un'indagine sugli immatricolati al I anno della facoltà di Economia di Firenze, Giacomo ha dichiarato di aver preso 28 all'esame di Statistica.

Popolazione

Studenti immatricolati al I anno a Economia

**Unità
statistica**

Giacomo, perché è un elemento che appartiene alla popolazione in esame

Carattere

Voto all'esame di Statistica

Modalità

28, cioè una delle modalità genericamente attribuibili agli elementi della popolazione esaminata

Terminologia

Frequenza

È il numero di volte in cui una determinata modalità si verifica nel collettivo di riferimento. La distribuzione delle frequenze evidenzia come il fenomeno si manifesta nella popolazione in rapporto ad un criterio di ordinamento delle frequenze. Quando la frequenza è un numero ≥ 0 si parla di **frequenza assoluta**, mentre quando tale frequenza è rapportata al totale delle unità statistiche della popolazione si parla di **frequenza relativa**.

La **distribuzione di frequenza** è un'organizzazione dei dati in forma tabellare tale che ad ogni modalità del carattere fa corrispondere la rispettiva frequenza. Essa *esplicita quante volte una determinata modalità si presenta nel collettivo in esame*. Si tratta, quindi, di un modo di presentare le unità statistiche che manifestano la stessa modalità sull'intera popolazione, enumerando le osservazioni per modalità coincidenti.

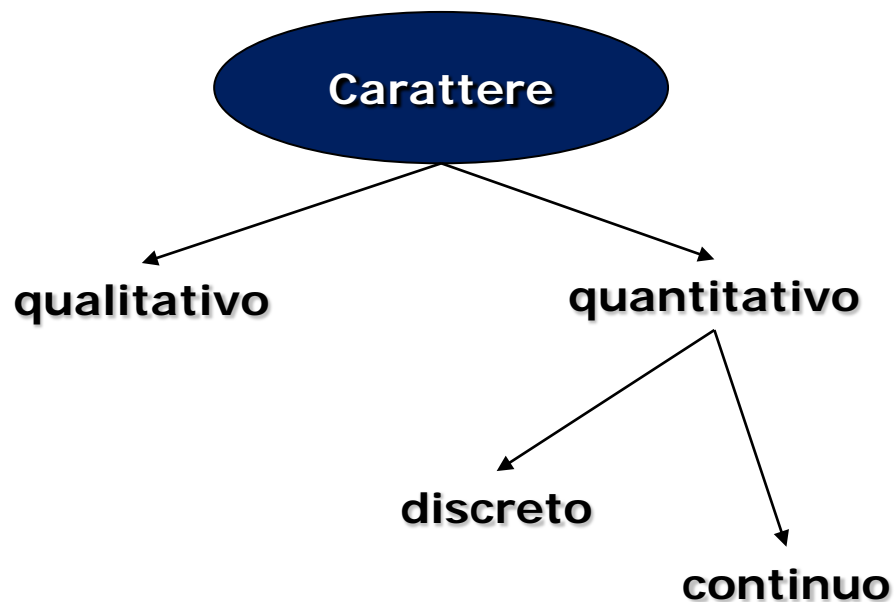
Esempio 2

Distribuzione dei docenti dell'Università degli Studi di Firenze per tipologia al 31/07/2012 (valori assoluti e %).

	Frequenza assoluta	Frequenza relativa (%)
Ordinari	553	$553/1.824 * 100 = 30,32$
Associati	623	$623/1.824 * 100 = 34,16$
Ricercatori	648	$648/1.824 * 100 = 35,53$
TOTALE	1.824	$1.824/1.824 * 100 = 100,00$

Tipologia delle informazioni statistiche

Per **informazione statistica** si intende ogni risultato ottenuto da un'indagine sui collettivi esaminati (**popolazione** o **campione**) rispetto ai loro costituenti (**unità statistiche**) e in rapporto ad uno o più fenomeni (**caratteri**).



Tipologia delle informazioni statistiche

Carattere qualitativo

Si definisce **mutabile** se assume per modalità attributi non numerici della più varia natura.

Il genere, la religione, il colore degli occhi, lo stato civile, la professione, la categoria sociale di appartenenza, il comune di nascita, la regione di residenza, la nazionalità, il gradimento di una canzone, la reazione di fronte a un quadro, ... costituiscono esempi di caratteri qualitativi, cioè di mutabili, perché per essi ogni associazione con numeri reali è arbitraria, se anche possibile o talvolta conveniente.

Tipologia delle informazioni statistiche

Carattere quantitativo

Si definisce **variabile** se assume per modalità numeri reali.

A loro volta, i caratteri quantitativi si distinguono in **continui** e **discreti**. Le variabili continue assumono qualsiasi valore contenuto in un intervallo reale predefinito, cioè le modalità possono essere poste in corrispondenza con l'insieme dei punti di un intervallo della retta reale; le variabili discrete, invece, possono assumere al più un numero discreto di modalità, finito o infinito, nel senso che possono essere poste in corrispondenza con l'insieme dei numeri naturali $\{1, 2, 3, \dots\}$ o un suo sottoinsieme.

Tipologia delle informazioni statistiche

Carattere quantitativo continuo

L'età, il peso, l'altezza, la temperatura di una stanza, il tempo impiegato per raggiungere la sede di lavoro, qualsiasi grandezza fisica misurabile costituiscono variabili continue, perché possono assumere qualsiasi valore in un intervallo predefinito, né esistono motivi concettuali per cui, ad es., fissati due pesi estremamente vicini, non sia possibile trovare una persona con peso intermedio a quei pesi. Tuttavia, la possibilità di valutare punti estremamente vicini dipende dagli strumenti di misura, che consentono misure accurate solo fino ad un certo punto.

Tipologia delle informazioni statistiche

Carattere quantitativo discreto

Il numero di clienti in attesa ad uno sportello postale, i nati vivi in provincia di Firenze, il numero di lanci di un dado necessari perché si verifichi [4] per la prima volta, il voto di laurea, il numero dei componenti di una famiglia ... costituiscono esempi di variabili discrete, perché concettualmente le possibili modalità di tali caratteri sono elencabili e, quindi, possono essere messe in corrispondenza con i numeri naturali.

Esempio 3

Stabilire la tipologia dei seguenti caratteri

Numero di aule del Polo di Novoli

quantitativo discreto

Colore fascia dei Presidi

qualitativo

Età dei docenti dell'Ateneo fiorentino

quantitativo continuo

Regione di residenza degli immatricolati

qualitativo

Posizione economica del personale T-A

qualitativo

Esito dell'esame di Statistica I

qualitativo

Voto di laurea di uno studente

quantitativo discreto

Soddisfazione del corso di studi

qualitativo

Durata del neon dell'ufficio

quantitativo continuo

Giorni di assenza dal lavoro

quantitativo discreto

Tipologia delle informazioni statistiche

La distinzione tra variabili discrete e continue è anzitutto **concettuale**, perché la limitatezza degli strumenti di misura e del tempo necessario per discriminare osservazioni estremamente vicine, rende qualsiasi variabile, anche teoricamente continua, di fatto discreta.

L'altezza di una persona adulta è una variabile continua, ma la convenzione suggerisce di attribuire il valore in cm più vicino a quello vero: pertanto una persona sarà alta 176 cm, se la sua altezza è compresa nell'intervallo $[175,5, 176,5)$.

Di contro, talvolta si operano estensioni al continuo per quelle variabili concettualmente discrete le cui modalità sono così numerose e vicine tra loro che è più semplice considerarle come continue.

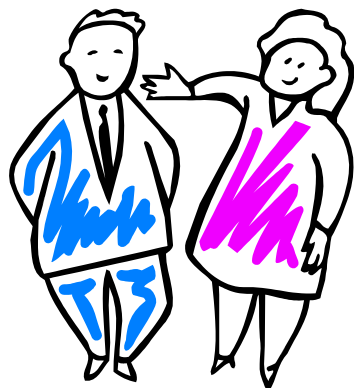
Il reddito dei lavoratori dipendenti in Italia è concettualmente una variabile discreta, perché sono sempre elencabili tutte le modalità che esso può assumere; in pratica, trattandosi di una popolazione con numerosità molto elevata, è più semplice studiarne le caratteristiche come variabile continua che assume valori in $[0, +\infty]$.

Scale di misurazione

Nell'ambito della teoria della misurazione, lo psicologo Stevens (1946) introdusse una classificazione dei caratteri statistici che ad oggi è ancora in uso.

I caratteri vengono distinti in funzione della **scala di misurazione: nominale, ordinale, ad intervallo, di rapporto**.

I caratteri nominali e ordinali sono qualitativi (sono, cioè, delle mutabili), mentre i caratteri che ammettono una scala ad intervallo o di rapporto sono quantitativi (sono, quindi, delle variabili in senso stretto).



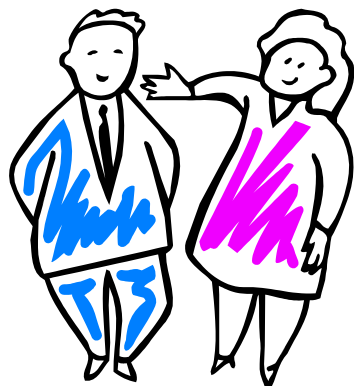
Scale di misurazione

Scala nominale

I caratteri con scala nominale (= caratteri nominali o “sconnessi”) costituiscono mutabili le cui modalità non assumono alcun ordine precostituito. L'unico confronto ammissibile tra due unità statistiche rispetto a caratteri nominali consiste nello stabilire se possiedono o meno lo stesso attributo, cioè se sono uguali oppure diversi rispetto a quella mutabile.

$$x_i = x_j \ ; \ x_i \neq x_j$$

Moda



Scale di misurazione

Scala nominale

Il genere, la religione, la professione, i giorni della settimana, la nazionalità costituiscono esempi di caratteri i cui attributi (maschio e femmina; cristiani, ebrei, musulmani, induisti, ...; lunedì, martedì, ..., domenica; italiana, francese, spagnola, inglese, tedesca, ...) non possiedono un ordine naturale, né esistono motivi logico-concettuali per preferire un ordinamento tra le modalità.

Un carattere si dice **dicotomo** (o dicotomico) se può assumere due sole modalità, convenzionalmente indicate con 0 e 1 (ad es. acceso – spento; sì – no; superato – non superato; ...).

Un carattere si dice **politomo** (o politomico) se può assumere un numero finito di modalità distinte (ad es. il colore preferito: rosso, arancio, giallo, verde, blu, viola, ...)



Scale di misurazione

Scala ordinale

I caratteri con scala ordinale (= caratteri ordinali) costituiscono mutabili che, pur non facendo riferimento a valori numerici, assumono modalità logicamente sequenziali, in ordine crescente o decrescente. Pur essendo ancora impossibile effettuare operazioni aritmetiche tra gli attributi delle singole unità statistiche, è possibile per essi stabilire non solo la diversità, ma anche la precedenza di una rispetto all'altra, cioè una relazione d'ordine.

$$x_i = x_j ; x_i \prec x_j ; x_i \succ x_j$$

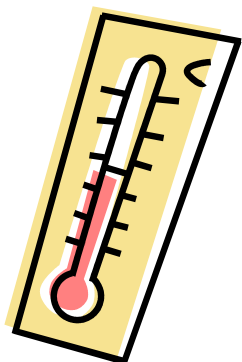
**Mediana
Moda**



Scale di misurazione

Scala ordinale

Il grado militare (sergente, maresciallo, tenente, capitano, maggiore, colonnello), un indice di gradimento (per niente, poco, abbastanza, molto), il livello di istruzione (nessun titolo, scuola elementare, scuola media inferiore, scuola media superiore, laurea) sono caratteri nei quali la gerarchia, il merito, la preferenza individuale stabiliscono un grado tra le unità statistiche, pur non potendo attribuire un valore numerico alla distanza tra le modalità.



Scale di misurazione

Scala ad intervallo

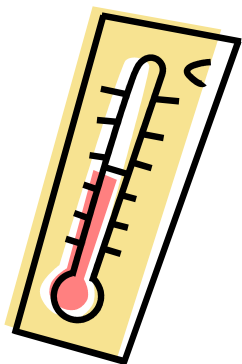
I caratteri con scala ad intervallo (= caratteri intervallari) sono variabili che consentono un confronto solo per differenza tra le modalità che i soggetti assumono, poiché essi fanno riferimento ad un'origine arbitraria. Infatti, lo zero della scala delle misure non è assoluto ma convenzionale.

$$(x_i - x_j) = (x_k - x_h)$$

$$(x_i - x_j) < (x_k - x_h)$$

$$(x_i - x_j) > (x_k - x_h)$$

Media
Mediana
Moda



Scale di misurazione

Scala ad intervallo

Variabili misurate su scala ad intervallo sono la temperatura in gradi Celsius (lo 0° si fa corrispondere a quella dell'acqua che gela), la latitudine e la longitudine (la cui origine arbitraria è l'intersezione tra l'Equatore e il meridiano che passa per Greenwich), gli anni misurati nel mondo occidentale (l'anno 0 corrisponde convenzionalmente alla nascita di Gesù Cristo).

Se a Bolzano ci sono 15°C e a Palermo 30°C, è corretto affermare che la temperatura di Palermo è più alta di quella di Bolzano, ma è sbagliato dire che è doppia, anche perché se fosse misurata in °F sarebbe superiore solo del 45,8%.

Analogamente, tra il 1982 e il 1992 ci sono gli stessi anni che tra il 2002 e il 2012, ma non ha senso dire che il 2000 è superiore del 25% al 1600.



Scale di misurazione

Scala di rapporto

I caratteri con scala di rapporto (= caratteri proporzionali) sono variabili per le quali è intrinseca e univoca la definizione dello zero assoluto (assenza del fenomeno): essi sono idonei per rappresentare la gran parte dei fenomeni fisici, temporali, spaziali, economici, ...

$$(x_i / x_j) = (x_k / x_h)$$

$$(x_i / x_j) < (x_k / x_h)$$

$$(x_i / x_j) > (x_k / x_h)$$

**Media
Mediana
Moda**

Scale di misurazione



Scala di rapporto

È evidente il significato del valore zero per il peso di un soggetto, così come la durata in secondi tra due spari (che è 0 se e solo se sono contemporanei) o per il guadagno di un imprenditore (costi e ricavi si equivalgono nell'unità di tempo presa a riferimento); per tali variabili sono lecite tutte le operazioni matematiche.

Se A pesa 50 kg e B pesa 100 kg, è corretto affermare che B pesa il doppio di A.

Analogamente, se nel 2012 un professore ordinario Unifi (P.ORD. t/p II pe cl.14 sc.9) costa all'Ateneo € 184.220,88 e un dipendente tecnico-amministrativo di categoria EP7 costa €61.898,27, è corretto affermare che il primo costa il triplo del secondo.

Esempio 4

Stabilire tipologia e scala di misurazione

Numero di aule del Polo di Novoli

Colore fascia dei Presidi

Età dei docenti dell'Ateneo fiorentino

Regione di residenza degli immatricolati

Posizione economica del personale T-A

Esito dell'esame di Statistica I

Voto di laurea di uno studente

Soddisfazione del corso di studi

Durata del neon dell'ufficio

Punteggio del test di accesso all'Università

quantitativo / rapporto

qualitativo / nominale

quantitativo / rapporto

qualitativo / nominale

qualitativo / ordinale

qualitativo / nominale

quantitativo / intervallo

qualitativo / ordinale

quantitativo / rapporto

quantitativo / intervallo

Giacomo Bulgarelli
Ufficio Servizi Statistici

SERVIZIO DAF: FONTI STATISTICHE



Mercoledì 3 ottobre 2012

FINE PARTE 2