



***Analisi dei dati per
le applicazioni di marketing***

Modulo MKT2
Analisi statistica, report tabellari, grafici e mappe con SPSS

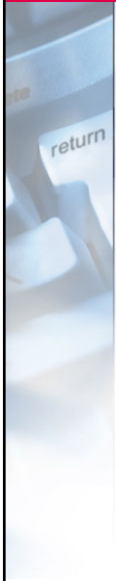
SPSS TRAINING



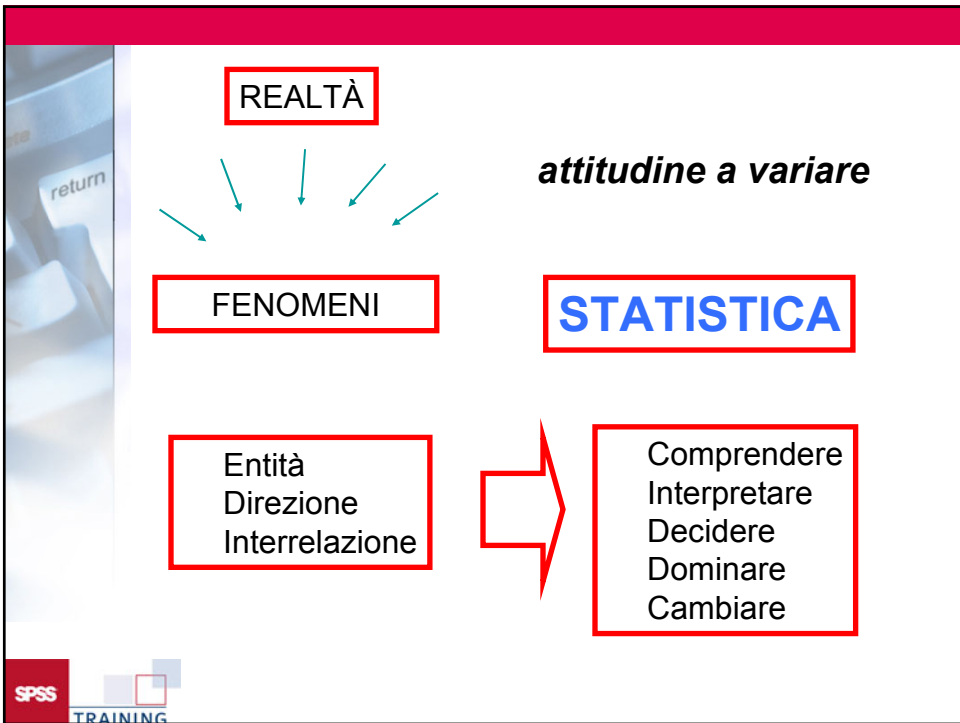

OBIETTIVI

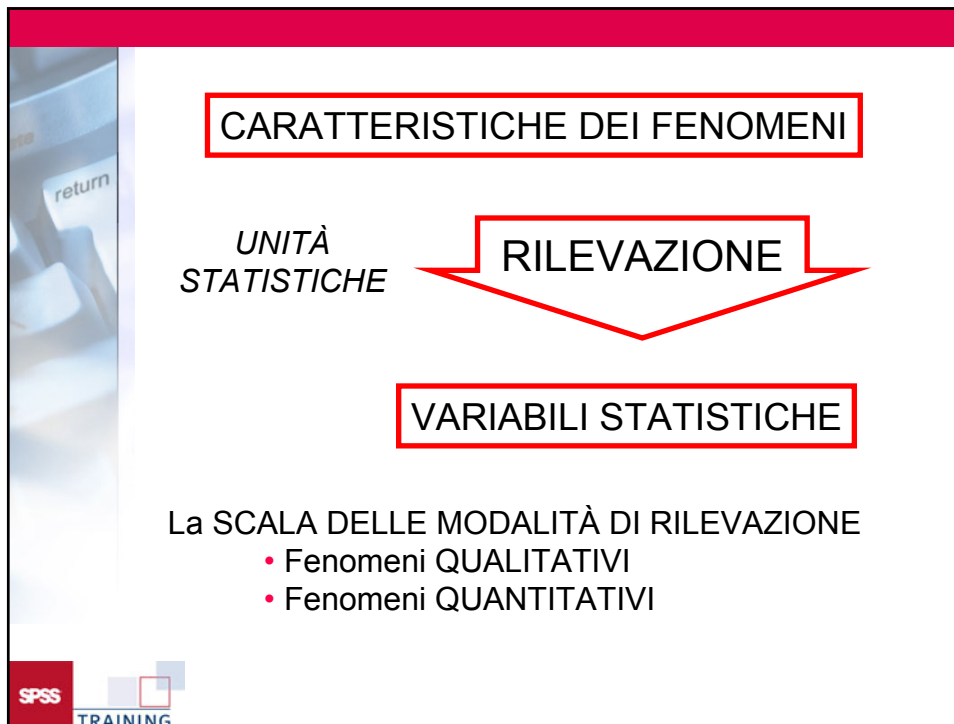
- Interpretare i risultati in modo opportuno
- Individuare la tecnica ottimale in relazione al tipo di dati e agli obiettivi dell'analisi
- Acquisire familiarità con le principali tecniche statistiche di sintesi e di indagine

SPSS TRAINING



1. INTRODUZIONE





La SCALA DELLE MODALITÀ DI RILEVAZIONE


- Fenomeni QUALITATIVI
- Fenomeni QUANTITATIVI

The slide is titled "I fenomeni qualitativi" in red. It explains that modalities are identified by attributes. It lists two types of scales: NOMINALI (categorical or disconnected) and ORDINALI (presenting a natural order). Examples are provided for each. To the left is a vertical image of a computer keyboard with a "return" key visible. In the bottom-left corner, there is a logo for "SPSS TRAINING" with a small red square icon.

I fenomeni qualitativi

le modalità si identificano in *attributi*



- Scale NOMINALI (o categoriali o sconnesse)
Le modalità non sono suscettibili ad alcun tipo di ordinamento
Ad es. colore dei capelli, religione, stato civile
- Scale ORDINALI
Le cui modalità presentano, in via naturale, un ordine
Ad es. titolo di studio, giudizio scolastico



I fenomeni quantitativi


le modalità si identificano in *numeri*


- Scale INTERVALLARI
Si può valutare la differenza tra due intensità, ma non è sensato stabilire rapporti; non sono sensibili a cambiamenti di origine
Es. anno di nascita, temperatura
- Scale RAPPORTO
Sono articolate in modalità ordinate la prima delle quali è in via naturale lo zero; consentono di valutare il rapporto esistente tra due modalità
Es. statura, peso, velocità, reddito



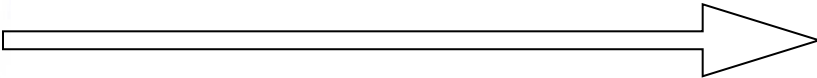
I fenomeni quantitativi

- Discreti
(Numero di figli, abitanti in una regione)
- Continui
(Superficie in kmq, età, altezza)






| SCALA DELLE MODALITÀ | | | | | |
|----------------------|----------------|-------------|----------|--------------|----------|
| | | QUALITATIVE | | QUANTITATIVE | |
| RELAZIONI | | Nominale | Ordinale | Intervallare | Rapporto |
| Uguaglianza | $x_i = x_j$ | ✓ | ✓ | ✓ | ✓ |
| Disuguaglianza | $x_i \neq x_j$ | ✓ | ✓ | ✓ | ✓ |
| Ordinamento | $x_i > x_j$ | | ✓ | ✓ | ✓ |
| Differenza | $x_i - x_j$ | | | ✓ | ✓ |
| Rapporto | x_i / x_j | | | | ✓ |



SPSS TRAINING



SCALA DICOTOMICA

**in cui le unità statistiche vengono ripartite
in base all'appartenenza o meno a una classe**


***Ad es. fumatore/non fumatore,
vivo/morto, pagato/non pagato***

SPSS TRAINING



La scala delle modalità è intrinseca al fenomeno

Sono le caratteristiche intrinseche del fenomeno a dare indicazioni sulla tecnica statistica di analisi più idonea



SPSS consente di attribuire la scala delle modalità di una variabile nel *dizionario dei dati*

•NOMINALE

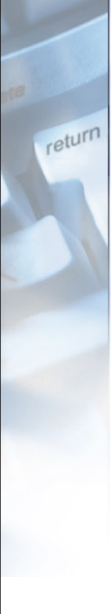
I valori rappresentano categorie distinte e non ordinate (ad esempio il settore di attività o l'area geografica). Le variabili nominali possono essere numeriche o stringhe.

•ORDINALE

I valori rappresentano categorie disposte secondo un ordine intrinseco (basso, medio, alto; primo, secondo, terzo). Le variabili ordinali possono essere numeriche o stringhe.

•SCALA

I valori sono numeri su scala per intervallo o per rapporto (età, reddito). Le variabili di scala devono essere numeriche.





SPSS distingue tre formati di variabili

- **Stringa lunga:** variabili alfanumeriche di lunghezza superiore a 8 caratteri
- **Stringa corta:** variabili alfanumeriche di lunghezza massima 8 caratteri
- **Numerica:** variabile costituita da numeri (nei vari formati: valuta, data, ecc.)

Sulle variabili numeriche
SPSS consente di effettuare qualsiasi tipo di elaborazione

Sulle variabili di stringa
possono essere eseguite un numero ristretto di procedure (conteggi, identificazione di sottogruppi) e sono utilizzabili come etichette per i casi





Molto spesso è comodo associare alle modalità qualitative codici numerici
es. codice ISTAT di comuni, codice identificativo cliente

Questa procedura è utile in quanto consente di eseguire elaborazioni di classificazione più agili

Ma

L'analista deve essere consapevole che, nonostante la ricodifica,
la variabile rimane connotata secondo la caratteristica intrinseca
del fenomeno di cui essa è rilevazione






Studiare un fenomeno statistico può voler dire effettuare una rilevazione di tutte le unità statistiche che ne sono determinazione o in alternativa solo di una parte di esse

UNIVERSO O POPOLAZIONE
È l'insieme esaustivo delle unità che è interesse del ricercatore considerare allorché intenda studiare un fenomeno

CAMPIONE
È un sottoinsieme dell'universo estratto seguendo una procedura probabilistica che lo rende *statisticamente rappresentativo* della popolazione di provenienza

SPSS TRAINING



La STATISTICA DESCRITTIVA
si occupa della raccolta integrale delle manifestazioni del fenomeno considerato e fornisce una sorta di descrizione delle caratteristiche del fenomeno medesimo

Ad esempio, una piccola azienda vuole conoscere alcune informazioni riguardanti i dipendenti

SPSS TRAINING

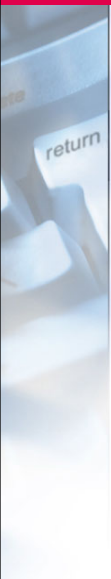


La **STATISTICA INDUTTIVA** o **INFERENZIALE** intende descrivere non tanto ciò che traspare dalle manifestazioni osservate (rilevazioni parziali), ma quello che emergerebbe qualora la rilevazione fosse estesa all'insieme di tutte le manifestazioni del fenomeno.


Ad esempio, il Ministero del Lavoro vuole fare un'indagine sui dipendenti delle piccole aziende

L'incertezza che deriva dalla parzialità della rilevazione è dominata dalla **TEORIA DELLE PROBABILITÀ**

Con i metodi della statistica inferenziale si può passare dalla conoscenza del campione a quella dell'universo



2. ANALISI DESCRITTIVA DI UN FENOMENO SINGOLARMENTE CONSIDERATO





Rappresentazione dei dati

***Una rappresentazione grafica
fornisce una visione sintetica ed intuitiva
del fenomeno che si vuole studiare***

**Esistono svariati metodi per rappresentare i singoli
fenomeni o gruppi di fenomeni considerati**

La rappresentazione grafica delle variabili
differisce per il livello di misurazione

SPSS

TRAINING



Descrizione dei dati

***Gli indicatori statistici hanno lo scopo
di fornire un'informazione sintetica e riassuntiva
di un fenomeno attraverso i valori osservati***

Indicatori di posizione

Indicatori di variabilità

Indicatori di forma

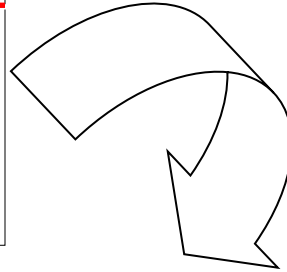
SPSS

TRAINING

Variabili qualitative

Es. Colore dei capelli

| Persona | Colore |
|---------|---------|
| Mario | Castano |
| Paolo | Biondo |
| Luigi | Nero |
| Sonia | Biondo |
| | |
| Franca | Rosso |



| Colore | Castano | Biondo | Nero | Rosso | |
|-----------|---------|--------|------|-------|----|
| Quantità? | | | | | |
| | 7 | 12 | 7 | 3 | 29 |

SPSS

TRAINING

f_i sono le FREQUENZE ASSOLUTE
 indicano quante unità fra le n considerate
 sono espressione della i -esima modalità di X .
 Naturalmente $\sum_i f_i = n$

FREQUENZE RELATIVE e FREQUENZE PERCENTUALI $p_i = \frac{f_i}{n} * 100$

FREQUENZE PERCENTUALI CUMULATIVE $P_i = \sum_{j=1}^i p_j$

TABELLA DI FREQUENZE

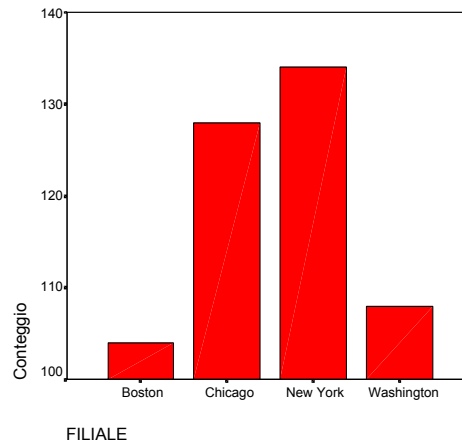
| | | Categoria lavorativa | | | |
|----------|---------------------|----------------------|-------------|--------------------|----------------------|
| | | Frequenza | Percentuale | Percentuale valida | Percentuale cumulata |
| Validi | Impiegato | 224 | 47,3 | 47,9 | 47,9 |
| | Impieg. special. | 136 | 28,7 | 29,1 | 76,9 |
| | Agente di secur. | 27 | 5,7 | 5,8 | 82,7 |
| | Impieg. laureato | 40 | 8,4 | 8,5 | 91,2 |
| | Funzionario | 30 | 6,3 | 6,4 | 97,6 |
| | Funzionario MBA | 5 | 1,1 | 1,1 | 98,7 |
| | Dirigente | 6 | 1,3 | 1,3 | 100,0 |
| Totale | | 468 | 98,7 | 100,0 | |
| Mancanti | Mancante di sistema | 6 | 1,3 | | |
| | Totale | 474 | 100,0 | | |

SPSS

TRAINING

Grafico a Barre

I grafici più adatti per rappresentare fenomeni qualitativi sono i *diagrammi a barre*



SPSS

TRAINING

Grafico a torta

Per rappresentare il peso relativo di ciascuna modalità sul totale è molto informativo il *grafico a torta*

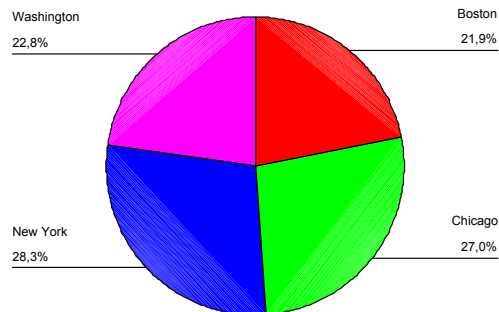


Grafico a punti, lineare, nastro ed area

SPSS

TRAINING

Gli indicatori di posizione

MODA

La modalità che è risultata privilegiata dal fenomeno ed è ricorsa più volte in sede di rilevazione

Fenomeni ordinali

Sezione A

Sezione B

| Giudizio | Frequenza | Freq. Cum | Giudizio | Frequenza | Freq. Cum |
|----------|-----------|-----------|----------|-----------|-----------|
| INSUFF | 3 | 3 | INSUFF | 25 | 25 |
| SCARSO | 7 | 10 | SCARSO | 30 | 55 |
| SUFF | 35 | 45 | SUFF | 35 | 90 |
| BUONO | 30 | 75 | BUONO | 7 | 97 |
| OTTIMO | 25 | 100 | OTTIMO | 3 | 100 |

MEDIANA modalità che occupa il posto centrale nella distribuzione di frequenza

$$\sum_{m_i \leq \text{Mediana}} p_i \geq 50\% \quad \sum_{m_i \geq \text{Mediana}} p_i \geq 50\%$$

Variabili quantitative

La distribuzione di frequenza in un fenomeno continuo non è molto indicativa

(MODA)

MEDIANA

MEDIA ARITMETICA

$$\bar{x} = (\mu) = \frac{1}{n} \sum_{i=1}^n x_i$$

Il valore atteso di una successiva rilevazione

**La parte del totale delle intensità
che spetta a ciascuna unità**

SPSS

TRAINING

MEDIA GEOMETRICA

$$\bar{x}_g = (\mu_g) = \exp\left(\frac{1}{n} \sum_{i=1}^n (\ln x_i)\right)$$

MEDIA ARMONICA

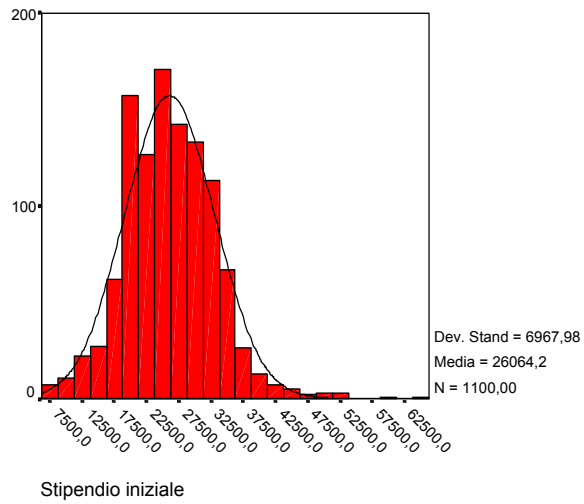
$$\bar{x}_h = (\mu_h) = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

SPSS

TRAINING

Grafico ad istogramma

La distribuzione di frequenza in un fenomeno continuo si può rappresentare mediante ISTOGRAMMI



SPSS TRAINING

VARIABILITÀ

Popolazione A

Popolazione B

| Altezza | | Altezza |
|---------|-------|---------|
| 160 | MIN | 175 |
| 170 | | 178 |
| 180 | MEDIA | 180 |
| 190 | | 182 |
| 200 | MAX | 185 |

RANGE DEVIAZIONE STANDARD

SPSS TRAINING

Indicatori di variabilità

Si vuole mostrare in quale misura le intensità si disperdono attorno alla media

MINIMO e MASSIMO valori osservati

L'INTERVALLO (o RANGE) di variazione

La VARIANZA

$$\bar{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

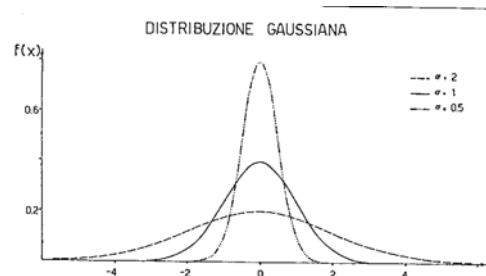
La radice quadrata della varianza è lo scarto quadratico medio o DEVIAZIONE STANDARD

SPSS

TRAINING

La distribuzione GAUSSIANA o Normale o a campana

- È la distribuzione teorica più utilizzata
- È la distribuzione degli errori casuali
- Tutte le distribuzioni con l'aumentare delle prove tendono ad assumere una distribuzione normale (teorema centrale del limite)
- È definita da due parametri: la media e la varianza

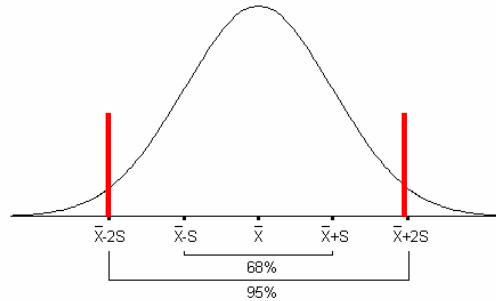


SPSS

TRAINING

In una **DISTRIBUZIONE NORMALE**

il **68%** dei casi cade nell'intervallo **Media±Deviazione Standard**
il **95%** dei casi nell'intervallo **Media±1,96 Deviazione Standard**
il **99,7%** nell'intervallo **Media±3 Deviazione Standard**.



Molte delle tecniche inferenziali sono basate sulla distribuzione della normale
Diventa importante capire se le nostre variabili sono normali

SPSS TRAINING

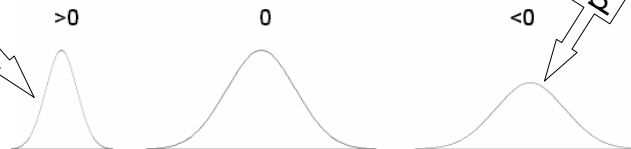
Indicatori di distribuzione

QUANTILI

Es 5° percentile $\sum_{m_i \leq 5^\circ \text{ perc}} p_i \geq 5\%$ e $\sum_{m_i \geq 5^\circ \text{ perc}} p_i \geq 95\%$

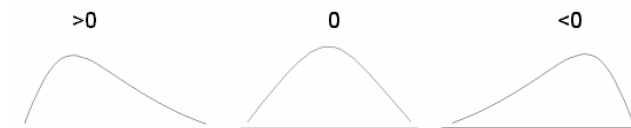
leptocurtica

CURTOSI



platycurtica

SIMMETRIA



SPSS TRAINING



La normale con media 0 e varianza 1 è detta
NORMALE STANDARD

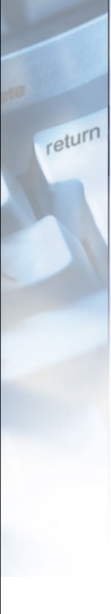
Esistono delle tavole che riportano
i valori e le corrispondenti probabilità sottese (quantili)

Ci si può sempre ricondurre ad un normale standard
attraverso il procedimento di STANDARDIZZAZIONE

Intensità standardizzate o PUNTEGGI Z $z_i = \frac{x_i - \mu}{\sigma}$



3. ANALISI PRELIMINARE DIAGNOSTICA ED ESPLORATIVA DEI DATI



È fondamentale prima di iniziare qualsiasi analisi GUARDARE i dati

- I dati possono essere sporchi e occorre fare pulizia
- La maggior parte delle analisi statistiche si basa su assunti, che non venendo soddisfatti possono seriamente alterare i risultati delle analisi
- Un'occhiata ai dati ci può fornire preziose indicazioni sulle tecniche statistiche idonee
- Un'attenta esplorazione è di per sé un primo metodo di indagine statistico e può perfino portare a fare considerazioni utili nella fase conclusiva del lavoro

SPSS TRAINING



Gli strumenti dell'analisi esplorativa (per "iniziare a guardare" i dati)

- **Controllo numerico dei dati**
 - Analisi di dati e incroci tra dati
 - Individuazione di valori estremi e fuorvianti
 - Analisi dei valori mancanti
 - Verifica sulla distribuzione di variabile
 - Verifica della robustezza della tendenza centrale
- **Controllo grafico dei dati**
 - Individuazione di valori estremi e fuorvianti
 - Addensamento sulle intensità
 - Verifica sulla distribuzione di variabile

SPSS TRAINING



4. LA TEORIA DELLA STIMA E DELL'INFERENZA STATISTICA

La conoscenza delle caratteristiche di un fenomeno può essere conseguita osservando tutta la popolazione o un suo campione

Inferenza o induzione:
l'estensione dal particolare al generale delle considerazioni emerse dall'esperimento casuale

È un *processo d'azzardo* a causa della presenza di incertezza

Esempio.

Un silo contiene 10.000.000 di semi di una pianta che dà fiori bianchi o rossi. Il contadino vuole sapere quanti sul totale faranno fiori bianchi.

Risposta.

L'unico modo di avere una risposta è piantare tutti i semi e contare i fiori bianchi. Il contadino però non è d'accordo perché vuole vendere i semi e comunque non ha tutto quel tempo da perdere. Lo statistico accorre in aiuto al contadino suggerendogli che è vero che non è possibile fare una previsione esatta ma è possibile costruire un'affermazione *probabilistica* attraverso la semina di un *campione*. Lo statistico illustra al contadino perplesso come può estendere in generalità la conclusione dell'*esperimento* sul totale dei 10.000.000 di semi partendo da un campione molto più piccolo. Il contadino può ora *verificare* statisticamente se il risultato emerso è coerente con l'idea che si era fatto

SPSS

TRAINING

LA STIMA DEI PARAMETRI

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{funzione di densità gaussiana}$$

La distribuzione è **PARAMETRICA**, cioè la sua forma è data una volta noti i parametri che la caratterizzano

Sulla base dei valori osservati (realizzazioni campionarie) si vuole *stimare* il valore dei parametri incogniti

Si ricorre a metodi statistici di stima che individuano delle statistiche (**STIMATORI**)

La STIMA PUNTUALE

è la determinazione della v.c. **STIMATORE** sulla base delle realizzazioni del campione estratto

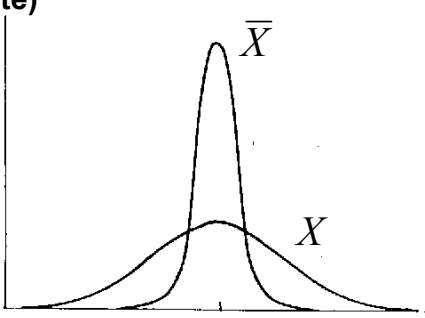
SPSS

TRAINING

La statistica che stima "meglio" la media μ di una popolazione è lo stimatore MEDIA CAMPIONARIA \bar{X} le cui realizzazioni sono le stime \bar{x} , media aritmetica dei valori campionati

La stima della varianza di popolazione σ^2 è data da \bar{s}^2

Le leggi di convergenza in probabilità
(Leggi dei grandi numeri, Teorema centrale del limite)
dimostrano che

$$\bar{X} \stackrel{n \rightarrow \infty}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$


SPSS TRAINING

L'ERRORE STANDARD

È la deviazione standard della variabile casuale media campionaria

Rappresenta l'unità di misura dell'errore casuale di stima commesso utilizzando la media campionaria come stimatore di μ

$$ES(\bar{X}) = \frac{\sigma}{\sqrt{n}} \quad ES(\bar{X}) \xrightarrow{n \rightarrow \infty} 0$$

SPSS TRAINING

In definitiva:

- Data una distribuzione di valori individuali (popolazione)
- Si applica una procedura statistica di selezione (campione)
- Si stimano i parametri della distribuzione (media campionaria)
- Si ottiene così la distribuzione di probabilità del fenomeno

È così possibile, una volta nota la distribuzione di probabilità, definire intervalli centrati intorno alla media campionaria del tipo:

$$\Pr\left\{\mu - z_{1-\alpha/2} ES(\bar{X}) < \bar{x} < \mu + z_{1-\alpha/2} ES(\bar{X})\right\} = 1 - \alpha$$

con z deviana gaussiana standardizzata e a livello prescelto di significatività

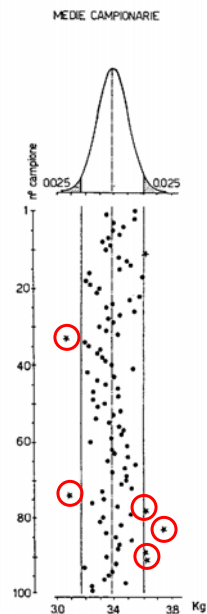
SPSS

TRAINING

$1 - \alpha$ è la probabilità che una media campionaria, scelta a caso fra i possibili campioni, cada nell'intervallo definito, ovvero la media campionaria non dista da μ più di

$$z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

In entrambe le direzioni



SPSS

TRAINING


In realtà, è inverosimile che due campioni estratti dalla stessa popolazione producano le stesse stime:
DISTRIBUZIONE CAMPIONARIA

In ogni stima è insito un certo margine di errore, di qui l'opportunità di presentare accanto alla stima puntuale una possibile misura dell'errore cui questa è soggetta

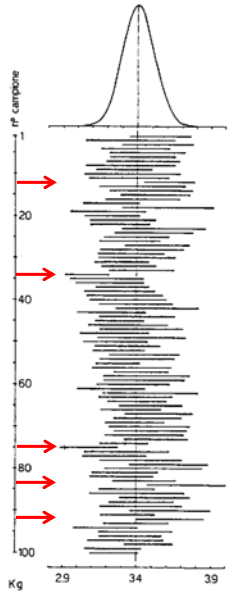

Si associa così ad un campione non un singolo valore, ma una stima intervallare:
l'INTERVALLO DI CONFIDENZA

$$\Pr\left\{\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

È un intervallo casuale che ha la probabilità $1 - \alpha$ di contenere la media della popolazione ed è definito come intervallo di confidenza al livello del $100(1 - \alpha)\%$



Non è vero che la probabilità che μ cada in questo intervallo è $1 - \alpha$, ma si può solo dire che si ha fiducia che l'intervallo espressione del presente campione contenga il valore incognito μ e questa fiducia è giustificata dall'alta probabilità $1 - \alpha$ che ci permette di fidare in un buon esito.

LA VERIFICA DI UNA IPOTESI STATISTICA

Un'IPOTESI STATISTICA

è un assunto circa un parametro della funzione di distribuzione di una variabile casuale

Il saggio di un'ipotesi statistica (ipotesi nulla H_0) si basa sulla dimostrazione per contraddizione

SPSS

TRAINING

LOGICA

Premessa

Argomentazioni logiche

Contraddizione (reductio ad absurdum)

Conclusione

Falsificazione della premessa

STATISTICA

Ipotesi nulla

Applicazione di un test

Risultato improbabile ($p < 0,05$)

Conclusioni

1. L'ipotesi è rifiutata (si è verificato un risultato altamente improbabile)
2. L'ipotesi è rifiutata (il risultato osservato è inconsistente con quanto specificato nell'ipotesi)

SPSS

TRAINING

Il saggio di H_0 avviene mediante una statistica test che viene confrontata con una distribuzione di riferimento valida sotto H_0 .

È auspicabile rigettare H_0 quando si può dire che il risultato che otterremo sarebbe altamente inverosimile (*cut-point* o *valore critico* 5%)

In questo caso non può essere accettata l'ipotesi nulla ma questo ragionamento non implica che sia vera l'ipotesi alternativa

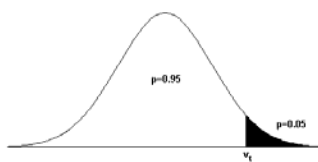
SPSS

TRAINING

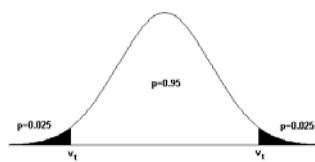
In definitiva:

- Definizione di un'ipotesi nulla e di un'alternativa
- Specificazione del parametro di interesse per la verifica di H_0
- Individuazione della statistica test e calcolo del valore numerico della stessa sui dati del campione
- Riferimento della statistica test a una distribuzione nota quando l'ipotesi nulla è vera
- Determinazione della probabilità di verificarsi (livello soglia)
- Confronto del valore empirico verso quello teorico e conclusioni

Test unidirezionali




Test bidirezionali



SPSS



TRAINING




Assunti per i test parametrici

- Le osservazioni campionarie devono essere indipendenti, ovvero il campione deve essere casuale
- Le osservazioni devono appartenere a popolazioni distribuite normalmente
- Le popolazioni devono avere la stessa varianza (omoschedasticità).
- Le variabili osservate devono essere misurabili su una scala per intervallo

Molti test parametrici sono abbastanza *robusti* da sopportare lievi deviazioni da alcune di questi postulati, soprattutto quando la numerosità campionaria è sufficientemente elevata.



5. ANALISI DI DUE FENOMENI CONGIUNTAMENTE CONSIDERATI



ANALISI DI DUE FENOMENI CONGIUNTAMENTE CONSIDERATI

| | Categoriale | Continua |
|-------------|--------------------------------|---|
| Categoriale | <i>Associazione statistica</i> | <i>Confronto tra medie entro gruppi</i> |
| Continua | | <i>Correlazione lineare</i> |

SPSS

TRAINING

L'ASSOCIAZIONE TRA DUE VARIABILI CATEGORIALI

Le unità di ciascuno dei due gruppi di X differiscono tra loro per l'influenza di Y?

Esempio. I clienti di sesso maschile acquistano di più delle donne?

Esempio. Il fatto di possedere una laurea "c'entra" col fatto di non essere soddisfatti del servizio?

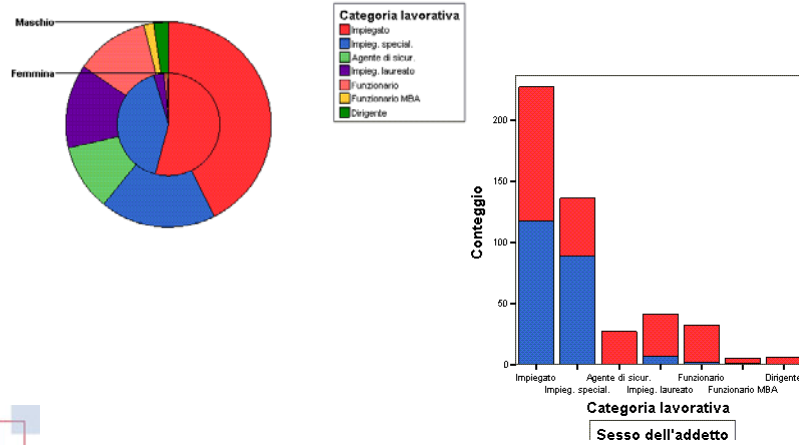
Esempio. Possiamo affermare che la campagna promozionale ha avuto più effetti al nord che al sud?

SPSS

TRAINING

Rappresentazione grafica

È possibile rappresentare graficamente gli incroci tra i livelli di due fattori ricorrendo a grafici bidimensionali



•Le tabelle di contingenza

| | Fumatore | Non fumatore |
|----------------------|----------|--------------|
| Consuma alcolici | | |
| Non consuma alcolici | | |

•L'indipendenza stocastica

– L'associazione analizza il mutuo comportamento di una coppia di variabili

Misura: CHI QUADRATO:
$$\chi^2 = \sum \frac{(\text{osservati} - \text{attesi})^2}{\text{attesi}}$$





Metodi per il calcolo della significatività statistica

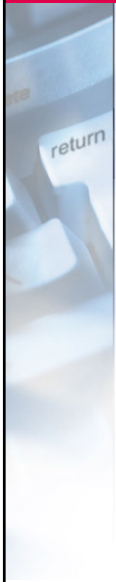
Il metodo di calcolo per default di SPSS è il metodo asintotico, ovvero i valori di *sign* sono stimati basandosi sull'assunto di conformità (almeno asintotica) del campione ad una distribuzione di riferimento

Qualora i dati campionari fossero di numerosità esigua, sbilanciati o mal distribuiti, il calcolo del livello di significatività sarebbe inficiato dal mancato rispetto di questo assunto e pertanto i risultati ottenuti non sono attendibili

I test ESATTI si basano sull'effettiva distribuzione della statistica test in modo da offrire una stima più accurata del valore di p

SPSS

TRAINING



Sebbene i test esatti siano sempre affidabili, essi richiedono un notevole sforzo computazionale e pertanto per basi dati troppo grandi non è possibile il calcolo


In queste situazioni si usa il metodo MONTE CARLO

Il metodo Monte Carlo fornisce una stima non distorta del valore esatto di *sign*.

È un metodo di campionamento ripetuto tra tutte le possibili combinazioni attese, in modo da ridurre sensibilmente il carico computazionale

SPSS


TRAINING



Test per la verifica sulla distribuzione di una variabile

- Test di Kolmogorov-Smirnov (unif., norm., Poisson, esp.)
- Test Chi Quadrato
Esempio: È vero che in un sacchetto di caramelle di frutta c'è la stessa proporzione di rosse, arancio, verdi e gialle?
È vero che invece c'è il 15% di rosse, il 20% di arancio, il 35% di verdi e il 30% di gialle?
- Test Binomiale

SPSS TRAINING



CONFRONTO TRA MEDIE DI UNA VARIABILE QUANTITATIVA PER I LIVELLI DI UNA VARIABILE CLASSIFICATORIA

Esempio. I salari medi dei lavoratori uomini sono più alti di quelli delle lavoratrici?

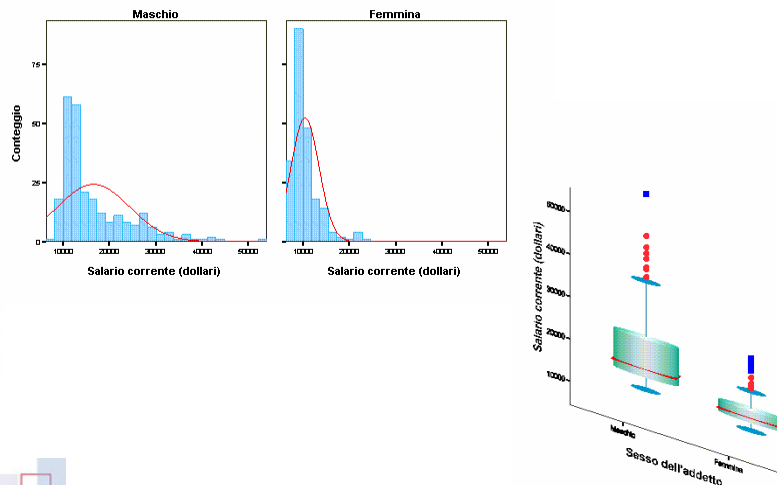
Esempio. La soddisfazione dei clienti è migliorata nei negozi in cui i commessi hanno seguito il corso di formazione?

Esempio. In quale stagione le vendite sono più alte?
Confrontando le medie nei gruppi si può avere una prima indicazione

Strumento: **MEDIE stratificate**

SPSS TRAINING

È possibile rappresentare le medie stratificate per i livelli di un fattore ricorrendo a grafici bidimensionali



SPSS
TRAINING

Test non parametrici

Un test non parametrico è un test il cui modello non precisa condizioni circa i parametri della popolazione da cui proviene il campione studiato


Vantaggi dei test non parametrici

- Assunti meno restrittivi
- Possibilità di impiego anche con piccoli campioni
- Analisi di popolazioni differenti
- Analisi di ranghi o dati nominali

Svantaggi dei test non parametrici

- Meno precisi a parità di informazione (in caso fossero verificati gli assunti)

SPSS
TRAINING




È possibile in primo luogo confrontare
la media di un fenomeno
con un valore prefissato in partenza

Es. Un'azienda vende batterie per pupazzetti meccanici
asserendo che la durata media delle stesse è di 12 ore.
Il cliente ne compra un tot e prende atto che nei suoi
pupazzetti durano in media 10 ore.
Deve ritenersi sfortunato o farsi ridare i soldi?

Strumento: TEST T PER UN CAMPIONE


Strumento: TEST BINOMIALE (per proporzioni)



Dalle medie stratificate, però non è possibile quantificare
se la differenza tra i gruppi è dovuta ad un errore di
campionamento (al caso che ha fatto estrarre un
campione "sfavorevole") o se questa differenza
effettivamente sussiste (*statisticamente significativa*).
Si verifica l'ipotesi che siano effettivamente uguali e se i
dati sono con essa coerenti.

Strumento: TEST T PER DUE CAMPIONI INDIPENDENTI

Strumento: TEST DEI RANGHI DI MANN-WHITNEY



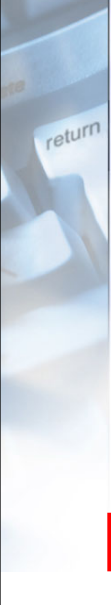

Può presentarsi un caso in cui il ricercatore sia interessato a valutare se le stesse unità statistiche riportino valori differenti dopo l'azione di un fattore nel tempo o nello spazio.

Es. Si vuole confrontare il peso di alcune signore prima e dopo una cura dimagrante.

Es. Si vuole valutare l'effetto di una campagna promozionale con un disegno *prima/dopo*

Strumento: TEST T PER DUE CAMPIONI APPAIATI

Strumento: TEST DEI RANGHI DI WILCOXON




I precedenti strumenti si riferiscono al confronto di due soli gruppi. È possibile confrontare n medie simultaneamente saggiando l'ipotesi nulla:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_n = \mu$$

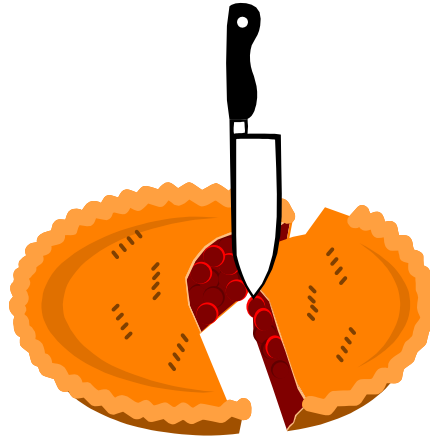
Es. Un'azienda divide la propria clientela in quattro gruppi differenti e li sottopone a quattro offerte diverse. La soddisfazione della clientela differirà significativamente?

Strumento: ANALISI DELLA VARIANZA

Strumento: TEST DEI RANGHI DI KRUSKAL-WALLIS



IL MODELLO DI ANALISI DELLA VARIANZA



SPSS

TRAINING

L'ANALISI DELLA VARIANZA (ANOVA)

È utilizzata per verificare se diversi gruppi indipendenti provengono da popolazioni con la stessa media

Esempio: si vuole valutare l'efficacia sulle vendite di una nuova confezione per il prodotto di punta di una famosa marca. I negozi a cui è stato fornito il nuovo prodotto hanno in media lo stesso aumento di vendite rispetto ai negozi a cui è stato fornito il prodotto con la confezione standard?



Pianificazione e disegno degli esperimenti (DOE)

SPSS

TRAINING

Modello di ANOVA ad un fattore

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

y è la variabile di risposta di ciascun negozio
(vendite rilevate nel j -esimo neg),

μ è la media generale
(vendite medie della popolazione
da cui è estratto il campione)

Ovvero il valore delle vendite rilevate in ciascun negozio
può differire dalla media della popolazione
per due componenti:

α è l'*effetto principale* del fattore sperimentale
(confezione nuova o standard),

ε è la componente casuale (errore di campionamento)

SPSS

TRAINING

Modello di ANOVA ad un fattore

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Il ricercatore è interessato
a valutare se l'effetto α è significativo

$$H_0 : \alpha = 0$$

$$H_0 : \mu_{\text{nuovo}} = \mu_{\text{standard}}$$

SPSS

TRAINING

Le vendite dei negozi differiscono tra loro per effetto della diversa confezione?

si scompone la devianza totale in due componenti:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 n_i + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Dev totale = devianza tra i gruppi + devianza entro i gruppi
(devianza spiegata + devianza residua)

Se le medie sono tutte uguali, la varianza tra i gruppi si annulla e tutta la varianza totale è dovuta alla variabilità presente all'interno dei gruppi (cioè alla componente casuale)

Pertanto per saggiare l'ipotesi nulla, è possibile costruire la statistica test:

$$F = \frac{\text{devianza spiegata} / (\text{n}^\circ \text{livelli} - 1)}{\text{devianza residua} / (\text{n}^\circ \text{unità} - \text{n}^\circ \text{livelli})} = \frac{\text{varianza spiegata}}{\text{varianza residua}}$$

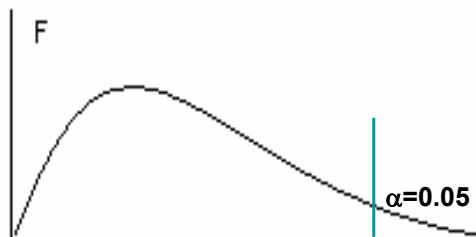


Esempio ANOVA

| | A | B | C | D | |
|------------------------|----------|----------|----------|----------|------------------------|
| 1 | 4 | 5 | 7 | 2 | |
| 2 | 4 | 5 | 8 | 1 | |
| 3 | 5 | 6 | 7 | 2 | |
| 4 | 5 | 6 | 9 | 3 | |
| 5 | 6 | 7 | 6 | 3 | |
| 6 | 3 | 6 | 3 | 4 | |
| 7 | 4 | 4 | 2 | 5 | |
| 8 | 4 | 5 | 2 | 4 | |
| 9 | 3 | 6 | 2 | 4 | |
| 10 | 4 | 3 | 3 | 3 | |
| tra | | | | | |
| media | 4.2 | 5.3 | 4.9 | 3.1 | media tot |
| (scarti) ² | 0.030625 | 0.855625 | 0.275625 | 1.625625 | 4.375 |
| Σ(scarti) ² | 27.875 | | | | gdl |
| gdl (tra) | 4-1 | | | | 40-1 |
| varianza (tra) | 9.292 | | | | |
| entro | | | | | varianza tot |
| Σ(scarti) ² | 7.6 | 12.1 | 68.9 | 12.9 | 2.94 |
| Σ(scarti) ² | 101.5 | | | | Σ(scarti) ² |
| gdl (entro) | 40-4 | | | | |
| varianza (entro) | 2.819 | | | | 129.38 |



La statistica F sotto H_0 segue la distribuzione F di Snedecor



Quanto più il valore empirico di F si allontana da 0 quanto più probabile sarà il rifiuto dell'ipotesi nulla di uguaglianza degli effetti delle due confezioni

Nel caso in cui i livelli del fattore sperimentale siano più di due, è interesse del ricercatore indagare quali tra i gruppi hanno portato al rigetto dell'ipotesi nulla

Confronti multipli (test post-hoc)

Questi test mettono in evidenza le coppie di medie statisticamente differenti

Per verificare particolari relazioni tra medie (multipli, trend) si saggiano le loro combinazioni lineari per mezzo dei

Contrasti

Studi sperimentali con struttura più complessa

Il ricercatore è interessato a valutare se nella determinazione dei valori delle vendite è discriminante, oltre alla confezione, anche la zona geografica di appartenenza dei negozi

Disegno fattoriale semplice (ANOVA a due vie)

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

α è l'effetto principale della confezione distribuita,
 β è l'effetto principale della zona geografica di appartenenza,
 $\alpha\beta$ è l'effetto di interazione dei due fattori

SPSS

TRAINING

- Effetti della confezione?
- Influenza della diversa zona geografica?
- Le vendite variano nei diversi possibili incroci tra una confezione e una zona geografica?

$$H_0 : \alpha = 0$$

$$H_0 : \mu_{\text{conf. A}} = \mu_{\text{conf. B}}$$

$$F_{\alpha} = \frac{\text{varianza conf.}}{\text{varianza residua}}$$

$$H_0 : \beta = 0$$

$$H_0 : \mu_{\text{zona geog. 1}} = \mu_{\text{zona geog. 2}}$$

$$F_{\beta} = \frac{\text{varianza zona geog.}}{\text{varianza residua}}$$

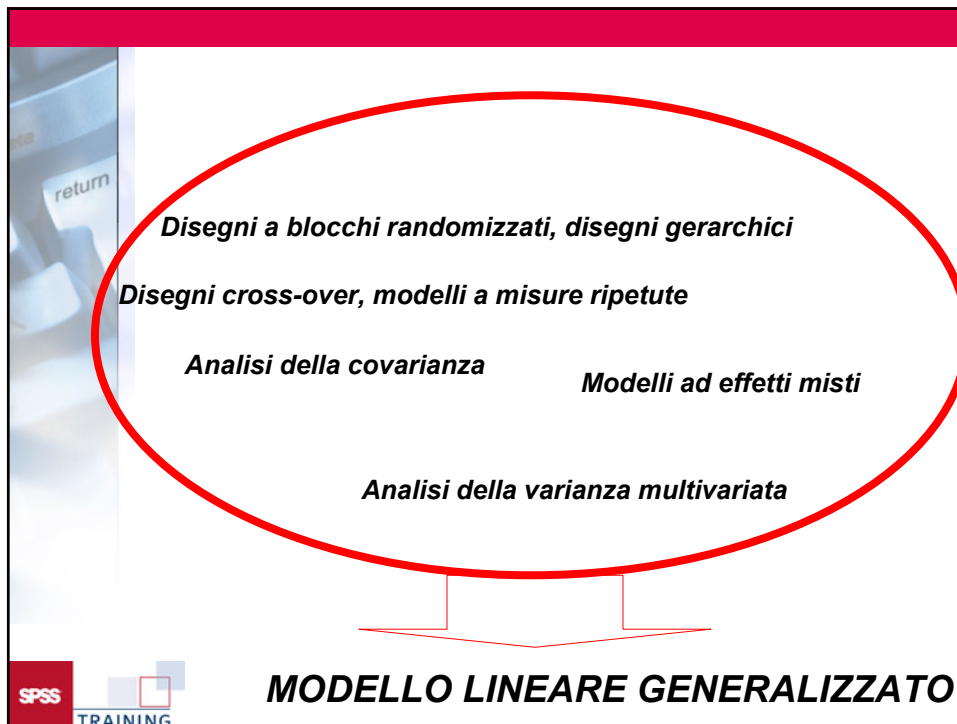
$$H_0 : \alpha\beta = 0$$

$$H_0 : \mu_{\text{conf. A} \cap \text{zona geog. 1}} = \mu_{\text{conf. A} \cap \text{zona geog. 2}} = \\ = \mu_{\text{conf. B} \cap \text{zona geog. 1}} = \mu_{\text{conf. B} \cap \text{zona geog. 2}}$$

$$F_{\alpha\beta} = \frac{\text{varianza conf.*zona geog.}}{\text{varianza residua}}$$

SPSS

TRAINING



Disegni a blocchi randomizzati, disegni gerarchici

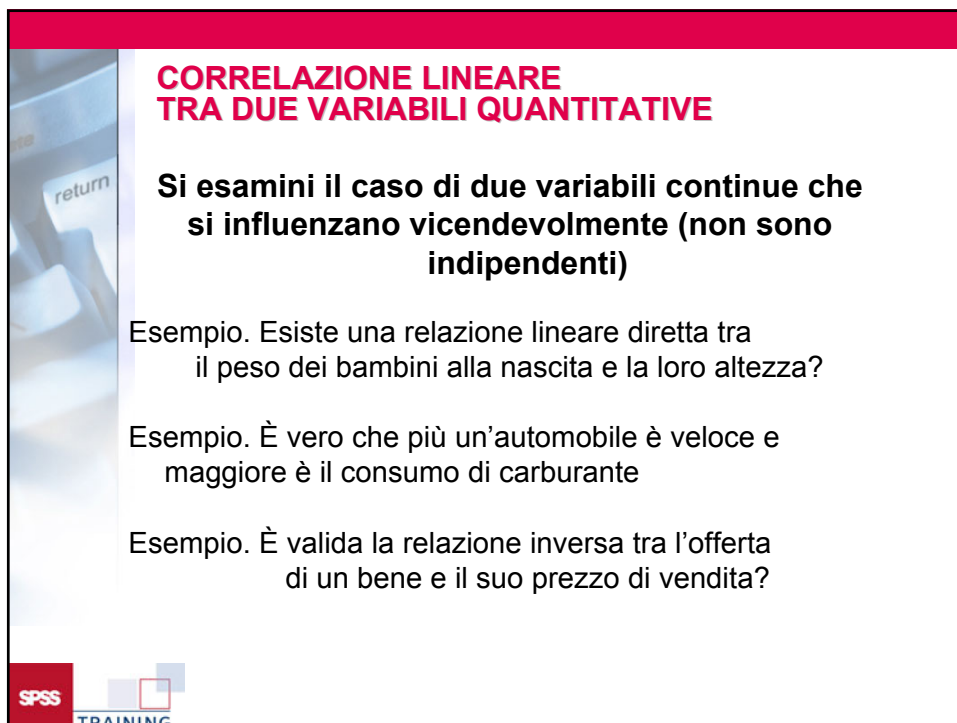
Disegni cross-over, modelli a misure ripetute

Analisi della covarianza *Modelli ad effetti misti*

Analisi della varianza multivariata

MODELLO LINEARE GENERALIZZATO

SPSS TRAINING



**CORRELAZIONE LINEARE
TRA DUE VARIABILI QUANTITATIVE**

**Si esamini il caso di due variabili continue che
si influenzano vicendevolmente (non sono
indipendenti)**

Esempio. Esiste una relazione lineare diretta tra
il peso dei bambini alla nascita e la loro altezza?

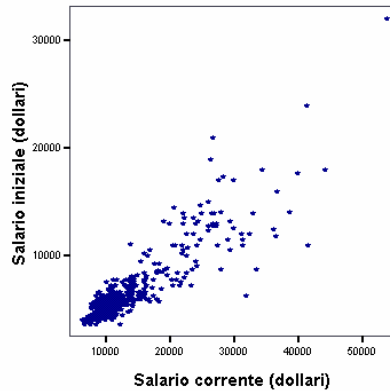
Esempio. È vero che più un'automobile è veloce e
maggiore è il consumo di carburante

Esempio. È valida la relazione inversa tra l'offerta
di un bene e il suo prezzo di vendita?

SPSS TRAINING

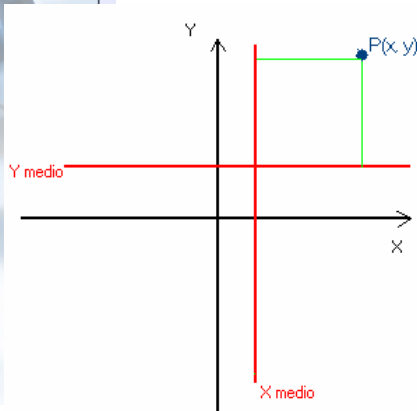
Rappresentazione grafica

Per fornire una rappresentazione grafica dell'andamento congiunto delle due variabili si ricorre al *diagramma a dispersione*



SPSS TRAINING

indicatore sintetico dell'attitudine di due fenomeni a essere suscettibili di variazioni in concomitanza



| X | Y | (X- μ_x) | (Y- μ_y) | |
|---|----|---------------|---------------|----|
| 3 | 6 | -2 | -3 | +6 |
| 4 | 8 | -1 | -1 | +1 |
| 5 | 9 | 0 | 0 | 0 |
| 6 | 10 | +1 | +1 | +1 |
| 7 | 12 | +2 | +3 | +6 |

$$\begin{aligned} \text{Codevianza} &= +14 \\ n-1 &= 4 \\ \text{Covarianza} &= + 3,5 \end{aligned}$$

SPSS TRAINING

COVARIANZA

$$\bar{s}_{xy} = (\sigma_{xy}) = \frac{1}{n-1} \sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{n-1} \sum_i \sum_j x_i y_j - \bar{x}\bar{y}$$

- Misura la variabilità congiunta di due variabili X e Y
- È un indice simmetrico che esprime il segno della direzione
- È compreso tra il prodotto delle due DevSt

$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq \sigma_x \sigma_y$$

- Risente della dimensione delle variabili

SPSS

TRAINING

COEFFICIENTE DI CORRELAZIONE LINEARE DI BRAVAIS-PEARSON

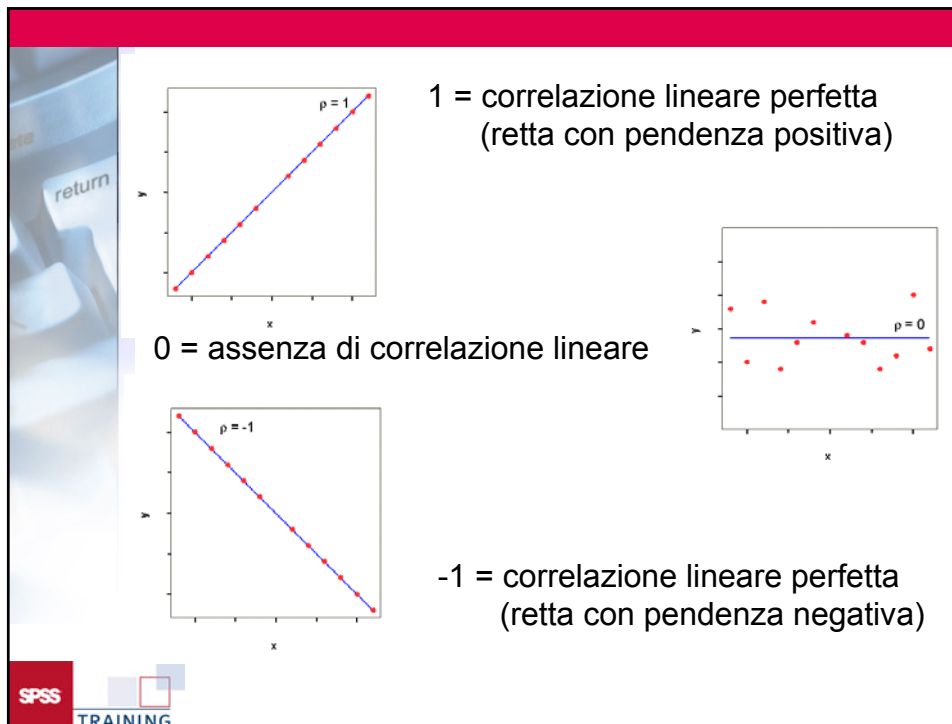
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \rightarrow \quad r_{xy} = \frac{\bar{s}_{xy}}{\bar{s}_x \bar{s}_y}$$

$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq \sigma_x \sigma_y \Rightarrow -1 \leq \rho \leq 1$$

- È simmetrico
- È un indice privo di unità di misura
- Misura l'intensità e il segno del legame lineare tra variabili

SPSS

TRAINING



return

Il modello di regressione lineare semplice

Traduce l'influenza di una variabile su un'altra in termini *funzionali*

- **Finalità descrittivo-esplicative:**
Esiste una relazione?
Qual è la forma analitica più appropriata?
In che maniera può essere spiegata la forma di questa relazione?
- **Finalità predittive:**
prevedere, una volta costruito il modello, il valore di un'informazione successiva

SPSS TRAINING

La regressione lineare

Traduce la dipendenza di Y da X in termini lineari

$$y = \alpha + \beta x$$

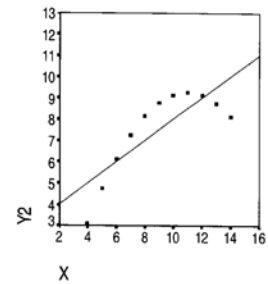
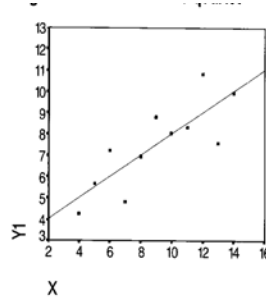
$$\hat{\beta} = \frac{\bar{s}_{xy}}{\bar{s}_x^2}$$

$$\hat{\alpha} = \bar{y} - \bar{x} \frac{\bar{s}_{xy}}{\bar{s}_x^2}$$

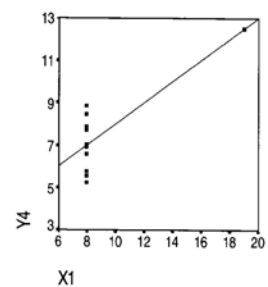
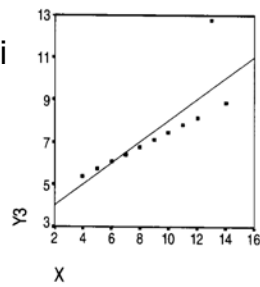
SPSS
TRAINING

Problemi di un modello di regressione:

Linearità



Valori aberranti



SPSS
TRAINING


Analisi dei residui

g p c

SPSS TRAINING

REPORTISTICA TABELLARE E GRAFICA

SPSS TRAINING



OLAP

- **Permette all'utente finale di costruire e manipolare report in tempo reale**
 - **Praticità**
 - **Efficienza**
 - **Flessibilità**
 - **Multiutenza**

SPSS TRAINING



CREAZIONE DI REPORT TABELLARI

- **CREAZIONE DI PROSPETTI**
- **LE TABELLE PERSONALIZZATE**
- **I CUBI OLAP**

SPSS TRAINING



GRAFICI STANDARD E GRAFICI INTERATTIVI

I **grafici standard** permettono all'utente di esplorare una particolare visualizzazione grafica dei dati

I **grafici interattivi** sono invece dinamici e consentono di esplorare interattivamente le informazioni dei dati cambiando punto di vista

SPSS

TRAINING



Grafici Interattivi

l'utente finale costruisce grafici inserendo progressivamente nuovi elementi e dimensioni di analisi

- alta risoluzione
- massima efficacia descrittiva
- potenza esplorativa

SPSS

TRAINING



Mappe tematiche e geomarketing

Il modulo Maps di SPSS rende possibile la rappresentazione cartografica dei dati dell'utente, in modo da ottenere mappe tematiche basate su una variabile geografica.

SPSS

TRAINING




Mappe tematiche

- **Mappe tematiche:** visualizzazione dei dati in una mappa, utilizzando colori, simboli o grafici.
- Le mappe sono composte di **strati**.
- Ogni strato viene sovrapposto ad un altro per creare una mappa dettagliata (come un lucido di proiezione).
- Gli strati vengono combinati secondo un'area geografica in un insieme denominato **geoset**.
- La creazione di geoset in SPSS avviene mediante il **gestore geoset**.



SPSS

TRAINING




Geoset

- I geoset contengono le informazioni necessarie a disegnare una mappa.
- È possibile creare nuovi geoset e aprire o salvare quelli esistenti con il Gestore di geoset. Gli strati di un geoset possono invece essere registrati e modificati nel Gestore geodizionario. L'opzione SPSS Maps utilizza i geoset nel formato MapInfo.
- Ulteriori geoset sono disponibili nella directory Geosets del CD-ROM di SPSS. Per installarli, copiare i file Geoset desiderati nella directory SPSS Maps del computer, quindi registrare i nuovi geoset con il Gestore di geoset.



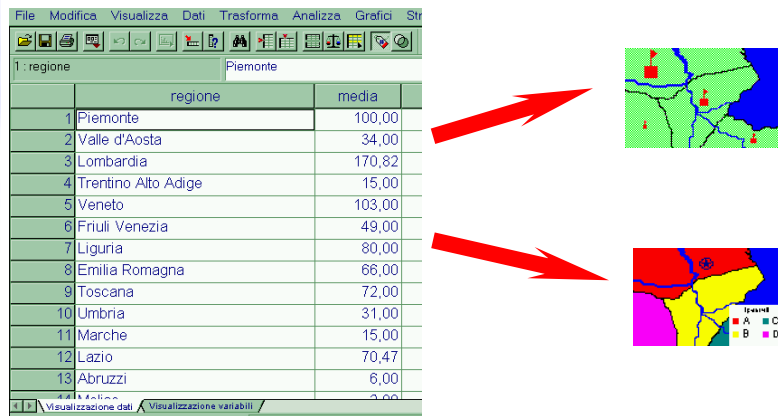
Mappe tematiche

- Per rappresentare dei dati sulla mappa è necessario un collegamento con la base di dati da rappresentare:
 - *Variabili geografiche.*
 - *Collegamento X, Y*
 - *Tabelle di riferimento punti*



Il geomarketing

Il geomarketing è uno strumento per stabilire il rapporto ottimale tra le proprie capacità di offerta e la localizzazione della domanda.



Nel geomarketing le mappe vengono utilizzate per

- Individuazione e localizzazione dei clienti e competitors attuali e potenziali
- localizzazione e ottimizzazione della rete distributiva e dei singoli punti vendita.
- definizione e analisi del bacino di utenza di un punto di vendita, in base alla costruzione di aree isocrone
- stima del potenziale espresso da una determinata area geografica per uno specifico settore merceologico
- definizione degli obiettivi di vendita, per agente o per area.
- pianificazione di campagne pubblicitarie e di azioni promozionali in funzione della localizzazione del proprio target.
- ottimizzazione del mix dei prodotti per singolo punto vendita, in funzione della localizzazione.

SPSS Maps è in grado di produrre le seguenti mappe:

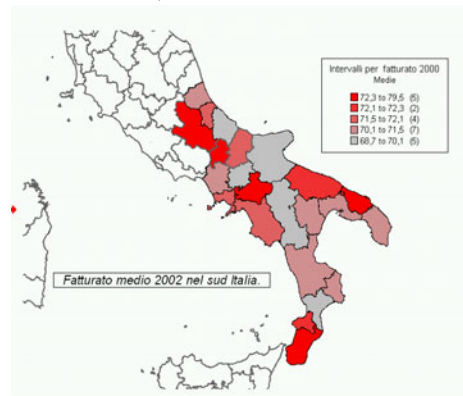
- intervallo di valori
- densità di punti
- simbolo graduato
- valori individuali
- grafici a barre
- grafici a torta
- mappe tematiche multiple rappresentate da una combinazione delle tematiche precedenti

SPSS

TRAINING

Mappa ad intervallo di valori

Un intervallo di valori di una statistica riassuntiva viene suddiviso in categorie e viene indicata la categoria di intervallo per ogni unità geografica, ad esempio 0-100, 101-200, 201-300 e così via.



SPSS

TRAINING

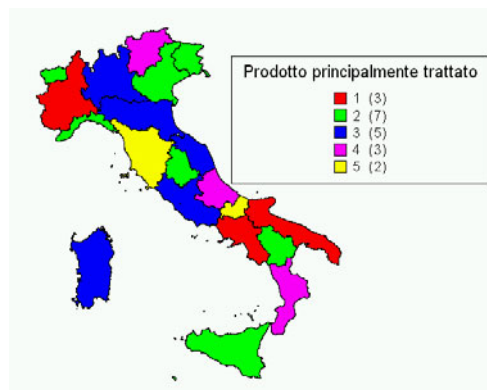
Mappa simbolo graduato

Vengono utilizzati i simboli per rappresentare i valori di una variabile di scala in base all'area geografica. Le dimensioni di ciascun simbolo sono proporzionali al valore per la regione specifica.



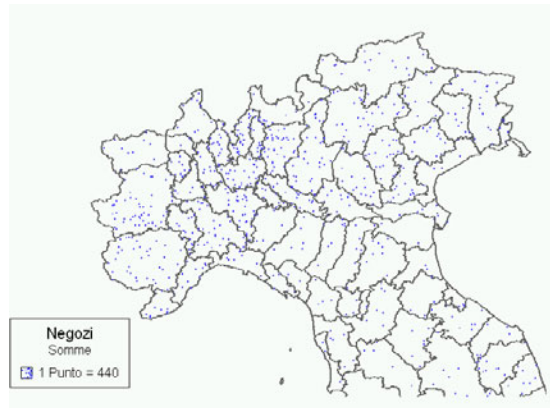
Mappa a valori individuali

Risulta utile per i file di dati nei quali ciascun caso rappresenta una singola area geografica, ad esempio una regione specifica



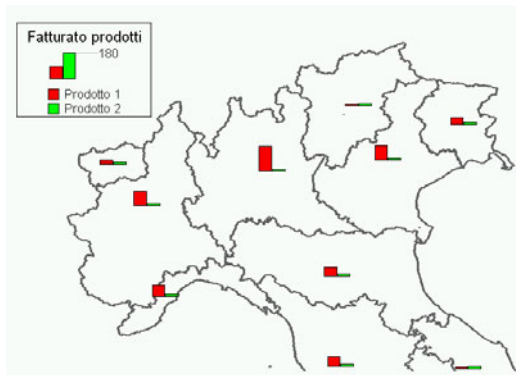
Mappa a densità di punti

Utilizzano punti per rappresentare un valore entro un'area o un limite geografico. La somma dei punti di un'area è proporzionale al valore relativo a quell'area.



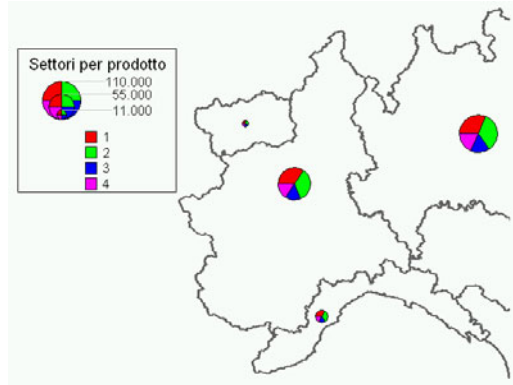
Mappa con grafici a barre

Utili se si desidera confrontare più variabili o categorie su un'area geografica. Ciascuna variabile o categoria è rappresentata da una barra inserita nel grafico mediante il confronto con una mappa in secondo piano.



Mappa con grafico a torta

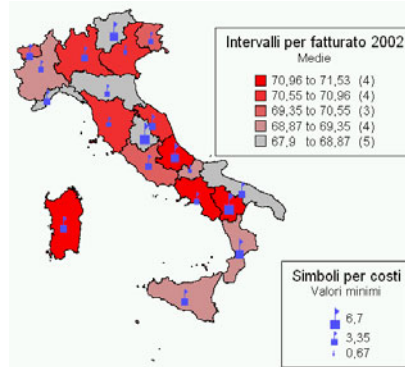
E' visualizzato il contributo delle parti a un intero. L'inserimento di torte su una mappa consente di verificare le differenze nella distribuzione delle categorie in base all'area geografica.



SPSS
TRAINING

Mappa tematica multipla

Utile per creare una mappa con più tipi di grafici o tematiche. Ad esempio, per illustrare il fatturato e i costi di distribuzione per area, è possibile utilizzare gli intervalli di valori per mostrare il fatturato e i grafici a barre per mostrare i costi di distribuzione.



SPSS
TRAINING




Campioni complessi con COMPLEX SAMPLES SPSS 12



Indagine Statistica

- i. progettazione**
- ii. rilevazione**
- iii. registrazione**
- iv. revisione e codifica**
- v. elaborazione**
- vi. validazione**
- vii. diffusione**

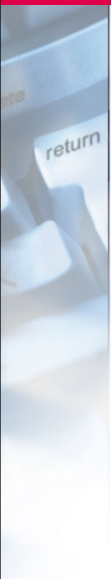




Popolazioni Finite

- **Population of inference (PI).** Popolazione sotto inferenza
- **Target population (PO).** Popolazione obiettivo
- **Frame population (PC).** Popolazione base per il campionamento.
- **Survey population (PE).** Popolazione effettivamente indagata.

SPSS TRAINING



Error Profile

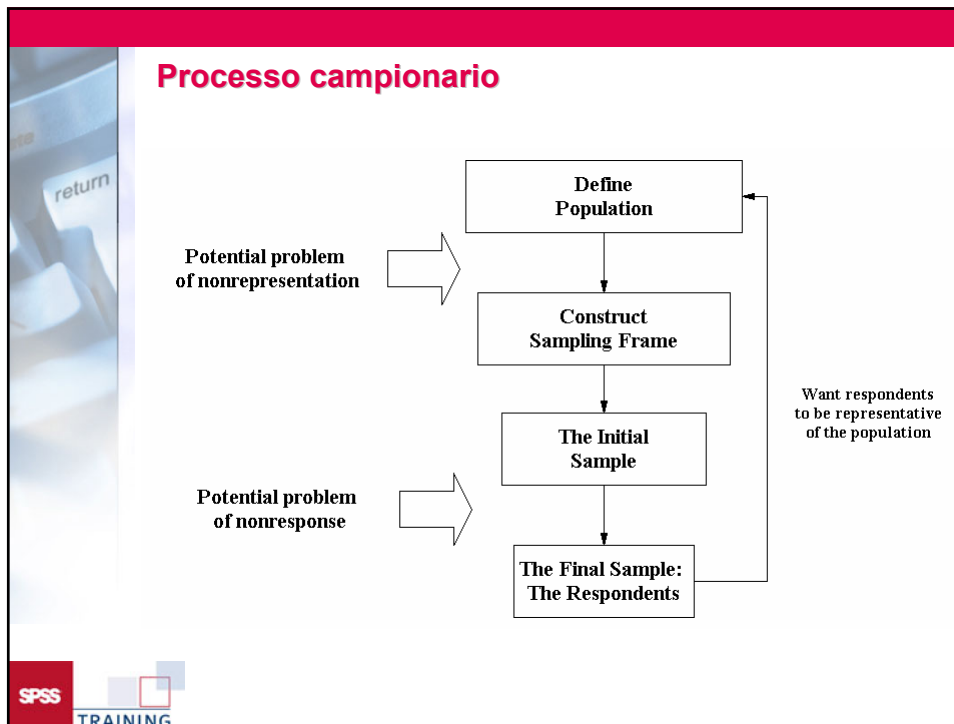
Errori Campionari

- **Errore di stima**

Errori non Campionari

- **Copertura**
- **Mancate risposte**
- **Errori di misura**

SPSS TRAINING



- ## Piani di campionamento
- ### Campionamenti non probabilistici
- a scelta ragionata (bilanciato, semi probabilistico, troncato)
 - campionamento per quota
 - campionamento tramite testimoni privilegiati
- ### Campionamenti probabilistici
- semplice con ripetizione (SCR)
 - semplice senza ripetizione (SSR)
 - stratificato (ST)
 - a grappoli (GR)
 - sistematico (SM)
 - a due stadi (DS)
- SPSS TRAINING



Stimatori del totale

SCR


- probabilità costanti → STIMATORE PER ESPANSIONE
- probabilità variabili → STIMATORE DI HANSEN-HURWITZ

SSR

- probabilità costanti → STIMATORE PER ESPANSIONE
- probabilità variabili → STIMATORE DI HORVITZ-THOMPSON

STRATIFICATO

- probabilità costanti → STIMATORE DEL TOTALE NELLO STRATIFICATO



Stimatori del totale


GRAPPOLI

- probabilità costanti
- probabilità variabili

DUE STADI

- probabilità costanti → STIMATORE PER ESPANSIONE NEL GRUPPO
- probabilità variabili

SISTEMATICO



Stimatori alternativi

- **Stimatore per quoziente**

$$\hat{Y}_q = \frac{\hat{Y}}{\hat{X}} \cdot X$$

- **Stimatore per regressione**

$$\hat{Y}_r = \hat{Y} + \beta(X - \hat{X})$$

SPSS

TRAINING

Effetto del disegno

$$Deff = \frac{V(\hat{Y})}{V_0(\hat{Y})}$$

Se $Deff > 1$ → il piano di campionamento utilizzato è **meno efficiente** del campionamento semplice,

Se $Deff = 1$ → ha la **stessa efficienza**,

Se $Deff < 1$ → è **più efficiente**.

SPSS

TRAINING



Applicazioni Pratiche

- [Demo 1: Campionamento Casuale Semplice – Stima della proporzione](#)
- [Demo 2: Campionamento Sistemático](#)
- [Demo 3: Campionamento Stratificato](#)
- [Demo 4: Campionamento Stratificato - Stima della Proporzione](#)
- [Demo 5: Campionamento a Grappoli – Intervalli di confidenza e test](#)
- [Demo 6: Campionamento a grappoli – Stimatori alternativi](#)
- [Demo 7: Campionamento a due Stadi](#)

