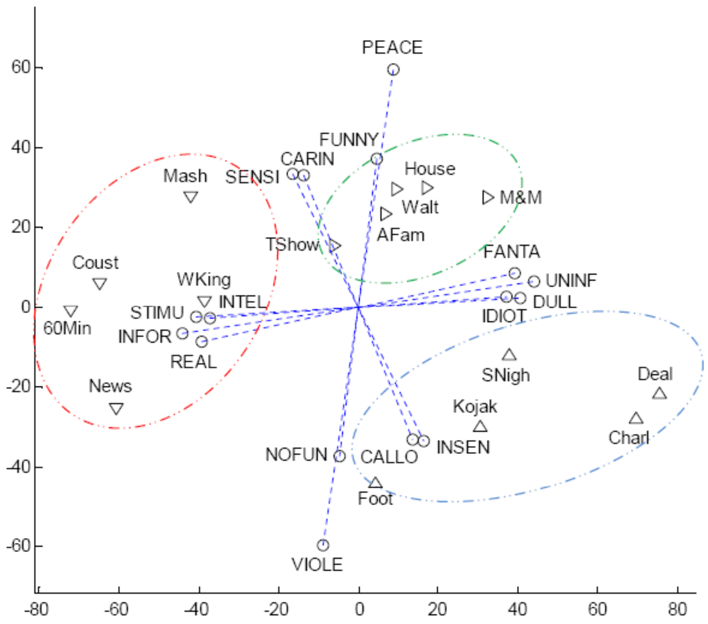


Metodi Statistici per il Management

Applicazioni: Marketing e Vendite

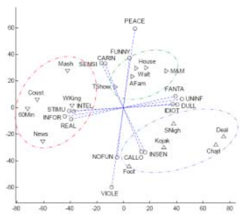
Simone Borra - Roberto Rocci

Le previsioni in azienda

- La complessità dell'ambiente esterno, la competizione concorrenziale, la globalizzazione dei mercati impongono alle aziende di utilizzare processi decisionali supportati da proiezioni di breve e di lungo periodo.
- Le previsioni costituiscono uno dei fatti fondamentali per i processi di pianificazione aziendale. In particolare, le previsioni riguardanti le *vendite* sono alla base di scelte operate dal management relativamente a investimenti, tecnologie, assetti organizzativi e commerciali dell'impresa.
- I metodi di previsione possono essere:

  - ***qualitativi.*** *Basati preminente su opinioni soggettive;*

  - ***quantitativi.*** *Basati su dati empirici analizzati con metodologie statistiche.*

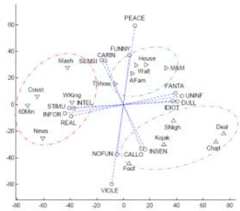


I metodi qualitativi

La varietà dei metodi qualitativi è molto ampia. Tra i metodi più diffusi ricordiamo:

- ***panel di esperti***;
- ***metodo Delphi.***

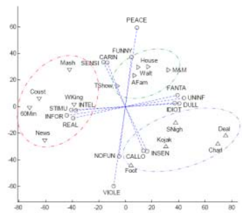
- Gli esperti interagiscono individualmente con il ricercatore (agiscono come se fossero in gruppo, ma senza il rischio di distorsione che provocherebbe il mutuo condizionamento)
- Tra un round e l'altro, il ricercatore fornisce informazioni mirate agli esperti (feedback)



I metodi quantitativi

I metodi quantitativi, pur con le differenze esistenti tra le molteplici tecniche, possono essere ricondotti al seguente schema logico:

- ***input***: dati necessari per il processo di previsione;
- ***modello di analisi***: sistema o algoritmo di elaborazione dei dati di input;
- ***output***: previsioni e stime finali.

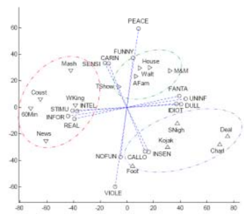


Modelli estrapolativi e causali

Nell’ambito dei metodi quantitativi si distinguono due famiglie:

- ***Modelli estrapolativi***: l’evoluzione della domanda dipende unicamente dalla variabile tempo, è quindi un fenomeno intrinseco. La previsione viene fatta basandosi unicamente sulla osservazione dei valori passati

- ***Modelli causali***: l’evoluzione della domanda è collegata all’andamento futuro di variabili socioeconomiche correlate. Si tratta di individuare le variabili che sono legate alla domanda di un bene e di applicare un modello di previsione adeguato.



Ricerca del Trend

Metodi per la ricerca del trend intrinseco, elencati in ordine di complessità:

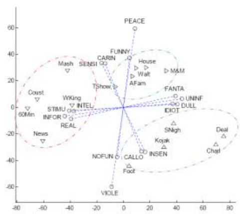
- ***Medie mobili***

  ad ogni valore della serie osservata viene sostituita la media di un certo numero di valori consecutivi
- ***Modello di regressione***

  la variabile dipendente è costituita dalla domanda, mentre il tempo è la variabile esplicativa.
- ***Analisi classica delle serie storiche***

  *l*a serie storica è decomposta nelle sue diverse componenti: *trend*, *stagionalità*, errore.
- ***Analisi Box-Jenkins***

  la serie storica è considerata una realizzazione di un processo stocastico.



Regressione lineare per dati temporali

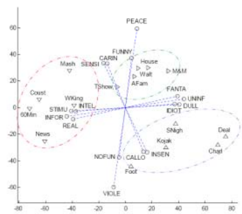
Le osservazioni sono rilevate nel tempo. Non possiamo assumere la loro indipedenza.

$$y_t = b_0 + b_1 x_{t1} + b_2 x_{t2} + \ldots + b_k x_{tk} + w_t$$

Equazione *Assunzioni*

$E(y_t \mid x_{t1},\ldots,x_{tk}) = \mu_t = b_0 + b_1 x_{t1} + \ldots + b_k x_{tk}$ $\quad y_t \mid \mathbf{x}_t \sim N(\mu_t, \sigma_w^2)$, dip.

Un possibile modello per la dipendenza delle osservazioni, o equivalentemente degli errori, è il modello autoregressivo del primo ordine (AR1).



reg: AR(1)

Il modello AR(1) per il termine di errore è

$w_t = \rho w_{t\text{-}1} + \varepsilon_t$ $\varepsilon_t \sim N(0, \sigma^2_\varepsilon)$, indipen.

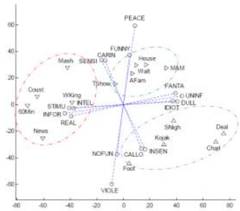
Si dimostra che:

- la varianza è costante (omoschedasticità)

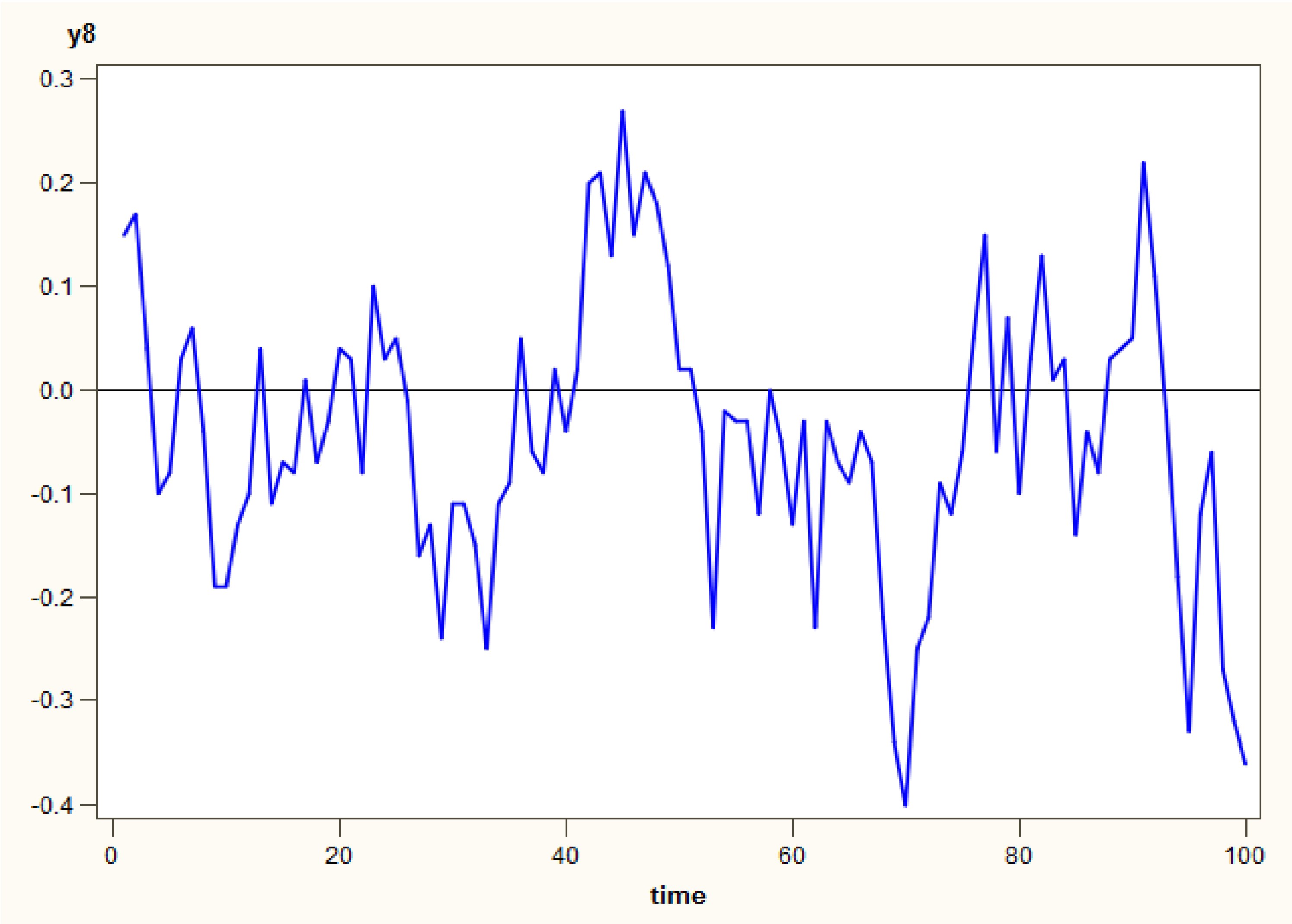
$$\mathrm{Var}(w_t) = \sigma^2_w = \sigma^2_\varepsilon / (1 - \rho^2)$$

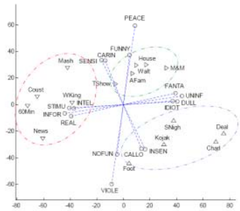
- la covarianza e la correlazione dipendono solo dal lag temporale e decadono al crescere di questo

$$\mathrm{Cov}(w_t, w_{t-r}) = \rho^r \sigma^2_w, \ \mathrm{Corr}(w_t, w_{t-r}) = \rho^r$$

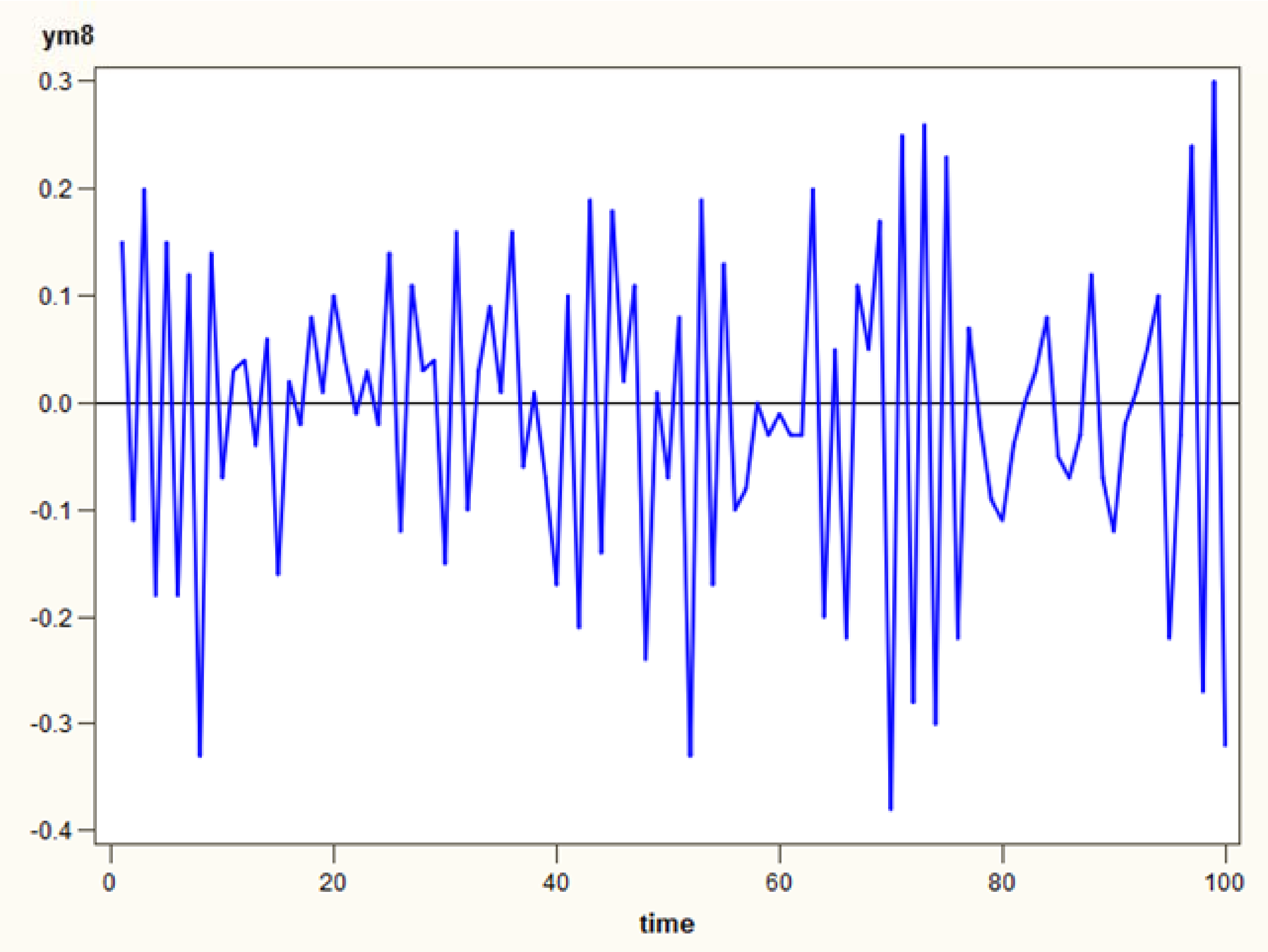


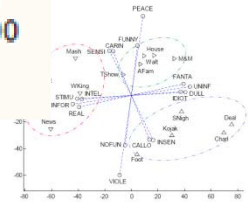
reg: esempio AR(1)





reg: esempio AR(1)



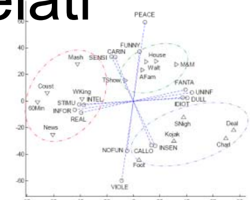


reg: diagnostica

Al fine di capire se c'è autocorrelazione tra gli errori è utile calcolare la seguente statistica utilizzando i residui della regressione OLS

$$DW = \frac{\sum_{t=2}^{n} (\hat{w}_t - \hat{w}_{t-1})^2}{\sum_{t=1}^{n} \hat{w}_t^2} \approx 2(1-\rho)$$

- un valore vicino a 2 indica assenza di autocorrelazione
- valori piccoli (<<2) indicano che i residui successivi sono, in media, vicini in valore l'uno all'altro, o correlati positivamente
- valori grandi (>>2) indicano che i residui successivi sono, in media, molto differenti in valore l'uno dall'altro, o correlati negativamente



reg: dt stima, C.I. e test

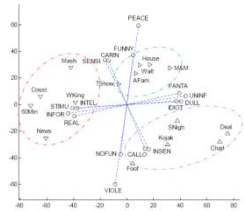
- la stima dei parametri si ottiene con il metodo della massima verosimiglianza assumendo una distribuzione normale per gli errori
- sfruttando la normalità asintotica degli stimatori di massima verosimiglianza si possono costruire intervalli di confidenza, test Z e di Wald.
- per confrontare modelli possiamo utilizzare:

Akaike's Information criterion

AIC = -2log(max *likelihood*) + 2(# parameters)

Bayesian Information criterion

BIC = -2log(max *likelihood*) + log(*n*)(# parameters)



reg: esempio

Number of Observations Used	100

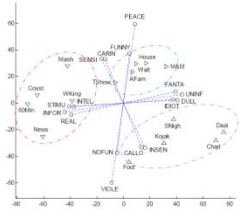
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.60	2.60	35.52	<.0001
Error	98	7.18	0.07		
Corrected Total	99	9.78			

Root MSE	0.27	R-Square	0.27
Dependent Mean	1.19	Adj R-Sq	0.26
Coeff Var	22.77		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	0.164	0.174	0.94	0.3473	-0.181	0.510
income	income	1	0.523	0.088	5.96	<.0001	0.349	0.697

Durbin-Watson D	0.63
Number of Observations	100
1st Order Autocorrelation	0.647

$$\rho \approx 1 - \frac{DW}{2} = 1 - \frac{0{,}63}{2} = 0{,}685$$



Reg AR(1): esempio

Maximum Likelihood Estimates			
SSE	3.928	DFE	97
MSE	0.040	Root MSE	0.201
SBC	-25.454	AIC	-33.270
MAE	0.164	AICC	-33.020
MAPE	15.430	HQC	-30.107
Log Likelihood	19.635	Regress R-Square	0.496
Durbin-Watson	2.053	Total R-Square	0.599
		Observations	100

È la stima della deviazione standard di ε

È la bontà di adattamento del modello AR(1)

Parameter Estimates						
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|	Variable Label
Intercept	1	0.152	0.123	1.23	0.2206	
income	1	0.527	0.054	9.78	<.0001	income
AR1	1	-0.693	0.077	-9.05	<.0001	

Durbin-Watson Statistics	
Order	DW
1	2.053

In SAS il coefficiente di autocorrelazione viene stimato con segno opposto a quello del modello

reg: esempio

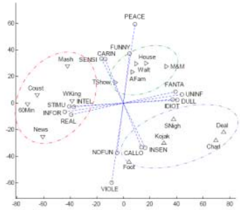
Modello di regressione

Stime dei minimi quadrati ordinari			
SSE	7.18	**DFE**	98
MSE	0.07	**Radice MSE**	0.2707
SBC	29.63	**AIC**	24.41862
MAE	0.21	**AICC**	24.54233
MAPE	19.82	**HQC**	26.52734
Durbin-Watson	0.63	**R-quadro regr.**	0.266
		R-quadro totale	0.266

Modello AR(1)

Stime della massima verosimiglianza			
SSE	3.93	**DFE**	97
MSE	0.04	**Radice MSE**	0.20123
SBC	-25.45	**AIC**	-33.2697
MAE	0.16	**AICC**	-33.0197
MAPE	15.43	**HQC**	-30.1066
Log verosimiglianza	19.63	**R-quadro regr.**	0.4964
Durbin-Watson	2.05	**R-quadro totale**	0.5985
		Osservazioni	100

BIC

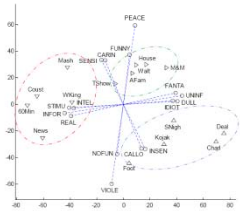


Segmentazione concorrenza: Cluster analysis

- Con il termine ***segmentazione*** si intende in generale la definizione di sottoinsiemi omogenei (cluster) su un campione di *n* unità (clienti, competitor, prodotti, zone).

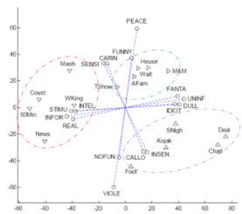
- Per ***omogeneità*** si intende il fatto che i cluster sono caratterizzati da elementi tra di loro vicini *simili* (vicini) rispetto ad un set di variabili di interesse.

- Gli obiettivi della segmentazione sono molteplici, in ogni caso vi è l'esigenza di ragionare e attuare strategie rispetto a *k* insiemi piuttosto che su un numero elevato (*n*) di oggetti.

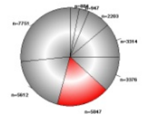


Il caso cartaviaggio trenitalia

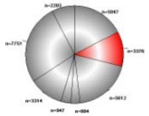
- Si analizzano le transazioni di acquisto effettuate da un campione di clienti cartaviaggio trenitalia (anno 2006) al fine di segmentare la clientela rispetto ai comportamenti di viaggio e acquisto.
- Si considerano le variabili legate alla:
  – Spesa (es: spesa media mensile; spesa media per ogni biglietto,..)
  – Frequenza di viaggio e di spesa (es: media biglietti mensili; numero biglietti;…)
  – Tipologia di treno preferita (es: % eurostar; %intercity; …)
  – Distribuzione dei viaggi nel tempo (es: indice di omogeneità,…)
- Alle variabili si applica una cluster analysis non gerarchica (k-means) e si individuano 8 gruppi



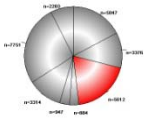
Descrizione dei gruppi



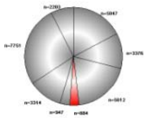
1. Occasionali Extra-comfort



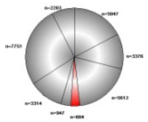
2. Gli allegri viaggiatori



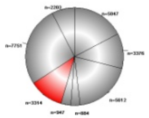
3. InterItaly solo andata



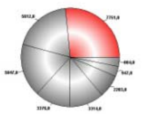
4. I fedelissimi



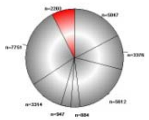
5. Occasionali lunghe tratte



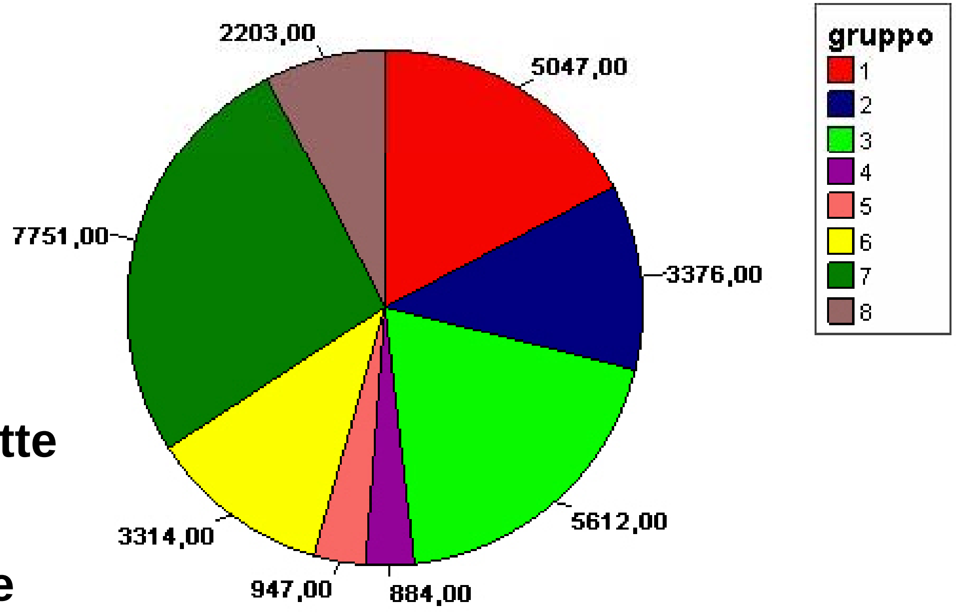
6. Occasionali brevi tratte

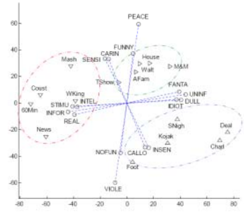


7. Gli amanti evasivi

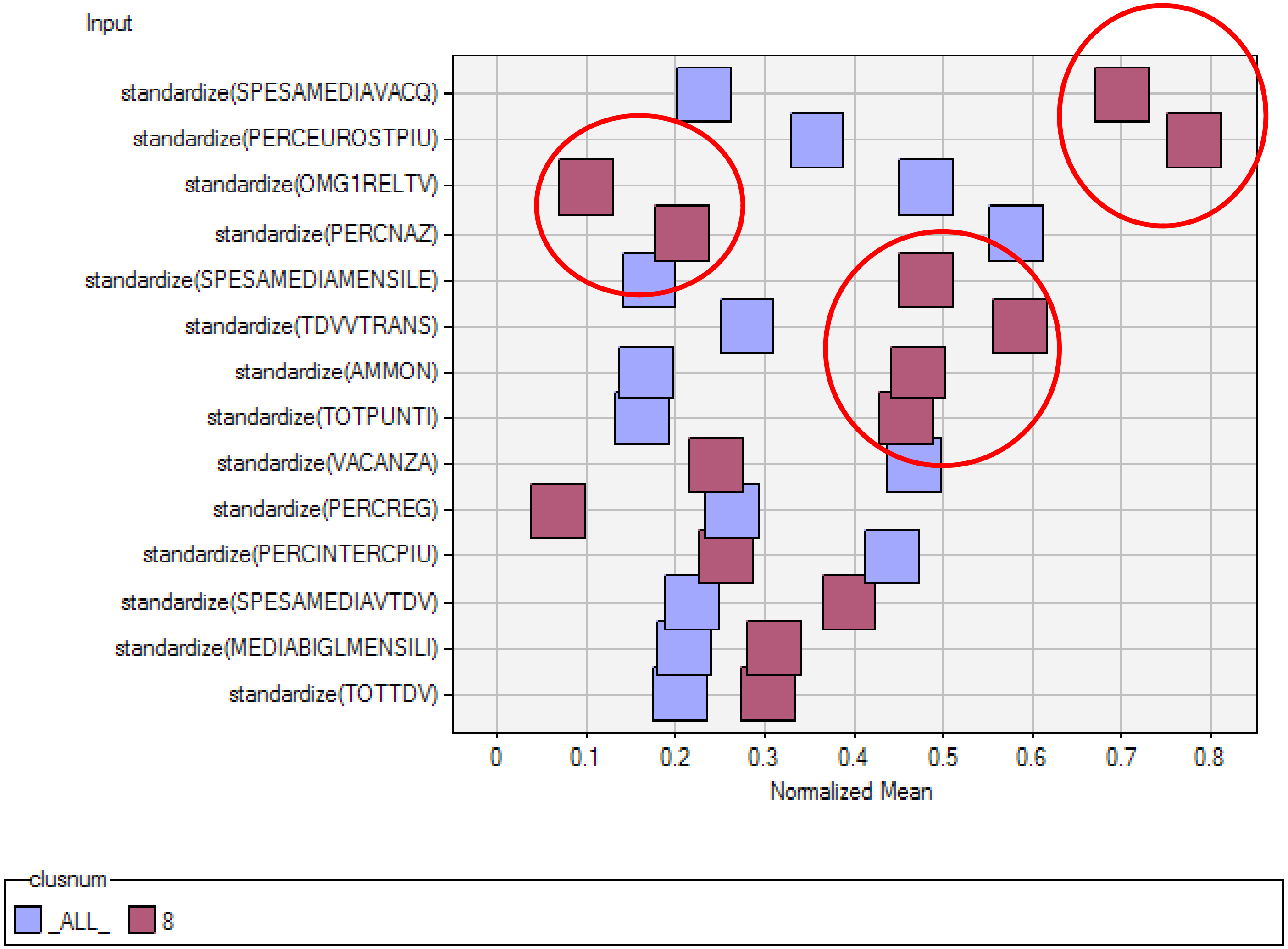


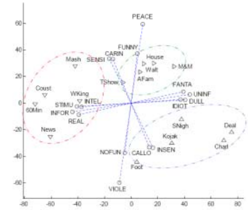
8. I business customers





Confronto gruppo 8 rispetto all'intero campione

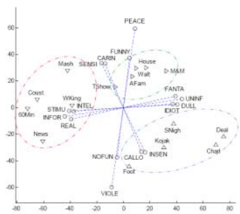




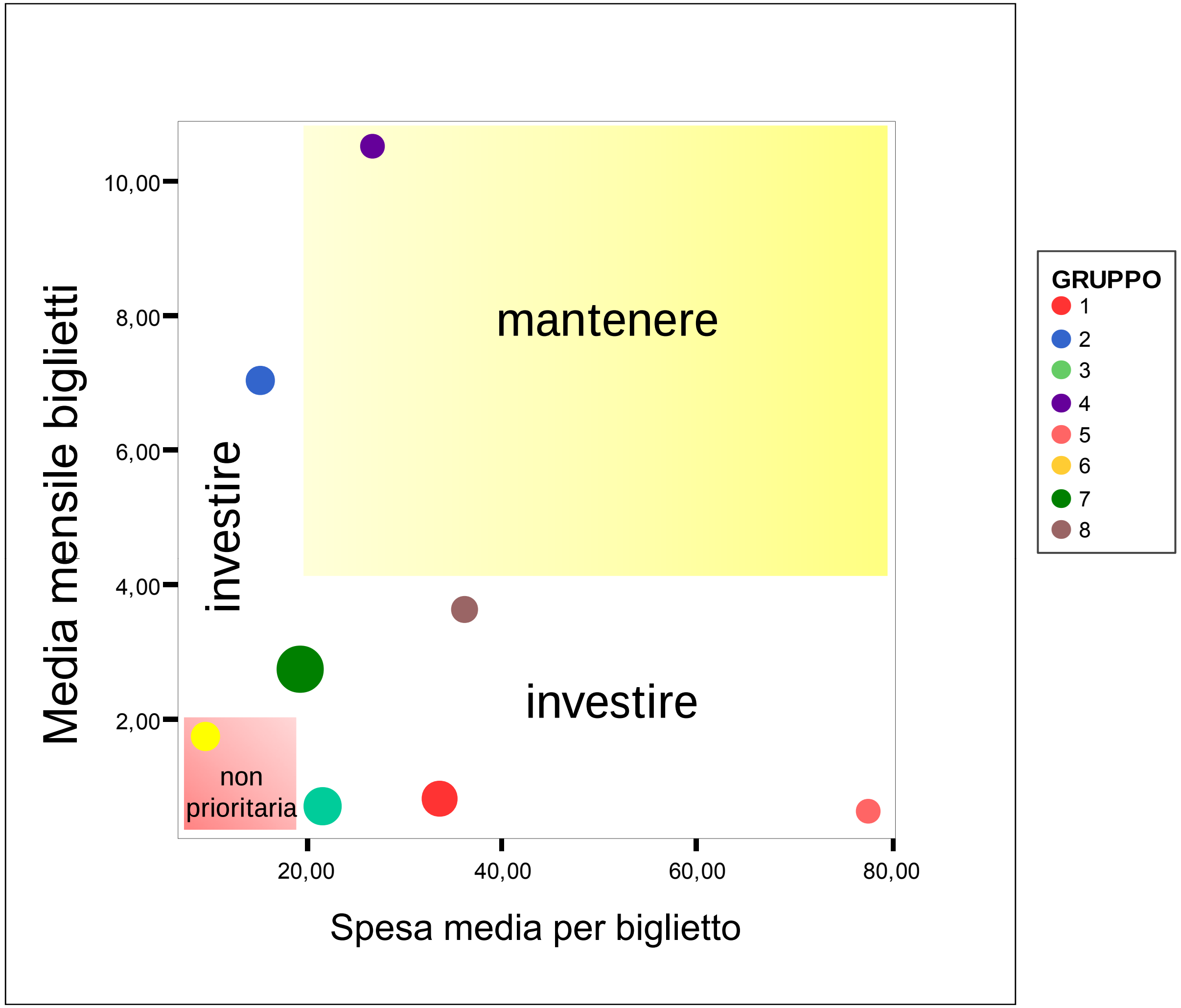
Esempio descrizione gruppo 8

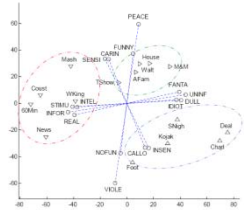
- *non può essere considerato un cliente occasionale* (visti i suoi livelli di spesa, di frequenza e l'indice di omogeneità);
- *preferisce viaggiare su treni di tipo Eurostar* (elevata percentuale di treni Eurostar presi) *ed esige quindi un servizio di buona qualità*;
- *non effettua in treno viaggi a lunga percorrenza*
- *percorre abbastanza costantemente le stesse tratte* (preferenza per un unico tipo di treno, basso indice di omogeneità);
- *è fondamentalmente un uomo di affari* (età media elevata, numero preponderante di dirigenti ed imprenditori)

In sintesi si tratta di coloro che viaggiano soprattutto per motivi di lavoro, che allo stesso tempo non si preoccupano del costo del biglietto, bensì si preoccupano di vedere soddisfatta nel migliore dei modi una propria necessità di trasporto ed esigono quindi dall'azienda fornitrice del servizio alte prestazioni (in termini di qualità, comodità e puntualità garantita).



Mappatura opportunità con sovrapposizione gruppi





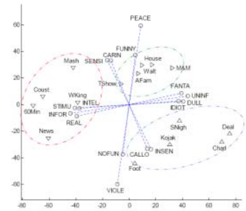
Ricerca di mercato

Obiettivi:

- previsione delle vendite;
- analisi della soddisfazione dei clienti;
- gradimento rispetto al lancio di nuovi prodotti;
- analisi comportamenti di acquisto.

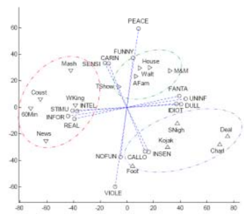
La ricerca può essere realizzata all'interno dell'azienda o commissionata ad agenzie specializzate.

Le ricerche di mercato possono essere molto differenti in base agli obiettivi, ai budget di spesa, al settore di riferimento.



Le fasi dell'indagine

1. Progettazione.
2. Somministrazione.
3. Rilevazione e input dati.
4. Elaborazione.
5. Output e rappresentazione.
6. Analisi e diffusione.

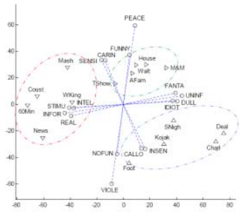


L’elaborazione

- L’***elaborazione*** è la fase più critica e importante dell’intero processo. Si tratta di trasformare *dati elementari* in *informazioni sintetiche*.

- Per ***dato*** si intende la singola unità informativa riferita alla singola unità: in pratica, può essere pensata come la risposta data da una persona ad una domanda.

- Per ***informazione*** intendiamo una sintesi di dati tramite processi di aggregazione o funzioni statistiche (somme, medie, …).



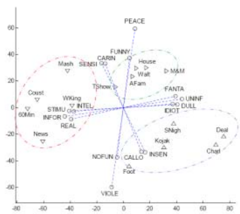
Tipi di output

A livello metodologico le elaborazioni consistono nell’applicazione di varie tecniche statistiche:

- tabelle e distribuzioni di frequenza;
- calcolo di misure di sintesi - media, moda, mediana;
- calcolo di misure di dispersione;
- verifica di dipendenza (correlazione, chi-quadro, tabelle bivariate);
- analisi multivariata.

Se si utilizza il foglio elettronico Excel gli strumenti più importanti sono:

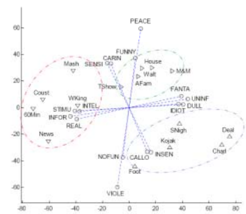
- funzioni statistiche;
- tabelle pivot (utilissime!);
- grafici.



Questionario e struttura dati

Il ***questionario*** è lo strumento fondamentale per realizzare una ricerca di mercato.

- Le domande possono essere di vari tipi:
  - Risposta aperta.
  - Binarie (V/F; S/N; …).
  - Risposta chiusa: una modalità su k possibili.
  - Risposta chiusa multipla: più modalità su k possibili.
  - Scale di punteggi.
- A livello di struttura dati:
  - Ogni risposta è una variabile.
  - Ogni questionario è una unità.



Strategia campionaria

Per ottenere informazioni di “buona qualità” il campione deve essere **rappresentativo**, ossia un’immagine abbastanza fedele di quegli aspetti della popolazione ritenuti rilevanti ai fini dell’indagine.

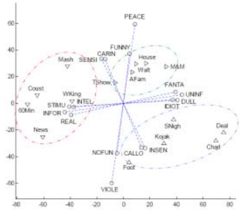
Formazione del campione: **non probabilistico – probabilistico**

Non probabilistico

Ad esempio, selezione del campione **per quote**. La scelte dell’unità è lasciata all’operatore e la dimensione del campione è fissata con criteri di convenienza.

Probabilistico

Equivale all’estrazione da un’urna di un certo numero di palline secondo una strategia o *piano di campionamento* che assegna una probabilità di estrazione ad ogni campione.



Strategia campionaria

non probabilistici

campioni

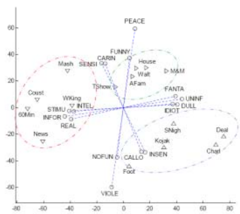
probabilistici

Campioni non probabilistici

1) A scelta ragionata
2) Per quote
3) Tramite testimoni privilegiati

Campioni probabilistici

1) Semplici con o senza ripetizione
2) Stratificato
3) A grappoli
4) Sistematico
5) A due stadi

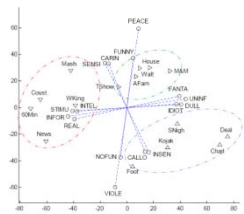


Strategia campionaria

Il campionamento **non probabilistico** non fornisce a ciascuna unità della popolazione la stessa possibilità di far parte del campione, pertanto alcune unità statistiche hanno maggiore probabilità di essere estratte dalla popolazione.

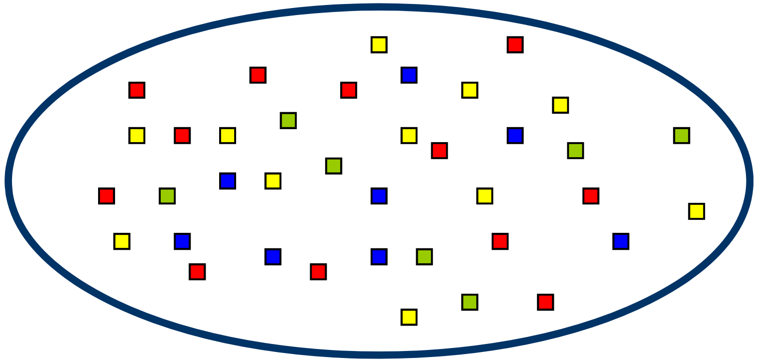
Questo metodo prevede l’estrazione del campione in base a **criteri di comodo o di praticità**: per esempio perché le unità da campionare sono più facilmente accessibili, o per ragioni di costo, o perché in una certa zona sono disponibili volontari ecc.

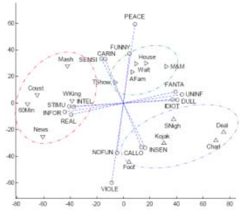
Un campione selezionato con questi criteri di comodo, sebbene abbia il vantaggio della rapidità, è soggetto ad un forte *bias*, fornisce dati poco affidabili e può essere facilmente viziato da errori sistematici.



Strategia campionaria

Mentre il campione non probabilistico pretende di essere rappresentativo, il campione probabilistico non ha simili pretese; esso è soltanto uno fra i possibili campioni estraibili dalla popolazione. Il vantaggio che presenta è, oltre all'eliminazione di un soggettivismo rischioso, nella possibilità di costruire un modello matematico per porre su basi razionali la scelta tra varie strategie possibili. La **teoria dei campioni** permette di valutare l'attendibilità dei risultati.



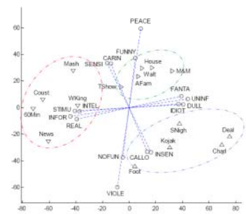


Campionamento probabilistico

- **Campionamento casuale**: insieme di tutte quelle tecniche di formazione del campione in cui la selezione delle unità è affidata a regole probabilistiche.

- **Campionamento casuale semplice**: i campioni della stessa dimensione estraibili da una popolazione hanno uguale probabilità di essere estratti.

- **Campionamento casuale stratificato**: la popolazione viene suddivisa in un certo numero di strati. Da ogni strato in maniera indipendente viene poi estratto un campione casuale semplice.



Dal campione alla stima di un parametro della popolazione – campione casuale semplice

Parametro della Popolazione Y è la media della popolazione: $\theta = \mu$

Sia *N* la numerosità della popolazione e *n* la dimensione del campione il rapporto *n/N=f* è detto frazione di sondaggio.

Stimatore corretto ed efficiente della media

- **Popolazione finita** (estrazione senza ripetizione)

Stimatore corretto della media è la media campionaria: $T = \overline{Y}$

Con varianza

$$Var\left(\overline{Y}\right) = \left(1 - f\right)\frac{\sigma^2}{n}$$

- **Popolazione infinita** (estrazione con ripetizione)

Stimatore corretto della media è la media campionaria: $T = \overline{Y}$

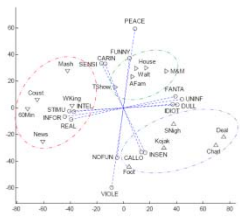
Con varianza

$$Var\left(\overline{Y}\right) = \frac{\sigma^2}{n}$$

Notiamo che:

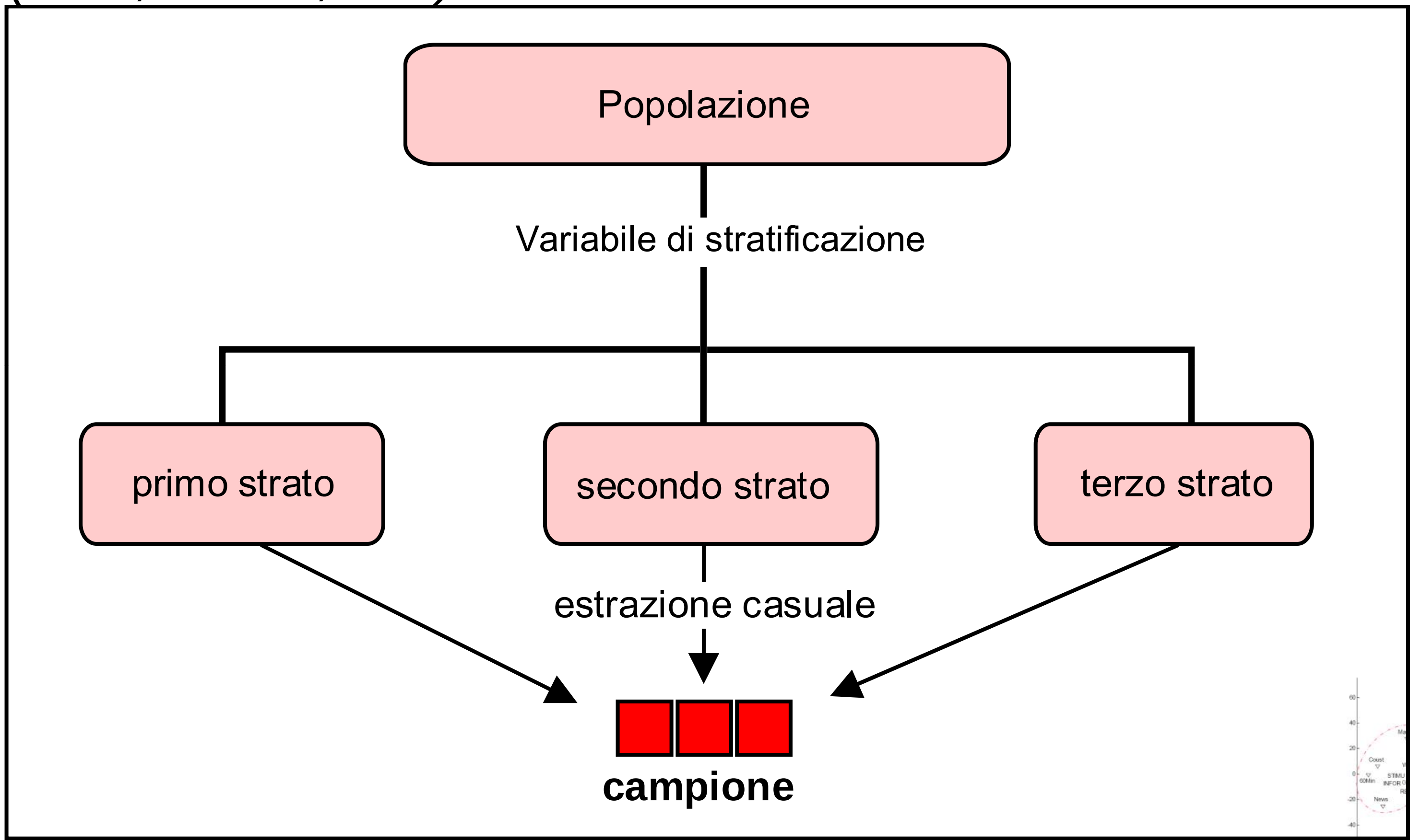
a) se N>>n i due stimatori si equivalgono, infatti $1 - f \approx 1$

b) Se non è nota σ^2 la si sostituisce con la sua stima s^2



Campionamento casuale stratificato

Nel **campionamento casuale stratificato** la popolazione viene suddivisa in strati. Da ogni strato vengono poi estratti, tramite un campionamento casuale semplice, le unità da inserire nel campione. *Esempio di variabile di stratificazione: Ripartizione territoriale con tre strati (Nord, Centro, Sud).*

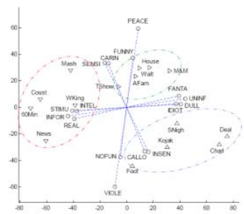


Vantaggi nell'uso del campione casuale stratificato rispetto a quello semplice

- Se sussiste una relazione fra la variabile di stratificazione e la variabile da stimare, l'operazione di stratificazione aumenta la **rappresentatività del campione**

- **Miglioramento della stima**, se gli strati sono stati ben scelti.(si riduce la variabilità dello stimatore ossia si aumenta la precisione)

- Possibilità di ottenere anche le **stime per le singole sottopopolazioni** o strati.



Come estrarre le unità dagli strati?

Il campionamento stratificato assicura la presenza nel campione di unità provenienti da ogni sub-popolazione (strato)

•Se la variabilità della Y all'interno degli L strati non si differenzia, si procede al **campionamento a frazione di sondaggio costante**, dove i campioni estratti in ogni strato risultano proporzionali alle rispettive sub-popolazioni (piano auto-ponderante)

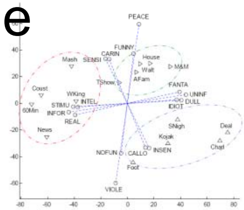
•Se la variabilità all'interno degli L strati si differenzia molto, si procede al **campionamento a frazione di sondaggio variabile**, che riduce la numerosità campionaria per gli strati con minore variabilità e l'aumenta a quelli più variabili

Problemi

Scelta degli strati; La numerosità campionaria di ogni strato

Vantaggi

Il guadagno in efficienza aumenta quanto più differiscono le medie della Y riferite ai singoli strati (ANOVA)



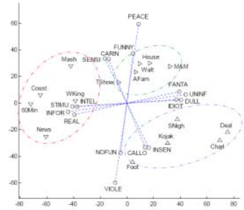
Stimatore della media attraverso il campione casuale stratificato

	Strato 1	Strato 2	…..	Strato L	Totale
Popolazione	N_1	N_2	…..	N_L	N
	σ_1^2	σ_2^2		σ_L^2	σ^2
Campione	n_1	n_2	…..	n_L	n
	$\bar{Y}_1$	$\bar{Y}_2$		$\bar{Y}_L$	$\bar{Y}$

Stimatore corretto ed efficiente della media (Pop. Finita)

$$\bar{Y}_{st} = \sum_{h=1}^{L} w_h \bar{Y}_h \quad \text{dove} \quad w_h = N_h / N$$

$$Var(\bar{Y}_{st}) = \sum_h w_h^2 (1 - f_h) \frac{\sigma_h^2}{n_h} \quad \text{dove} \quad f_h = \frac{n_h}{N_h}$$



Esempio: Spesa media annuale dei possessori di CartaViaggio di Trenitalia

Obiettivo: Stimare la spesa media annuale attraverso un campionamento stratificato con allocazione proporzionale e confrontare l’efficienza del campionamento stratificato rispetto a quello casuale semplice

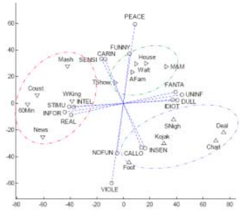
N=29.206 clienti

n=1455 campione pari a circa 0,5% della popolazione

L=3 in funzione della condizione occupazionale:

- Imp=imprenditori+dirigenti
- Dip=impiegati+insegnanti+commercianti+operai+altro
- Noccup=pensionati+casalinghe+studenti+in cerca di+disocc

	Strati		
L	**Imp**	**Dip**	**Noccup**
Dim. N_h	**7136**	**14654**	**7416**



Esempio: Spesa media annuale dei possessori di CartaViaggio di Trenitalia

I parametri ignoti della popolazione

N=29.206 clienti (in realtà è il 10% della Popolazione)

μ = Spesa Media = 509

σ^2 = Varianza della spesa = 367918

σ = Deviazione standard = 607

	Imp	Dip	Noccup
Dim. N_h	7136	14654	7416
%	24,4	50,2	25,4
media	642	525	350
varianza	454996	427191	124334
Dev. Stand.	675	654	353

Esempio: Spesa media annuale dei possessori di CartaViaggio di Trenitalia

1° caso: Camp. Casuale semplice (0,5%)

f = *n*/N =1455/29206= 0,05

e *n*=1455

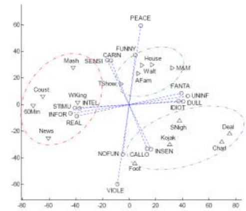
Stime del campione		
Media	**Varianza**	**Dev.Stand.**
521	**486093**	**697**

$$\hat{\bar{y}} = \sum_{i=1}^{n} y_i = \boxed{521}$$

Varianza dello stimatore $$V(\hat{\bar{y}}) = \frac{(1-f)}{n}\sigma^2 = \frac{0{,}95}{1455}367918 = \boxed{240}$$

Stima della varianza $$v(\hat{\bar{y}}) = \frac{(1-f)}{n}s^2 = \frac{0{,}95}{1455}486093 = 317$$

Stima dell'errore standard $$e.r.(\hat{\bar{y}}) = \sqrt{v(\hat{\bar{y}})} = 17{,}8$$



2° caso: Camp. stratificato con allocazione proporzionale

f = *n*/N =1455/29206= 0,05

e *n*=1455

Stime del campione

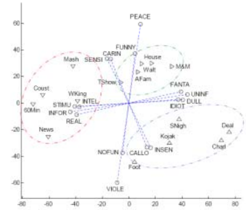
L	Imp	Dip	Noccup
Dim. N_h	7136	14654	7416
W_h	0,244	0,502	0,254
n_h	356	730	369
media	632	558	321
varianza	371148	341479	92033
Dev.stand.	609.2	584.4	303.4

$$\hat{\bar{y}}_{str} = \sum_{h=1}^{L} w_h \hat{\bar{y}}_h = \sum_{h=1}^{L} \frac{N_h}{N} \hat{\bar{y}}_h = \boxed{516}$$

$$V(\hat{\bar{y}}_{str}) = \sum_{h=1}^{L} w_h^2 V(\hat{\bar{y}}_h) = \sum_{h=1}^{L} w_h^2 \frac{(1-f_h)}{n_h} \sigma_h^2 = \boxed{233}$$

$$v(\hat{\bar{y}}_{str}) = \sum_{h=1}^{L} w_h^2 v(\hat{\bar{y}}_h) = \sum_{h=1}^{L} w_h^2 \frac{(1-f_h)}{n_h} s_h^2 = 186$$

Stima dell'errore standard $e.r.(\hat{\bar{y}}_{st}) = \sqrt{v(\hat{\bar{y}}_{st})} = 13,6$



Confronto

Confronto					
			Int. Conf. 95%		
	Stima	**Err. Stand.**	**Estr. Inf.**	**Estr. Sup.**	**Amp.**
Camp.cas. Sempl.	**521**	**17,8**	**486,1**	**555,9**	**69,8**
Camp. Cas. Str.	**516**	**13,6**	**489,3**	**542,7**	**53,3**
Popolazione	**media**	**509**			

- Gli strati ottenuti dalla variabile di stratificazione individuano livelli medi di spesa diversi
- La stima ottenuta dal campione casuale stratificato è più efficiente di quella ottenuta dal campione casuale semplice
- Si ottiene un intervallo di confidenza più piccolo a parità di livello di confidenza.