Metodi Quantitativi per Economia, Finanza e Management

Lezione n°6

Analisi Bivariata: discussione di un caso; Analisi Fattoriale: le ipotesi del modello e il metodo delle component principali

Bivariate Analysis

Objective

To jointly describe the relationship between two variables.

- qualitative variables: Analysis of Connection
- quantitative variables: Analysis of Correlation
- mixed variables: Analysis of Variance

Bivariate Analysis

	Descriptive Tools	Descriptive Indexes	Statistical Test	Null Hypothesis
Connection	Contingency Table	Chi-Square Kramer's V	Chi-Square test	Statistical Indipend.
Correlation	Scatter Plot	Linear Correlation Coeffcient	t-Test	No linear relation
ANOVA	Means by Classes	Spearman Coefficient	F-Test	Indipend. by mean

Caso Caffè

2.2. ANALISI BIVARIATA

Considerando la relazione tra la compagnia con cui si preferisce consumare caffè e la concezione che si ha del momento di consumo è possibile avere un'idea più completa sull'immagine da trasmettere con la campagna pubblicitaria. Seguendo le analisi precedenti la maggior parte del campione considera il caffè un'abitudine e, secondariamente, un rito.

TABELLA COMPAGNIA/DESCRIZIONE CONSUMO					
Compagnia	**Descrizione consumo**				
	ABITUDINE	**BEVANDA**	**ESIGENZA**	**RITO**	**Totale**
AMICI	51 24.52 52.04 53.13	9 4.33 9.18 36	14 6.73 14.29 41.18	24 11.54 24.49 45.28	98 47.12
COLLEGHI	9 4.33 37.5 9.38	6 2.88 25 24	6 2.88 25 17.65	3 1.44 12.5 5.66	24 11.54
FAMIGLIA	19 9.13 40.43 19.79	4 1.92 8.51 16	8 3.85 17.02 23.53	16 7.69 34.04 30.19	47 22.6
SOLO	17 8.17 43.59 17.71	6 2.88 15.38 24	6 2.88 15.38 17.65	10 4.81 25.64 18.87	39 18.75
Totale	96 46.15	25 12.02	34 16.35	53 25.48	208 100

Delle 96 persone che hanno scelto “abitudine” **51** preferisce berlo con gli AMICI, **19** con la FAMIGLIA, **17** da SOLO e **9** coi COLLEGHI.

Delle 53 persone che hanno scelto “rito” **24** preferiscono berlo con gli AMICI, **16** con la FAMIGLIA, **10** da SOLO e **3** coi COLLEGHI.

Risultati che indicano comunque come si preferisca consumare una tazzina di caffè con gli amici e, in secondo luogo, con la famiglia.

Infatti in generale la distribuzione marginale della variabile “AMICI” è **98 su un totale di 208**, seguita da un **47** della “FAMIGLIA”.

Questo porta l’azienda a conoscere le preferenze del consumatore e quindi trasmettere l’immagine di un prodotto non solo vissuto come un’abitudine o un rito, ma da consumarsi circondato da amici o famigliari. La campagna pubblicitaria dovrebbe basarsi su questi elementi in modo da dare al consumatore ciò che vuole e colpire la sua attenzione.

TEST CHI QUADRO

Per testare l'ipotesi di indipendenza statistica tra le due variabili qualitative *luogo di consumo* e *compagnia* si deve fare il "test chi quadro".
Il "chi quadro" risulta essere 0,0001. Si considera un livello di significatività di 0,05

Statistic	DF	Value	Prob
Chi-Square	6	27.225	0.0001
Likelihood Ratio Chi-Square	6	26.893	0.0002
Mantel-Haenszel Chi-Square	1	5.088	0.0241
Phi Coefficient		0.361	
Contingency Coefficient		0.340	
Cramer's V		0.255	

0,0001<<0,05 → si rifiuta, quindi, l'ipotesi nulla di indipendenza statistica e si può affermare che *le due variabili sono statisticamente dipendenti*.
L'azienda dovrebbe considerare questo aspetto, durante la campagna pubblicitaria, in modo da offrire un messaggio coerente (es. creare l'immagine di un bar insieme a degli amici).

TABELLA LUOGO DI CONSUMO/COMPAGNIA					
Luogo di consumo	Compagnia				
	AMICI	COLLEGHI	FAMIGLIA	SOLO	Totale
BAR	45 21.63 66.18 45.92	7 3.37 10.29 29.17	7 3.37 10.29 14.89	9 4.33 13.24 23.08	68 32.69
CASA	29 13.94 35.37 29.59	5 2.4 6.1 20.83	29 13.94 35.37 61.7	19 9.13 23.17 48.72	82 39.42
DISTRIBUTORE	24 11.54 41.38 24.49	12 5.77 20.69 50	11 5.29 18.97 23.4	11 5.29 18.97 28.21	58 27.88
Totale	98 47.12	24 11.54	47 22.6	39 18.75	208 100

Anche da questa tabella risulta evidente che in qualsiasi luogo si beva il caffè prevale l'opzione **AMICI**. Solo a **CASA** è rilevante anche la compagnia della **FAMIGLIA** con una frequenza di **29** pari a quella degli amici.

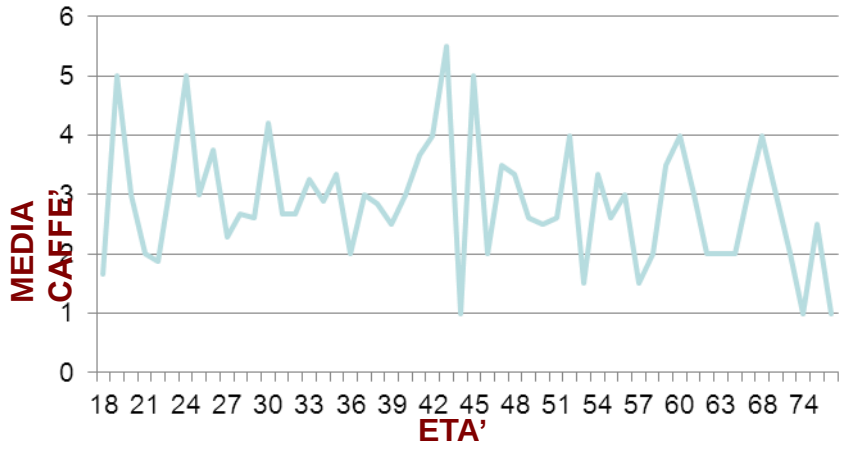
Quindi, dalle tabelle di contingenza analizzate, risulta come qualsiasi sia la concezione che si ha del momento del consumo di caffè e in qualsiasi luogo lo si beva, la compagnia preferita sia la stessa.

ANALISI DI CORRELAZIONE TRA LA VARIABILE *ETA`* E LA VARIABILE *NUMERO DI CAFFE` CONSUMATI*

L'analisi è svolta per capire se esiste una relazione tra le due variabili e se è di tipo positivo o negativo.

In questo modo l'azienda, attraverso il risultato ottenuto, può concentrarsi su un'eventuale target di clienti divisi per fascia di età.

Bisogna considerare il coefficiente di correlazione per capire che tipo di relazione intercorre tra le due variabili quantitative.
In questo caso il suo valore è pari a 0.03451. E' un coefficiente positivo, ma molto prossimo allo 0 e quindi si può affermare che non esiste relazione tra le due variabili.



Pearson Correlation Coefficients, N = 208 Prob > |r| under H0: Rho=0		
	ETA'	NUMERO CAFFE'
ETA'	1	0.03451 0.6207
NUMERO CAFFE'	0.03451 0.6207	1

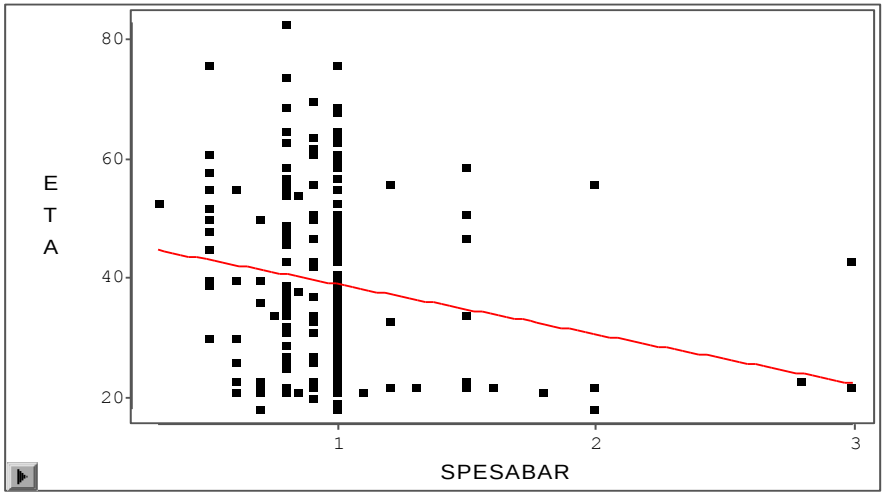
Come si può osservare nel grafico, non vi è una correlazione lineare, né tantomeno positiva, tra le variabili età e numero di caffè bevuti.

ANALISI DI CORRELAZIONE TRA LE VARIABILI QUANTITATIVE *ETA`* E *PROPENSIONE ALL'ACQUISTO*

Questa analisi può essere utile per capire se la *sensibilità al prezzo* possa variare con il variare dell'età

Coefficienti di correlazione di Pearson, N = 208 **Prob > |r| con H0: Rho=0**		
	ETA	**SPESABAR**
ETA ETA	1	-0.19727 0.0043
SPESABAR SPESABAR	-0.19727 0.0043	1

Nel caso in questione il coefficiente di correlazione risulta essere -0.19727, valore negativo che ci porta a dire che esiste una relazione lineare negativa tra le due variabili: all'aumentare dell'età diminuisce la disponibilità a pagare.



L'azienda può usare questi dati per capire in che modo l'età influisca sulla sensibilità al prezzo e, di conseguenza, decidere che strategie di prezzo assumere in base al target su cui ci si focalizzerà.

TEST T

Infatti, eseguendo il test t, considerando il valore 0.0043 e prendendo come livello di significatività il valore 0.05 risulta essere 0.0043<0,05.
Si rifiuta quindi l'ipotesi nulla di indipendenza lineare. Le due variabili *età* e *spesa al bar* sono dipendenti.

Coefficienti di correlazione di Pearson, N = 208 **Prob > |r| con H0: Rho=0**		
	ETA	**SPESABAR**
ETA ETA	1	-0.19727 0.0043
SPESABAR SPESABAR	-0.19727 0.0043	1

TEST F

Col test F si può considerare la relazione tra variabili indicanti le caratteristiche del campione (età, professione) e le abitudini di consumo del caffè.

La professione potrebbe influenzare il numero di caffè bevuti giornalmente: per esempio una persona che svolge turni di notte potrebbe bere caffè per l'esigenza di mantenersi sveglio.

Anche l'età è un fattore rilevante che potrebbe spingere le persone ad avere diverse abitudini e diverse preferenze.

Col test F si può capire se sussiste realmente questa relazione accettando o rifiutando l'ipotesi nulla di uguaglianza tra medie. All'azienda è utile per avere idee chiare e prendere decisioni relative al consumatore target e alla comunicazione più idonea da farsi.

Test F tra le variabile qualitativa *professione* e la variabile quantitativa *numero di caffè bevuti in un giorno*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	30.3848377	6.0769675	2.68	0.0225
Error	202	457.3026623	2.2638746		
Corrected Total	207	487.6875			

R-Square	Coeff Var	Root MSE	NUMCAF Mean
0.062304	53.49752	1.504618	2.8125

Possiamo constatare il valore di 0,0225. Valore che è minore del livello di significatività 0,05:

0,0225<0,05.

Questo porta a rifiutare l'ipotesi nulla e ad affermare l'esistenza di una relazione di dipendenza in media tra le due variabili.

Il valore di *Eta quadro*, 0.06, è positivo quindi indica dipendenza in media, ma risulta essere debole in quanto il dato è molto prossimo allo zero.

Test F tra la variabile quantitativa *età* e la variabile qualitative *marca preferita*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1673.44718	239.06388	1.1	0.3618
Error	200	43298.2259	216.49113		
Corrected Total	207	44971.67308			

R-Square	Coeff Var	Root MSE	ETA Mean
0.037211	37.58827	14.71364	39.14423

Possiamo constatare il valore di 0.3618. Valore che è maggiore del livello di significatività 0.05:

0.3618>0,05.

Questo porta ad accettare l'ipotesi nulla e ad affermare l'inesistenza di una relazione di dipendenza in media tra le due variabili.

Percorsi di Analisi

Tipo di analisi	Cosa è?	Strumenti
ANALISI UNIVARIATA	La statistica descrittiva **univariata** ha come obiettivo lo studio della distribuzione di ogni variabile, singolarmente considerata, all'interno della popolazione. Fornisce strumenti per la lettura dei fenomeni osservati di rapida ed immediata interpretazione.	- DISTRIBUZIONI DI FREQUENZA - INDICI DI POSIZIONE (MISURE DI TENDENZA CENTRALE E MISURE DI TENDENZA NON CENTRALE) - INDICI DI DISPERSIONE - MISURE DI FORMA DELLA DISTRIBUZIONE
ANALSI BIVARIATA E TEST STATISTICI PER LO STUDIO DELL'ASSOCIAZIONE TRA VARIABILI	La statistica descrittiva **bivariata** si occupa dello studio della distribuzione di due variabili congiuntamente considerate. I **test statistici** per lo studio dell'associazione tra variabili ci permettono di formulare delle ipotesi e verificarle tramite i dati campionari. I dati campionari sono utilizzati per stabilire se tale ipotesi è ragionevolmente accettabile o rifiutabile.	**Due variabili qualitative o quantitative discrete:** TABELLA DI CONTINGENZA E INDICI CHI QUADRO E V DI CRAMER TEST CHI QUADRO PER L'INDIPENDENZA STATISTICA **Due variabili quantitative continue:** INDICE DI CORRELAZIONE DI PEARSON (ρ) E COVARIANZA TEST t PER L'INDIPENDENZA LINEARE **Una variabile qualitativa e una quantitativa continua:** INDICE η^2 TEST F PER L'INDIPENDENZA IN MEDIA
ANALISI MULTIVARIATA	L'analisi statistica **multivariata** e' l'insieme di metodi statistici usati per analizzare simultaneamente più variabili. Esistono molte tecniche diverse, usate per risolvere problemi anche lontani fra loro.	- ANALISI FATTORIALE - REGRESSIONE LINEARE - REGRESSIONE LOGISTICA - SERIE STORICHE

Factor Analysis

Factor Analysis

12 Assign to each one of the following options a percentage of: Where do you usually use internet? (Sum percentages =100)

a. Home	
b. Work	
c. University	
	0%

13 What do you use internet for?

	Low			Medium			High		
	1	2	3	4	5	6	7	8	9
a. Sources of information									
b. Work									
c. Friendship									
d. Buy and sell									
e. University									
f. Organizing events									

Factor Analysis

Correlations

		Internet_ Home	Internet_ Work	Internet_ University	Internet for_ Information	Internet for_ Work	Internet for_ Friendship	Internet for_ Buy&Sell	Internet for_ University	Internet for_Organize Events
Internet_Home	Pearson Correlation	1	-.541**	-.746**	-.094	-.188**	-.031	-.085	.028	-.074
	Sig. (2-tailed)		.000	.000	.194	.009	.671	.239	.695	.309
	N	193	193	193	193	193	193	193	193	193
Internet_Work	Pearson Correlation	-.541**	1	-.156*	.077	.195**	-.031	.095	-.272**	.015
	Sig. (2-tailed)	.000		.030	.288	.007	.666	.188	.000	.836
	N	193	193	193	193	193	193	193	193	193
Internet_University	Pearson Correlation	-.746**	-.156*	1	.048	.067	.061	.025	.182*	.077
	Sig. (2-tailed)	.000	.030		.511	.357	.397	.735	.011	.289
	N	193	193	193	193	193	193	193	193	193
Internet for_Information	Pearson Correlation	-.094	.077	.048	1	.367**	.085	.154*	.135	.045
	Sig. (2-tailed)	.194	.288	.511		.000	.242	.033	.061	.534
	N	193	193	193	193	193	193	193	193	193
Internet for_Work	Pearson Correlation	-.188**	.195**	.067	.367**	1	.113	.031	-.029	.088
	Sig. (2-tailed)	.009	.007	.357	.000		.119	.669	.688	.222
	N	193	193	193	193	193	193	193	193	193
Internet for_Friendship	Pearson Correlation	-.031	-.031	.061	.085	.113	1	.025	.227**	.314**
	Sig. (2-tailed)	.671	.666	.397	.242	.119		.732	.001	.000
	N	193	193	193	193	193	193	193	193	193
Internet for_Buy&Sell	Pearson Correlation	-.085	.095	.025	.154*	.031	.025	1	.063	.208**
	Sig. (2-tailed)	.239	.188	.735	.033	.669	.732		.383	.004
	N	193	193	193	193	193	193	193	193	193
Internet for_University	Pearson Correlation	.028	-.272**	.182*	.135	-.029	.227**	.063	1	.167*
	Sig. (2-tailed)	.695	.000	.011	.061	.688	.001	.383		.020
	N	193	193	193	193	193	193	193	193	193
Internet for_Organize Events	Pearson Correlation	-.074	.015	.077	.045	.088	.314**	.208**	.167*	1
	Sig. (2-tailed)	.309	.836	.289	.534	.222	.000	.004	.020	
	N	193	193	193	193	193	193	193	193	193

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Factor Analysis

Nel caso in cui l'informazione disponibile per eseguire un'analisi è distribuita tra molte variabili tra loro correlate:

$\Rightarrow$ *Possono insorgere diversi problem di tipo logico/applivcativo.*

- Informazione solo apparente;
- Missunderstanding;
- Difficoltà nella fase interpretativa dei fenomeni;
- Robustezza dei risultati;
- Efficienza delle stime;
- Gradi di libertà;
- …..

Factor Analysis

Quando le variabili considerate sono numerose spesso risultano tra loro correlate => numerosità e correlazione tra variabili porta a difficoltà di analisi

Perché sintetizzare?

- Se l'informazione è condivisa tra più variabili correlate tra loro, è ridondante utilizzarle tutte.
- La sintesi semplifica le analisi successive ma comporta una perdita di informazione, si deve evitare, di perdere informazioni rilevanti.

Analisi fattoriale

Perché sintetizzare mediante l'impiego della tecnica?

Se l'informazione è "dispersa" tra più variabili correlate tra loro, le singole variabili faticano da sole a spiegare il fenomeno oggetto di studio, mentre combinate tra loro risultano molto più esplicative.

Esempio: l'attrattività di una città da cosa è data? Dalle caratteristiche del contesto, dalla struttura demografica della popolazione, dalla qualità della vita, dalla disponibilità di fattori quali capitale, forza lavoro, know-how, spazi, infrastrutture, ecc.

I fattori latenti sono "concetti" che abbiamo in mente ma che non possiamo misurare direttamente.

Factor Analysis

We used the Factor Analysis in order to summarize and reduce the different variables into a lower number trying to loose the least number of information possible.

VARIABLES OF ANALYSIS

- Reasons that drive you to check facebook?
  - Make new friends
  - Keep in touch with friends
  - Reconnect with old classmates
  - Have news about products
  - Share photos and videos
  - Curiosity
  - Discuss interest and hobbies
  - Plan Parties and events
- Which features do you use?
  - Wall
  - Photo & Video
  - Private Messaging
  - Events Creation
  - Group Affiliation

Number of starting variables= 13

Factor Analysis

Final Factors

	Spying	Broadening	Keeping Up	Public Relations
Features_PhotoVideo	0,799803449			
Check_Share	0,721346404			0,362558752
Check_Curiosity	0,64474328	0,381412457		
Features_Wall	0,54919374		0,35168123	
Check_Discuss		0,757302403		
Check_Products		0,752178894		
Check_NewFriends	0,356266658	0,695178418		
Features_PrivateMsg			0,721205388	
Check_OldClassmates		0,333749139	0,717594045	
Check_KeepFriends			0,701607063	
Features_Events				0,854398211
Check_Plan				0,795770255

Analisi fattoriale

Quando le variabili considerate sono numerose spesso risultano tra loro correlate.

Numerosità e correlazione tra variabili porta a difficoltà di analisi => ridurre il numero (semplificando l'analisi) evitando, però, di perdere informazioni rilevanti.

L'Analisi Fattoriale è una tecnica statistica multivariata per l'analisi delle correlazioni esistenti tra variabili quantitative.

A partire da una matrice di dati : $X_{(nxp)}$, con "n" osservazioni e "p" variabili originarie, consente di sintetizzare l'informazione in un set ridotto di variabili trasformate (i fattori latenti).

Analisi fattoriale

Le ipotesi del Modello Fattoriale

Variabili Quantitative $x_1, x_2, ..., x_i, ... x_p$

Info	x_i	=	Info condivisa +	Info specifica
Var	x_i	=	Communality +	Var specifica
	x_i	=	$f(CF_1, ...,CF_k)$ +	UF_i

$i = 1, ..., p$
$k << p$

CF_i = Common Factor$_i$
UF_i = Unique Factor$_i$

$Corr (UF_i , UF_j) = 0$ per $i \neq j$
$Corr (CF_i , CF_j) = 0$ per $i \neq j$
$Corr (CF_i , UF_j) = 0$ per ogni i,j

Analisi fattoriale

Factor Loadings & Factor Score Coefficients

$x_i = l_{i1}CF_1 + l_{i2}CF_2 + ... + l_{ik}CF_k + UFi$

$l_{i1}, l_{i2},.....,l_{ik}$ factor loadings

$i = 1, ..., p$ *significato fattori*

$CF_j = s_{j1}x_1 + s_{j2}x_2 + ... + s_{jp}x_p$

$s_{j1}, s_{j2},.....,s_{jp}$ factor score coeff.

$j = 1, ..., k << p$ *costruzione fattori*

Analisi fattoriale

Metodo delle Componenti Principali

- I fattori calcolati mediante il metodo delle CP sono combinazioni lineari delle variabili originarie

$$CP_j = s_{j1}x_1 + s_{j2}x_2 + \ldots\ldots\ldots\ldots + s_{jp}x_p$$

- Sono tra loro ortogonali (non correlate)
- Complessivamente spiegano la variabilità delle p variabili originarie
- Sono elencate in ordine decrescente rispetto alla variabilità spiegata

Analisi fattoriale

Metodo delle Componenti Principali

Il numero massimo di componenti principali è pari al numero delle variabili originarie (p).

La prima componente principale è una combinazione lineare delle p variabili originarie ed è caratterizzata da varianza più elevata, e così via fino all'ultima componente, combinazione sempre delle p variabili originarie, ma a varianza minima.

Se la correlazione tra le p variabili è elevata, un numero k<<p (k molto inferiore a p) di componenti principali è sufficiente rappresenta in modo adeguato i dati originari, perché riassume una quota elevata della varianza totale.