Statistica per l’impresa

Introduzione al corso

Aggiornata al 27.10.2025

Prerequisiti

- ✓ **Statistica descrittiva**
  - **Indici di posizione** (media, mediana, …)
  - **Indice di variabilità** (varianza, scarto quadratico medio,…)
  - **Indici di commovimento** (covarianza)
- ✓ **Distribuzioni di frequenza e probabilità**
- ✓ **Nozioni di base della statistica inferenziale**
  - Stima di parametri incogniti (variabili aleatorie, …)
  - Verifica di ipotesi
  - Intervalli di confidenza

La statistica in azienda

- ✓ Scienza che studia con **metodi matematici** dei **fenomeni collettivi**
  - **Statistica descrittiva**: si limita a descrivere i fenomeni attraverso indici, grafici, rappresentazioni tabellari
  - **Statistica inferenziale**: avvalendosi di metodi probabilistici, permette di trarre conclusioni generali a partire dall'esame di un campione
- ✓ **L'impresa**, essa stessa origine di dati, ha la necessità di:
  - **organizzare, sintetizzare (capire) e presentare l'informazione** relativa alla vita **dell'azienda, del contesto** in cui opera e dei **concorrenti**
  - usare le informazioni osservate per «**inferire**» le caratteristiche di alcuni **aspetti del ciclo di creazione di valore** e/o **prevederne l'evoluzione**
  - utilizzare le informazioni ottenute per **prendere decisioni**

Strumenti informatici

- ✓ Questo non è un corso sul *software*, che è solo uno strumento per il calcolo e la visualizzazione. Esistono molti pacchetti o ambienti capaci di eseguire i calcoli descritti nel corso
  - Microsoft Excel: un foglio elettronico
  - R/GRETL: ambienti statistici

- ✓ La competenza di utilizzo dei *software* non è strettamente necessaria ai fini dell'esame, ma lo sarà sicuramente nella futura pratica professionale

Testo di riferimento

- ✓ Il **testo consigliato** è:
  - L. Biggeri, M. Bini, A. Coli, L. Grassini, M. Maltagliati, seconda edizione, **«Statistica per le decisioni aziendali»**, Pearson Italia (2023)

- ✓ Può essere utile la consultazione di:
  - B. Brancaleone, A. Mulas, M. Cossignani (2009), «Statistica Aziendale», MC Graw Hill Education

- ✓ Le slides sono solo un ausilio alla presentazione. Si studia dal libro.

Orario e struttura delle lezioni

- ✓ **Aule e orari**:
  - **Lunedì 17 – 19** - Aula 3_B - Edificio D
  - **Martedì 18 – 19** - Aula 3_B - Edificio D
  - **Venerdì 12 – 13** - Aula 4_B - Edificio D

- ✓ **Ricevimento** studenti:
  - Martedì 19 – 20
  - Su appuntamento

Modalità della verifica

- ✓ **Esame scritto** alla fine del corso:
  - Domande teoriche
  - Esercizi numerici da svolgere
  - No *software*, solo una calcolatrice
  - No telefoni cellulari

Statistica per l'impresa

2. Le informazioni statistiche: Disponibilità e produzione

GENERALITA'

- ✓ Concetto di impresa
- ✓ Concetto di dato statistico

LE FONTI DEI DATI STATISTICI

- ✓ Fonti interne vs esterne, dati primari e secondari
- ✓ Fonti esterne
  - I conti nazionali
  - Le informazioni sulle imprese
  - Le informazioni sui consumatori

LA QUALITA' DEI DATI STATISTICI

L'INDAGINE CAMPIONARIA AD HOC

- ✓ Popolazione obiettivo, di selezione e di indagine
- ✓ Le fasi dell'indagine campionaria
- ✓ Il piano di campionamento
- ✓ Campioni probabilistici vs non probabilistici
- ✓ Metodi di selezione dei campioni probabilistici
- ✓ Stima dei parametri di una popolazione
- ✓ Stime intervallari
- ✓ Tecniche di rilevazione
- ✓ Il questionario

Statistica per l'impresa

3. Interpretazione e comparazione dei dati

RAPPORTI
- ✓ Differenze assolute e relative
- ✓ Rapporti di composizione
- ✓ Rapporti di coesistenza
- ✓ Rapporti di densità e derivazione
- ✓ Aggregazione e scomposizioni dei rapporti

NUMERI INDICI
- ✓ Base fissa e Base mobile, cambio di base e proprietà
- ✓ Numeri indici sintetici: indici di Laspeyres, Paasche e Fisher

TASSI DI VARIAZIONE
- ✓ Tassi medi
- ✓ Variazioni congiunturali e tendenziali
- ✓ Variazioni nominali e reali

ANALISI DELLA MOBILITA'
- ✓ Tassi di entrata e di uscita
- ✓ Rapporti di rinnovo e durata e turnover
- ✓ Matrici di transizione

Statistica per l'impresa

6. Misura delle relazioni tra variabili per le decisioni aziendali

RELAZIONI TRA VARIABILI AZIENDALI

- ✓ Le relazioni tra variabili aziendali per prendere decisioni in azienda
- ✓ Correlazione e regressione
- ✓ Analisi grafica della correlazione
- ✓ L'indice di correlazione di Pearson e la verifica delle ipotesi

REGRESSIONE SEMPLICE

- ✓ Il modello di regressione semplice
- ✓ Il criterio dei minimi quadrati ordinari (OLS)
- ✓ La bontà di adattamento di un modello: l'indice R^2
- ✓ Le proprietà degli stimatori OLS
- ✓ Test di ipotesi sui coefficienti di regressione e intervalli di confidenza
- ✓ Analisi dei residui

REGRESSIONE MULTIPLA

- ✓ Il modello di regressione semplice
- ✓ Analogie con il modello di regressione semplice
- ✓ l'indice R^2 adjusted
- ✓ L'indice F di significatività congiunta

Statistica per l’impresa

7. L’analisi delle serie storiche per la programmazione delle attività

L’ANALISI DELLE SERIE STORICHE IN AZIENDA
- ✓ Serie storiche univariate
- ✓ Scopo dell’analisi delle serie storiche in azienda
- ✓ Componenti della serie storica e fasi dell’analisi

ANALISI PRELIMINARI
- ✓ Rappresentazione grafiche
- ✓ Coefficiente di autocorrelazione e correlogramma
- ✓ Valutazione della capacità previsiva

METODI DI SCOMPOSIZIONE E STIMA DELLE COMPONENTI
- ✓ Metodi di stima delle componenti
- ✓ Le medie mobili
- ✓ Stima analitica del trend

LIVELLAMENTO ESPONENZIALE
- ✓ Livellamento esponenziale semplice
- ✓ Metodo di Holt e Winters

Statistica per l'impresa

8. Valutazione delle prestazioni economiche-finanziarie delle imprese

IL BILANCIO
- ✓ Le principali componenti e la sua riclassificazione
- ✓ Indici di bilancio

ANALISI STATISTICA DEGLI INDICI DI BILANCIO
- ✓ Valori medi
- ✓ Benchmarking

ANALISI MULTIVARIATA
- ✓ Analisi in componenti principali
- ✓ Analisi cluster

APPLICAZIONI
- ✓ Diagnosi precoce dell'insolvenza aziendale con l'analisi discriminante
- ✓ Modelli Logit

Statistica per l’impresa

Richiami di statistica descrittiva

- ✓ Mediana
- ✓ Quartili
- ✓ Distribuzione di frequenza
- ✓ Moda
- ✓ Media
- ✓ Varianza
- ✓ Scarto quadratico medio
- ✓ Covarianza

Mediana e quartili 1/2

- ✓ **Sono indici di posizione di una distribuzione**
- ✓ La **Mediana** rappresenta il punto centrale della distribuzione della serie
  - Data una serie di 5 quantità (unità di prodotti venduti in un intervallo di tempo in 5 comuni)

5 4 7 2 3

  - Devo ordinarli

2 3 4 5 7

  - La mediana è l'elemento che lascia a destra e a sinistra il 50% delle osservazioni

- ✓ In generale, data una distribuzione $x_1, x_2, \ldots, x_n$, con n numero di osservazioni:
  - **Se *n* è dispari** la mediana è l'elemento:

$$x_{(n+1)/2}$$

  - **Se *n* è pari** la mediana è la media dei due elementi centrali:

$$(x_{n/2} + x_{n/2+1})/2$$

Si veda parte pratica in excel

Mediana e quartili 2/2

- ✓ **Una distribuzione ha tre quartili che la suddividono in quattro parti, ognuna delle quali contiene il 25% delle osservazioni**
  - Il primo quartile ha alla sua «sinistra» il 25% delle osservazioni e alla sua destra il 75%
  - Il secondo quartile ha alla sua «sinistra» il 50% delle osservazioni e alla sua destra il 50% (la mediana)
  - Il terzo quartile ha alla sua «sinistra» il 75% delle osservazioni e alla sua destra il 25% (la mediana)

- ✓ Anche in questo caso devo ordinare la serie (15 osservazioni)

2 3 4 **5** 7 8 10 **12** 13 15 16 **18** 19 21 22

- ✓ In generale, data una distribuzione $x_1, x_2, \ldots, x_n$, con n numero di osservazioni:
  - **Se *n+1* è divisibile per quattro:**

Il primo quartile $x_{(n+1)*(1/4)}$, Il secondo quartile $x_{(n+1)*(1/2)}$, Il terzo quartile $x_{(n+1)*(3/4)}$

  - **Se *n+1* non è divisibile per quattro** (i quartili vengono con la virgola) devo fare le medie tra l'intero precedente e quello successivo

- ✓ I quartili sono dei particolari **quantili** (termine che tornerà nell'ambito del corso). Data un distribuzione, il **quantile di ordine *p*** (con p compreso tra zero e uno) è il **valore della distribuzione** che ha la proprietà di **dividerla in due** parti che **contengono** rispettivamente ***p* e *1-p* valori**. Il primo quartile è il quantile di ordine 0,25

Si veda parte pratica in excel

Distribuzione di frequenza e moda

✓ Una **distribuzione di frequenza** è una tabella dove **in corrispondenza delle modalità** viene riportato il numero di volte che quelle stesse modalità si sono verificate (**frequenza**)

DISTRIBUZIONE DI FREQUENZA	
Fatturato (€x1000	N. agenti
71,0	1
73,0	1
74,0	2
76,0	1
76,5	1
77,0	1
78,0	3
82,0	2
87,0	1
88,0	1
89,0	1
89,5	1
90,0	1
96,0	1
98,0	1
99,0	1
Totale complessivo	**20**

La classe di fatturato che corrisponde alla maggior frequenza si dice **moda**

Si veda parte pratica in excel

Media, varianza e scarto quadratico medio

- ✓ La **media semplice** è calcolata sommando tutte le osservazioni di una serie e dividendola per il numero delle osservazioni stesse. Data una distribuzione $x_1, x_2, \ldots, x_n$, con n numero di osservazioni

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

- ✓ La **varianza** rappresenta una misura di quanto le osservazioni «si muovono» attorno alla media della serie che stiamo studiando. E' espressa nell'unità di misura al quadrato della variabile che stiamo osservando. A seconda dei campi di analisi può assumere un significato economico con accezioni differenti: nella finanza per esempio la varianza dei rendimenti rappresenta la rischiosità (volatilità) insita nell'investimento stesso

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

  - Per facilità di calcolo a volte si usa la seguente formula indiretta, che deriva dalla precedente con alcuni passaggi algebrici che non vediamo:

$$var(x) = E(x^2_i) - E(x)^2$$

- ✓ Lo **scarto quadratico medio** (deviazione standard) è la **radice quadrata della varianza** ed è espressa nell'unità di misura della variabile che stiamo misurando

Si veda parte pratica in excel

Media e varianza pesate

- ✓ La **media pesata** è calcolata sommando tutte le osservazioni di una serie moltiplicate per il loro peso e dividendola per la somma dei pesi. Data una distribuzione $x_1, x_2, \ldots, x_n$, con n numero di osservazioni e w_i il peso assegnato ad ogni osservazione:

$$E_p(x) = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

- ✓ In presenza di pesi anche la varianza si può calcolare «pesata»

$$var_p(x) = \frac{\sum_{i=1}^{n} (x_i - E(x))^2 w_i}{\sum_{i=1}^{n} w_i}$$

- ✓ Lo **scarto quadratico medio** (deviazione standard) è la **radice quadrata della varianza**

Si veda parte pratica in excel

La covarianza

- ✓ La **covarianza** tra due variabili *x* e *y* rappresenta quanto la loro variabilità è correlata. E' alla base del calcolo dell'indice di correlazione e dei coefficienti della retta di regressione con il metodo dei minimi quadrati che vedremo nel Capitolo 6 del libro. La covarianza, come la varianza, ha come unità di misura il quadrato di quella delle osservazioni di partenza

$$\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(x))\ (y_i - E(y))$$

- ✓ Anche per la covarianza abbiamo a disposizione la **formula indiretta** che deriva dalla precedente con alcuni passaggi algebrici: la media dei prodotti tra *x* e *y* meno il prodotto delle medie

$$\text{cov}(x, y) = E(xy) - E(x)\,E(y)$$

Si veda parte pratica in excel

Statistica per l'impresa

2. Le informazioni statistiche: Disponibilità e produzione

GENERALITA'

- ✓ Concetto di impresa
- ✓ Concetto di dato statistico

LE FONTI DEI DATI STATISTICI

- ✓ Fonti interne vs esterne, dati primari e secondari
- ✓ Fonti esterne
  - I conti nazionali
  - Le informazioni sulle imprese
  - Le informazioni sui consumatori

LA QUALITA' DEI DATI STATISTICI

L'INDAGINE CAMPIONARIA AD HOC

- ✓ Popolazione obiettivo, di selezione e di indagine
- ✓ Le fasi dell'indagine campionaria
- ✓ Il piano di campionamento
- ✓ Campioni probabilistici vs non probabilistici
- ✓ Metodi di selezione dei campioni probabilistici
- ✓ Stima dei parametri di una popolazione
- ✓ Stime intervallari
- ✓ Tecniche di rilevazione
- ✓ Il questionario

Imprese, addetti e unità locali

- ✓ **L'impresa è un soggetto giuridico-economico** che produce **beni e/o servizi** destinabili alla **vendita**.
  Possono essere:
  - **organizzate in varie forme**, per esempio imprese individuali, società di persone, capitali, cooperative. Anche lavoratori autonomi e liberi professionisti sono da considerarsi imprese
  - **unilocalizzate o plurilocalizzate** a seconda svolgano l'attività in una sola sede territoriale oppure in più luoghi
- ✓ **L'unità locale rappresenta il luogo fisico** dove l'impresa esercita l'attività economica (per esempio un albergo, un negozio, un'officina, eccetera)
- ✓ **L'addetto indica la persona occupata** presso l'impresa

N° di imprese, di unità locali e di addetti alle imprese e alle unità locali, nei settori trasporti e delle attività finanziarie in Toscana

Attività economica	Imprese	Addetti imprese	Unità locali	Addetti unità locale
Trasporto terrestre e trasporto mediante condotte (49)	5.412	27.247	5.921	28.945
Attività di servizi finanziari (64)	589	29.287	2.480	22.982

Fonte: elaborazione su dati ISTAT

- ✓ Per **prendere decisioni** l'impresa ha bisogno di **dati e strumenti** per la loro elaborazione

- ✓ Il **dato statistico** è il risultato di:
  - una misurazione
  - su un'unità statistica
  - appartenente ad una popolazione (collettività)

- ✓ Rispetto alla **fonte dei dati**, l'impresa può disporre di:
  - **Dati Interni**: le informazioni sono presenti all'interno del sistema informativo aziendale (contabilità, controllo di gestione, dati tecnici, …, spesso gestiti attraverso *data warehouse*)
  - **Dati Esterni**: le informazioni relative:
    - al comportamento e ai prodotti offerti dai concorrenti
    - all'andamento del proprio mercato di riferimento
    - al contesto macroeconomico e finanziario

- ✓ **A volte i dati necessari non sono disponibili** da fonti esistenti. L'impresa diventa allora essa stessa produttrice di dati (es. un'analisi campionaria sulla soddisfazione dei propri clienti). Si parla in questo caso di **dati primari**, per distinguerli da quelli **secondari** (che esistono da fonti esterne o come conseguenza dell'attività di impresa

- ✓ Il **metadato** è «un dato che descrive altri dati»
  - Definisce le unità di analisi e i caratteri osservati
  - Specifica la classificazione adottata
  - Descrive la metodologia adottata

I principali metadati delle statistiche sulle imprese

✓ Per classificare le attività delle imprese, delle loro unità locali e dei loro addetti sul territorio si impiegano:

- **Classificazione delle attività economiche** delle imprese (**ATECO**)
  - Si tratta di un **codice alfanumerico**, in cui la lettera indica la macro attività economica e poi un codice numerico di sei cifre, che ne indica la sottocategoria
  - I **primi quattro numeri** «classificatori» dell'attività economica corrispondono ad una **classificazione internazionale (NACE** in Europa)
  - Il codice ATECO viene **gestito dall'ISTAT** , attualmente in uso la versione 2007, con aggiornamento 2025
- **Nomenclatura delle Unità Territoriali** a fini statistici (**NUTS**)
  - Si tratta di un **codice alfanumerico**, che ha lo scopo di suddividere il territorio dei Paesi dell'Unione Europea e del Regno Unito in unità territoriali gerarchiche e comparabili
  - Ci sono **tre livelli** di classificazione demografica. Dal più grande al più piccolo: **NUTS1, NUTS2 e NUTS3**

Livello	Popolazione Limite inferiore	Popolazione Limite superiore	Esempi
NUTS1	3 mln	7 mln	Stati federali Germania, Regioni del Belgio, Galles, Scozia. Per l'Italia Nord Est, Nord Ovest, Centro, Sud e Isole
NUTS2	0,8 mln	3 mln	Entità amministrative -> Regioni
NUTS3	0,15 mln	0,8 mln	Entità amministrative -> Comuni

I principali metadati delle statistiche sulle imprese: alcuni esempi

L	**ATTIVITÀ IMMOBILIARI**
68	**ATTIVITÀ IMMOBILIARI**
68.1	**COMPRAVENDITA DI BENI IMMOBILI EFFETTUATA SU BENI PROPRI**
68.10	**Compravendita di beni immobili effettuata su beni propri**
68.10.0	Compravendita di beni immobili effettuata su beni propri
68.10.00	Compravendita di beni immobili effettuata su beni propri
68.2	**AFFITTO E GESTIONE DI IMMOBILI DI PROPRIETÀ O IN LEASING**
68.20	**Affitto e gestione di immobili di proprietà o in leasing**
68.20.0	Affitto e gestione di immobili di proprietà o in leasing
68.20.01	Locazione immobiliare di beni propri o in leasing (affitto)
68.20.02	Affitto di aziende
68.3	**ATTIVITÀ IMMOBILIARI PER CONTO TERZI**
68.31	**Attività di mediazione immobiliare**
68.31.0	Attività di mediazione immobiliare
68.31.00	Attività di mediazione immobiliare
68.32	**Gestione di immobili per conto terzi**
68.32.0	Amministrazione di condomini e gestione di beni immobili per conto terzi
68.32.00	Amministrazione di condomini e gestione di beni immobili per conto terzi

ATECO[1] ⟷ NACE[2]

Codice NACE Rev. 2	Titolo NACE Rev. 2	Codice Ateco 2007 aggiornamento 2022
L	REAL ESTATE ACTIVITIES	L
68	Real estate activities	68
68.1	Buying and selling of own real estate	68.1
68.10	Buying and selling of own real estate	68.10
68.10	Buying and selling of own real estate	68.10.0
68.10	Buying and selling of own real estate	68.10.00
68.2	Rental and operating of own or leased real estate	68.2
68.20	Rental and operating of own or leased real estate	68.20
68.20	Rental and operating of own or leased real estate	68.20.0
68.20	Rental and operating of own or leased real estate	68.20.01
68.20	Rental and operating of own or leased real estate	68.20.02
68.3	Real estate activities on a fee or contract basis	68.3
68.31	Real estate agencies	68.31
68.31	Real estate agencies	68.31.0
68.31	Real estate agencies	68.31.00
68.32	Management of real estate on a fee or contract basis	68.32
68.32	Management of real estate on a fee or contract basis	68.32.0
68.32	Management of real estate on a fee or contract basis	68.32.00

NUTS[3] →

Denominazione (Italiana e straniera)	Denominazione in italiano	Codice Ripartizione Geografica	Ripartizione geografica	Denominazione Regione	Denominazione dell'Unità territoriale s	Codice Comune formato numerico	Codice NUTS1 2024	Codice NUTS2 2024 (3)	Codice NUTS3 2024
Aiello del Friuli	Aiello del Friuli	2	Nord-est	Friuli-Venezia	Udine	30001	ITH	ITH4	ITH42
Amaro	Amaro	2	Nord-est	Friuli-Venezia	Udine	30002	ITH	ITH4	ITH42
Ampezzo	Ampezzo	2	Nord-est	Friuli-Venezia	Udine	30003	ITH	ITH4	ITH42
Aquileia	Aquileia	2	Nord-est	Friuli-Venezia	Udine	30004	ITH	ITH4	ITH42
Arta Terme	Arta Terme	2	Nord-est	Friuli-Venezia	Udine	30005	ITH	ITH4	ITH42

[1] Attività Economiche

[2] Nomenclature statistique des Activités économiques dans la Communauté Européenne

[3] Nomenclatura comune delle unità territoriali statistiche

Fonte: ISTAT

METADATI ESEMPIO IMPRESE ESPORTATRICI

Le fonti esterne: i conti nazionali 1/2

- ✓ L'attività dell'impresa risente del **contesto macroeconomico** in cui opera, dunque rilevante citare i **conti nazionali tra le fonti di dati esterni**
- ✓ Essi **registrano i flussi monetari relativi alle fasi del processo economico**, ovvero:
  - **Fase di produzione**: le diverse unità produttive, grazie all'utilizzo di beni intermedi e fattori produttivi, generano nuove risorse
  - **Fase di distribuzione e re-distribuzione**: i detentori di capitale e lavoro vengono remunerati. L'operatore pubblico, attraverso prelievi e trasferimenti, redistribuisce parte delle risorse, determinando così il reddito disponibile di ciascun operatore
  - **Fase del consumo**: gli operatori economici scelgono quanta parte del reddito destinare ai consumi o al risparmio
  - **Fase di accumulazione**: il reddito non «consumato» viene destinato ad investimenti
- ✓ Tra i conti nazionali ricordiamo:
  - Il **Conto delle risorse e degli impieghi**: descrive le operazioni di scambio che avvengono sul mercato, bilanciando offerta (PIL + Importazioni) e domanda (Consumi + Investimenti + Esportazioni)
  - I **Conti per settore istituzionale**: analizzano il comportamento economico di gruppi di operatori omogenei, divisi in cinque settori (Società non finanziarie, Società finanziarie, Amministrazioni pubbliche, Famiglie e Istituzioni senza scopo di lucro)

Le fonti esterne: i conti nazionali 2/2

PRINCIPALI AGGREGATI PER SETTORE ISTITUZIONALE

Anni 2021-2023, Milioni di euro, variazioni percentuali

		2021	2022	2023	2022/2021	2023/2022
Famiglie consumatrici	Valore aggiunto	171.427	174.107	184.467	1,6	6,0
	Risultato lordo di gestione	148.122	151.313	161.526	2,2	6,7
	Reddito disponibile	1.179.835	1.247.137	1.305.853	5,7	4,7
	Accreditamento/Indebitamento netto	102.929	31.219	33.593		
Famiglie produttrici	Valore aggiunto	283.138	296.705	313.231	4,8	5,6
	Risultato lordo di gestione	255.632	268.500	282.675	5,0	5,3
	Accreditamento/Indebitamento netto	-2.400	-7.380	-5.983		
Società non finanziarie	Valore aggiunto	871.043	956.766	1.016.028	9,8	6,2
	Risultato lordo di gestione	379.787	434.602	455.386	14,4	4,8
	Accreditamento/Indebitamento netto	46.422	54.872	86.361		
Società finanziarie	Valore aggiunto	68.701	79.739	105.120	16,1	31,8
	Risultato lordo di gestione	29.562	38.981	63.928	31,9	64,0
	Accreditamento/Indebitamento netto	57.455	66.618	57.269		
Amministrazioni Pubbliche	Valore aggiunto	239.019	253.275	253.704	6,0	0,2
	Risultato lordo di gestione	51.586	54.712	56.213	6,1	2,7
	Accreditamento/Indebitamento netto	-159.169	-167.958	-149.475		

Fonte: ISTAT

Territorio	Italia			
Valutazione	valori concatenati con anno di riferimento			
Correzione	dati destagionalizzati			
Edizione	Mag-2024			
Seleziona periodo	T2-2023	T3-2023	T4-2023	T1-2024
Tipo aggregato (milioni di euro)				
conto economico delle risorse e degli impieghi	..	..	..	..
risorse	**590.653**	**589.034**	**589.831**	**589.001**
prodotto interno lordo ai prezzi di mercato	446.614	448.298	448.920	450.461
importazioni di beni (fob) e servizi	144.681	141.815	141.995	139.597
importazioni di beni (fob)	113.330	111.032	110.255	108.368
importazioni di servizi	31.756	31.191	32.346	31.688
impieghi	**590.653**	**589.034**	**589.831**	**589.001**
spesa per consumi finali nazionali	344.065	346.215	343.102	343.939
spesa per consumi finali delle famiglie e delle istituzioni sociali private senza scopo di lucro al servizio delle famiglie (isp) concetto nazionale	262.501	264.570	260.901	261.656
spesa per consumi finali delle amministrazioni pubbliche	81.674	81.768	82.286	82.393
investimenti fissi lordi	98.053	99.402	101.356	101.850
esportazioni di beni (fob) e servizi	148.260	149.853	151.716	152.637
esportazioni di beni (fob)	119.283	120.363	121.353	120.755
esportazioni di servizi	29.204	29.738	30.657	32.307

Dati estratti il 07 Aug 2024 13:31 UTC (GMT) da I.Stat

Conto economico Risorse ed Impieghi dal sito dell'ISTAT

(https://www.istat.it/statistiche-per-temi/economia/conti-nazionali/#Accesso-ai-dati)

Le fonti esterne: le informazioni sulle imprese

✓ Le fonti sulle imprese riguardano:

- Le **caratteristiche strutturali** delle imprese, quali
  - **Settore**
  - **Dimensione**
  - **Territorio**

- **Censimento decennale dell'Industria e dei Servizi** (CIS), fino al 2011
- **Censimento permanente Industria e Servizi** (triennale) su base campionaria
- **ASIA**: archivio statistico delle imprese attive, **basato su dati amministrativi**
  https://www.istat.it/tavole-di-dati/registro-statistico-delle-imprese-attive-anno-2023/

- I **Risultati economici** delle imprese:
  - **Indagini ISTAT:** la Rilevazione sul sistema dei conti delle grandi imprese (SCI), la rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni (PMI) e gli indicatori congiunturali (serie storiche su fatturato, ordinativi, costi delle materie prime, …
  - **Banche dati di bilanci aziendali:** si tratta di raccolte di bilanci di imprese, che consentono funzioni di interrogazione e/o di analisi. Tra le più citate CERVED, Centrale dei Bilanci. A livello settoriali associazioni di categoria e/o autorità pubbliche di controllo spesso raccolgono i bilanci delle imprese associate/sottoposte a controllo

Le fonti esterne: le informazioni sui consumatori

- ✓ La **comprensione e la previsione dei comportamenti dei consumatori** è sempre oggetto di studio da parte delle imprese. Sui consumatori utile ricordare:
  - Indagine sui **consumi delle famiglie italiane** (ISTAT)**:**
    - annuale, su un campione di 28.000 famiglie
    - basata su un diario di spesa compilato dalla famiglia stessa ed un'intervista finale sulle abitudini di consumo
    - fornisce indicazioni anche a livello regionale
    - comparabile a livello europeo
  - **Indagini panel** di fornitori privati (es. Nielsen)
    - Seguono un campione nel corso del tempo e ne colgono le variazioni
    - Sui consumatori (consumer panel)
    - Sui prodotti (retail panel)
  - Indagine sui **Bilanci delle famiglie italiane della Banca d'Italia**: condotta a partire dagli anni '60 per raccogliere informazioni su redditi, ricchezza e risparmi delle famiglie italiane.

Le fonti esterne: le informazioni sui consumatori - esempi

statistiche today

25 MARZO 2024

Istat | Istituto Nazionale di Statistica

ECONOMIA

Resta stabile la povertà assoluta, la spesa media cresce ma meno dell'inflazione

Spesa delle famiglie in valori correnti ancora in aumento per l'inflazione

Povertà assoluta familiare e individuale stabile

BANCA D'ITALIA EUROSISTEMA

Statistiche

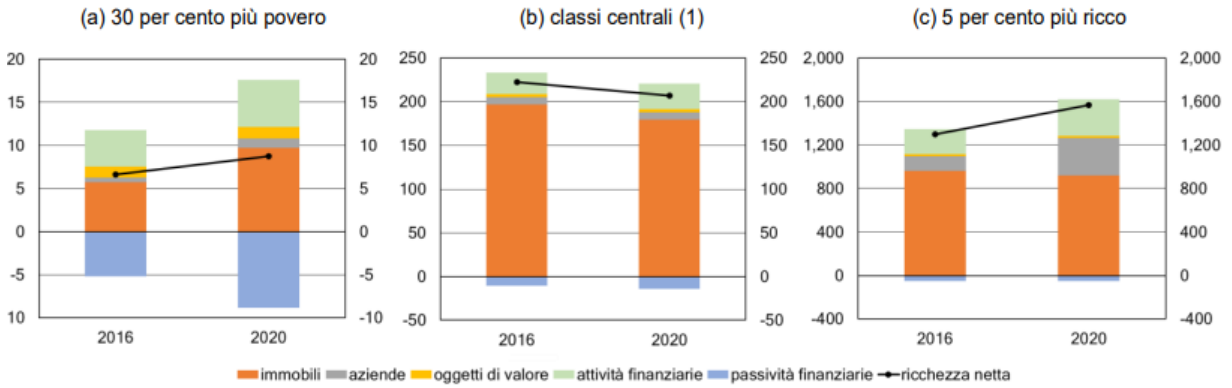
Indagine sui bilanci delle famiglie

22 luglio 2022

Per informazioni: statistiche@bancaditalia.it
www.bancaditalia.it/statistiche/index.html

Figura 6

Valore medio della ricchezza familiare netta e delle sue componenti lungo la distribuzione della ricchezza
(migliaia di euro; prezzi 2020)



Fonte: Elaborazioni sull'archivio storico dell'Indagine sui bilanci delle famiglie italiane, versione 11.1.
1) Tra il 30° e il 95° percentile della distribuzione della ricchezza familiare netta.

La qualità delle informazioni statistiche

- ✓ Qualsiasi sia la fonte dati che si va ad utilizzare è bene tenere presente i sei criteri guida dei sistemi statistici (individuati da Eurostat), anche indicati come **dimensione della qualità**
  - **Rilevanza:** capacità del sistema statistico di rispondere alle esigenze degli utilizzatori
  - **Accuratezza:** statisticamente è la vicinanza tra la stima ed il valore vero di una caratteristica misurata (es. differenziale tra dati preliminari e definitivi)
  - **Puntualità e tempestività:** la puntualità fa riferimento alla corrispondenza tra l'effettiva data di diffusione delle statistiche e quella prevista, mentre la tempestività fa riferimento al tempo che intercorre tra la data di diffusione ed il periodo a cui fanno riferimento
  - **Accessibilità e chiarezza**: la prima è la capacità del sistema di rendere chiare e semplici le procedure per l'acquisizione dei dati, la seconda la capacità di rendere comprensibili e correttamente interpretabili le statistiche fornite
  - **Comparabilità**: molte analisi confrontano l'evoluzione nel tempo o nello spazio di fenomeni. E' importante che le differenze che si ravvisano dipendano da veri fenomeni economici e non cambi di metodologia statistica
  - **Coerenza**: quando il complesso delle informazioni desumibile dalle diverse statistiche disponibili fornisce un quadro in sé coerente e non contraddittorio

La produzioni di statistiche ad hoc: l'indagine campionaria

- ✓ Quando le statistiche disponibili non sono adeguate l'impresa può decidere di svolgere **un'indagine ad hoc**
- ✓ In genere è difficile e costoso acquisire le informazioni necessarie su tutte le unità che compongono la popolazione che intendo osservare → **si osserva** una parte di essa: un **campione**
- ✓ Un po' di terminologia che ci sarà utile:
  - **Popolazione obiettivo** (universo): la popolazione che intendo osservare (le famiglie, le imprese artigiane, …)
  - **Il campione**: la parte della popolazione obiettivo di cui vengono raccolte le informazioni
  - **Le unità di rilevazione**: l'unità a cui rivolgono le domande (l'individuo, l'artigiano, …)
  - **Le unità di analisi**: l'unità che intendo analizzare. Può coincidere con quella di rilevazione o essere diversa (rivolgo le domande ad un individuo, ma l'oggetto dell'analisi è la sua abitazione (di proprietà, in affitto, metratura, ..)

Le fasi dell'indagine campionaria

✓ Il **processo** di realizzazione di **un'indagine campionaria** attraversa alcune **fasi**:

- Definizione degli **obiettivi**
- Identificazione della **popolazione di riferimento**
- Scelta dei **criteri di selezione** del campione (piano di campionamento)
- Scelta della **metodologia di stima**
- Scelta della **modalità di raccolta dati**
- Messa a punto del **questionario**
- **Organizzazione** della fase di rilevazione dei dati
- **Valutazione dei costi** di realizzazione dell'intera indagine **vs il margine di errore tollerabile**

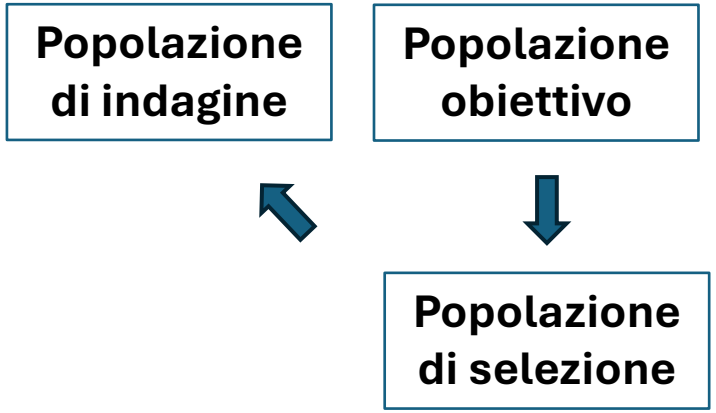
Popolazione obiettivo, popolazione di selezione e popolazione di indagine

- ✓ Le **popolazioni oggetto di analisi** possono avere dimensioni **finite** (es. residenti) o **infinite** (i pezzi di un prodotto che esce da un processo produttivo per valutarne la qualità)
- ✓ **Ci occuperemo di popolazioni finite di numerosità *N***
- ✓ Per definire la **popolazione obiettivo** occorre stabilire le unità da osservare (es. le famiglie), ma anche le coordinate temporali (*sentiment* dopo un evento per esempio) e spaziali (totale Italia vs Nord, Centro e Sud, ..)
- ✓ Stabilita la popolazione obiettivo è necessario reperire **la lista di campionamento** delle unità che ne fanno parte. Per esempio:
  - Anagrafe
  - Le liste elettorali
  - Elenchi telefonici/raccolte via web
  - Registro delle imprese presso le Camere di Commercio
  - Liste di associati ad associazioni di categoria
  - …
- ✓ **Ottengono così la popolazione di selezione** e su questa effettuo la mia indagine
- ✓ Svolgo poi la mia indagine (interviste telefoniche, questionari via *web*, …) ed ottengo **la popolazione di indagine** che è quella **su cui posso generalizzare le informazioni** raccolte sul campione. Attenzione che potrebbe essere diversa da quella obiettivo

Problemi di selezione della popolazione

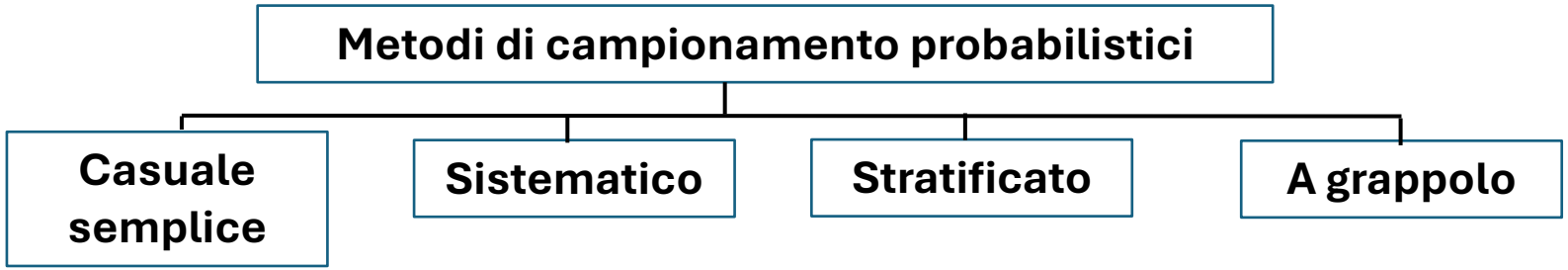
- ✓ Se la **popolazione di indagine non rappresenta quella obiettivo**, le informazioni raccolte non possono essere generalizzate a quest'ultima

- ✓ Può succedere quando:
  - **Manca** perfetta **corrispondenza tra popolazione di selezione e popolazione obiettivo** (es. voglio osservare i maggiori di 14 anni, usando le liste elettorali che hanno solo i maggiorenni)
  - **Mancata risposta** per *drop out*, rifiuto o risposte contraddittorie (es. questionario di soddisfazione dei miei punti vendita sul territorio, alcuni dei quali, magari tutti appartenenti ad un'area territoriale, non mi hanno mandato i risultati)



La formazione del campione/Il piano di campionamento

- ✓ **L'obiettivo** di **un'indagine campionaria** è quello di **stimare** alcuni **parametri** della **popolazione obiettivo**. Per esempio il tempo medio di attesa dei clienti in un certo esercizio commerciale, la proporzione dei clienti soddisfatti rispetto a quelli insoddisfatti
- ✓ Attraverso l'indagine campionaria calcoliamo **otteniamo una stima del parametro** che vogliamo conoscere. La **differenza tra parametro vero** della popolazione **e la sua stima** è **l'errore campionario**
- ✓ E' **importante conoscere le caratteristiche dell'errore campionario**: è il 5% del valore che voglio stimare? E' il 25%? Quando varia l'errore rispetto alla sua media? Se l'errore è troppo «grande» in media o variabilità, la mia indagine potrebbe essere inutile
- ✓ **L'errore può essere stimato** (e si dice errore statistico) **se il campione** è stato selezionato con un metodo **casuale** (campione probabilistico). NOI **CI CONCENTREREMO SUL CAMPIONE PROBABILISTICO**
- ✓ Se invece abbiamo tra le mani un campione non probabilistico, non potremo misurare statisticamente l'errore (per esempio raccolgono opinioni di esperti, *influencer*, ..)



- ✓ La regola con cui vengono selezionate le unità del campione è il **Piano di campionamento**

Campioni probabilistici 1/5

- ✓ I **campioni probabilistici**:
  - sono selezionati con un meccanismo **casuale**
  - tutte le unità hanno una stessa **probabilità nota** di essere estratte

- ✓ Sia ***Y* una variabile** che rappresenta il carattere di interesse nella **popolazione di *N* unità**. Estraiamo casualmente il campione:

$$\{ y_1, y_2, y_3, \ldots\ldots, y_n \}$$

- ✓ Definiamo:

$n \rightarrow$ la dimensione campionaria

$f = \frac{n}{N} \rightarrow$ la frazione di campionamento

- ✓ *Esempio: voglio stimare l'altezza media degli studenti delle scuole superiori in Italia (Y=2,6 mln). Estraggo un campione di 1.000 studenti (n). La frazione di campionamento è f=1.000/2.600.000=0,038 (3,8%)*

Campioni probabilistici 2/5

✓ Il **campionamento casuale semplice (CCS):**

- **Ogni singola unità** ha la **stessa probabilità di essere estratta**. Si parla allora di probabilità di inclusione, che è pari alla frazione di campionamento:

$$\pi_i = \frac{n}{N}$$

- La **procedura** di selezione del campione è analoga all'estrazione di pallina da un'urna, **con o senza reimmissione**

✓ Il **campionamento sistematico:**

- Come nel caso precedente ogni unità ha la stessa probabilità di essere estratta
- Operativamente seleziono casualmente un numero *j* compreso tra *1* e k, dove *k = int (N/n)*, si dice anche il passo di campionamento
- *Per esempio: ho 250 indirizzi (N=250). Ad ognuno è associato un numero progressivo da 1 a 250. Voglio selezionare un campione in modo sistematico di numerosità 25 (n=25). Genero un numero casuale intero tra 1 e k, che in questo esempio è k=10 (N/n ovvero 250/25). Ottengo per esempio 7. Estraggo il 7° indirizzo dalla lista di campionamento. Poi aggiungo alla prima osservazione (7) il valore di k che era 10. La seconda osservazione sarà allora la 17 (7+10). Proseguo fino ad aver selezionato 20 indirizzi*

Si veda parte pratica in excel

Campioni probabilistici 3/5

✓ Il **campionamento stratificato: un esempio**

- **Voglio che il campione** che vado ad estrarre casualmente presenti delle **caratteristiche simili** a quelle della **popolazione obiettivo** (es. proporzione uomini e donne, distribuzione Nord, Centro, Sud e Isole)
- **Le caratteristiche che voglio «preservare» sono gli strati**: per esempio la percentuale di donne è il 60%, quella degli uomini del 40% a livello di popolazione obiettivo. Voglio che anche il campione che andrò ad estrarre abbia le medesime caratteristiche
- **La stratificazione consente** usualmente di **essere più precisi** nelle stime, a parità di numerosità campionaria, **oppure a ridurre le dimensioni del campione** per raggiungere il livello di precisione desiderato (con conseguente **risparmio di costi** per la rilevazione)
- Facciamo **un esempio:**
  - **N=3.000**, di cui **N_1=1.800 donne** e **N_2=1.200 uomini**
  - La proporzione delle donne sul totale **W_1=0,6** (cioè 1.800/3000), mentre quella degli uomini è **W_2 = 0,4.** Ovviamente **$W_1 + W_2 = 1$**
  - Voglio estrarre un campione di numerosità **n=300** (f=0,1, ovvero 300/3000), in modo che la proporzione di donne e uomini nel campione sia quella che posso misurare nella popolazione obiettivo (**stratificazione proporzionale**)
  - La **numerosità campionaria** sarà costituita allora dalla somma di **$n_1 + n_2$**, dove **$n_1 = f*N_1$ e $n_2 = f*N_2$**, cioè **n_1=180** (cioè 1.800*0,1), **n_2=120** (cioè 1,200*0,1). Ovviamente n_1+n_2=**n=300**

si veda parte pratica in excel

Campioni probabilistici 4/5

✓ Il **campionamento stratificato in termini analitici**

- Sia h_i **il singolo strato e** H **il numero degli strati**, per *cui h = 1, ...H*. Nell'esempio uomini/donne *H=2*, dunque ci sono due strati h_1 (le donne) e h_2 (gli uomini). Sia inoltre:

N_h la dimensione dell'h-esimo strato della popolazione

$W_h = \frac{N_h}{N}$ la **proporzione della popolazione nello strato h**, dove $\sum_{h=1}^{H} W_h = 1$

(nell'esempio la W_1 è la proporzione delle donne sul totale e W_2 la proporzione degli uomini, W_1 + e W_2 =1 o in alternativa il 100%)

n_h la dimensione del campione estratto dall'**h-esimo strato**, dove $\sum_{h=1}^{H} n_h = n$

$$f_h = \frac{n_h}{N_h} \quad \text{ovvero} \quad n_h = f_h * N_h$$

- Nel campionamento con **stratificazione proporzionale**

$$f_h = f = \frac{n}{N} \quad \text{dunque } n_h = f * N_h$$

si veda parte pratica in excel

Campioni probabilistici 5/5

- ✓ Il **campionamento a grappoli**
  - La lista degli *N* elementi è suddivisa in cosiddetti grappoli
  - Seleziono prima i grappoli dai quali estrarre il campione e poi da questi il campione su cui misurare le caratteristiche della popolazione che ho come obiettivo dell'indagine
  - Per esempio:
    - devo fare un'indagine sulle esportazioni delle imprese artigiane del Nord Est. Devo trovare ragione sociale ed indirizzo e poi recuperare i bilanci
    - prendo a riferimento gli elenchi dei soci delle associazione artigiane a livello provinciale
    - potrebbe richiedere molto lavoro manuale avere tutti gli elenchi o costare molto. Allora considero le province/associazioni i miei grappoli, ne selezioni alcune con i metodi già visti (devono essere casuali!)
    - Fatta questa prima scrematura, per i grappoli recupero le informazioni che servono e finalmente selezioni il campione con i metodi visti
- ✓ Il **campionamento a stadi**
  - Può essere visto come una variante del campionamento a grappoli
  - Una volta selezionati i grappoli (unità di primo stadio), seleziono le unità elementari che compongono il grappolo (unità di secondo stadio)
  - Nell'esempio visto prima, potrei selezionare prima le associazioni provinciali, poi quelle comunali

si veda parte pratica in excel

Campioni non probabilistici

✓ I campioni sono selezionati in base a considerazioni di ordine pratico

- **Campionamento di comodo**: scelta arbitraria (l'intervistatore va nel posteggio dell'Università ed intervista tutti gli studenti che passano tra le 11 e le 12)
- **Campionamento ragionato**: selezione non casuale basata su informazioni a priori (intervisto esperti o *influencer* a seconda del caso)
- **Campionamento per quote**: la popolazione viene suddivisa in base a caratteristiche note, come nel campionamento stratificato, ma poi all'interno degli strati si campiona in modo non casuale (si lascia per esempio che l'intervistatore decida chi coinvolgere nell'indagine, pur rispettando le proporzioni degli strati)

✓ Si tratta di soluzioni subottimali che non offrono alcuna garanzia di rappresentatività della popolazione. E' molto probabile che qualche tipo di involontaria selezione renda la popolazione di indagine diversa dalla popolazione obiettivo

Stima dei parametri 1/2

- ✓ Una volta **selezionato il campione**, vogliamo **stimare un parametro ignoto della popolazione *Θ*** (teta, un valore medio, una proporzione, …) **utilizzando** appunto **il campione estratto**. Prendiamo familiarità con i seguenti concetti:
  - Il **parametro è il vero valore di una caratteristica** della popolazione (per esempio l'altezza media dei residenti in Italia)
  - **Lo stimatore è la variabile (aleatoria) che serve per risalire al valore vero della popolazione** (che tipicamente è ignoto). In una misura un po' imprecisa, possiamo dire che è la «formula» che ci serve per calcolare la stima del parametro a partire dal campione estratto
  - **La stima è il valore assunto dallo stimatore** in corrispondenza del **campione osservato**

Stima dei parametri 2/2

- ✓ In generale:
  - Il **parametro** viene indicato con una lettera greca, nel nostro caso *Θ (teta)*
  - Lo **stimatore** viene indicato con la **lettera maiuscola *T*** e dipende (è **funzione) di tutti i possibili campioni** che possiamo estrarre dalla popolazione obiettivo, per questo scriviamo

$$T = f(Y_1, Y_2, \ldots, Y_n)$$ dove *n* è la numerosità campionaria

  - **La stima** viene indicata con la **lettera minuscola** ed è **funzione del singolo campione selezionato**

$$t = f(y_1, y_2, \ldots, y_n)$$ dove *n* è la numerosità campionaria

- ✓ Uno **stimatore è corretto** se la **media di tutte le possibili stime di *Θ* è uguale al parametro da stimare**

$$E(T) = \Theta$$

  - Vedremo con un esempio che **la media campionaria è uno stimatore corretto della media** della popolazione. Dunque se calcoliamo le medie di tutti i possibili campioni estraibili dalla popolazione e di queste facciamo la media, otteniamo il parametro vero della popolazione
  - **Se la media dello stimatore non è uguale al parametro** vero della popolazione, si dice che **è distorto**

Stima dei parametri: la media e la varianza campionaria

✓ La **media campionaria** è uno **stimatore corretto** della media della popolazione ed è:

$$T_{\overline{Y}} = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (1)$$

✓ La varianza campionaria è uno stimatore distorto della varianza della popolazione. Si usa allora la seguente formula, che costituisce uno **stimatore corretto della varianza della popolazione**

$$S^2_Y = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})2 \qquad (2)$$

✓ La formula dello stimatore della varianza campionaria, dopo semplici trasformazioni algebriche, può essere riscritta nella seguente formula, che è a volta comoda in fase di calcolo

$$S^2_Y = \sigma^2_Y \frac{n}{n-1} \qquad (3)$$

Stima dei parametri: un esempio

- ✓ Ipotizziamo che la popolazione obiettivo sia ***U = {1, 2, 3}***. La media sarà ***μ = 2***, la varianza ***σ^2 = 0,67***
- ✓ Ipotizziamo che la numerosità campionaria sia ***n=2***. Posso allora avere nove campioni (con re-immissione)
- ✓ Ogni campione ha la sua media, la sua varianza. Siccome sappiamo che la varianza campionaria è uno stimatore distorto della varianza della popolazione, allora calcoliamo lo stimatore corretto che abbiamo imparato nella slide precedente (2)

Campione	y1	y2	E(y)	var.p(y)	var.c(y)		
1	1	1	1,00	0,00	0,00		
2	1	2	1,50	0,25	0,50	E(y) =	(y1+y2)/2
3	1	3	2,00	1,00	2,00	var.p(y) =	((y1-E(y))+(y2-E(y))/2
4	2	1	1,50	0,25	0,50	var.c(y) =	((y1-E(y))+(y2-E(y))/1
5	2	2	2,00	0,00	0,00		
6	2	3	2,50	0,25	0,50		
7	3	1	2,00	1,00	2,00		
8	3	2	2,50	0,25	0,50		
9	3	3	3,00	0,00	0,00		
		E(T)	2,00	0,33	0,67		

- ✓ Calcoliamo la media delle medie campionarie ed otteniamo ***E(T)=2*** che è proprio la media ***μ della popolazione***
- ✓ Facciamo lo stesso per la varianza, ma vediamo che la media della varianza campionaria (***0,33***) è molto diversa dalla varianza della popolazione (***σ^2 = 0,67***). **Infatti è uno stimatore distorto. Usiamo quello corretto S^2_Y** *(si veda la formula 2 della slide precedente) è vediamo che la media dello stimatore corrisponde alla varianza campionaria*

si veda parte pratica in excel

Stima dei parametri: l'errore standard

✓ **L'errore standard** della **media campionaria** (nel caso di CCS senza ripetizione) è pari a:

$$ES(T_{\bar{Y}}) = \sqrt{(1-f)\frac{S^2_Y}{n}} \qquad (1)$$

- Ricordiamoci che *f* è la frazione di campionamento (*n/N*), mentre *n* è la numerosità campionaria
- **(1-f)** è il **fattore di correzione per popolazione finite** (trascurabile per popolazione «infinite»)
- Guardando la (1) **possiamo concludere che la precisione dello stimatore $T_{\bar{Y}}$**
- **Più piccolo è l'errore standard, maggiore è la precisione** (efficienza) dello stimatore
  - **Aumenta all'aumentare di *f*** (che al *max* può essere 1)**.** Per piccole frazioni di campionamento (la maggior parte dei casi) questo fattore è trascurabile: infatti *f* tende a zero e *(1-f)* tende a *1*
  - **Aumenta al diminuire della varianza σ^2** della popolazione e del suo stimatore corretto S^2_Y
  - **Aumenta al crescere di *n*** (maggiore è la numerosità campionaria, *ceteris paribus*, minore l'errore *standard*

Stima dei parametri della popolazione: la proporzione 1/3

- ✓ Spesso nelle indagine campionarie è utile conoscere quanta parte della popolazione che sto osservando presenta un certo carattere. Per esempio: voglio capire se le persone con figli acquistano il mio prodotto più degli altri. Sto cercando in altre parole di capire **la proporzione** dei miei clienti con figli rispetto a quelli senza. Così fosse potrei indirizzare un'azione di *marketing* a questo segmento di mercato
- ✓ Facciamo un esempio (immaginiamo di avere solo 10 clienti):

I dati del problema

Clienti	Figli (Sì=1, No=0)
Anna	0
Giacomo	1
Federica	0
Paolo	1
Maria	0
Carlotta	1
Francesca	0
Roberto	1
Vittoria	0
Sofia	1
E(y) =	0,5
var.corr(y) =	0,28

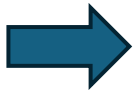


Tabella della frequenze assoluta e relativa

z	F(z)	f(z)	
1	5	0,5	
0	5	0,5	
E(y) =	f(1)	0,5	π

La proporzione di clienti con figli è uguale a quella senza figli (50-50). Dunque chi ha figli non ha una particolare propensione all'acquisto del mio prodotto, però …… Mi sembra di vedere molti nomi di donne….

si veda parte pratica in excel

Stima dei parametri della popolazione: la proporzione 2/3

- ✓ Ripetiamo allora l'esempio dando il numero *1* alle donne e *0* agli uomini. Sto usando una variabile dicotomica di valore 0 e 1. Per tale variabile, la media è la proporzione con cui il carattere *1* si presenta nella popolazione, che abbiamo chiamato π, mentre la varianza è il prodotto di $\pi\ (1-\pi)$ (su tutta la popolazione, se invece stiamo usando dati campionari devo usare la varianza campionaria corretta – vedi slide successiva!)

I dati del problema

Clienti	Donne (Sì=1, No=0)
Anna	1
Giacomo	0
Federica	1
Paolo	0
Maria	1
Carlotta	1
Francesca	1
Roberto	0
Vittoria	1
Sofia	1
E(y) =	0,7
var.corr(y) =	0,23

Tabella della frequenze assoluta e relativa

z	F(z)	f(z)
1	7	0,7
0	3	0,3
E(y) =	π	0,7

Ne 70% (π espresso in %) dei casi i miei clienti sono donne, dunque meglio concentrare gli investimenti in *marketing* verso questo segmento perché ha un'alta propensione dall'acquisto

si veda parte pratica in excel

Stima dei parametri della popolazione: la proporzione 3/3

- ✓ **La proporzione $\pi = N_k / N$** individui che nella popolazione **presentano la caratteristica *k*** può essere **stimata attraverso l'equivalente proporzione nel campione ($p=n_k/n$)**, questo perché sappiamo che la media campionaria è uno stimatore corretto della media della popolazione
- ✓ Posta *Z* la variabile dicotomica che vale *1* se è presenta la caratteristica *k*, oppure *0* in caso contrario, gli **stimatori corretti rispettivamente della proporzione e della varianza della proporzione sono**:

$$T_p = \frac{1}{n} \sum_{i=1}^{n} Z_i = \frac{n_k}{n} \qquad S^2_P = \frac{n\,p\,(1-p)}{n-1}$$

- ✓ **L'errore standard della proporzione campionaria, stimatore corretto della proporzione nella popolazione,** (nel caso di CCS senza ripetizione) con π incognita, è pari a:

$$ES\,(T_p) = \sqrt{(1-f)\,\frac{p\,(1-p)}{n-1}}$$

  - Come abbiamo detto, in questo caso e nella pratica succede che non sappiamo a priori quanto sia π nella popolazione. Possiamo allora:
    - Porre $\pi = p$ dove p è la proporzione campionaria
    - Porre $\pi = 0{,}5$ perché si può dimostrare che è la situazione che massimizza l'errore standard della proporzione. Dunque vogliamo essere prudenti

Stima della media nel campionamento stratificato 1/2

- ✓ La **media campionaria nel campionamento stratificato** non è che la **media ponderata con i pesi degli strati**. Vediamo un esempio. Analogamente vale lo stesso concetto anche per la varianza campionaria corretta e per l'errore standard campionario

- ✓ Vediamo un **esempio**:
  - Sia data una popolazione di ***N=20*** individui: lo stratifico per femmine (11) e maschi (9)
  - Il peso dello strato 1 è dunque W_1***=0,55*** *(11/20)*, mentre quella dello strato 2 W_2***=0,45*** *(9/20)*
  - Sia ***f=0,3*** La frazione campionaria (uguale nei due strati → campionamento proporzionale
  - Estraggo casualmente un campione di ***n=5*** unità, stratificato di tre donne (***h_1=int(5x0,55))*** e uomini (***h_2=int(5*0,45))***

Campione h1			Statistiche	
4	Sara	1,55	$E(y_1)$	*1,52*
11	Teresa	1,51	S^2_{y1}	*0,001*
11	Teresa	1,51	$ES(y_1)$	*0,012*

Campione h2			Statistiche	
5	Giuseppe	1,8	$E(y_2)$	*1,71*
8	Giorgio	1,61	S^2_{y2}	*0,018*
			$ES(y_2)$	*0,082*

  - **Calcolo la media campionaria**: $E(y) = W_1*E(y_1)+W_2*E(y2)=0,55*1,52+0,45*1,71=1,61$ m

Stima della media nel campionamento stratificato 2/2

✓ Nel caso del campionamento stratificato, lo stimatore della **media campionaria** sarà dunque la **media ponderata** degli **stimatori** ottenuti **per ciascun strato:**

$$T_{\overline{Y}_{ST}} = \sum_{h=1}^{H} W_h \overline{Y}_h = \sum_{h=1}^{H} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$$

✓ **L'errore standard dello stimatore della media** sarà:

$$ES\left(T_{\overline{Y}_{ST}}\right) = \sqrt{\sum_{h=1}^{H} W^2{}_h \left(1 - f_h\right) \frac{S^2{}_h}{n_h}}$$

- Dove:
  - N_h è la dimensione dello strato nella popolazione
  - $W_h = \frac{N_h}{N}$ è la proporzione dello strato nella popolazione
  - n_h è la dimensione del campione estratto dallo strato *h*
  - $f_h = \frac{n_h}{N_h}$ è la frazione di campionamento dello strato *h*
  - $S^2{}_h = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Yih - \overline{Y}h)2$ stimatore corretto di var(Y) nello strato *h*

Stima della proporzione nel campionamento stratificato

- ✓ **La stima della proporzione** nel campionamento stratificato è analoga a quanto appena visto. Sarà

$$T_{P_{ST}} = \sum_{h=1}^{H} W_h P_h = \sum_{h=1}^{H} W_h \frac{r_h}{n_h}$$

  - Dove r_h è il numero degli elementi dello strato che presenta la caratteristica in esame

- ✓ Per calcolare l'errore standard della proporzione nel caso del campionamento stratificato **devo avere la varianza della popolazione del singolo strato *h***:

$$S^2_{P_h} = \frac{n_h P_h (1-P_h)}{n_h - 1}$$

- ✓ **L'errore standard della proporzione campionaria** sarà**:**

$$ES\,(T_{P_{ST}}) = \sqrt{\sum_{h=1}^{H} W^2_h (1 - f_h) \frac{P_h (1-P_h)}{(n_h - 1)}}$$

Stima intervallari 1/3

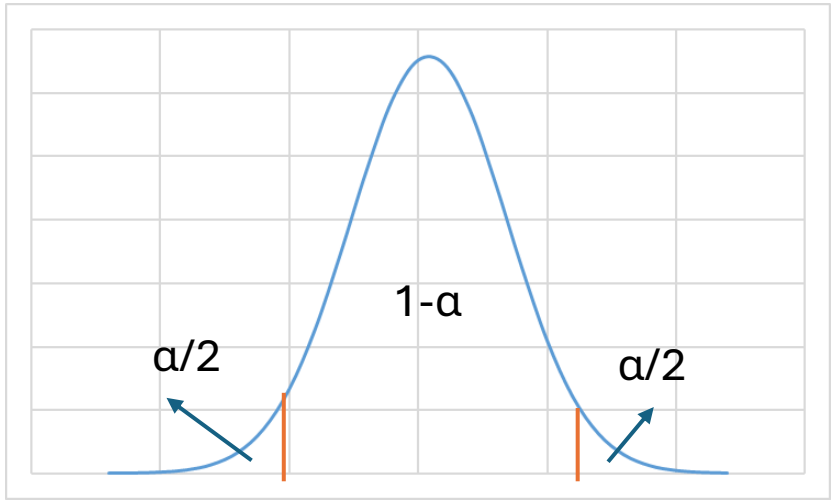
- ✓ Ci siamo sinora occupati di stime campionarie puntuali
- ✓ Un approccio differente è quello di costruire attorno alla stima puntuale **un intervallo che contenga il valore vero della popolazione con un certo livello di probabilità** decisa a priori:
  - L'intervallo che contiene il parametro ignoto della popolazione con un certo livello di probabilità assegnata a priori, si dice **intervallo di confidenza** (o fiducia)
  - La **probabilità** decisa a priori è detta **livello di confidenza** (o fiducia) ed è indicata con ***1-α***
  - Chiamiamo infine livello di **significatività** il valore ***α*** che ci indica la probabilità che l'intervallo non contenga il parametro ignoto della popolazione: in altre parole è **l'errore** che mi attendo dalla stima
- ✓ Facciamo un **esempio**: un'impresa vuole sapere quanti **giorni in media rimane in giacenza** un certo prodotto, in modo da ottimizzare la produzione. Lo statistico dell'impresa estrae un **campione (**per esempio 100 osservazioni) e calcola il **valore medio campionario, pari a E(y)=45 giorni e la deviazione standard corretta, pari a 5 giorni**. Vogliamo adesso sapere **qual è l'intervallo in cui al 95%** *(1-α)* sarà **compresa la giacenza media del prodotto** a partire dal valore campionario calcolato: in altri termini quello che cerchiamo di calcolare è:

$$P(45 - err \leq \mu \leq 45 + err) = 95\%$$

  - Non ci rimane che **calcolare l'errore.** Sappiamo che dipenderà dalle caratteristiche della distribuzione dei giorni di giacenza del prodotto e dalla dimensione del campione

Stima intervallari 2/3

✓ **Per «grandi» campioni** (sul libro si dice >30, ma non c'è una vera regola aurea) **la media campionaria si distribuisce come una Normale** a prescindere dalla distribuzione della popolazione (**Teorema del Limite Centrale**)



✓ In questo caso[1] **l'intervallo di confidenza per la media campionaria al *livello 1-α*** è dato da, rispettivamente per *1-α=95%, 1-α=99% e 1-α=90*:

$$I_{\bar{Y},\,95\%} = T_{\bar{Y}} \text{ +/- } 1{,}96 * ES(T_{\bar{Y}})$$
$$I_{\bar{Y},\,99\%} = T_{\bar{Y}} \text{ +/- } 2{,}58 * ES(T_{\bar{Y}})$$
$$I_{Y,\,90\%} = T_{\bar{Y}} \text{ +/- } 1{,}64 * ES(T_{\bar{Y}})$$

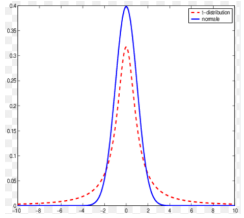
$\Rightarrow$ **$I_{\bar{Y},\,1-\alpha\%} = T_{\bar{Y}} \text{ +/- } z_{\alpha/2} * ES(T_{\bar{Y}})$**

dove, per <u>popolazioni</u> «grandi» $ES(T_{\bar{Y}}) = \frac{\sigma_Y}{\sqrt{n}}$ perché *(1-f) ~ 1*

✓ Calcoliamo allora l'intervallo nel **nostro esempio** sapendo che $err_{2,5\%} = 45 \text{ +/-} 1{,}96 * \frac{5}{\sqrt{100}} = 1$ → **P (44 ≤ μ ≤ 46) = 95%**

[1] Normale standardizzata

Stima intervallari 3/3



- ✓ **Nel caso di campioni «piccoli»** ed è **nota la varianza** della caratteristica della popolazione, la media campionaria si distribuisce **sempre come una normale** (si tratta di un caso un po' teorico)
- ✓ **Nel caso di campioni «piccoli»,** in cui è **sconosciuta** a priori **la varianza** della caratteristica della popolazione, la media campionaria si distribuisce **come una *t* di *Student* con *n-1* gradi di libertà**, dove n è la grandezza del campione**.** Si tratta di una normale con le «code» un po' più alte e dunque con dei i valori critici più elevati
  - Facciamo un altro esempio (libro pag. 67 tab.2.8). Un'impresa di assistenza idraulica vuole stimare il tempo medio di attesa dei clienti nell'ultimo mese. La lista dei clienti (la popolazione obiettivo è di *N=115* unità)
  - Viene estratto un campione di *n=25* unità, dunque *f=0,22* (frazione campionaria pari a *n/N*), con media *y=2,24* ore ed una varianza campionaria corretta di $S^2_y=2,11$
  - Fisso un livello di confidenza al *95%* e costruisco l'intervallo di confidenza corrispondente. Sarà:
  - Si tratta di un campione piccolo, la cui varianza è sconosciuta → la media campionaria si distribuirà come una *t* di *Student* con *n-1=24* gradi di libertà. In questo caso specifico l'intervallo di confidenza sarà «delimitato» dai punti calcolati con la seguente formula

$$I_{Y,\,95\%} = T_{\bar{Y}} +/- 2,064 * ES(T_{\bar{Y}})$$ ⇒ $$\boxed{I_{\bar{Y},\,1-\alpha\%} = T_{\bar{Y}} +/- t_{\alpha/2} * ES(T_{\bar{Y}})}$$

$$ES(T_Y) = \sqrt{(1-0,22)\frac{2,11}{25}} = 0,26 \rightarrow I_{Y,\,95\%} = 2,24 +/- 2,064 * 0,26 \rightarrow P(1,7 \leq \mu \leq 2,8) = 95\%$$

Errore e numerosità campionarie 1/2

- ✓ Riprendiamo la formula che calcola gli estremi **dell'intervallo di confidenza** per il generico livello di confidenza pari a *1-α (n>30)*

$$I_{Y,\alpha} = T_Y +/- z_{\alpha/2} * ES(T_Y)$$

- ✓ Definiamo **errore campionario** la semi-ampiezza dell'intervallo di confidenza:

$$e = z_{\alpha/2} * ES(T_Y)$$

- ✓ Dal momento che il **campione è «grande»**, *(1-f) ~ 1* e dunque **media campionaria che si distribuisce come una normale**, si ottiene:

$$e = z_{\alpha/2} \frac{\sigma_y}{\sqrt{n}}$$

- ✓ Da cui si ricava:

$$\mathbf{n = \frac{z^2_{\alpha/2}\sigma^2_y}{e^2}}$$

- ✓ Si tratta di un risultato molto importante, perché **consente di calcolare l'errore atteso in funzione di vari livelli della numerosità campionaria (e dunque di costi della rilevazione), del livello di significatività tollerato e della varianza della popolazione (o di una sua stima) . La scelta finale della numerosità campionaria starà dunque nell'impresa, che bilancerà precisione dei risultati e relativi costi**

Errore e numerosità campionarie 2/2

✓ Riprendiamo l'esempio che avevamo fatto sulle giacenze di magazzino

- Il campione era n=100, dunque «grande»
- La deviazione campionaria corretta era pari a 5 giorni
- L'intervallo di confidenza era stato fissato al 95%, dunque $z_{\alpha/2} = 1{,}96$
- Vogliamo capire quanta «precisione» perdiamo nelle stime se riduciamo la numerosità campionaria
- Sappiamo che la relazione che lega numerosità campionaria ed errore campionario è la seguente:

$$n = \frac{z^2_{\alpha/2}\sigma^2_y}{e^2}$$

- Fissiamo allora vari valore di errore *e* calcoliamo le corrispondenti numerosità campionarie, da cui dipenderanno anche i relativi costi di rilevazione (nell'esempio €10 a unità campionaria):

Errore giorni di giacenza (input)	Numerosità campionaria	Errore giorni di giacenza (% della media)	Costo indagine campionaria (€)
1,50	42,7	3,3%	427
1,25	61,5	2,8%	615
1,00	96,0	2,2%	960
0,75	170,7	1,7%	1.707
0,50	384,1	1,1%	3.841
0,25	1536,6	0,6%	15.366

si veda parte pratica in excel

Tecniche di rilevazione

- ✓ **Intervista diretta** (rapporto umano permette spiegazioni, ma crea anche barriere emotive e permette arbitrarietà). L'attività dell'intervistatore è spesso supportata dalla tecnica CAPI (*Computer Assisted Personal Interview*), con la quale l'intervista stessa è compilata direttamente al computer, consentendo anche dei controlli in tempo reale (su errori/coerenza risposte)

- ✓ **Autocompilazione** dei questionari (quando ci si attende una buona collaborazione dell'intervistato, ma presenta spesso alti tassi di mancata o incompleta risposta)

- ✓ **Intervista telefonica** (economica e automatizzabile, ma bassi tassi di copertura e possibile selezione degli intervistati, che dunque possono rischiare di non rappresentare la popolazione obiettivo). In questo caso le interviste sono supportate dalla tecnica CATI. Oltre al supporto per l'intervistatore, i colloqui possono essere registrati/seguiti da un supervisore, che può intervenire e dare indicazioni

- ✓ **Indagine WEB** (molto economica e facile da somministrare, ma possibili seri effetti di selezione)

Il questionario – Formulazione delle domande

- ✓ **Domande chiuse** (costringono l'intervistato a scegliere tra modalità predefinite; facili da processare ma limitanti per l'espressione del pensiero dell'intervistato)
  - Come definirebbe la situazione finanziaria della sua famiglia?
    1. Peggiore dell'anno scorso
    2. Equivalente a quella dello scorso anno
    3. Migliore a quello dello scorso anno
- ✓ **Domande aperte** (permettono all'intervistato di esprimere compiutamente il proprio pensiero; richiedono tempi lunghi di elaborazione e digitalizzazione)
  - Come definirebbe la situazione finanziaria della sua famiglia?

  .....
- ✓ **Domande filtro** (domande chiuse che hanno la funzione di indirizzare l'intervistato verso una o l'altra sezione successiva del questionario)
  - La condizione economica della sua famiglia è peggiorata rispetto allo scorso anno?
    - ❑ SI'
    - ❑ NO

Il questionario – Misurazione delle risposte

- ✓ La **misurazione delle modalità di risposta è un aspetto cruciale dell'indagine**. In **alcuni casi** le **risposte** sono già espresse **in unità di misura** (kg, anni, metri**,..). Per percezioni, atteggiamenti, opinioni e caratteri qualitativi (maschio/femmina) la misurazione può avvenire con l'ausilio di scale *ad hoc***

  - **La scala nominale** si limita ad attribuire un codice a caratteri qualitativi non ordinabili (maschio/femmina, castano/biondo, …) senza che si possa alcun ordinamento
  - **La scala ordinale** attribuisce dei codici che sono tra loro ordinabili (es. buono>cattivo, Ferrari>Fiat, …), ma non si possono calcolare differenze, distanze e valori medi. Si possono però calcolare indici di posizione (mediana e quartili)
  - **La scala a intervallo** permette di definire sia l'ordine, sia una distanza tra modalità diverse, come ad esempio in *1=molto negativo, 2=negativo, 3=neutro, 4=positivo e 5=molto positivo*

Il questionario – Valutazione dei risultati

- ✓ L'indagine perfetta sarebbe affetta esclusivamente da errore campionario, dipendente dalla natura esaustiva dell'indagine stessa
- ✓ Imperfezioni nel processo danno luogo ad errori di diverso tipo, detti errori non campionari. Questi vengono distinti in:
  - **Errori di copertura:** causati da difetti della lista di campionamento
  - **Errori da mancata risposta:** dovuti all'impossibilità di osservare parte delle unità campionarie
  - **Errori di misurazione:** domande non chiara o volontà dell'intervistato di non dare risposte veritieri

Statistica per l'impresa

3. Interpretazione e comparazione dei dati

RAPPORTI
- ✓ Differenze assolute e relative
- ✓ Rapporti di composizione
- ✓ Rapporti di coesistenza
- ✓ Rapporti di densità e derivazione
- ✓ Aggregazione e scomposizioni dei rapporti

NUMERI INDICI
- ✓ Base fissa e Base mobile, cambio di base e proprietà
- ✓ Numeri indici sintetici dei prezzi: indici di Laspeyres, Paasche e Fisher

TASSI DI VARIAZIONE
- ✓ Tassi medi
- ✓ Variazioni congiunturali e tendenziali
- ✓ Variazioni nominali e reali

ANALISI DELLA MOBILITA'
- ✓ Tassi di entrata e di uscita
- ✓ Rapporti di rinnovo e durata e turnover
- ✓ Matrici di transizione

Differenze assolute e relative

✓ Obiettivo di questi indicatori è **evidenziare nel tempo o nello spazio le differenze di intensità** (es. il fatturato) **o frequenza** (es. il numero di pezzi venduti) di un certo fenomeno *X*. Sia t il tempo e x_t le osservazioni sul fenomeno *X*

- **Variazione (differenza) assoluta** sarà:

$$d_a = x_t - x_{(t-1)}$$

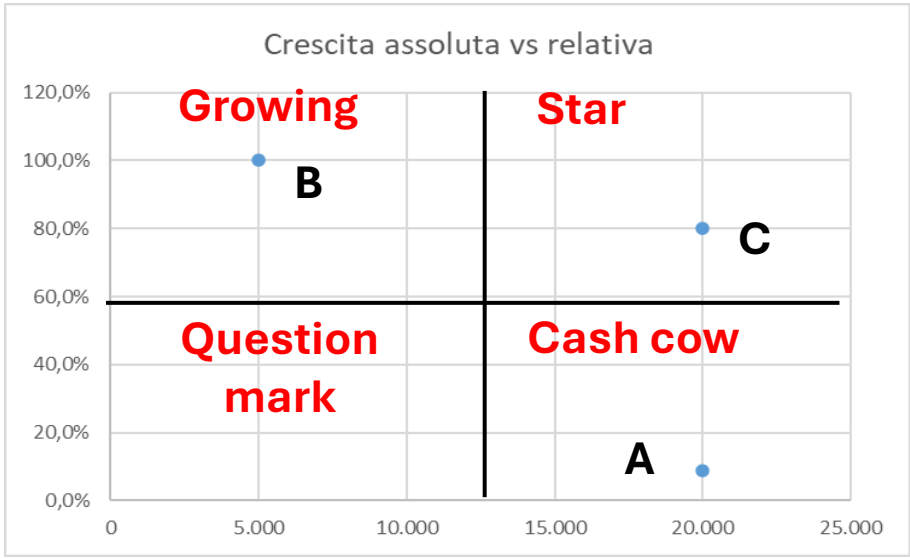
- La variazione **assoluta** è espressa **nell'unità di misura «originaria»**

- **Variazione (differenza) relativa** sarà:

$$d_r = \frac{x_t - x_{(t-1)}}{x_{(t-1)}} = \frac{x_t}{x_{(t-1)}} - 1$$

- La variazione **relativa** è **numero puro**

in €	2023	2022	Var. assoluta	Var. relative
Fatturato prodotto A	250.000	230.000	20.000	8,7%
Fatturato prodotto B	10.000	5.000	5.000	100,0%
Fatturato prodotto C	45.000	25.000	20.000	80,0%
Fatturato Totale	**305.000**	**260.000**	**45.000**	**17,3%**



si veda parte pratica in excel

Tabelle a doppia entrata e Rapporti di composizione

- ✓ Utile per la rappresentazione dei fenomeni le **tabelle a doppia entrata**, che rappresentano la **distribuzione congiunta di due variabili**
- ✓ Per esempio pongo in una tabella a doppia entrata la distribuzione dei dipendenti per classe di età e qualifica. Le **somme** di **riga e di colonna** rappresentano le **distribuzioni marginali** di ciascuna caratteristica

Qualifica/Età	15-35	36-55	56-70	Totale
Dirigenti	2	10	11	**23**
Quadri	10	28	20	**58**
Impiegati	101	177	78	**356**
Operai	98	180	34	**312**
Totale	**211**	**395**	**143**	**749**

- ✓ Ottenuta la tabella posso essere interessato a sintetizzarne le caratteristiche attraverso **rapporti di composizione**

A)
Voglio capire come di distribuiscono per età

Qualifica/Età	15-35	36-55	56-70	Totale
Dirigenti	0,3%	1,3%	1,5%	**3,1%**
Quadri	1,3%	3,7%	2,7%	**7,7%**
Impiegati	13,5%	23,6%	10,4%	**47,5%**
Operai	13,1%	24,0%	4,5%	**41,7%**
Totale	**28,2%**	**52,7%**	**19,1%**	**100,0%**

B)
Voglio capire la distribuzione per età per qualifica professionale

Qualifica/Età	15-35	36-55	56-70	Totale
Dirigenti	8,7%	43,5%	47,8%	**100,0%**
Quadri	17,2%	48,3%	34,5%	**100,0%**
Impiegati	28,4%	49,7%	21,9%	**100,0%**
Operai	31,4%	57,7%	10,9%	**100,0%**
Totale	**28,2%**	**52,7%**	**19,1%**	**100,0%**

C)
Voglio capire la distribuzione per qualifica professionale nelle diverse classi di età

Qualifica/Età	15-35	36-55	56-70	Totale
Dirigenti	0,9%	2,5%	7,7%	**3,1%**
Quadri	4,7%	7,1%	14,0%	**7,7%**
Impiegati	47,9%	44,8%	54,5%	**47,5%**
Operai	46,4%	45,6%	23,8%	**41,7%**
Totale	**100,0%**	**100,0%**	**100,0%**	**100,0%**

Rapporti di composizione

- ✓ Per la generica cella di una tabella a doppia entrata di riga *i* e colonna *j* posso calcolare i rapporti di composizione come calcolati nel precedente caso A)

$$c_{ij} = \frac{x_{ij}}{\sum i \sum j x_{ij}}$$

- ✓ I rapporti di composizione che compaiono nell'ultima riga e nell'ultima colonna (colonne se avessi più di una distribuzione) sono i seguenti (si veda rispettivamente l'ultima riga e l'ultima colonna del precedente caso A)

$$c_{i.} = \frac{\sum j x_{ij}}{\sum i \sum j x_{ij}} = \frac{x_{i.}}{x_{..}} \qquad c_{.j} = \frac{\sum i x_{ij}}{\sum i \sum j x_{ij}} = \frac{x_{.j}}{x_{..}}$$

- ✓ In una distribuzione doppia possiamo calcolare i rapporti di composizione condizionati ad una particolare riga (valore di *i* – vedi caso B) o di colonna (valore di *j* – vedi caso C)

$$c_{jli\bullet} = \frac{x_{ij}}{\sum j x_{ij}} = \frac{x_{ij}}{x_{i.}} \qquad c_{ilj\bullet} = \frac{x_{ij}}{\sum i x_{ij}} = \frac{x_{ij}}{x_{.j}}$$

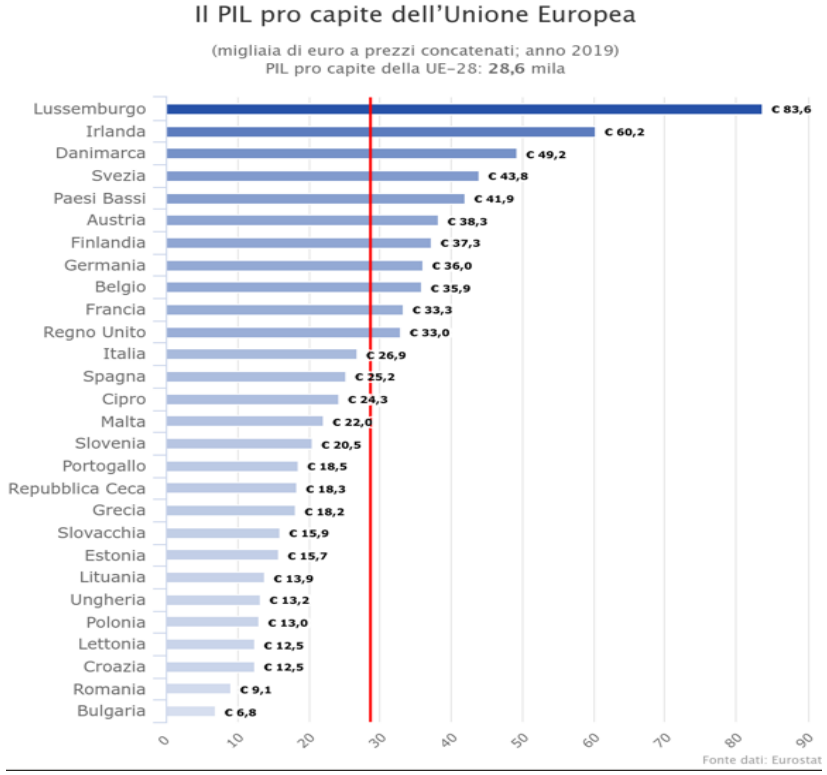
si veda parte pratica in excel

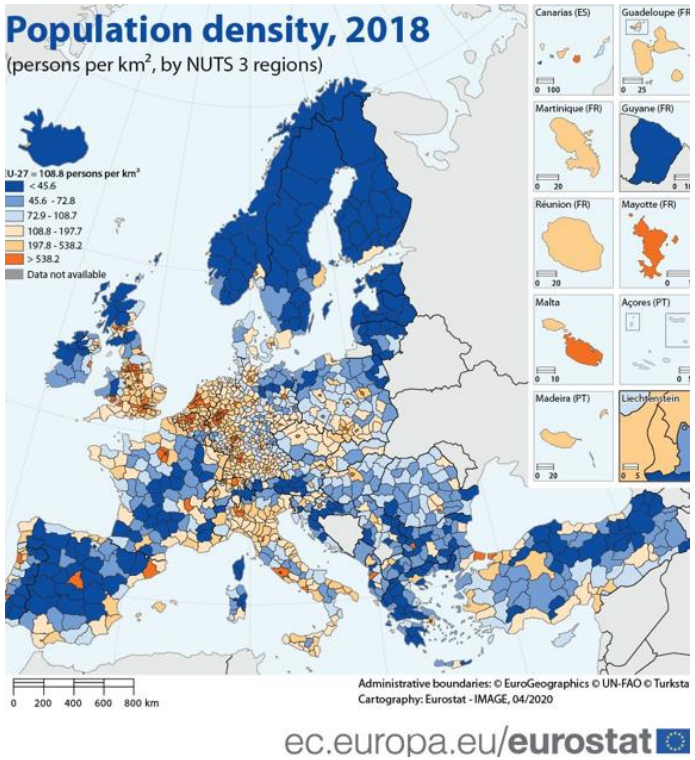
Rapporti di coesistenza

- ✓ I **rapporti di coesistenza** servono **confrontare due grandezze «collegate»** evidenziando eventuali squilibri. Per esempio:
  - Il rapporto tra importazioni ed esportazioni indica un'evoluzione equilibrata se non si discosta troppo da *1*
  - Il *leverage ratio* (rapporto tra ammontare dei debiti ed capitale proprio) confronta il livello di indebitamente di un'impresa, evidenziandone la capacità o meno di fare fronte agli impegni che maturano nel corso del tempo

Rapporti di densità

- ✓ **Rapporti di densità**: quando si vogliono **confrontare caratteristiche** che appartengono a **popolazioni di «dimensioni» diverse**, per standardizzare e rendere comparabili i dati si può calcolare questa tipologia di rapporto. Si tratta per esempio dei rapporti cosiddetti pro capite (prodotto interno lordo diviso la popolazione)





Rapporti di derivazione 1/2

- ✓ Quando i **dati** che si vogliono analizzare sono il **risultato di un fenomeno che ne è il presupposto**, allora si parla di **rapporti di derivazione**. Viene rapportato un dato di flusso al nominatore, con un dato di consistenza (*stock*) al denominatore. I quozienti demografici sono gli esempi classici di derivazione. Per esempio il numero delle nascite rapportate al numero di donne di età compresa tra 20 e 45 anni. In ambito aziendale potrebbero essere le imprese fallite sul totale delle imprese attive (o la media tra le imprese attive negli ultimi due periodi)
- ✓ Vediamo un esempio: raccolgo il numero di infortuni e di operai presenti al momento dell'infortunio, suddivisi per classe di ore lavorate e reparto. Guardando solo gli infortuni risulta che il reparto *B* è più «pericoloso», ma se costruisco i rapporti di derivazione (infortuni/numero operai presenti) mi accorgo che gli infortuni sono più alti per la classi di ore lavorate maggiori a 41, che hanno interessato maggiormente nel periodo il Reparto *B*

(1)

Numero di infortuni	Reparto		
Classe di ore lavorate a settimana	A	B	Totale
fino a 35	5	7	**12**
35-40	7	5	**12**
41 e oltre	6	11	**17**
Totale	**18**	**23**	**41**

(2)

Numero di operai presenti	Reparto		
Classe di ore lavorate a settimana	A	B	Totale
fino a 35	200	300	**500**
35-40	600	400	**1000**
41 e oltre	100	200	**300**
Totale	**900**	**900**	**1800**

(3)

Infortuni/Numero operai	Reparto		
Classe di ore lavorate a settimana	A	B	Totale
fino a 35	2,50%	2,33%	**2,40%**
35-40	1,17%	1,25%	**1,20%**
41 e oltre	6,00%	5,50%	**5,67%**
Totale	**2,00%**	**2,56%**	**2,28%**

si veda parte pratica in excel

Rapporti di derivazione 2/2

- ✓ Generalizziamo quanto abbiamo visto a partire dalle distribuzioni doppie, dove le *x* equivalgono al numero degli operai presenti e le *y* agli infortuni nella generica riga *i=1,..,n* e colonna *j=1, ...,m*

$$q_{ij} = \frac{y_{ij}}{x_{ij}}$$ ➡ Corrispondono alle celle centrali (quelle per reparto nella tabella 3 della slide precedente). Si dicono anche rapporti specifici, rispetto agli altri che sono generici perché fanno riferimento a somme o valori medi

$$q_{i.} = \frac{\sum_{j=1}^{m} y_{ij}}{\sum_{j=1}^{m} x_{ij}}$$ ➡ Corrisponde al totale per riga della tabella 3 vista in precedenza

$$q_{.j} = \frac{\sum_{i=1}^{n} y_{ij}}{\sum_{i=1}^{n} x_{ij}}$$ ➡ Corrisponde al totale per colonna della tabella 3 vista in precedenza

$$q_{..} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}}$$ ➡ Corrisponde al totale per riga e colonna (la cella in basso a dx nella tabella 3 vista in precedenza)

Rapporti generici: aggregazione e scomposizione

- ✓ I rapporti di composizione sono facilmente aggregabili o scomponibili per somma/sottrazione (il denominatore è lo stesso)

- ✓ I rapporti generici di densità e di derivazione, che si ottengono come somma dei rapporti specifici, non sono facilmente interpretabili e scomponibili. Possono essere interpretabili come media ponderata dei rapporti specifici

  - Infatti il generico rapporto specifico

$$q_{ij} = \frac{y_{ij}}{x_{ij}} \quad \text{può essere riscritto come } y_{ij} = q_{ij}\, x_{ij}$$

  - Allora un rapporto generico

$$q_{.j} = \frac{\sum_{i=1}^{n} y_{ij}}{\sum_{i=1}^{n} x_{ij}} = \frac{\sum_{i=1}^{n} q_{ij}\, x_{ij}}{\sum_{i=1}^{n} x_{ij}}$$

che non è altro che la media ponderata dei rapporti specifici sulla determinata colonna j (o riga i), dove la ponderazione è la struttura della popolazione di riferimento

Confronto tra rapporti generici 1/3

- ✓ Come abbiamo visto confrontare due rapporti generici non è immediato. Differenze tra i due rapporti possono dipendere dal numeratore, così come dal denominatore. Nell'esempio degli infortuni su lavoro per esempio vedo che il reparto «più pericoloso» è il *B*, perché presenta un rapporto generico di infortuni su ore lavorate pari a 2,56%, a fronte del 2% del reparto *A*. Ma è proprio così? Perché succede?
- ✓ Utilizzo allora i **metodi di scomposizione della popolazione tipo e dei quozienti tipo**. Che significa? Nello sostanza stiamo per fare una sorta di «standardizzazione». Vediamolo nell'esempio.

  - Calcoliamo i rapporti generici dei due reparti, come media ponderata, usando però i medesimi pesi delle ore lavorate nel reparto *A*, che ho deciso essere la popolazione tipo:

Infortuni/Numero operai	Reparto		
Classe di ore lavorate a settimana	A	B	Totale
fino a 35	2,50%	2,33%	**2,40%**
35-40	1,17%	1,25%	**1,20%**
41 e oltre	6,00%	5,50%	**5,67%**
Totale	**2,00%**	**2,56%**	**2,28%**

Numero di operai presenti	Reparto		
Classe di ore lavorate a settimana	A	B	Totale
fino a 35	200	300	**500**
35-40	600	400	**1000**
41 e oltre	100	200	**300**
Totale	**900**	**900**	**1800**

$$q^{p*}_{.1} = \frac{0{,}025*200+0{,}0117*600+0{,}06*100}{200+600+100} = 2{,}00\% \qquad q^{p*}_{.2} = \frac{0{,}0233*200+0{,}0125*600+0{,}055*100}{200+600+100} = 1{,}96\%$$

Confronto tra rapporti generici 2/3

- Calcoliamo i rapporti generici dei due reparti, come media ponderata prendendo i quoziente tipo (ovvero i rapporti specifici della prima colonna), pesati per le diverse popolazioni:

Infortuni/Numero operai	Reparto		
Classe di ore lavorate a settimana	A	B	Totale
fino a 35	2,50%	2,33%	**2,40%**
35-40	1,17%	1,25%	**1,20%**
41 e oltre	6,00%	5,50%	**5,67%**
Totale	**2,00%**	**2,56%**	**2,28%**

Numero di operai presenti	Reparto		
Classe di ore lavorate a settimana	A	B	Totale
fino a 35	200	300	**500**
35-40	600	400	**1000**
41 e oltre	100	200	**300**
Totale	**900**	**900**	**1800**

$$q^{q^*}_{.1} = \frac{0{,}025*200+0{,}0117*600+0{,}06*100}{200+600+100} = 2{,}00\%$$

$$q^{q^*}_{.2} = \frac{0{,}025*300+0{,}0117*400+0{,}06*200}{300+400+200} = 2{,}69\%$$

- Il fatto che i due rapporti calcolati con il metodo della popolazione tipo sono molto simili, mentre quelli calcolati con il metodo dei quozienti tipo sono piuttosto diversi, mi spingono a dire che il motivo del diverso livello di infortuni nei due reparti dipende dalla distribuzione dei lavoratori per classi di ore lavorate

Confronto tra rapporti generici 3/3

- ✓ Esprimiamo quanto abbiamo visto in modo più generale
- ✓ Il **metodo della popolazione tipo** applica ai **rapporti specifici, la struttura dei pesi** presa da una **popolazione standard x^*** (di seguito la rappresentazione generale nel caso dei totali rispettivamente per riga e per colonna)

$$q^{p*}_{i.} = \frac{\sum_{j=1}^{m} q_{ij}\, x^*_{ij}}{\sum_{j=1}^{m} x^*_{ij}} \qquad q^{p*}_{.j} = \frac{\sum_{i=1}^{n} q_{ij}\, x^*_{ij}}{\sum_{i=1}^{n} x^*_{ij}}$$

- ✓ Il **metodo dei quozienti tipo** assume invece un insieme di **rapporti specifici standard q^*** (di seguito la rappresentazione generale nel caso dei totali rispettivamente per riga e per colonna)

$$q^{*}_{i.} = \frac{\sum_{j=1}^{m} q^*_{ij}\, x_{ij}}{\sum_{j=1}^{m} x_{ij}} \qquad q^{*}_{.j} = \frac{\sum_{i=1}^{n} q^*_{ij}\, x_{ij}}{\sum_{i=1}^{n} x_{ij}}$$

- ✓ I risultati di una scomposizione basata su popolazione o quozienti tipo dipenderanno dalla bontà delle ipotesi di partenza

Numeri indici

- ✓ I **numeri indici** sono un rapporto statistico che serve a misurare le variazioni relative nel tempo o nello spazio. Ci occuperemo solo del caso del tempo perché è quello di uso più comune

- ✓ Dato un fenomeno *X* e due osservazioni x_0 e x_1 in due stadi successivi:
  - Sia *0* la situazione di base
  - Sia *1* la situazione corrente
  - Definiamo numero indice di base *0*, al tempo *1* il seguente rapporto:

$$ {}_0I_1 = \frac{x_1}{x_0} $$

  - Esso esprime di quanto è variate l'intensità del fenomeno tra *0* e *1*. A partire dal numero indice si può calcolare anche la variazione relativa:

$$ \frac{x_1 - x_0}{x_0} = {}_0I_1 - 1 $$

Base fissa e base mobile

✓ Consideriamo una sequenza di osservazioni

$$x_0, x_1, x_2, \ldots\ldots, x_n$$

✓ Rapportando tutte ad una di esse, per esempio x_0, si ottiene la serie dei numeri indici a base fissa

$$_0I_0 = \frac{x_0}{x_0},\ _0I_1 = \frac{x_1}{x_0},\ _0I_2 = \frac{x_2}{x_0},\ \ldots\ldots,\ _0I_n = \frac{x_n}{x_0}$$

✓ Rapportando invece ciascuna osservazione alla precedente, si ottiene la serie a base mobile

$$_0I_1 = \frac{x_1}{x_0},\ _1I_2 = \frac{x_2}{x_1},\ _2I_3 = \frac{x_3}{x_2},\ \ldots\ldots,\ _{n-1}I_n = \frac{x_n}{x_{n-1}}$$

✓ Vediamo un esempio

Numeri Indice dei prezzi

Anni	Tempo	Prezzi	Numeri indice a base fissa, 2018=100	Numeri indice a base mobile t-1=100
2018	0	1,65	100,0%	-
2019	1	1,68	101,8%	101,8%
2020	2	1,55	93,9%	92,3%
2021	3	1,65	100,0%	106,5%
2022	4	2,10	127,3%	127,3%

si veda parte pratica in excel

Proprietà dei numeri indici

✓ **Proprietà** dei **numeri indici** elementari

- **Identità:** $_tI_t = \frac{x_t}{x_t} = 1$
- **Reversibilità delle basi:** $_tI_s = \frac{1}{_sI_t} = \frac{1}{\frac{x_t}{x_s}} = \frac{x_s}{x_t}$
- **Transitività delle basi** (circolarità): $_tI_r = {}_tI_q * {}_qI_r = \frac{x_q}{x_t} * \frac{x_r}{x_q} = \frac{x_r}{x_t}$
- **Commensurabilità**: il numero indice non varia se varia l'unità di misura impiegata per esprimere il fenomeno
- **Scomposizione delle cause**: se un fenomeno può essere ottenuto come prodotto di due «cause», il relativo numero indice è anch'esso rappresentabile come prodotto dei numeri indici delle cause. Pensiamo per esempio al valore di una merce, espresso come prodotto tra prezzo e quantità in un certo istante

$$v_t = p_t * q_t$$

Il numero indice del valore può essere scomposto nei numeri indici dei prezzi e delle quantità

$$\frac{v_t}{v_s} = \frac{p_t}{p_s} * \frac{q_t}{q_s}$$

Cambiamento di base

- ✓ Il **cambio di base** (per esempio da *h* a *k)* per un indice in base fissa avviene dividendo ogni termine della ${}_hI_t$ della serie per l'indice del periodo «nuova base»: ${}_hI_k$. Infatti:
- ✓ Vediamo un esempio: supponiamo di voler cambiare la base da *t=0* (2018=100) a *t=2* (2020=100)

Numeri Indice dei prezzi				
Anni	**Tempo**	**Prezzi**	**Numeri indice a base fissa, 2018=100**	**Numeri indice a base fissa, 2020=100**
2018	0	1,65	100,0%	106,5%
2019	1	1,68	101,8%	108,4%
2020	2	1,55	93,9%	100,0%
2021	3	1,65	100,0%	106,5%
2022	4	2,10	127,3%	135,5%

- Per le proprietà di reversibilità e di transitività delle basi:

$${}_2I_0 = \frac{1}{{}_0I_2} = \frac{1}{0{,}939} = 1{,}065$$

$${}_2I_1 = {}_2I_0 * {}_0I_1 = \frac{{}_0I_1}{{}_0I_2} = \frac{1{,}018}{0{,}939} = 1{,}084$$

$${}_2I_3 = {}_2I_0 * {}_0I_3 = \frac{{}_0I_3}{{}_0I_2} = \frac{1{,}000}{0{,}939} = 1{,}065$$

$${}_2I_3 = {}_2I_0 * {}_0I_4 = \frac{{}_0I_4}{{}_0I_2} = \frac{1{,}273}{0{,}939} = 1{,}355$$

- Come si vede nell'esempio basta dividere ogni elemento dell'indice nella **vecchia base** per il **valore che lo stesso assume in corrispondenza della nuova base**, chiamato **coefficiente di conversione**

Da base fissa a base mobile

- ✓ Il passaggio **da base fissa a base mobile avviene dividendo ogni termine** (tranne il primo che non è definito) per **l'indice in base fissa relativo al periodo precedente**

$$ {}_{t-1}I_t = \frac{{}_kI_t}{{}_kI_{t-1}} $$

- ✓ Il passaggio **da base mobile ad un base fissa *k*** per un qualsiasi indice ${}_{t-1}I_t$ in un indice a base fissa ${}_0I_t$, *(dove 0 è il primo termine della serie considerata)* avviene **moltiplicando l'indice per tutti i precedenti indici a base mobile da *k* a *t* .** A partire da questa serie possiamo poi passare a serie espresse in una base diversa da 0 dividendo per il coefficiente di conversione

Numeri Indice dei prezzi

Anni	Tempo	Prezzi	Numeri indice a base fissa, 2018=100	Numeri indice a base fissa, 2020=100	Numeri indice a base mobile
2018	0	1,65	100,0%	106,5%	-
2019	1	1,68	101,8%	108,4%	101,8%
2020	2	1,55	93,9%	100,0%	92,3%
2021	3	1,65	100,0%	106,5%	106,5%
2022	4	2,10	127,3%	135,5%	127,3%

Da base fissa a mobile
100,0%/108,4%

Da base mobile a fissa
101,8% * 92,3%

Tassi medi di variazione 1/2

- ✓ Quando confrontiamo un fenomeno nel tempo può essere utile misurare la sua variazione complessiva nel periodo considerato o quella media, se l'intervallo temporale copre più unità di tempo

- ✓ **Data una serie storica** $x_0, x_1, x_2, ..., x_n$:

  - La **variazione relativa complessiva** nell'arco temporale i_c sarà quella che abbiamo già visto con riferimento al calcolo delle variazioni in generale (*slide 46*)

$$i_c = \frac{x_t - x_0}{x_0} = \frac{x_t}{x_0} - 1$$

  - La variazione relativa media annua dipenderà da come si assume funzioni il meccanismo della «capitalizzazione» delle variazioni annue (seguiamo l'approccio della matematica finanziaria):

    - **Tasso medio semplice**:

$$x_t = x_0\ (1+it)$$

$$x_t = x_0 + x_0\, it$$

$$x_0\, it = x_t - x_0$$

$$i = \frac{1}{t}\ \frac{x_t - x_0}{x_0}$$

- Tasso **medio composto** (il più usato, specie quando si usano grandezze monetarie):

$$x_t = x_0\ (1+i')^t$$

$$x^{1/t}_t = x^{1/t}_0\ (1+i')$$

$$x^{1/t}_t = x^{1/t}_0\ + i'x^{1/t}_0$$

$$i'\ x^{1/t}_0 = x^{1/t}_t - x^{1/t}_0$$

$$i'\ x^{1/t}_0 = x^{1/t}_t - x^{1/t}_0$$

$$i' = \frac{x^{1/t}}{x^{1/t}_0} - 1$$

$$\mathbf{i' = \left(\frac{x_t}{x_0}\right)^{1/t} - 1}$$

CAGR (Compound Annual Growth Rate)

si veda parte pratica in excel

Numeri indici sintetici

- ✓ Un **indice sintetico** sintetizza la **variazione di un aggregato** anziché un valore elementare
- ✓ Quando per esempio si calcola l'indice dei prezzi per il calcolo dell'inflazione per le famiglie, l'aggregato in questione è il paniere medio di consumo delle famiglie stesse. Di questo ci interessa valutare la variazione del prezzo (non del valore che è prezzo *x* quantità), mantenendo fisse le quantità
- ✓ Si **calcola** pertanto **l'indice** (il rapporto) tra il **valore del paniere ai prezzi in *t*** e quello dello **stesso paniere in *0***

$$\frac{v_t}{v_0} = \frac{\sum_i p_{it} q_i}{\sum_i p_{i0} q_i}$$

- ✓ q_i è dunque il paniere fisso di riferimento. Cosa consideriamo?
  - q_{i0} cioè la quantità del bene *i* al tempo *0*?
  - q_{it} cioè la quantità del bene *i* al tempo *t*?

Numeri indice sintetici dei prezzi di Laspeyres e di Paasche

✓ A seconda della scelta del sistema di pesi (cioè delle quantità viste nella slide precedente) si ottengono:

- L'**Indice di Laspeyres** se le **quantità** vengono fissate al **tempo base**

$$ {}_0^{p}I_t^{L} = \frac{\sum_i p_{it} q_{i0}}{\sum_i p_{i0} q_{i0}} $$

- L'**Indice di Paasche** se le **quantità** vengono fissate al **tempo corrente *t***

$$ {}_0^{p}I_t^{P} = \frac{\sum_i p_{it} q_{it}}{\sum_i p_{i0} q_{it}} $$

si veda parte pratica in excel

Gli indici di Laspeyres e di Paasche come medie pesate

- ✓ Gli indici di Laspeyres e Paasche possono essere ottenuti anche come medie degli indici semplici dei prezzi di ogni bene pesati per l'importanza della spesa dedicata al bene nella spesa complessiva:

- ✓ **Indice di Laspeyres:**

Peso

$$ {}_0^{p}I_t^{L} = \frac{\sum_i \frac{p_{it}}{p_{i0}} p_{i0} q_{i0}}{\sum_i p_{i0} q_{i0}} = \frac{\sum_i p_{it} q_{i0}}{\sum_i p_{i0} q_{i0}} $$

- ✓ **Indice di Paasche:**

Peso

$$ {}_0^{p}I_t^{P} = \frac{\sum_i \frac{p_{it}}{p_{i0}} p_{i0} q_{it}}{\sum_i p_{i0} q_{it}} = \frac{\sum_i p_{it} q_{it}}{\sum_i p_{i0} q_{it}} $$

Le proprietà degli indici sintetici e l'indice di Fisher

- ✓ Anche per gli indici sintetici sono desiderabili le proprietà viste per gli indici semplici viste alla *slide 77*. Per esempio la transitività delle basi ci serve per fare cambi di base ad una serie di indici

- ✓ Per gli indici sintetici, alle proprietà appena citate, si aggiungono anche:
  - **Proporzionalità**: se i prezzi dei *k* prodotti variano di un fattore *a*, anche l'indice deve variare in proporzione
  - **Determinatezza**: l'indice sintetico non deve tendere a infinito o diventare indeterminato se si annulla un termine compreso nella formula

- ✓ Gli indici di Paasche e Laspeyres non soddisfano la proprietà della transitività e della scomposizione delle cause

- ✓ Le proprietà sono invece soddisfatte **dall'indice sintetico di Fisher**, pari alla **media geometrica**[1] degli indici di **Paasche e Laspeyres** e che dunque rappresenta il numero indice ideale

$$ {}_0^p I^F_t = \sqrt{{}_0^p I^L_t \; {}_0^p I^P_t} $$

[1] La media geometrica della serie $x_1, x_2, \ldots, x_n$ è pari a $\sqrt[n]{x_1 \, x_2 \, \ldots xn}$

Numeri indici sintetici valore

- ✓ Alla *slide 82* avevamo parlato degli indici sintetici per calcolare la variazione dei prezzi di un paniere di beni in un intervallo di tempo, tenendo fermo le quantità dei singoli beni.

$$\frac{v_t}{v_s} = \frac{\sum_i p_{it} q_i}{\sum_i p_{i0} q_i}$$

- ✓ **Indici sintetici** possono peraltro catturare anche la **variazione di valore**, inteso come **prodotto tra prezzi e quantità** nei due differenti momenti di osservazione. Sarà allora:

$$ {}_0^{v}I_t = \frac{\sum_i p_{it} q_{it}}{\sum_i p_{i0} q_{i0}} $$

- ✓ Potrei essere interessato anche ad un **indice sintetico** che rappresenta la **variazione delle quantità** in un periodo. In questo caso variano le quantità, mentre **rimane fisso il riferimento dei prezzi**:

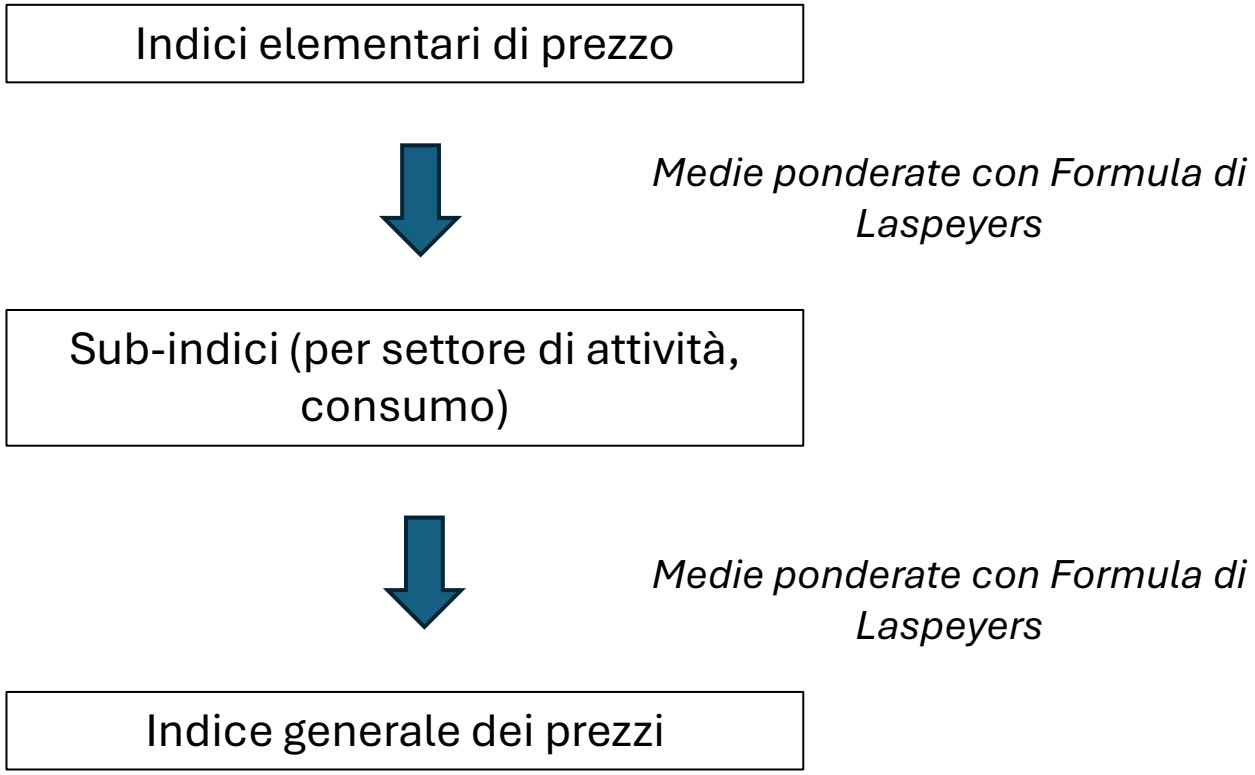
$$ {}_0^{q}I_t = \frac{\sum_i p_{i0} q_{it}}{\sum_i p_{i0} q_{i0}} $$

Alcuni numeri indici rilevanti

- ✓ L'ISTAT pubblica numerosi numeri indici, che ripercorrono le categorie descritte nelle pagine precedenti
  - **Indice di valore** (visti alla *slide 86*)
    - Fatturato e ordinativi dell'industria
    - Fatturato dei servizi
    - Valore delle vendite del commercio
  - **Indice dei prezzi** (visti alle slides *83-85*)
    - Prezzi al consumo (formula di Laspeyres)
      - NIC: indice nazionale dei prezzi al consumo per l'interna collettività. L'indice con cui si calcola l'inflazione, che non è altro che la variazione relativa dell'indice rispetto al periodo precedente
      - FOI: indice dei prezzi al consumo per le famiglie con capofamiglia operaio o impiegato (è un sottoinsieme del NIC)
      - IPCA: indice armonizzato dei prezzi per i paesi dell'Unione Europea, in modo che l'Eurostat possa calcolare un indicatore europeo. Stessa popolazione del NIC, ma esclude lotterie, lotto e concorsi pronostici. Inoltre considera i prezzi effettivamente pagati, tenendo in considerazione gli sconti, mentre NIC e FOI fanno riferimento ai prezzi pieni di vendita
  - **Indici delle quantità** (visti alla *slide 86*)
    - Produzione Industriale
    - Volume dell'export e dell'import

Numeri indici sintetici: scomposizione 1/3

✓ Può essere utile **scomporre gli indici sintetici in sub-indici**. L'indice generale può essere ottenuto anche media ponderata dei sub-indici



Numeri indici sintetici: scomposizione 2/3

- ✓ Formalmente per il generico gruppo *g* contenente i prodotti *i=1, …., S e dato* v_{i0} che rappresenta la spesa dell'*i-esimo* prodotto nel gruppo *g,* il relativo sub-indice dei prezzi sarà:

$$ {}_0I^g_t = \frac{\sum_{i=1}^{S} \frac{p_{it}}{p_{i0}} v_{i0}}{\sum_{i=1}^{S} v_{i0}} = \sum_{i=1}^{S} \frac{p_{it}}{p_{i0}} w_{i0} \quad \text{con} \quad w_{i0} = \frac{v_{i0}}{\sum_{i=1}^{S} v_{i0}} $$

→ Peso del generico prodotto *i* all'interno del sottogruppo *g*

- ✓ L'indice generale sarà la somma pesata delle variazioni dovute ad ogni singolo gruppo (*g=1,…, k*), ovvero:

$$ {}_0I^G_t = \sum_{g=1}^{k} {}_0I^g_t \, w_{g0} \quad \text{con} \quad w_{g0} = \frac{v_{g0}}{\sum_{g=1}^{k} v_{g0}} $$

→ Peso del generico sub-indice all'interno dell'indice generale

- ✓ Il contributo del singolo gruppo *g* alla dinamica dell'indice generale è dato dall'indice di Gruppo moltiplicato il suo peso sul totale:

$$ C_g = {}_0I^g_t \, w_{g0} $$

Numeri indici sintetici: scomposizione 3/3

- ✓ L'ISTAT mensilmente pubblica i dati dell'inflazione evidenziando il contributo ai sub-indici relativi a principali gruppi merceologici che compongono l'indice dei generali dei prezzi

PROSPETTO 2. INDICI DEI PREZZI AL CONSUMO NIC PER DIVISIONE DI SPESA,
Luglio 2024, pesi, variazioni percentuali congiunturali e tendenziali (base 2015=100) e contributi alla variazione tend. dell'indice generale

DIVISIONI DI SPESA	Pesi	Variazioni congiunturali		Variazioni tendenziali		Contributo alla variazione tendenziale dell'indice generale	Inflazione acquisita a luglio
		lug-24 / giu-24	lug-23 / giu-23	lug-24 / lug-23	giu-24 / giu-23		
Prodotti alimentari e bevande analcoliche	171.945	-0,5	0,0	+0,9	+1,4	0,163	+1,9
Bevande alcoliche e tabacchi	29.033	0,0	-0,3	+2,4	+2,2	0,070	+2,3
Abbigliamento e calzature	59.553	0,0	-0,1	+1,1	+1,0	0,066	+1,1
Abitazione, acqua, elettricità e combustibili	112.550	+2,9	-1,4	-2,2	-6,2	-0,244	-5,9
Mobili, articoli e servizi per la casa	69.621	-0,1	+0,2	+0,3	+0,5	0,016	+0,8
Servizi sanitari e spese per la salute	82.746	0,0	+0,1	+1,5	+1,6	0,133	+1,4
Trasporti	147.401	+0,5	+0,5	+1,5	+1,6	0,225	+1,2
Comunicazioni	21.835	-0,4	-0,9	-5,2	-5,6	-0,115	-5,2
Ricreazione, spettacoli e cultura	81.071	+0,9	+0,2	+2,0	+1,2	0,156	+1,6
Istruzione	8.932	0,0	0,0	+1,8	+1,8	0,017	+1,3
Servizi ricettivi e di ristorazione	117.950	+0,5	+0,4	+4,3	+4,2	0,529	+4,5
Altri beni e servizi	97.363	+0,1	+0,3	+2,3	+2,6	0,227	+2,4
Indice generale	**1.000.000**	**+0,4**	**0,0**	**+1,3**	**+0,8**		**+1,0**

Variazioni tendenziali e congiunturali

«……la crescita congiunturale del Pil diffusa il 30 aprile 2024
era stata anch'essa dello 0,3%, mentre quella tendenziale era stata dello 0,6%……

Dal sito dell'ISTAT

- ✓ Preso un fenomeno misurato con cadenza infrannuale, tale per cui nell'anno ci sono k periodi (per esempio per i dati trimestrali *k=4*, mensili *k=12*, …) si parla di variazione:
  - **Congiunturale** quando si rapporta il **dato corrente x_t al dato precedente x_{t-1}**
  - **Tendenziale** quando si rapporta **il dato corrente x_t al dato corrispondente dell'anno precedente**. Nel caso di dati trimestrali ci confrontiamo con x_{t-4}
- ✓ Per esempio, le immatricolazioni di nuove autovetture in Italia sono state pari a 139.509 a maggio 2024. La variazione congiunturale (sul mese precedente) è stata del +3,1%, mentre quella tendenziale (vs maggio 2023) è stata del -6,6%: stiamo andando peggio dello scorso anno, ma nell'ultimo mese si è registrata una ripresa

ANFIA Associazione Nazionale Filiera Industria Automobilistica

	2023-05	2023-06	2023-07	2022-08	2023-09	2023-10	2023-11	2023-12	2024-01	2024-02	2024-03	2024-04	2024-05
Autovetture	149.308	139.031	119.120	79.725	136.227	139.040	139.211	111.147	141.909	147.071	161.923	135.339	139.509

si veda parte pratica in excel

Variazioni nominali e reali 1/4

- ✓ Un aggregato monetario (misurato dunque in valore) può variare, sia per effetto di variazioni nel volume dei beni e servizi sottostanti, sia per effetto di una variazione nei prezzi
- ✓ Dato un generico aggregato:

$$A_t = \sum_i p_{it}\, q_{it}$$

- Si indica con **variazione relativa nominale o variazione a prezzi correnti, la crescita in valore di A tra il tempo t e il tempo 0:**

$$\text{Var}\left(\frac{A_t}{A_0}\right) = \frac{\sum_i p_{it}\, q_{it}}{\sum_i p_{i0}\, q_{i0}} - 1$$

- Si indica con **variazione reale relativa o variazione a prezzi costanti, la variazione in quantità dell'aggregato:**

$$\text{Var}\left(\frac{A_{t(0)}}{A_0}\right) = \frac{\sum_i p_{i0}\, q_{it}}{\sum_i p_{i0}\, q_{i0}} - 1$$

Variazioni nominali e reali 2/4

- ✓ L'aggregato $A_{t(0)}$ può essere calcolato anche in via indiretta ricorrendo a numeri indici di prezzo e quantità

  - Dividendo il valore corrente dell'aggregato per un indice dei prezzi di Paasche (in questo caso si chiama deflazionamento)

$$A_{t(0)} = \sum_i p_{i0}\, q_{it} = \sum_i p_{it}\, q_{it} \frac{\sum_i p_{i0}\, q_{it}}{\sum_i p_{it}\, q_{it}} = \frac{A_t}{{}_0I^P_t}$$

  - Moltiplicando il valore corrente dell'aggregato al tempo 0 per un indice di quantità di tipo Laspeyres (in questo caso si chiama estrapolazione)

$$A_{t(0)} = \sum_i p_{i0}\, q_{it} = \sum_i p_{i0}\, q_{i0} \frac{\sum_i p_{it}\, q_{i0}}{\sum_i p_{i0}\, q_{i0}} = A_0\, {}_0I^L_t$$

Variazioni nominali e reali 3/4

✓ Supponiamo che la produzione sia di tre beni, di cui si riportano quantità e prezzi ai tempi *t* e *0*

	Bene 1		Bene 2		Bene 3	
Periodo	q	p	q	p	q	p
0	1000	1,2	800	2,5	500	10
t	1100	1,6	850	2,3	600	8

✓ Il numero indice che esprime gli aggregati a valori nominali $_{20}^{p}I_{21}$ è ottenuto dal rapporto tra la somma dei prezzi e delle quantità dei tre beni al tempo *t* e *0*

$$_{20}^{p}I_{21} = \frac{1100*1,6+850*2,3+600*8}{1000*1,2+800*2,5+500*10} = 1,038$$

$\Rightarrow$ $Var(_{20}^{p}I_{21}) = 1,038-1=3,8\%$

✓ Il numero indice che esprime gli aggregati a valori reali $_{20}^{p}I^{*}_{21}$ è ottenuto dal rapporto tra la somma dei delle quantità dei tre beni al tempo *t* e *0* moltiplicate i prezzi al tempo *0*

$$_{20}^{p}I^{*}_{21} = \frac{1100*1,2+850*2,5+600*10}{1000*1,2+800*2,5+500*10} = 1,15$$

$\Rightarrow$ $Var(_{20}^{p}I'_{21}) = 1,15-1=15\%$

si veda parte pratica in excel

Variazioni nominali e reali 4/4

- ✓ Spesso ci troviamo a dovere calcolare valori e variazioni **a prezzi costanti**, **senza conoscere i valori dei prezzi e delle quantità** degli aggregati sottostanti
- ✓ Quello che **tipicamente si conosce è l'aggregato a valori correnti** nel tempo ed un **indice dei prezzi** (l'ISTAT pubblica mensilmente delle tavole per le rivalutazioni monetarie)
- ✓ Per ottenere gli aggregati espressi ad un medesimo livello dei prezzi:
  - **Deflazioniamo** l'aggregato a valori correnti (i ricavi di vendita del prodotto di consumo *A* nell'esempio che segue) dividendolo per l'indice dei prezzi in %
  - **Otteniamo** così i ricavi di vendita **a prezzi costanti** (nell'esempio a prezzi del 2015 in Italia)
  - La **variazione** 23/22 relativa dei ricavi **a prezzi costanti**, corrisponde alla variazione in termini reali

Anno	Ricavi di vendita del prodotto di consumo A (€ x1.000)	Variazione corrente dei ricavi di vendita 23 vs 22	Indice FOI 2015=100*	Ricavi di vendita del prodotto di consumo A a prezzi 2015	Variazione reale dei ricavi di vendita 23 vs 22
2021	10.500	-	104,2%	10.077	-
2022	12.650	20,5%	112,6%	11.234	11,5%

* ISTAT tavole per le rivalutazione monetarie

si veda parte pratica in excel

L'analisi della mobilità 1/2

- ✓ Con **l'analisi della mobilità** si analizza il cambio di stato delle unità di un insieme nel tempo
  - Le giacenze di magazzino
  - Le carriere del personale
  - **.....**
- ✓ Consideriamo le giacenze di magazzino. Sia C_0 la giacenza iniziale, E_1 la quantità entrata nell'intervallo di tempo e *U1* quella uscita, da cui la giacenza finale:

$$C_1 = C_0 + E - U$$

- ✓ I **tassi di entrata** e, rispettivamente, **uscita** vengono ottenuti rapportando i flussi alla media dello *stock*

$$\mathbf{e} = \frac{E}{(C_0+C_1)/2} \qquad \mathbf{u} = \frac{U}{(C_0+C_1)/2}$$

- ✓ Per esempio: ad inizio anno avevamo uno *stock* di merci in magazzino di C_0=*100* unità. Nel corso dell'anno sono entrate *e=25* nuove unità e ne sono uscite *u=5 unità*. Lo *stock* a fine periodo è dunque di C_1=*120* unità

$$e = \frac{25}{(100+120)/2} = 0{,}057\ (5{,}7\%) \qquad u = \frac{5}{(100+120)/2} = 0{,}023\ (2{,}3\%)$$

L'analisi della mobilità 2/2

- ✓ Può essere interessante, a prescindere dalla variazione nelle giacenze totali, misurare quanta parte delle unità in giacenza sia stata rinnovata nel periodo
- ✓ Calcoliamo allora il **rapporto di rinnovo**: quanta parte dello *stock* è cambiato. Calcoliamo il rapporto tra il flusso è calcolato come semisomma di entrate e di uscite e lo *stock*, pari alla giacenza media già vista nella slide precedente.

$$\frac{(E+U)/2}{(C_0+C_1)/2} = \frac{(\boldsymbol{E+U})}{(\boldsymbol{C_0+C_1})}$$

- ✓ Il reciproco misura il **rapporto di durata**: quanti periodi ci vogliamo affinché tutto lo *stock* si rinnovi

$$\frac{(C_0+C_1)/2}{(E+U)/2} = \frac{(\boldsymbol{C_0+C_1})}{(\boldsymbol{E+U})}$$

- ✓ Nell'ambito della gestione del personale (nuovi assunti vs uscite per dimissioni/licenziamento/pensione) questi rapporti sono chiamati i **tassi di *turnover***
- ✓ Nell'esempio della slide precedente:

Rapporto di rinnovo $= \frac{25+5}{100+120} = 0{,}136$ (13,6%)

Rapporto di durata $= \frac{100+120}{25+5} = 7{,}3$

Possiamo interpretare il risultato che il 13,6% dei prodotti in magazzino si rinnovano ogni anno e che ci vogliono 7,3 anni affinché tutta la merce in magazzino di rinnovi (non a caso 13,6% x 7,3 dà il 100%)

L'analisi della mobilità – Matrice di transizione

✓ Per analizzare la mobilità di un collettivo si può costruire una **matrice di transizione**

- Le **righe *i*** indicano gli **stati** (es. livello professionale, clienti per possesso di prodotti,…) al **periodo *t-1***
- Le **colonne** *j* indicano gli **stati** al periodo ***t***

		Stato t							
		S_1	S_2	..	S_j	..	S_k	Uscite	Totale
Stato t-1	S_1	n_{11}	n_{12}		n_{1j}		n_{1k}	U_1	$n_{1.}$
	S_2	n_{21}	n_{22}		n_{2j}		n_{2k}	U_2	$n_{2.}$
	..								
	S_i	n_{i1}	n_{i2}		n_{ij}		n_{ik}	U_i	$n_{i.}$
	..								
	S_k	n_{k1}	n_{k2}		n_{kj}		n_{kk}	U_k	$n_{k.}$
	Entrate	E_1	E_2		E_j		E_k		
	Totale	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.k}$	$n_{..(t)}/n_{..(t-1)}$	

✓ Nell'esempio dei livelli professionali, n_{11} misura quante persone erano al livello *1* (S_1) al periodo *t-1* e vi sono rimaste anche nel periodo *t*. n_{12} misura invece quanti erano al livello *2* in *t-1* e sono passati al livello *2* al periodo *t* e così via. Il totale di riga (esempio $n_{1.}$ ci dice quante persone c'erano al livello *1* al tempo *t-1*, mentre il totale di colonna, per esempio $n_{.1}$, ci dice quante persone ci sono al livello *1* al tempo *t*

L'analisi della mobilità : tassi di permanenza, transizione, uscita ed entrata

✓ Sulla base dei dati nella matrice, si possono calcolare alcuni indicatori per verificare la proporzione di unità che rimangono in uno stato nell'intervallo di tempo .....

- Il **tasso di permanenza** nello stato *i=j* (gli elementi che si trovano sulla diagonale della matrice):

$$p_{ii} = \frac{n_{ii}}{n_{i.(t-1)}}$$

- Il **tasso di transizione** dallo stato *i* allo stato *j (i ≠ i)* :

$$p_{ij} = \frac{n_{ij}}{n_{i.(t-1)}}$$

- Il **tasso di uscita** dallo stato *i* :

$$u_i = \frac{U_i}{n_{i.(t-1)}}$$

- Il **tasso di entrata** nello stato *i*:

$$e_i = \frac{E_i}{n_{i.(t-1)}}$$

Prospetto dei tassi di transizione o permanenza

- ✓ Sulla base dei tassi di transizione possiamo costruire un prospetto che rappresenti la frequenza di
  - permanenza in uno stato
  - transizione verso un altro stato
- ✓ Vediamolo con un esempio (ipotizzando assenza di Entrate ed Uscite):
  - È data una matrice di transizione tra *t* e *t-1*. Siccome il passaggio è solo tra livelli contigui e crescenti, non tutte le celle della matrice sono presenti
  - Calcolo il prospetto di permanenza e transizione con i tassi definiti alla slide precedente

Matrice di transizione tra diversi stati professionali degli occupati tra il tempo t e t-1

	Cat. Professionale t				
Cat. Professionale t-1	**1**	**2**	**3**	**4**	**Totale (t-1)**
1	400	50	-	-	**450**
2	-	270	30		**300**
3	-	-	70	30	**100**
4	-	-	-	25	**25**
Totale t	**400**	**320**	**100**	**55**	**875**

Prospetto dei tassi di permanenza e transizione

	Cat. Professionale t				
Cat. Professionale t-1	**1**	**2**	**3**	**4**	**Totale (t-1)**
1	0,89	0,11	-	-	**1,00**
2	-	0,90	0,10		**1,00**
3	-	-	0,70	0,30	**1,00**
4	-	-	-	1,00	**1,00**
Totale t	**0,46**	**0,37**	**0,11**	**0,06**	**1,00**

- ✓ Possiamo utilizzare il prospetto dei tassi di permanenza e transizione per prevedere la distribuzione al tempo *t+1*

Previsione dei diversi stati professionali al tempo t

	Cat. Professionale t+1				
Cat. Professionale t	**1**	**2**	**3**	**4**	**Totale (t-1)**
1	356	44	-	-	**400**
2	-	288	32		**320**
3	-	-	70	30	**100**
4	-	-	-	55	**55**
Totale t	**356**	**332**	**102**	**85**	**875**

si veda parte pratica in excel

Statistica per l'impresa

6. Misura delle relazioni tra variabili per le decisioni aziendali

RELAZIONI TRA VARIABILI AZIENDALI

- ✓ Le relazioni tra variabili aziendali per prendere decisioni in azienda
- ✓ Correlazione e regressione
- ✓ Analisi grafica della correlazione
- ✓ L'indice di correlazione di Pearson e la verifica delle ipotesi

REGRESSIONE SEMPLICE

- ✓ Il modello di regressione semplice
- ✓ Il criterio dei minimi quadrati ordinari (OLS)
- ✓ La bontà di adattamento di un modello: l'indice R^2
- ✓ Le proprietà degli stimatori OLS
- ✓ Test di ipotesi sui coefficienti di regressione e intervalli di confidenza
- ✓ Analisi dei residui

REGRESSIONE MULTIPLA

- ✓ Il modello di regressione semplice
- ✓ Analogie con il modello di regressione semplice
- ✓ l'indice R^2 adjusted
- ✓ L'indice F di significatività congiunta

Relazioni tra variabili aziendali per prendere decisioni

- ✓ Nella gestione operativa di un'impresa capita spesso, **prima di prendere determinate decisioni**, di voler verificare **l'esistenza e l'intensità delle relazioni tra le variabili** di interesse
- ✓ Per esempio due ambiti rilevanti sono l'analisi e dunque la comprensione delle dinamiche dei costi di produzione da un lato e delle vendite dall'altro
    - Voglio per esempio capire come variano i miei **costi di produzione** al crescere del **volume di produzione** (la relazione potrebbe non essere lineare per la presenza di economie di scala)
    - Voglio capire la relazione tra volume delle **vendite di un prodotto** e **prezzo** dello stesso. In altre parole come cambiano i volumi delle vendite se faccio degli sconti. Potrei anche indagare la relazione tra le **vendite del mio prodotto rispetto a quello dei concorrenti**: potrei accorgermi che all'aumento delle vendite del concorrente si associa una diminuzione delle mie. Ancora potrei indagare la relazione delle **vendite** con gli investimenti in **pubblicità**

- ✓ Gli ambiti di **applicazione** sono peraltro **molto ampli**:
    - Assenze dal lavoro vs qualifiche professionali o anzianità di servizio
    - Incidenti sul lavoro vs numero di ore lavorate o età del lavoratore
    - ……..

Relazioni tra variabili aziendali: correlazione e regressione

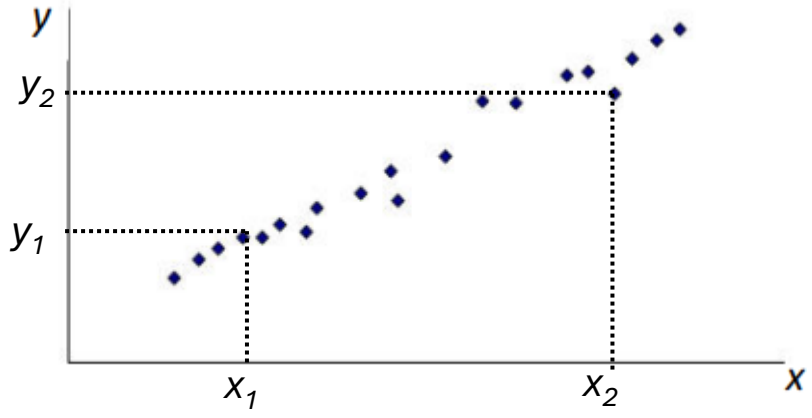
✓ In questa parte del corso affrontiamo l'analisi delle relazioni tra variabili seguendo due approcci:

- **Analisi della correlazione**: andremo a visualizzare l'intensità e la forma che assume il legame tra due o più variabili di interesse
- **Analisi di regressione**: andremo a spiegare l'andamento di una variabile obiettivo, mediante informazioni su una (regressione semplice) o più variabili esplicative (regressione multivariata)

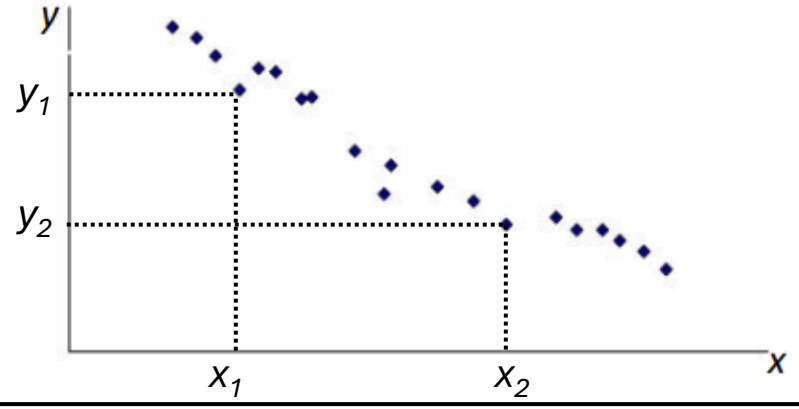
Lo studio della correlazione: analisi grafica

- ✓ La **correlazione tra due variabili** può essere in prima battuta **analizzata graficamente**
- ✓ Si tratta di visualizzare i dati delle due variabili **attraverso un diagramma di dispersione** (*scatterplot*), dove ogni punto rappresenta nel piano il punto definito dalle coordinate x_i e y_i, che corrispondono alle osservazioni nella popolazione di due caratteristiche *X* e *Y*
  - Potremmo avere una **correlazione positiva** tra *X* e *Y*: se cresce *X* cresce anche *Y*. Per esempio *X* potrebbe essere l'età dei clienti dell'azienda, mentre *Y* il volume di vendite dell'impresa a quel cliente
  - Potremmo avere una **correlazione negativa** tra *X* e *Y*: se cresce *X* diminuisce *Y*. Per esempio *X* potrebbe essere il prezzo del prodotto in diversi momenti, mentre *Y* il volume di vendite
  - Potremmo avere una **correlazione assente** tra *X* e *Y*: le due caratteristiche sono indipendenti: volumi delle vendite di creme per il viso e giornali

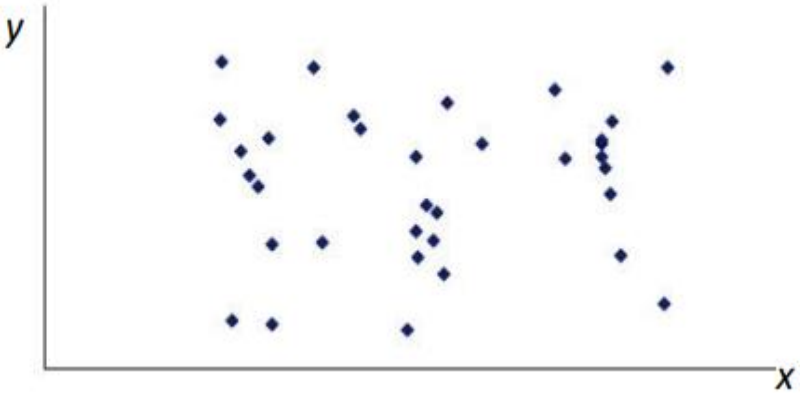
CORRELAZIONE POSITIVA



CORRELAZIONE NEGATIVA



CORRELAZIONE ASSENTE



Analisi e misura della correzione: l'indice di Pearson 1/2

- ✓ Per avere una misura dell'intensità della relazione tra le due variabili di interesse, si calcola **l'indice parametrico di correlazione di Pearson**

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

- ✓ E' un numero puro (senza dimensioni), compreso tra -1 e 1
  - Se $\rho_{xy}=1$ ***c'è una correlazione lineare positiva «perfetta» tra x e y*** *(se x o y cresce per esempio del 10%, y o x crescerà della stessa misura)*
  - Se $\rho_{xy}=-1$ ***c'è una correlazione lineare negativa «perfetta» tra x e y*** *(se x o y cresce per esempio del 10%, y o x diminuirà della stessa misura)*
  - Se $\rho_{xy}=0$ ***non c'è correlazione***
- ✓ Ricordando le definizioni di covarianza e scarto quadratico medio, il coefficiente di correlazione di Pearson nella popolazione sarà

$$\rho_{xy} = \frac{\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{n}}{\sqrt{\frac{\sum_i (x_i-\bar{x})^2}{n}} \cdot \sqrt{\frac{\sum_i (y_i-\bar{y})^2}{n}}}$$

$Cov(x,y)$ *covarianza tra x e y*

σ_y *scarto quadratico medio di y*

σ_x *scarto quadratico medio di x*

si veda parte pratica in excel

Analisi e misura della correzione: l'indice di Pearson 2/2

- ✓ Quando **l'indice di correlazione viene calcolato su dati provenienti da un campione**, piuttosto che da un'intera popolazione obiettivo, **occorre disporre di uno stimatore campionario e su di esso fare inferenza con un test di ipotesi** sulla significatività del suo valore
- ✓ **L'inferenza statistica** è il procedimento per cui si **deducono le caratteristiche di una popolazione dall'osservazione di una parte di essa** (detta "campione"), selezionata solitamente mediante un esperimento casuale (aleatorio)
- ✓ Quando abbiamo di fronte un campione di osservazioni, la **correlazione** può essere stimata con il seguente **stimatore campionario corretto**:

$$r_{xy} = \frac{\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n-1}}}$$

- ✓ r_{xy} è dunque una **variabile aleatoria**, funzione del campione bivariato $((x_1,y_1), (x_2, y_2), \ldots, (x_n,y_n))$
- ✓ Nel caso in cui $\rho_{xy} = 0$, **l'errore standard**[1] della **variabile aleatoria correlazione campionaria** r_{xy} è:

$$ES(r_{xy}) = \sqrt{\frac{1-r^2_{xy}}{n-2}}$$

- ✓ Sotto opportune ipotesi, **l'errore standard** può essere **usato per verificare ipotesi di l'indice di correlazione di Pearson della popolazione ρ_{xy}** , a partire da un'osservazione campionaria r_{xy} (dunque fare inferenza)

[1] Si veda *la slide 47*. L'errore standard misura la «precisione» dello stimatore in questo caso dell'indice di correlazione di Pearson

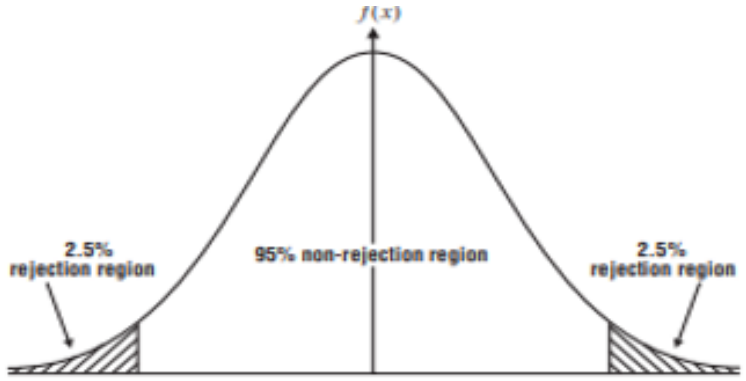
La verifica delle ipotesi 1/2

- ✓ La **verifica (*test*) delle ipotesi** statistiche consiste nel:
  - formulare un'ipotesi su uno o più parametri (incogniti) della popolazione
  - estrarre opportunamente un campione da questa popolazione
  - verificare se l'ipotesi di partenza è supportata dai dati osservati

- ✓ Se il **fenomeno** che stiamo studiando si può rappresentare con una **distribuzione di probabilità** definita da **parametri** (si pensi a quanto abbiamo detto per la media campionaria), a questo punto:
  - Si **specificano le ipotesi** da sottoporre a verifica. Saranno:
    - H_0 ipotesi di interesse (detta **ipotesi nulla** - per esempio il reddito medio delle famiglie del FVG è stato pari a €25.000 nel 2022)
    - H_A **ipotesi alternativa** (per esempio il reddito medio delle famiglie del FVG è stato stato diverso da €25.000 nel 2022)
  - Si considera una **statistica test**, la cui **distribuzione sia nota nel caso sia vera l'ipotesi nulla H_0**
  - Si **estrae un campione**, si calcola il **valore assunto dalla statistica test** e **se ne valida la coerenza con l'ipotesi di partenza**

La verifica delle ipotesi 2/2

- ✓ La verifica di coerenza tra **ipotesi di partenza ed evidenza campionaria** si basa sulla **distribuzione di probabilità** che assume la **statistica test *τ*** nel caso in cui ***H₀* sia vera**
- ✓ Data questa statistica test:
  - **Fisso un livello di confidenza *1-α*** del test. ***α* è il livello di errore probabile tollerabile**. Si tratta di una «probabilità» ragionevolmente «piccola». Spesso *α=5%*
  - Sulla base della **distribuzione della statistica test *τ*, si calcolano i valori critici** (i confini) per:
    - **Regione di accettazione**: dove sotto l'ipotesi che sia vera H_0 cade con probabilità *1-α* il valore vero della popolazione
    - **Regione di rifiuto**: dove sotto l'ipotesi che sia vera H_0, il valore vero della popolazione cade con una probabilità molto «piccola» (*α*)
  - **Si estrae il campione**, si calcola il **valore della statistica test *τ*** e lo **confronto con i valori critici** visti sopra:
    - Se il **valore campionario** cade nella **regione di accettazione**, allora **non si rifiuta l'ipotesi *H₀***
    - Se il **valore campionario** cade nella regione di rifiuto, allora **si rifiuta l'ipotesi *H₀***



  - I **valori critici** dipendono ovviamente dalla **forma della distribuzione** della statistica

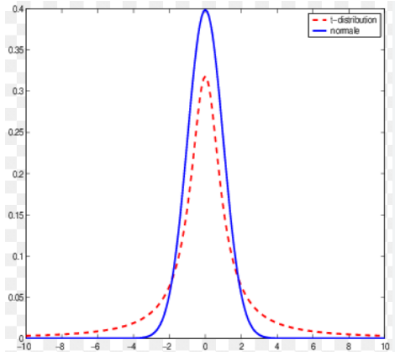
Il test *t*

- ✓ **Nella pratica la statistica *test* viene standardizzata**, ovvero:
  - si sottrae il valore atteso che corrisponde all'ipotesi nulla (sub H_0), in modo da centrare la distribuzione sullo *0*
  - si divide per l'errore standard (stimato in modo di ottenere una varianza pari a *1*)
- ✓ **Si ottiene così una statistica nota come *t – test***. Se è vera l'ipotesi H_0: $\mu = m^*$:

$$t = \frac{\hat{\mu} - m^*}{ES(\hat{\mu})}$$

Per campioni «grandi» ~ N (0, 1)

Per campioni «piccoli» ~ t_{n-1} t-Student con n-1 gradi di libertà



- ✓ Conoscendo la distribuzione della t-test posso **individuare i valori critici** entro i quali si distribuirà al livello 1-α la distribuzione della statistica. **Confronto poi il valore della statistica corrispondente a H_0. Se** per questo valore la statistica test **finisce fuori dall'area di accettazione rifiuto l'ipotesi, altrimenti «non rifiuto» H_0**

$$t_{crit} \leq \frac{\hat{\mu} - m^*}{ES(\hat{\mu})} \leq t_{crit}$$

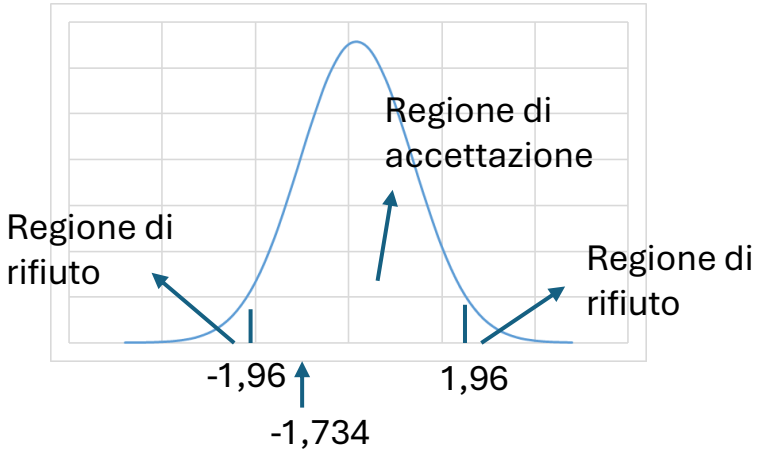
- ✓ Si noti che se pongo in evidenza m* dalla espressione precedente, posso trovare l'intervallo al livello di confidenza desiderato della media della popolazione a partire da media ed errore standard campionari (vedi *slides 54-56*)

$$\hat{\mu} - t_{crit}\, ES(\hat{\mu}) \leq m^* \leq \hat{\mu} + t_{crit}\, ES(\hat{\mu})$$

La verifica delle ipotesi - il caso della media campionaria

- ✓ Vediamo un esempio applicato alla media campionaria
- ✓ Consideriamo un caso in cui non siano note media e varianza della popolazione, ma il campione sia «sufficientemente grande» (es. *n=100*), con media *175* ed errore *standard 3*
- ✓ Voglio verificare, sulla base del campione estratto, H_0: $\mu = 180$, con un livello di confidenza del *95% (1-α)*
- ✓ Utilizziamo la statistica *t-test*, che per campioni sufficientemente grandi si distribuisce come una normale di media zero e varianza uno
- ✓ Posso calcolare i valori critici del test al *95%*, risulterà $t_{crit} = 1,96$
- ✓ Calcolo adesso il valore della statistica t-test in corrispondenza della media e dell'errore standard campionari

$$t' = \frac{\overline{Y} - \mu}{ES(\overline{Y})} = \frac{175 - 180}{3} = -1,734$$



- ✓ Confronto il valore di *t'* con i valori critici e vedo che esso cade all'interno della regione di accettazione, dunque «non rifiuto» l'ipotesi H_0,
- ✓ Posso altresì andare a costruire l'intervallo di confidenza al 95% per la media della popolazione. Risulterà

$$\overline{Y} - t_{crit}\, ES(\overline{Y}) \leq \mu \leq \overline{Y} + t_{crit}\, ES(\overline{Y}) \Rightarrow 175 - 1,96*3 \leq \mu \leq 175 + 1,96*3 \Rightarrow 169 \leq \mu \leq 181$$

Verifica delle ipotesi e indice di Pearson

- ✓ Quando abbiamo di fronte un campione di osservazioni, l'indice di correlazione di Pearson può essere stimato con il seguente **stimatore campionario corretto**:

$$r_{xy} = \frac{\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n-1}}}$$

  - Con errore *standard* pari a:

$$ES\,(r_{xy}) = \sqrt{\frac{1 - r^2_{xy}}{n-2}}$$

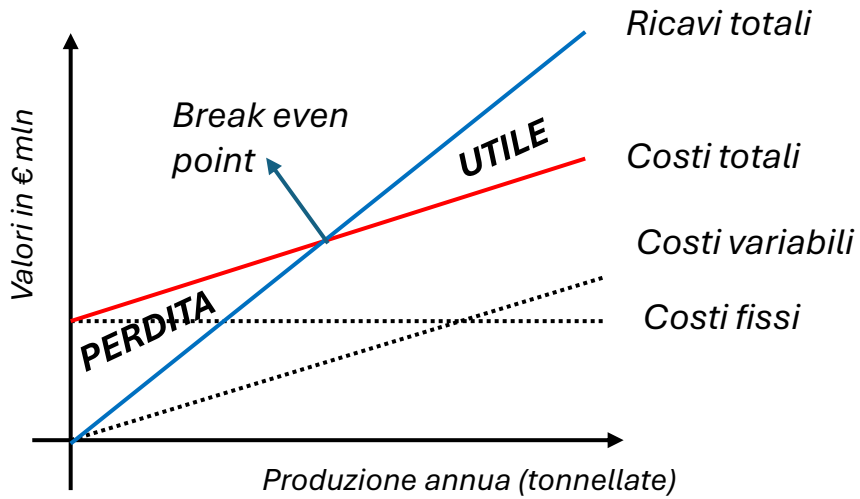
- ✓ Se ***X*** e ***Y*** sono normalmente distribuite e sotto l'ipotesi che l'indice di correlazione di Pearson della popolazione sia zero (***H_0: $\rho_{xy} = 0$),*** la statistica t-test che useremo è la seguente

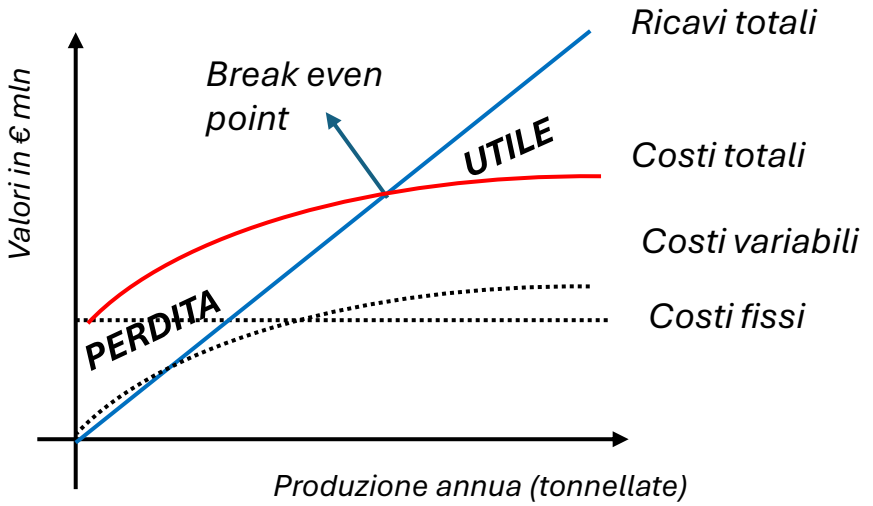
$$\frac{r_{xy} - 0}{ES\,(rxy)} = r_{xy} \sqrt{\frac{n-2}{1 - r^2_{xy}}} \sim t_{n-2}$$

- ✓ Possiamo allora effettuare un test *t* dell'ipotesi di incorrelazione tra *X* e *Y* come segue con modalità analoghe a quanto visto in precedenza
- ✓ Ipotesi più generali del tipo $H_0: r_{xy} = r^*$ *non sono testabili con questa statistica, poiché vale solo per il caso ρ=0*

Esempio studio della relazione tra costi e volumi di produzione 1/5

- ✓ Affrontiamo a titolo esemplificativo lo studio della relazione tra costi e volumi di produzione, che in termini aziendali è detta *break-even point analysis*
- ✓ Si tratta di un tema molto importante, che ritorna spesso nella valutazione di *business case*
- ✓ Sono interessato a capire la relazione tra ricavi e costi, per capire qual è il livello di vendite che mi assicura il punto di pareggio (efficienza), ovvero il *break even point*. Superato tale punto, l'impresa sarà in utile. Siccome per arrivare ad un certo livello atteso di ricavi ci vogliono degli anni, il *break even* può avere anche una connotazione temporale. La forma della curva dei costi e dunque la relazione con i ricavi, è determinante per concludere questa analisi. Lo possiamo vedere semplicemente per via grafica





Esempio studio della relazione tra costi e volumi di produzione 2/5

- ✓ Abbiamo detto dunque dell'importanza per l'impresa di studiare la relazione tra costi e produzione. Vediamo un esempio concreto

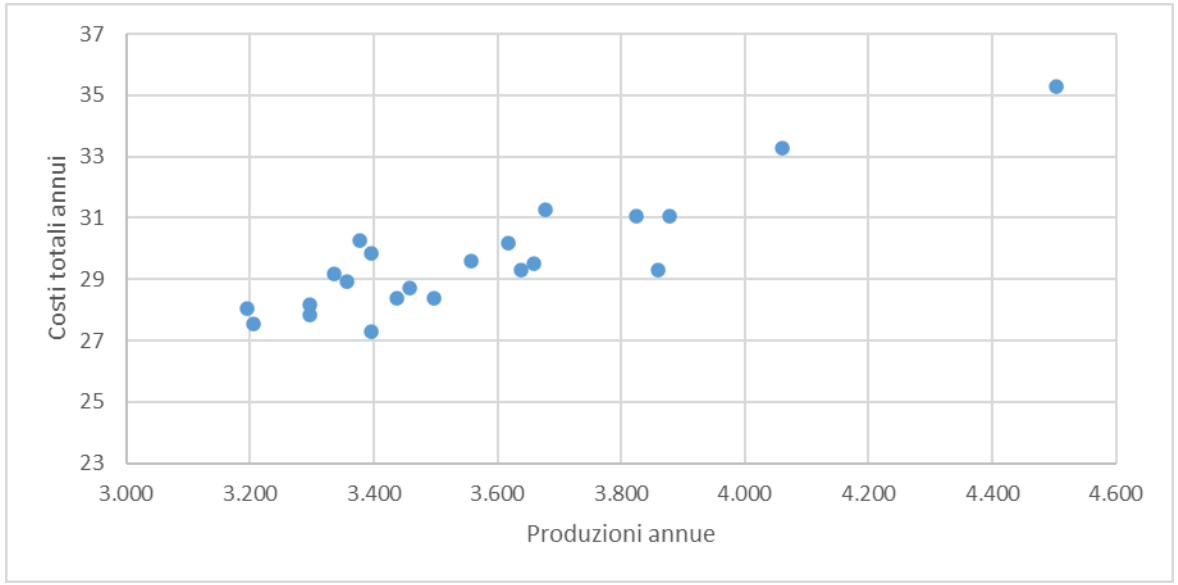
- ✓ Supponiamo che un'azienda del settore alimentare abbia diversi stabilimenti (oppure molti reparti all'interno di uno stabilimento) di dimensioni diverse secondo il numero di addetti, che producono lo stesso tipo di prodotto con volumi di produzione e di vendite differenti, e che sono localizzati in tutto il territorio nazionale, ad esempio nei capoluoghi di regione Italiani. L'azienda ha rilevato, attraverso un campione casuale, i dati sui volumi di produzione e costi totali riportati nella tabella

- ✓ L'obiettivo dell'analisi è individuare e misurare la relazione tra i costi e volume della produzione, che sono le due variabili in gioco.

	Produzione annua in tonnellate	Costi totale annui in mln euro
Ancona	3.558	30
Aosta	3.296	28
Bari	3.437	28
Bologna	3.337	29
Campobasso	3.618	30
Catanzaro	3.377	30
Firenza	3.196	28
Genova	4.060	33
L'Aquila	3.859	29
Milano Nord	3.658	30
Milano Sud	3.678	31
Napoli	3.825	31
Palermo	3.397	30
Perugia	3.497	28
Potenza	3.296	28
Roma Nord	3.638	29
Roma Nord	3.879	31
Sud	4.502	35
Torino	3.397	27
Trento	3.457	29
Trieste	3.206	28
Venezia	3.357	29

Esempio studio della relazione tra costi e volumi di produzione 3/5

✓ La prima cosa da fare per la verifica dell'esistenza della relazione è la costruzione del diagramma dispersione



✓ Come si vede, all'aumentare del livello di produzione, aumenta il valore del costo totale: esiste pertanto una relazione positiva tra le due variabili

si veda parte pratica in excel

Esempio studio della relazione tra costi e volumi di produzione 4/5

- ✓ Tuttavia al manager interessa conoscere anche l'intensità di questo legame, che si ottiene calcolando l'indice di correlazione di Pearson

$$\bar{x}\,\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{78526,68}{22} = 3569,39 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{652,39}{22} = 29,65$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 500,52 \qquad \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y} = \frac{1}{22} * 2339644,8\text{-}(3569,39 * 29,65) = 500,52$$

$$corr(x,y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{\sigma_x \cdot \sigma_y} = \frac{500,52}{\sqrt{92531,01 * 3,42}} = \frac{500,52}{304,19 * 1,85} = \frac{500,52}{562,14} = 0,89$$

- ✓ Per l'azienda di prodotti alimentari il valore dell'indice di correlazione tra le due variabili è molto alto. Si può concludere, pertanto, che la relazione tra il volume dei costi indicato con y, e il volume della produzione (x) è quasi del 90%

si veda parte pratica in excel

Esempio studio della relazione tra costi e volumi di produzione 5/5

- ✓ Essendo i dati ottenuti attraverso un campione si esegue un test a due code dell'ipotesi nulla di assenza di correlazione ad un livello di significatività pari, ad esempio, a $\alpha=0,05$. Il valore della statistica *t* è:

$$t = 0,89\sqrt{\frac{22-2}{1-(0,89)^2}} = 8,74$$

- ✓ Per una distribuzione *t* di *Student* con 20 gradi di libertà, per un livello di confidenza del 95%, i valori critici sono -2,4 e +2,4. t=8,74 cade fuori da tale intervallo (regione di accettazione dell'ipotesi che la correlazione sia nulla) e dunque «non rifiutiamo» l'ipotesi che vi sia correlazione

- ✓ Alla stessa conclusione si poteva arrivare verificando che, per una distribuzione *t* di *Student* con 20 gradi di libertà, la probabilità di osservare un valore maggiore di |8,74|, è p=0,000000028, inferiore al livello di significatività del 5% che abbiamo fissato. Dunque rifiutiamo l'ipotesi che non vi sia correlazione

si veda parte pratica in excel

Regressione, correlazione e modello

- ✓ Nell'analisi della **correlazione** vista sinora abbiamo trattato le **variabili *x* e *y*** in maniera totalmente simmetrica, **senza** cioè che vi sia un **rapporto causa-effetto** tra i due fenomeni che stiamo misurando

- ✓ Nella **regressione** invece **ipotizziamo che vi sia una relazione causa ed effetto** e vogliamo in prima battuta **verificare** se l'ipotesi che stiamo facendo è vero o meno ed in seconda vogliamo **quantificare le relazioni**

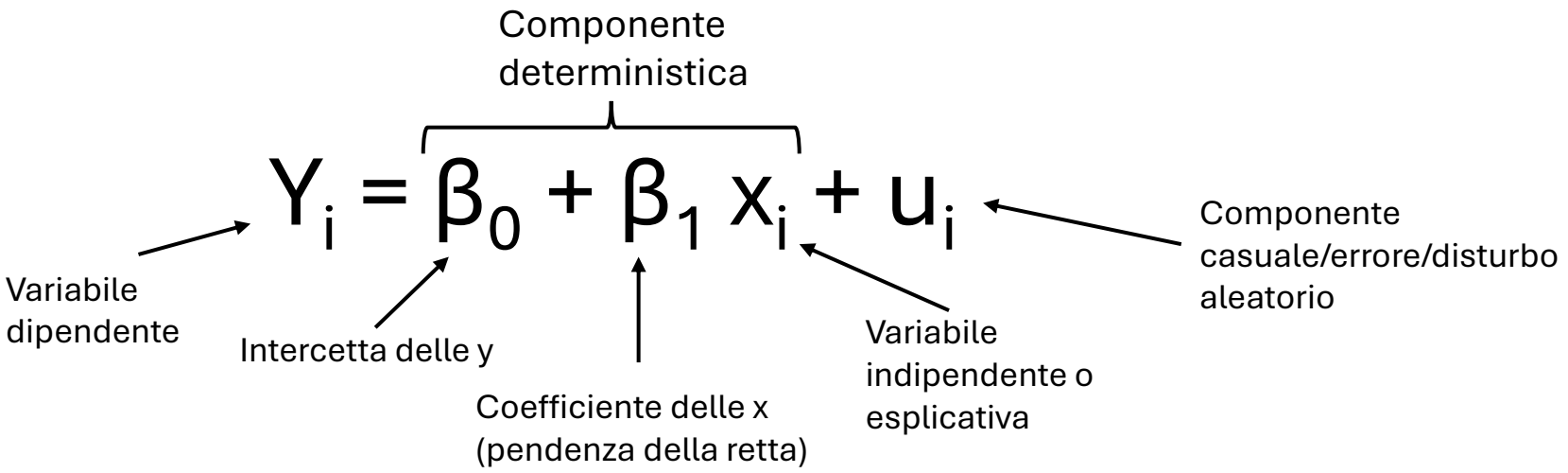
- ✓ L'idea di base **dietro la regressione** è che dietro le osservazioni campionarie che abbiamo di fronte, **esista un modello**, ovvero un processo generatore di dati (GDP), che per noi avrà la forma lineare

- ✓ Un modello è una **rappresentazione semplificata della realtà** che stiamo osservando, che riflette le nostre conoscenze sulla base della teoria o del nostro *know how*/conoscenza del settore che stiamo analizzando. Una volta che siamo riusciti a specificare il modello, saremo capaci di:

  - Interpretare la realtà che abbiamo rappresentato
  - Riprodurla sotto condizioni diverse per:
    - previsione
    - *what-if analysis*

Il modello di regressione lineare semplice

✓ Ci focalizzeremo per il momento sul modello di **regressione lineare semplice (solo due variabili *x* e *y*)**



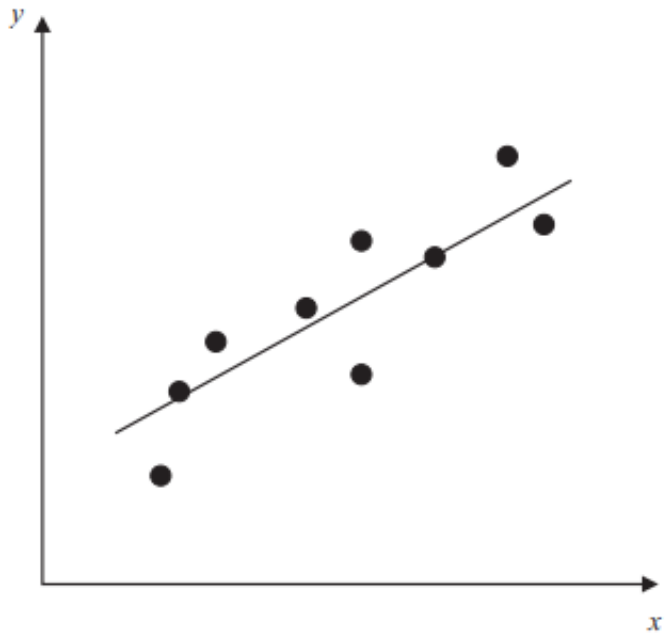
✓ Alcune considerazioni sui termini dell'equazione:

- La **componente deterministica** rappresenta la **retta di regressione**
- Il termine $\boldsymbol{\beta_0}$ è, nella componente deterministica, il **valore di *Y* quando *x=0***
- Il **termine $\boldsymbol{\beta_1}$** è la derivata prima della componente deterministica rispetto alla *x* e rappresenta la **pendenza della retta.** Inoltre rappresenta anche una sorta di **elasticità di *Y* rispetto ad *x***
- Il termine $\boldsymbol{u_i}$ può rilevare diversi fenomeni (errori di misura non modellizzabili di *Y*, variabili esplicative di *Y* importanti, ma che non abbiamo considerato nella specificazione del modello, ….

Determinare o coefficienti del modello

- ✓ Come determinare β_0 *e* β_1?

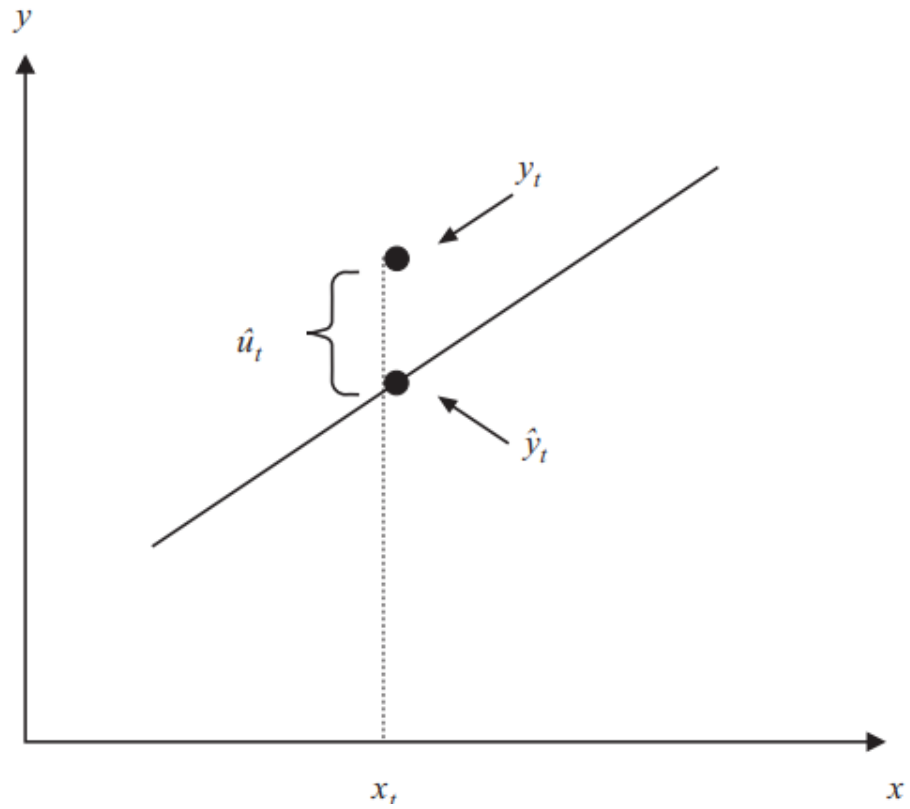
  - ✓ Cerchiamo β_0 e β_1 tali da rendere minime le distanze (verticali) tra i punti rappresentativi dei dati osservati e la retta stimata:



Ordinary Least Squares

- ✓ Il metodo di stima più comune è noto come OLS (*ordinary least squares* o minimi quadrati ordinari)
- ✓ Si minimizzano i quadrati delle distanze indicate in figura (da cui il nome).
- ✓ Formalmente, siano:
  - y_t i valori osservati per ogni t
  - $\hat{y}_t$ i valori osservati per ogni t
  - $\hat{u}_t$ i residui, $\hat{u}_t = y_t - \hat{y}_t$
- ✓ Operativamente devo cercare **quei valori di $\hat{\beta}_0$ e $\hat{\beta}_1$,** tali da rendere **minima la somma dei quadrati dei residui** (la ns funzione di perdita)

$$\min L(\hat{\beta}_0, \hat{\beta}_1) = \sum (y_t - \hat{y}_t)^2 = \sum (y_t - \hat{\beta}_0 + \hat{\beta}_1 x_{it})2$$



Gli stimatori OLS

- ✓ I valori di $\hat{\beta}_0$ e $\hat{\beta}_1$ che minimizzano la funzione definita alla slide precedente sono i seguenti:

$$\hat{\beta}_1 = \frac{\sum X_i Yi - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- ✓ Ricordando che $E(X_iY_i) = \frac{\sum X_i Yi}{n}$ e che $E(X_i^2) = \frac{\sum X_i^2}{n}$ la formula di $\hat{\beta}_1$ può essere scritta anche come:

$$\hat{\beta}_1 = \frac{\sum X_i Yi - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{n\,E(X_iY_i) - n\bar{X}\bar{Y}}{n\,E(X_i^2) - n\bar{X}^2} = \frac{n\left[E(X_iY_i) - \bar{X}\bar{Y}\right]}{n\left[E(X_i^2) - \bar{X}^2\right]} = \frac{Cov\,(XY)}{Var\,(X)}$$

- ✓ Si osservi che la formula per $\hat{\beta}_1$ può essere scritta anche come rapporto tra devianza di *xy* e devianza di *x*

$$\hat{\beta}_1 = \frac{ss\,(XY)}{ss\,(X)}$$

- ✓ Dove:

$$ss\,(XY) = \sum (x_t - \bar{x})\,(y_t - \bar{y}) \qquad ss\,(XX) = \sum (x_t - \bar{x})^2$$

La bontà di adattamento del modello – L'indice R^2 1/2

- ✓ Siamo interessati alla bontà di adattamento del nostro modello ai dati. Per valutarla usiamo un'altra statistica: il cosiddetto R^2
- ✓ La variabilità di y_t attorno alla sua media y (devianza) è la somma dei quadrati totali, o *total sum of squares* (TSS):

$$TSS = \sum(y_t - \bar{y})^2$$

- ✓ Tale devianza può essere divisa in due parti: la devianza spiegata (ESS) e quella residua (RSS)

$$TSS = ESS + RSS$$

$$\sum(y_t - \bar{y})^2 = \sum(\hat{y}_t - \bar{y})^2 + \sum(y_t - \hat{y}_t)^2$$

- ✓ Misureremo la bontà di adattamento con il rapporto R^2 = ESS/TSS
- ✓ Riesce

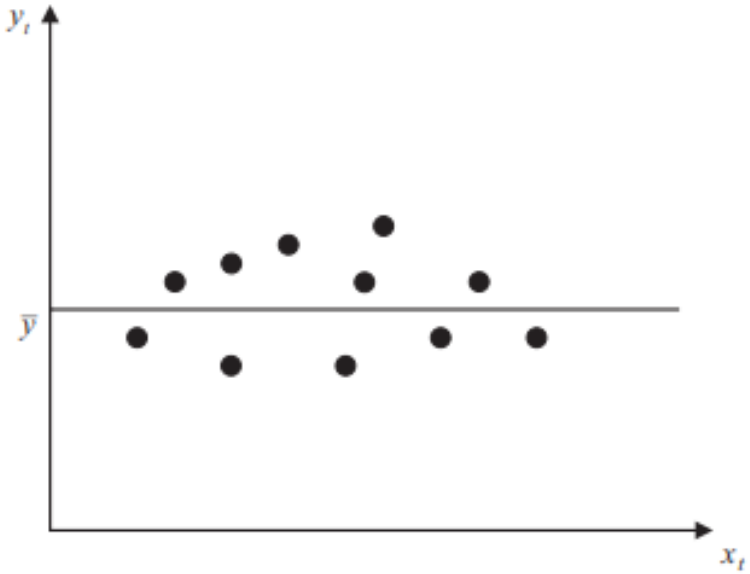
$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$R^2 = 1 - \frac{\sum(y_t - \hat{y}_t)^2}{\sum(y_t - \bar{y})^2}$$

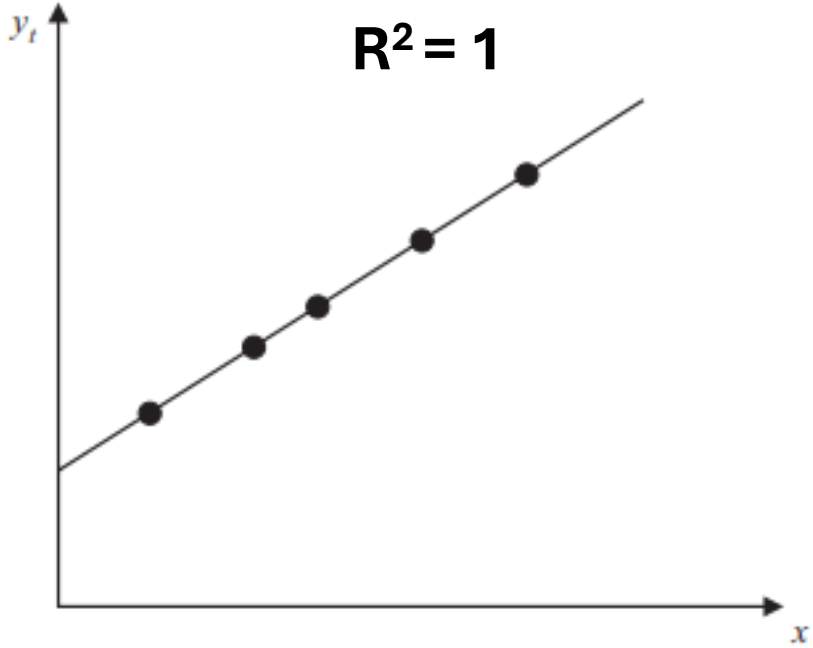
La bontà di adattamento del modello – L'indice R^2 2/2

- ✓ L'indice R^2 è compreso tra 0 e 1
- ✓ I casi limite:
  - $R^2 = 0$ → ESS = 0 → RSS = TSS
  - $R^2 = 1$ → RSS = 0 → ESS = TSS

$R^2 = 0$



$R^2 = 1$



si veda parte pratica in excel (regressione esempio 1)

Le proprietà degli stimatori OLS

✓ Facciamo le seguenti ipotesi su u_t

	Notazione	**Interpretazione**
(1)	$E(u_t) = 0$	Gli errori hanno valore atteso nullo
(2)	$Var(u_t) = \sigma^2$	La varianza degli errori è costante e finita (errore omoschedastici)
(3)	$Cov(u_i; u_j) = 0$	Gli errori non sono correlati tra loro
(4)	$Cov(u_t; x_t) = 0$	Non c'è correlazione tra l'errore in *t* e la corrispondenti variabile x_t

✓ Una quinta ipotesi è richiesta per fare inferenza sui parametri della popolazione (i veri β_0 e β_1)

(5) u_t è normalmente distribuito

Gli stimatori OLS sono BLUE

✓ **Sotto le ipotesi da (1) a (4), gli stimatori OLS sono detti BLUE** (Best Linear Unbiased Estimators)

- **Linear**: $\hat{\beta}_0$ e $\hat{\beta}_1$ sono funzioni lineari (dei dati del campione)
- **Unbiased**: i valori attesi di $\hat{\beta}_0$ e di $\hat{\beta}_1$ sono uguali ai parametri «veri» valori di β_0 e di β_1
- **Best**: significa che lo stimatore OLS ha varianza minima nella classe degli stimatori lineari corretti; questo risultato prende il nome di Teorema di Gauss–Markov.

Le proprietà degli stimatori OLS

✓ **Consistenza**

- Gli stimatori dei minimi quadrati $\hat{\beta}_0$ e di $\hat{\beta}_1$ sono consistenti, ovvero le stime convergeranno ai veri valori dei parametri al divergere a infinito della dimensione del campione (*T*). **Servono le ipotesi Cov(u_t; x_t) = 0 e Var(u_t) = σ^2** per dimostrarlo

$$\lim_{T\to\infty} \Pr\left[|\hat{\beta} - \beta| > \delta\right] = 0 \quad \forall\, \delta > 0$$

✓ **Correttezza**

- Gli stimatori dei minimi quadrati $\hat{\beta}_0$ e di $\hat{\beta}_1$ sono corretti, ovvero $E(\hat{\beta}_0) = \beta_0$ *e* $E(\hat{\beta}_1) = \beta_1$. Pertanto, in media le stime saranno uguali ai veri valori. **Serve che E(u_t) = 0**

✓ **Efficienza**

- Uno stimatore $\hat{\beta}$ del parametro **β è detto efficiente se è corretto ed ha varianza minima tra gli stimatori corretti**. Il teorema di Gauss-Markov ci dice che gli stimatori dei minimi quadrati sono i più efficienti tra gli stimatori corretti se **E(u_t) = 0** (stimatore corretto), **Cov(u_i; u_j) = 0** (errori incorrelati) e **Var(u_t) = σ^2** *(errori omoschedastici)*

Gli stimatori OLS 1/2

- ✓ Riprendiamo alcuni concetti
- ✓ Quando applichiamo una **regressione lineare** su dati campionari **per fare inferenza**, ricavando un modello lineare semplice, usiamo una gamma di stimatori che sono necessari per stime puntuali ed intervallari. Per costruire la statistica *t-test* su cui basare la verifica delle ipotesi e costruire degli intervalli di confidenza, **ci servono stimatori e relativi errori *standard***
- ✓ I primi **stimatori** che abbiamo incontrato sono quelli che ci servono per calcolare i coefficienti della retta di regressione:

$$\hat{\beta}_1 = \frac{\sum X_t Yt - T\bar{X}\bar{Y}}{\sum X^2_t - T\bar{X}^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- ✓ Ci serve poi una misura dell'affidabilità/precisione (gli **errori standard**) degli stimatori $\hat{\beta}_0$ e $\hat{\beta}_1$. Si può dimostrare che date le assunzioni 1-4 della slide 125, gli errori standard dei due coefficienti sono rispettivamente:

$$ES(\hat{\beta}_0) = s\sqrt{\frac{\sum x_i^2}{T\left(\left(\sum x_i^2\right) - T\bar{x}^2\right)}} \qquad ES(\hat{\beta}_1) = s\sqrt{\frac{1}{\sum x_i^2 - T\bar{x}^2}}$$

- Dove *s* è l'errore *standard* stimato degli errori (non conosciamo la varianza «vera» degli errori)

- ✓ **L'errore standard dello stimatore degli errori** (residui) di previsione è dato dalla radice quadrata della somma dei quadrati dei residui, diviso la dimensione *T* del campione ridotto per il numero dei regressori (2 nella regressione semplice):

$$s = \sqrt{\frac{\sum \hat{u}_i^2}{T-2}}$$

- ✓ Alcune considerazioni sugli stimatori visti in queste ultime due *slides*
  - Sia $ES(\hat{\beta}_0)$, sia $ES(\hat{\beta}_0)$ dipendono da *s*, cioè dall'errore *standard* degli errori stimato su base campionaria
  - Maggiore è la dimensione del campione *T,* minori saranno gli errori standard degli errori e dei coefficienti

Test di ipotesi sui coefficienti di regressione 1/2

- ✓ Come abbiamo detto alla *slide 124*, **per fare inferenza sui coefficienti della regressione** (*test* di ipotesi/calcolo intervalli di confidenza) è necessario che gli **errori si distribuiscano come una normale di media *0* e varianza costante σ^2** , ovvero

$$U_t \sim N(0, \sigma^2)$$

- ✓ **Se l'ipotesi di partenza è vera, allora anche gli stimatori dei coefficienti di regressione** si distribuiscono come una **normale** (essendo riconducibili a somme di variabili Normali). **Se non vale l'ipotesi della normalità** dei residui, i parametri **si distribuiranno come una normale** solo **se valgono la ipotesi da 1 a 4** della slide 125 ed **il campione è sufficientemente grande**

$$\hat{\beta}_0 \sim N(\beta_0, var(\hat{\beta}_0)) \qquad \hat{\beta}_1 \sim N(\hat{\beta}_1, var(\hat{\beta}_1))$$

- ✓ Da $\hat{\beta}_0$ e $\hat{\beta}_1$ posso costruire le **variabili normali standard**, che a loro volta si distribuiranno come una normale standard, di media 0 e varianza 1

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{var(\hat{\beta}_0)}} \sim N(0,1) \qquad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{var(\hat{\beta}_1)}} \sim N(0,1)$$

- ✓ **Le varianze dei degli stimatori dei coefficienti non sono note, perciò si usano le stime ES($\hat{\beta}_0$) eES($\hat{\beta}_1$)**

$$\boxed{\frac{\hat{\beta}_0 - \beta_0}{ES(\hat{\beta}_0)}} \sim t_{T-2} \qquad \boxed{\frac{\hat{\beta}_1 - \beta_1}{ES(\hat{\beta}_1)}} \sim t_{T-2}$$

Test di ipotesi sui coefficienti di regressione 2/2

- ✓ **Soddisfatte le condizioni viste nelle slide precedenti** e con riferimento al modello di regressione $y_t = \beta_0 + \beta_1 x_t + u_t$, **possiamo testare l'ipotesi statistica $H_0 : \beta = \beta^*$ con β^* una costante** (che potrebbe essere *0* o un'ipotesi economica di interesse)
- ✓ Si può procedere come già descritto alla *slides 108*
- ✓ Per una specifica ipotesi $H_0 : \beta = \beta^*$ **con** β^*
- ✓ Si calcola la statistica test: $\frac{\hat{\beta} - \beta^*}{ES(\hat{\beta})}$
- ✓ Si considera la distribuzione con cui **confrontare il valore assunto dalla statistica *test*.** Come abbiamo visto, sotto l'ipotesi H_0 **la statistica test si distribuisce come una t di Student con *T-2* gradi di libertà** (dove *T* è la dimensione del campione)
- ✓ **Si sceglie un livello di significatività** α (di solito il *5%* o l'*1%*)
- ✓ Dalle tavole della *t* **si ottengono i valori critici** che delimitano le regioni di rifiuto
- ✓ **Se la statistica *test* finisce nella regione di accettazione allora non si rifiuta H_0** (vedi sotto), **altrimenti non la si rifiuta**

$$-t_{crit} \leq \frac{\hat{\beta} - \beta^*}{ES(\hat{\beta})} \leq t_{crit}$$

Intervalli di confidenza dei coefficienti di regressione

✓ **Partendo dalla statistica *test*** e dai suoi valori critici .....

$$-t_{crit} \leq \frac{\hat{\beta} - \beta*}{ES(\hat{\beta})} \leq t_{crit}$$

✓ ... **posso riordinare i fattori** e **ricavare l'intervallo di confidenza** al fissato di significatività

$$\hat{\beta} - t_{crit} * s \sqrt{\frac{1}{\sum x_i^2 - Tx^2}} \leq \beta* \leq \hat{\beta} - t_{crit} * s \sqrt{\frac{1}{\sum x_i^2 - Tx^2}}$$

✓ ... che è la regola nell'approccio dell'intervallo di confidenza

✓ Vediamo *test* ed intervalli applicati ai coefficienti di regressione con un semplice esempio

Intervalli di confidenza dei coefficienti di regressione – esempio 1/3

- ✓ Siamo un'impresa e vogliamo stimare la relazione che c'è tra i nostri ricavi per cliente (variabile dipendente *y*) ed il relativo numero di componenti della famiglia (variabile esplicativa *x*). Mi aspetto una relazione diretta e crescente (se cresce il numero dei componenti aumenta il ricavo medio per famiglia)

	N° componenti famiglia clienti (x)	Ricavi (€x1.000) y	xy	x^2	y'	u'	u'^2	(y-E(y))^2
	1	12	12	1	14	-2	4	324
	1	15	15	1	14	1	1	225
	2	25	50	4	22	3	9	25
	3	28	84	9	30	-2	4	4
	5	45	225	25	46	-1	1	225
	6	55	330	36	54	1	1	625
Totale	**18**	**180**	**716**	**76**	**180**	**0**	**20**	**1.428**

- ✓ Voglio stimare la retta di regressione con il metodo dei minimi quadrati per il campione selezionato riportato in tabella fatto da 6 unità

si veda parte pratica in excel

Intervalli di confidenza dei coefficienti di regressione – esempio 2/3

	N° componenti famiglia clienti (x)	Ricavi (€x1.000) y	xy	x^2	y'	u'	u'^2	(y-E(y))^2
	1	12	12	1	14	-2	4	324
	1	15	15	1	14	1	1	225
	2	25	50	4	22	3	9	25
	3	28	84	9	30	-2	4	4
	5	45	225	25	46	-1	1	225
	6	55	330	36	54	1	1	625
Totale	**18**	**180**	**716**	**76**	**180**	**0**	**20**	**1.428**

$\bar{x} = 3$

$\bar{y} = 30$

✓ Ricordiamoci gli stimatori da utilizzare sui nostri dati campionari:

$$\hat{\beta}_1 = \frac{\sum X_t Y_t - T\bar{X}\bar{Y}}{\sum X^2_t - T\bar{X}^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

✓ Applicati ai dati del nostro problema risulta:

$\beta_1 = \frac{716 - 6*3*30}{76 - 6*3\text{^}2} = 8$ $\beta_0 = 30 - 8*3 = 6$ → $y' = 6 + 8x$ → $R^2 = 1 - \frac{20}{1428} = 0{,}986$

si veda parte pratica in excel

Intervalli di confidenza dei coefficienti di regressione – esempio 3/3

	N° componenti famiglia clienti (x)	Ricavi (€x1.000) y	xy	x^2	y'	u'	u'^2	(y-E(y))^2
	1	12	12	1	14	-2	4	324
	1	15	15	1	14	1	1	225
	2	25	50	4	22	3	9	25
	3	28	84	9	30	-2	4	4
	5	45	225	25	46	-1	1	225
	6	55	330	36	54	1	1	625
Totale	**18**	**180**	**716**	**76**	**180**	**0**	**20**	**1.428**

$\bar{x} = 3$

$\bar{y} = 30$

$t_{0,025,\,4} = 2,78$

T-Student per α/2 e 4 gradi di libertà (t-2)

✓ Vediamo adesso di costruire un intervallo di confidenza per il parametro β_1

$$s = \sqrt{\frac{\sum \hat{u}_i^2}{T-2}} \qquad \hat{\beta} - t_{crit} * s \sqrt{\frac{1}{\sum x_i^2 - Tx^2}} \leq \beta^* \leq \hat{\beta} - t_{crit} * s \sqrt{\frac{1}{\sum x_i^2 - Tx^2}}$$

✓ Pertanto sul campione selezionato:

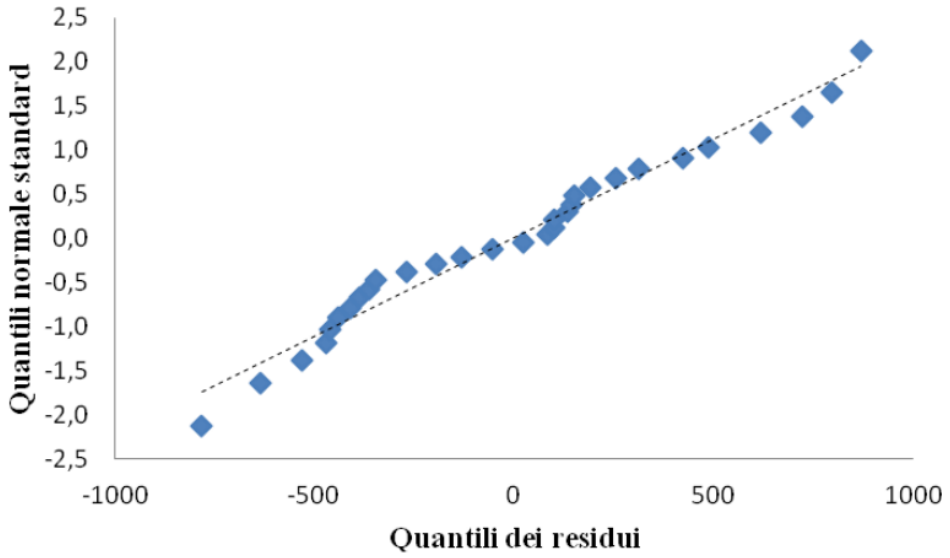
$$s = \sqrt{\frac{20}{4}} = 2,24 \quad \Rightarrow \quad 8 - 2,74 * 2,24 * \sqrt{\frac{1}{76-6*3^2}} \leq \beta_1 \leq 8 + 2,74 * 2,24 * \sqrt{\frac{1}{76-6*3^2}} \quad \Rightarrow \quad 6,69 \leq \beta_1 \leq 9,31$$

✓ Siccome β_1, parametro della popolazione, varia tra *6,69* e *9,31* con un livello di confidenza del *95%* e siccome β_1 misura di quanto aumenta il ricavo medio per famiglia se la stessa aumenta i propri componenti di *1* unità, allora posso concludere che (al *95%*) se una famiglia aumenta di *1* i propri componenti il relativo ricavo medio aumenterà tra *€ 6.690* e *€ 9.310*

si veda parte pratica in excel

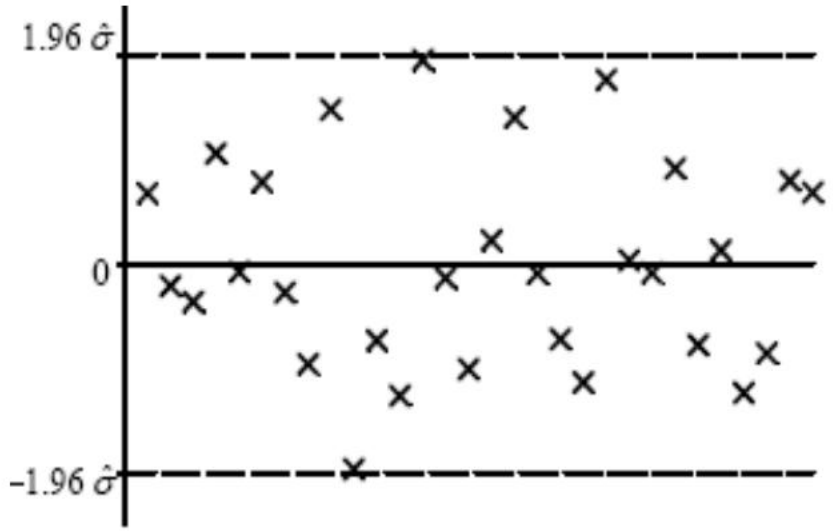
Analisi grafica dei residui 1/5

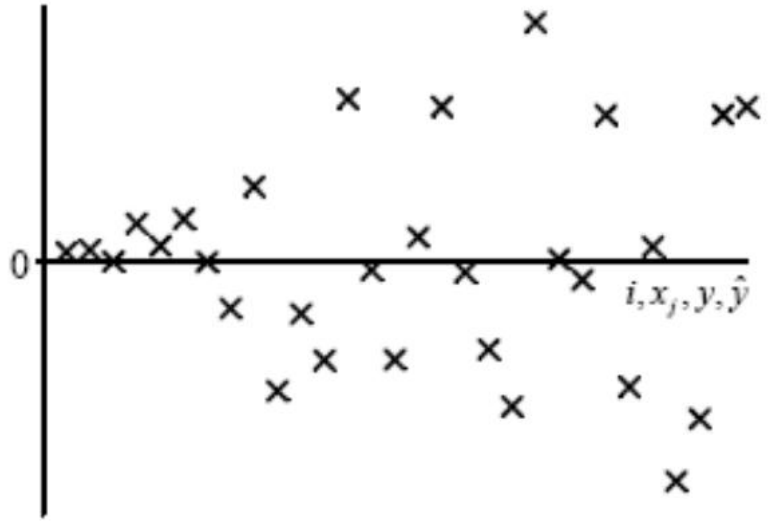
- ✓ Nei *software* statistici è disponibile un'amplia diagnostica per verificare le caratteristiche dei residui, al fine di poter fare inferenza sui dati campionari
- ✓ In questa sede ci accontentiamo di analizzare graficamente i residui, per poter comprendere le caratteristiche della loro distribuzione e c'è qualche segnale di «*warning*»
- ✓ Per verificare la **normalità dei residui si può guardare il grafico *q-q normal plot*** (che confronta i quantili della distribuzione empirica dei residui standardizzati con quelli di una normale *standard*). Se le distribuzioni sono simili la condizione di normalità è rispettata



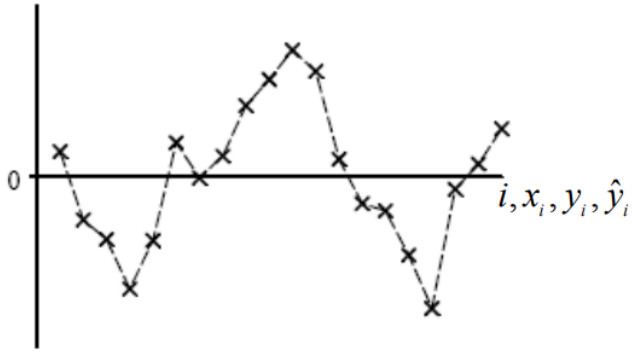
Analisi grafica dei residui 2/5

- ✓ **Un altro modo per verificare la normalità** della distribuzione empirica dei residui è **porre i residui standardizzati su un grafico**. Dobbiamo aspettarci si distribuiscano in modo casuale attorno allo zero, compresi tra -/+ 1,96 * errore standard dei residui
- ✓ **Se i residui dovessero distribuirsi in modo non casuale (**per esempio più grande è il valore della variabile dipendente, maggiore è l'errore standardizzato), allora siamo probabilmente in presenza di **eteroschedasticità, che «viola» la condizioni di errori con varianza costante**

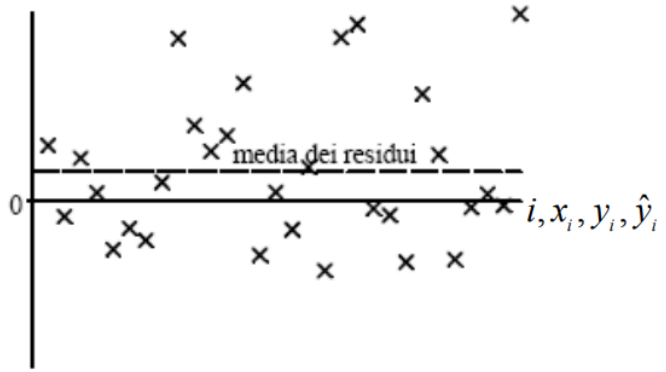




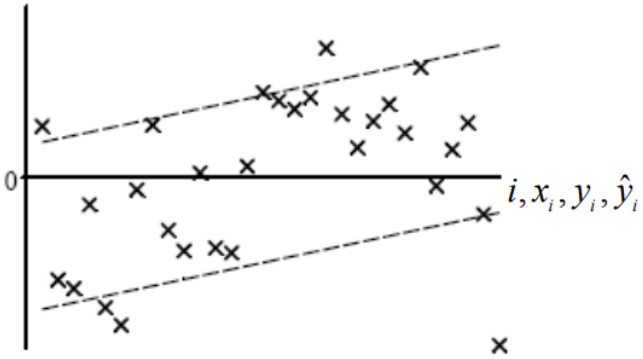
Analisi grafica dei residui 3/5



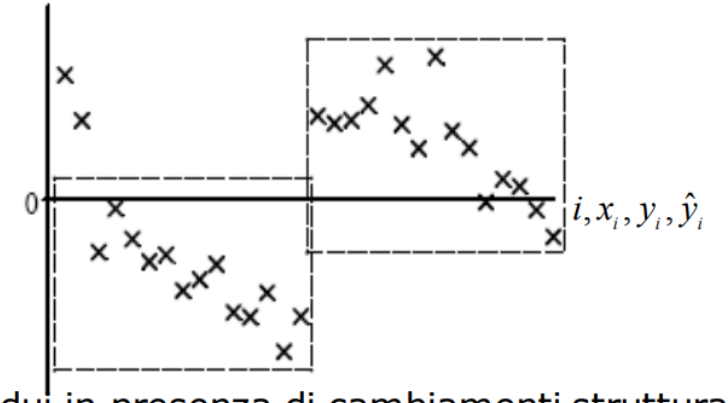
Residui in presenza di autocorrelazione



Residui nel caso di una stima del modello senza intercetta

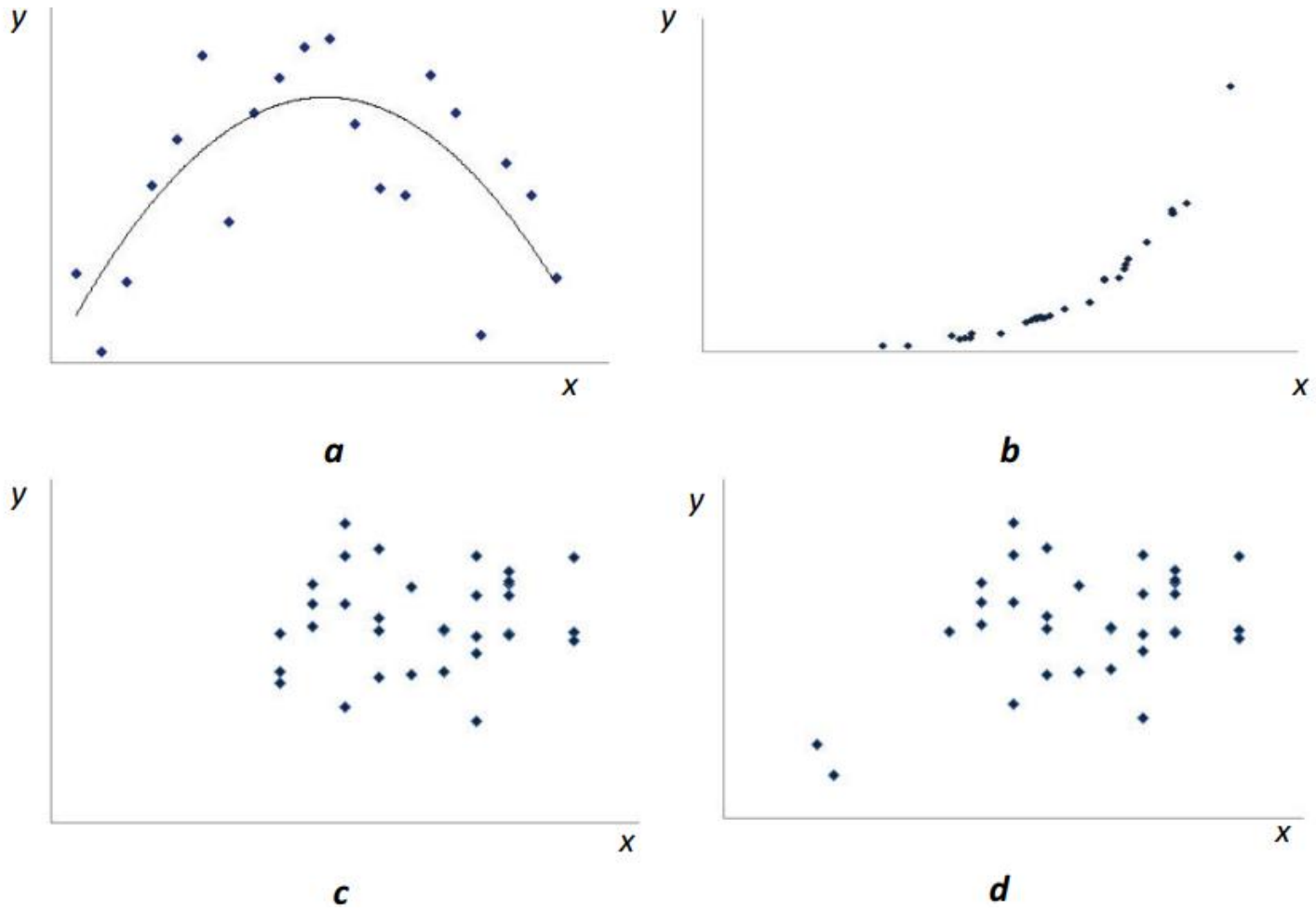


Residui nel caso di una stima del modello con una variabile esplicativa omessa



Residui in presenza di cambiamenti strutturali

Analisi grafica dei residui 4/5



Legenda:

(a) relazione quadratica

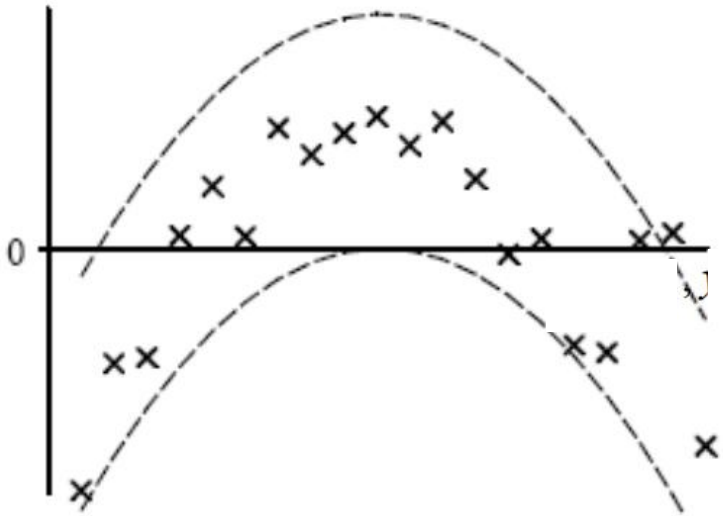
(b) relazione esponenziale

(c) relazione quasi nulla

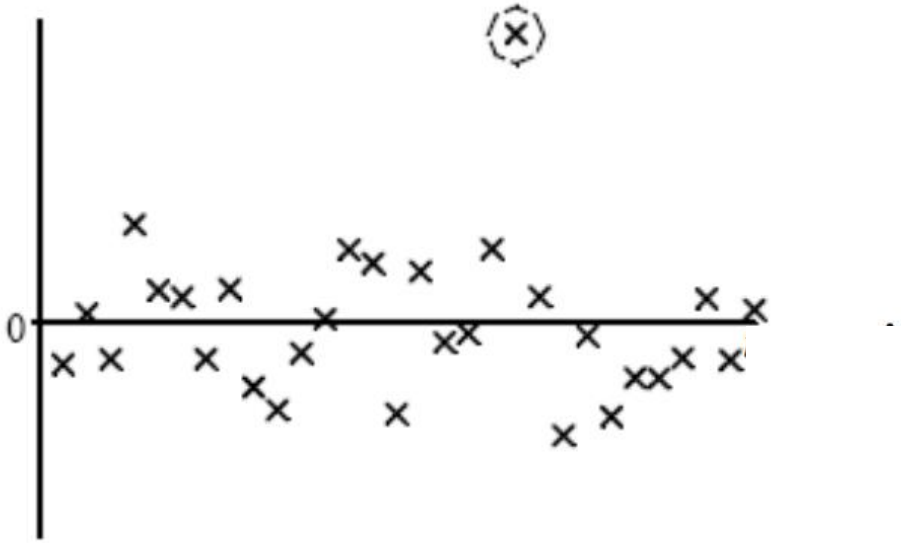
(d) relazione quasi nulla con la presenza anche di valori anomali

Analisi grafica dei residui 5/5

Residui in presenza di una relazione non lineare



Valori anomali (outliers)



Correlazione e regressione con più di due variabili

- ✓ Abbiamo sin qui analizzato l'interazione tra due variabili. Quello che abbiamo visto sinora può essere esteso anche a più di due variabili
- ✓ La correlazione multipla non è che l'insieme di tutte le correlazioni tra variabili, prese due a due
- ✓ La **regressione multipla** ha, invece, caratteristiche più utili:
  - **L'effetto di tutte le variabili esplicative** incluse viene **valutato congiuntamente**
  - I **coefficienti** pertanto rappresentano **effetti parziali**, nel senso che rappresentano l'effetto di quella variabile tenendo costanti tutte le altre, ovvero, nel modello lineare:

$$y_t = \alpha + \beta x_t + \gamma z_t + u_t$$

  - I coefficienti $\beta = \frac{dy}{dx}$ e $\gamma = \frac{dy}{dz}$ rappresentano gli effetti parziali di ogni variabile esplicativa
  - Si dice pertanto che **nella regressione multipla l'effetto di ogni variabile viene valutato controllando per quello di ogni altro**

Regressione semplice e regressione multipla

- ✓ Possiamo scrivere il **generico modello di regressione lineare**, con un numero di variabili esplicative che va da 1 (regressione semplice) a *k*

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_k x_{kt} + u_t$$

- ✓ La **specificazione della regressione multipla è un'estensione di quella della regressione lineare semplice**. Le proprietà ipotizzate per il termine di errore sono, *mutatis mutandis*, le stesse
- ✓ La **bontà di adattamento** si valuta sempre con R^2, ma vedremo che assumerà importanza una versione *adjusted*
  - Siamo tentati di scegliere il modello che ha il valore di R^2 maggiore
  - In un modello di regressione multipla l'R^2 tende a crescere ogni volta che aggiungiamo una variabile, anche se essa non significativa per spiegare la variabilità del fenomeno che stiamo analizzando
  - Si definisce allora un **R^2 corretto (*adjusted*)**. Indicando con *n* il numero di osservazioni e *k* il numero delle variabili esplicative risulta:

$$\bar{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

  - Una nuova variabile può aumentare RSS, ma al tempo stesso aumenta il valore *(n-k-1)*. **L'R^2 corretto aumenterà solo se il guadagno in precisione delle stime è superiore alla correzione per la riduzione dei gradi di libertà**
- ✓ La significatività di ogni variabile X_j può essere valutata con il *t – test* visto per la regressione semplice

Ordinary Least Squares

- ✓ Come nel mondo univariato, il metodo di stima più comune è quello dei minimi quadrati (OLS - *ordinary least squares*)
- ✓ Si minimizzano i quadrati delle distanze
- ✓ Più formalmente, siano:
  - y_t i valori osservati per ogni t
  - $\hat{y}_t$ i valori osservati per ogni t
  - $\hat{u}_t$ i residui, $\hat{u}_t = y_t - \hat{y}_t$

- ✓ Operativamente devo cercare quei valori di β_0 e $\beta_1, \beta_2, \ldots, \beta_k$ tali da rendere minima la somma dei quadrati dei residui (la ns funzione di perdita)

$$L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k) = \sum (y_t - \hat{y}_t)^2$$

- ✓ Questo calcolo richiede l'algebra delle matrici per risolvere il sistema, che non affronteremo in questa sede

Stima di parametri con il metodo dei minimi quadrati nella regressione multipla

- ✓ A livello indicativo, in un modello con due variabili

$$y_t = b_0 + b_1 x_1 + b x_2 + u_t$$

- ✓ Le espressioni degli stimatori dei parametri β_0, β_1 e β_2 sono le seguenti:

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

$$b_1 = \frac{Cov(x_1, y)Var(x_2) - Cov(x_2, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \left[Cov(x_1, x_2)\right]^2}$$

$$b_2 = \frac{Cov(x_2, y)Var(x_1) - Cov(x_1, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \left[Cov(x_1, x_2)\right]^2}$$

- ✓ Mentre lo stimatore corretto della varianza degli errori è:

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-3}$$

Variabili omesse nella regressione multipla

- ✓ **Se il «vero» processo** di generazione dei dati (**DGP**) **contiene una certa variabile**

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

- ✓ **Ma il modello che andiamo a stimare omette una variabile esplicativa**:

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \hat{\beta}_3 x_{3t\ t}$$

- ✓ **Quest'ultimo risulta incompleto**. Le conseguenze:
  - Gli stimatori dei parametri relativi ai regressori inclusi $\hat{\beta}_1$ e $\hat{\beta}_2$ sono distorti e inconsistenti (si veda *slide 125*) a meno che la variabile omessa x_2 sia incorrelata con entrambe x_1 e x_3, cosa che nella pratica succede raramente
  - I relativi ES sono a loro volta distorti e inconsistenti
  - Tutte le statistiche diagnostiche sono inconsistenti
  - … insomma il modello è da buttare

Variabili ridondanti nella regressione multipla

✓ **Se il «vero» processo** di generazione dei dati (**DGP) non contiene una certa variabile**

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$$

✓ **Ma il modello che andiamo a stimare la include**:

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + \hat{\beta}_3 x_{3t}$$

✓ **Quest'ultimo risulta sovraparametrizzato**. Le conseguenze:

- Sub 1)-4) (vedi *slide 125*) gli stimatori $\hat{\beta}_j$ sono *BLUE*
- Il coefficiente della variabile ridondante verrà quindi stimato per quello che è, ovvero zero
- **Tutte le statistiche diagnostiche sono valide**, incluso il *t-test* di significatività della variabile ridondante
- … sulla base di tali statistiche **potremo semplificare il modello**

Significatività congiunta delle variabili 1/2

- ✓ Una volta stimato il modello è importante **verificare l'esistenza o meno di un legame lineare tra la variabile dipendente *y* e le variabili indipendenti $x_1, x_2, \ldots, x_k$**
- ✓ Quello che facciamo (**test *F* globale**[1]) è verificare l'ipotesi:

$$H_0: \beta_1 = \beta_2 = \ldots. = \beta_k = 0$$

- ✓ Contro l'ipotesi alternativa, che almeno un β_i sia diverso da zero

$$H_1: \text{almeno un } \beta_i \neq 0$$

- ✓ La **statistica test per questa verifica di ipotesi** è la seguente (con *n* numero di osservazioni e *k* numero di regressori), che si distribuisce come una *F* di Fisher con *k, n-k-1* gradi di libertà:

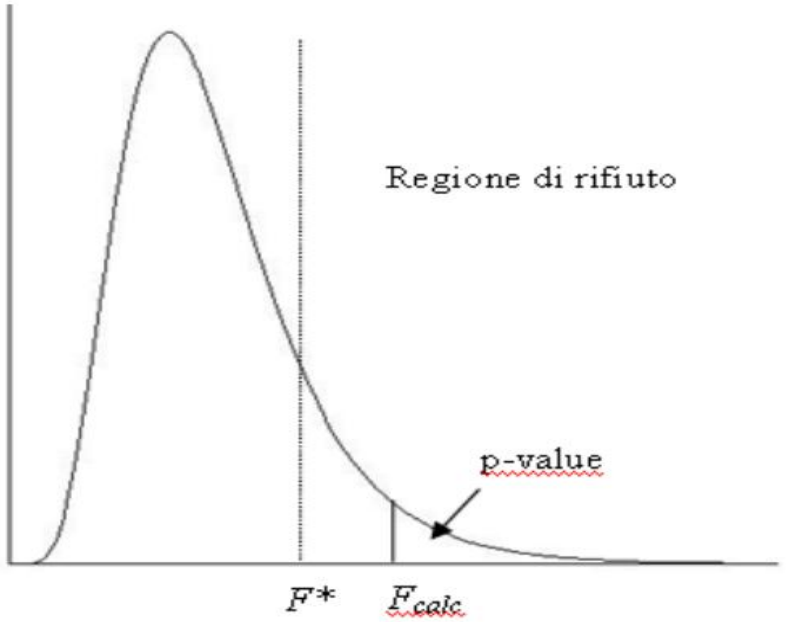
$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \quad \sim F_{k,\, n-k-1}$$

- ✓ Si può osservare che al crescere di R^2, la statistica *F* aumenta (e come vedremo la probabilità di rifiutare l'ipotesi H_0 aumenta)

[1] Si tratta di un caso particolare del test F parziale che confronta la bontà di modelli annidati per testare l'opportunità o meno di considerare variabili aggiuntive

Significatività congiunta delle variabili 2/2

- ✓ A questo punto (fissato il nostro α) calcoliamo il valore della nostra statistica F_{calc} e la confrontiamo con il valore critico della distribuzione $F_{k, n-k-1}$. Se finisce a destra del valore critico, rifiuteremo l'ipoteso H_0 e dunque almeno un valore β_i sarà diverso da zero
- ✓ La verifica delle ipotesi può anche essere fatta osservando il *p-value* a destra del valore calcolato della statistica, confrontandolo con il valore di α (così spesso i *software*)



Statistica per l’impresa

7. L’analisi delle serie storiche per la programmazione delle attività

L’ANALISI DELLE SERIE STORICHE IN AZIENDA
- ✓ Serie storiche univariate
- ✓ Scopo dell’analisi delle serie storiche in azienda
- ✓ Componenti della serie storica e fasi dell’analisi

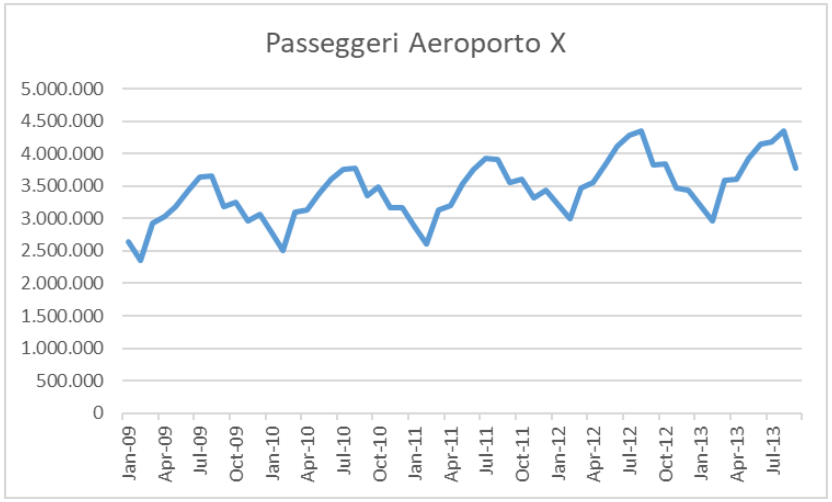
ANALISI PRELIMINARI
- ✓ Rappresentazione grafiche
- ✓ Coefficiente di autocorrelazione e correlogramma
- ✓ Valutazione della capacità previsiva

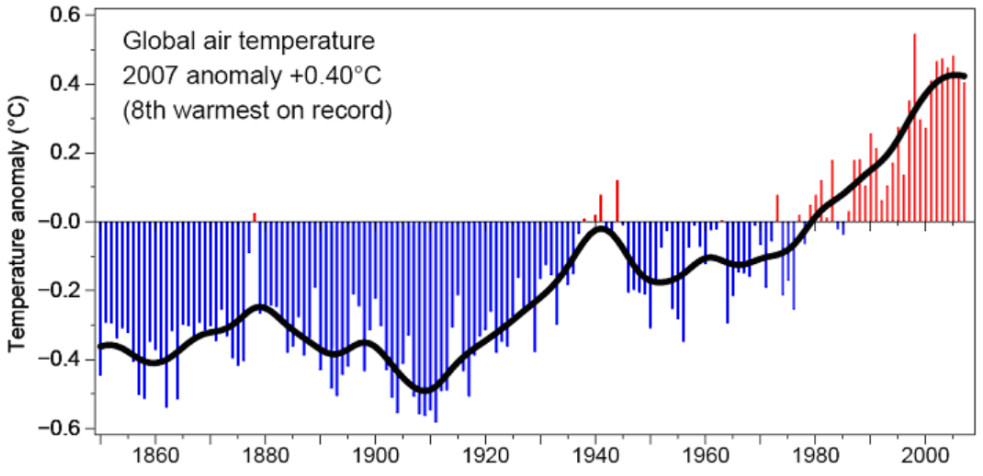
METODI DI SCOMPOSIZIONE E STIMA DELLE COMPONENTI
- ✓ Metodi di stima delle componenti
- ✓ Le medie mobili
- ✓ Stima analitica del trend

LIVELLAMENTO ESPONENZIALE
- ✓ Livellamento esponenziale semplice
- ✓ Metodo di Holt e Winters

Serie storiche univariate

- ✓ Una **successione di dati osservati su una variabile *Y* nel tempo** (i passeggeri mensili di un aeroporto, le vendite trimestrali di un prodotto, i rendimenti settimanali di un titolo azionario,....) :





Source: University of EA

- ✓ I metodi per le **serie storiche** affrontano la **modellazione**, generalmente a fini previsivi, dell'evoluzione di una **variabile** di interesse **nel tempo**
- ✓ Mentre nel caso della regressione lineare la variazione nella variabile obiettivo, *Y*, veniva «spiegata» sulla base della variazione si una o più variabile esplicative X_1, ... X_k, qui viene «**spiegata dal suo andamento passato**»
- ✓ Ci occuperemo di serie storiche relative ad **una sola variabile (univariate)**.
- ✓ L'analisi delle serie storiche multivariate è parente stretto della regressione lineare, in cui la serie Y_t dipende da un'altra serie X_t e dal proprio passato Y_{t-1}, ..., Y_{t-h}

Scopo dell'analisi delle serie storiche

- ✓ Comprendere l'andamento di una serie storica può essere importante ai **fini interpretativi**, ma spesso è essenziale ai fine della **previsione**. In azienda sono regolarmente oggetto di previsione:
  - La domanda di prodotti finiti
  - Il fabbisogno di risorse umane
  - Il fabbisogno di materie prime
  - Le scorte
  - Le principali voci di bilancio
  - Budget
  - .....
- ✓ La pianificazione e controllo in un'impresa passa inevitabilmente la proiezione delle principali grandezze che esprimono le varie fasi che caratterizzano la creazione di valore (le vendite, i costi, ...)
- ✓ Le previsioni possono essere:
  - **breve termine** (fino a 12 mesi)
  - **medio termine** (da 12 a 24 mesi)
  - **lungo termine** (oltre24 mesi)
- ✓ Inutile dire che l'affidabilità delle previsioni si riduce all'aumentare del loro orizzonte temporale. Le serie univariate che vedremo sono utili in chiave di previsioni di breve periodo

Cosa è una serie storica

- ✓ Una **successione di dati osservati su una variabile *Y* nel tempo** (i passeggeri mensili di un aeroporto, le vendite trimestrali di un prodotto, i rendimenti settimanali di un titolo azionario,....) :

$$Y_t, \quad t=1, \ldots\ldots, T$$

- ✓ I dati possono essere misurati:
  - in un istante (serie di stato, per esempio il numero di dipendenti alla fine di ogni mese)
  - in un intervallo (serie di flusso, per esempio le vendite di ogni settimana)

- ✓ In una serie storica è **comune la presenza di dipendenza tra ogni osservazione e le sue passate determinazioni** ed è questo «**modello di comportamento» che cerchiamo di catturare**, anche se può succedere che manifestazioni successive di un fenomeno siano tra loro indipendenti o con una scarsa «correlazione» (*random walk*)

(Possibili) componenti di una serie storica

✓ Le serie storiche possono presentare le seguenti **componenti:**

- ***Trend***: movimento tendenziale di fondo dovuto all'evoluzione di lungo periodo del fenomeno
- **Ciclo**: oscillazione congiunturale di carattere ricorrente, spesso dovuto all'oscillare di un sistema economico attorno alle condizioni di equilibrio
- **Stagionalità**: regolarità empirica legata ai periodi dell'anno e dovuta a fattori climatici (alternanza delle stagioni) oppure organizzativi (ferie, festività)
- **Accidentalità**: componente residuale rispetto alle cause strutturali 1)-3), in genere relativa a molte influenze di piccola entità o comunque non chiaramente identificabili né suscettibili di modellazione esplicita (v. errori del modello *OLS*)

✓ Le **prime tre componenti** costituiscono la cosiddetta ***parte sistematica*** della serie

I possibili approcci e le fasi dell'analisi

✓ Si distinguono **due approcci** all'analisi delle serie storiche a fini previsivi:

- **Classico: scomposizione della serie nelle componenti** sopra descritte (sola parte sistematica) e proiezione di ciascuna separatamente
- **Moderno: considera il processo Y_t come un tutt'uno** di carattere stocastico da modellare con tecniche probabilistiche

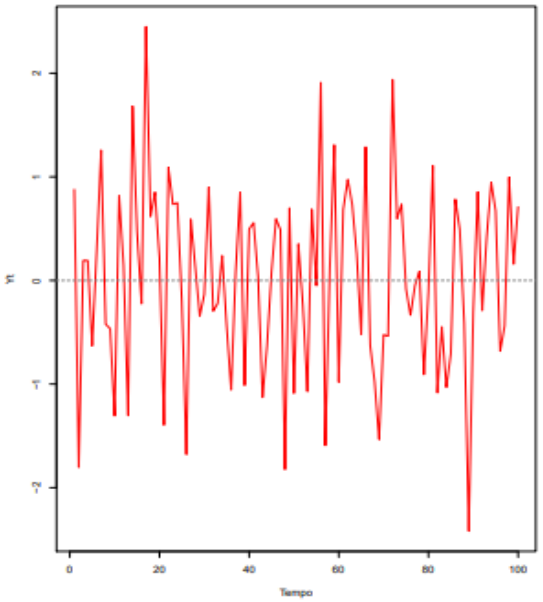
✓ Le fasi di un'analisi volta alla previsione saranno:

- analisi del problema
- raccolta dati
- analisi preliminare della struttura della serie storica
- scelta e stima del modello
- valutazione della bontà del modello a fini previsivi (a breve, a medio e a lungo termine)
- messa in «produzione» (utilizzo in pratica)

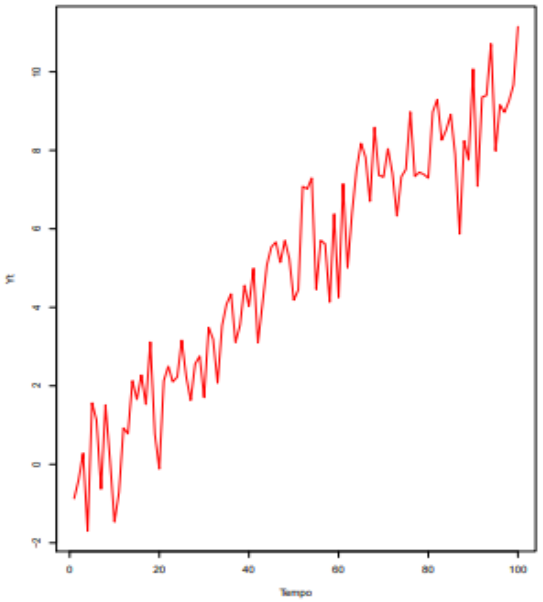
Rappresentazione grafica delle serie storiche 1/2

- ✓ La **prima cosa** da fare quando vogliamo analizzare la serie storica, per verificare l'andamento e l'eventuale presenza di oscillazioni e dele componenti descritte in precedenza, è la **rappresentazione grafica**
- ✓ In genere si usa costruire un grafico su tutta la serie (con il tempo sull'asse delle *x* o **Time Plot**)
- ✓ Se nel periodo osservato il livello della serie oscilla **attorno ad un valore** che rimane grosso modo lo stesso, la serie si dice **stazionaria in media**. Se invece si vede un ***trend***, allora si dice che la serie è **evolutiva**

Stazionaria

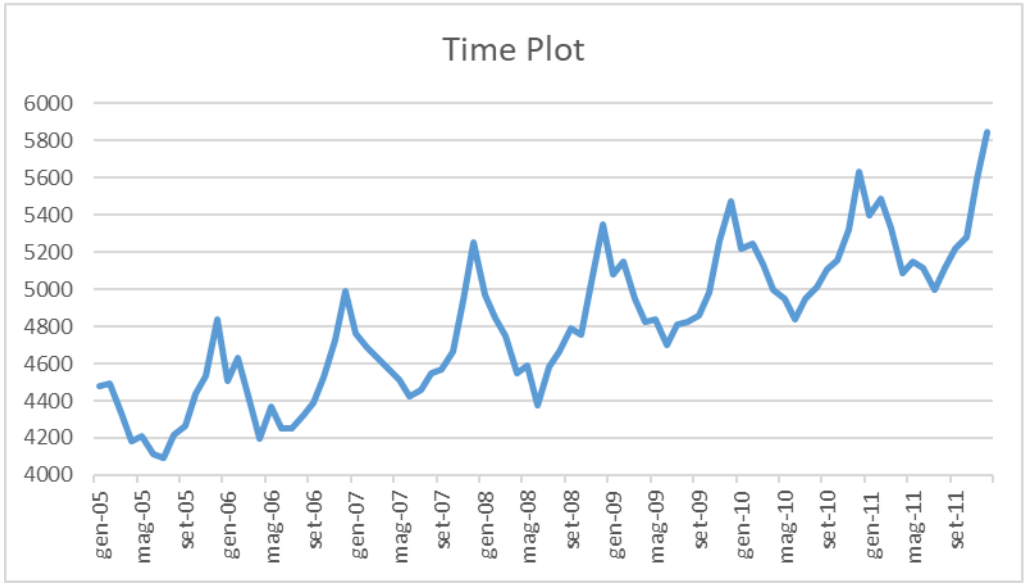


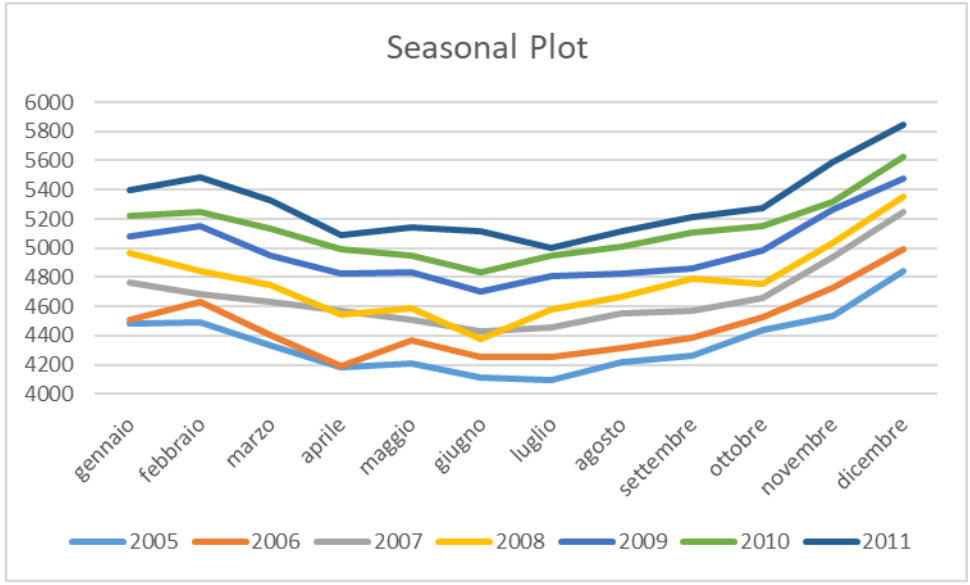
Evolutiva



Rappresentazione grafica delle serie storiche 2/2

- ✓ Per poter vedere meglio se vi è una **regolarità stagionale** si costruisce il cosiddetto ***Seasonal Plot***, che riporta, una linea per anno, sull'asse delle *X* la periodicità stagionale (settimanale, mensile, trimestrale, …) e sulle *Y* i valori delle serie





Grafici tratti dall'esempio 7.1 pag. 301 libro. Si veda parte pratica in excel

Coefficiente di autocorrelazione 1/2

- ✓ Per valutare compiutamente l'esistenza di componenti stagionali (eventualmente cicliche) è utile calcolare i **coefficienti di autocorrelazione** e la relativa stima della **funzione di autocorrelazione**
- ✓ Introduciamo prima il concetto di **ritardo (*lag*):** in ogni istante *t* il ritardo *k-esimo* di Y_t è Y_{t-k}
- ✓ Se per esempio consideriamo un ritardo di *k=1* sulla serie osservata:

$$y = y_1, y_2, y_3, \ldots, y_t$$

- ✓ Diamo «origine» ad una nuova serie:

$$y_{-1} = na, y_1, y_2, \ldots, y_{t-1}$$

- ✓ Il coefficiente di autocorrelazione di Y_t a ogni ritardo k viene stimato come **correlazione campionaria di Y_t** e $\boldsymbol{Y_{t-k}}$

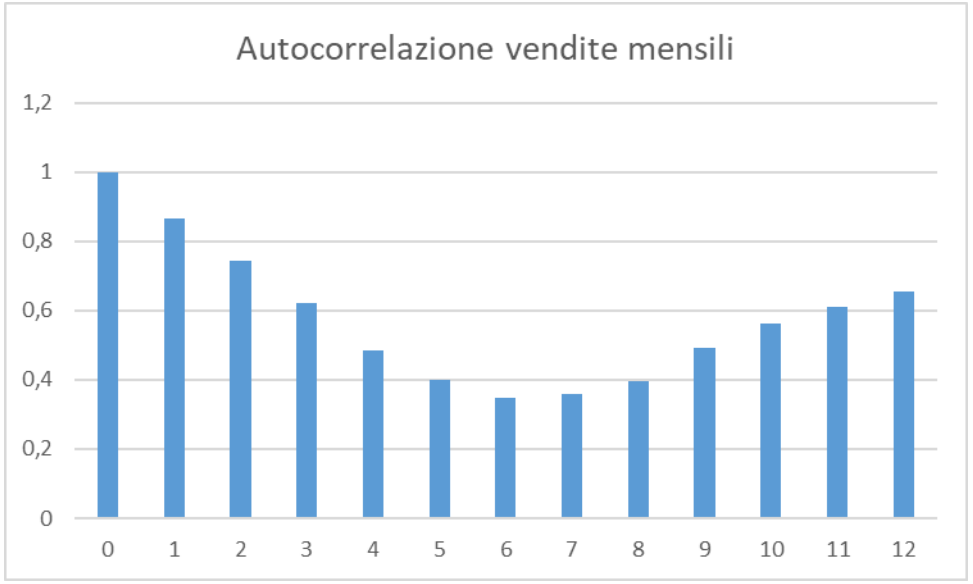
$$\rho_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

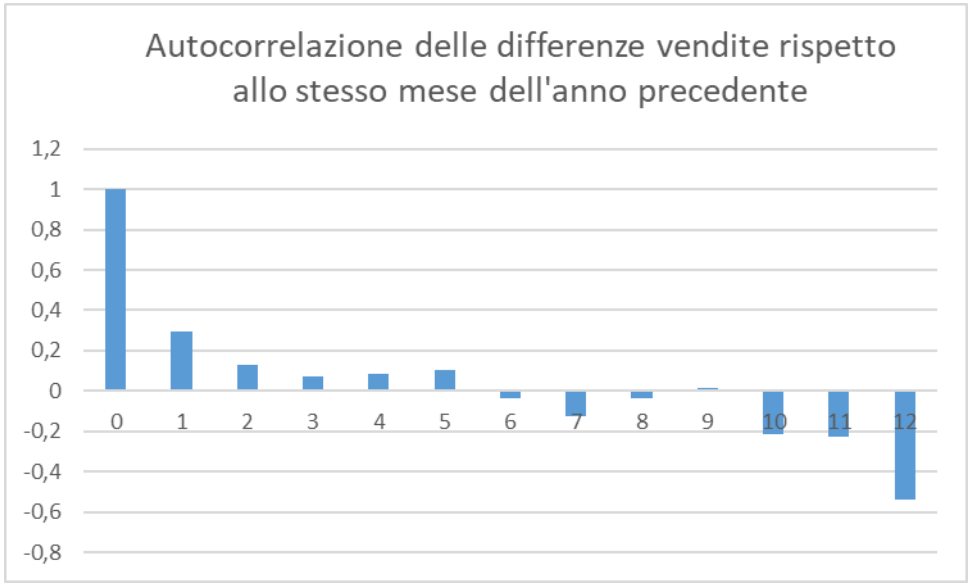
- ✓ La **stima dei vari ρ_k per *k=1,..., T*** dà luogo alla **funzione di Autocorrelazione (ACF)** empirica

Si veda parte pratica in excel

Coefficiente di autocorrelazione 2/2

- ✓ Sotto due esempi di funzione di autocorrelazione empirica, tratte dall'esempio 7.1 pag. 301 del libro
- ✓ Nel primo caso osserviamo che ogni ritardo appare significativo, con intensità via via decrescenti: è verosimilmente un indizio che siamo di fronte ad una serie con un *trend*
- ✓ Nel secondo caso invece vediamo che dopo il ritardo *1*, l'autocorrelazione si avvicina allo zero, per poi crescere nuovamente a *-12*. Si tratta di una serie probabilmente stazionaria in media





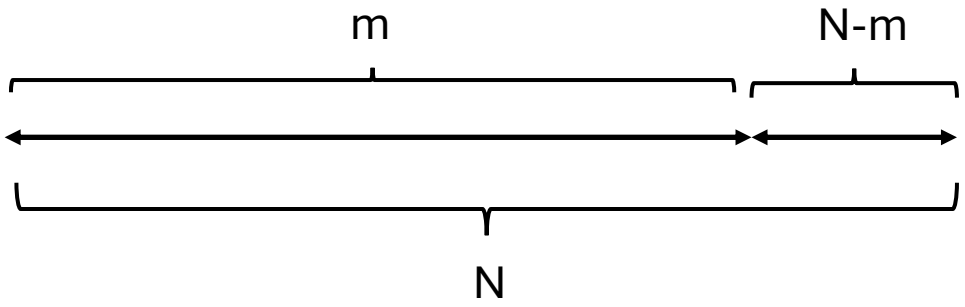
Si veda parte pratica in excel

La valutazione della capacità previsiva 1/2

- ✓ Supponendo di aver stimato un **modello univariato**

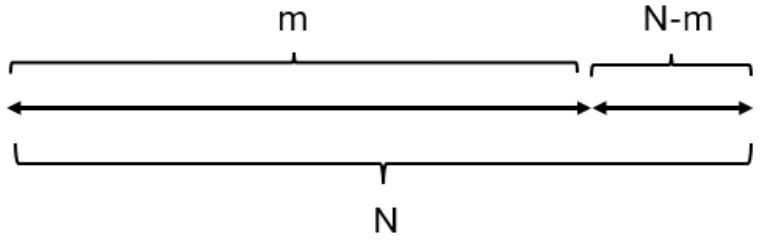
$$Y_t = f(Y_1, \ldots., Y_{t-1})$$

- ✓ L'analisi della **bontà di adattamento** del modello confronta **i valori stimati (o previsti) $\hat{y}_t$, con quelli effettivamente osservati y_t**
- ✓ I due aspetti della bontà di adattamento che vogliamo catturare sono:
  - ***Goodness of fit*:** la capacità del modello di riprodurre i dati storici
  - ***Goodness of forecast***: la capacità del modello di prevedere i dati futuri
- ✓ Se ci poniamo su tutta la dimensione della serie storica (*N*), non abbiamo possibilità di valutare la *goodness of forecast*, se non quando saranno disponibili i nuovi dati (*N+1, N+2, ….).* Quello che si fa nella pratica è troncare il periodo di osservazione ad un certo periodo (*m*) e da quel punto fare le previsioni sino a *N* (*ex post*)



La valutazione della capacità previsiva 2/2

- ✓ Quello che faremo è dunque:
  - Per valutare la ***goodness of fit* (*in-sample***) si confronteranno le stime $\hat{y}_1, \ldots, \hat{y}_m$, con i valori osservati $y_1, \ldots, y_m$
  - Per valutare la ***goodness of forecast* (*out-of-sample*)** si confronteranno le stime $\hat{y}_{m+1}, \ldots, \hat{y}_N$, con i valori osservati $y_{m+1}, \ldots, y_N$



- ✓ Per valutare **l'adattamento *in-sample*** si possono usare vari indicatori:
  - *Mean Error* (ME): $ME = \frac{1}{m}\sum_{t=1}^{m}(y_t - \hat{y_t})$
  - *Mean Square Error* (MSE): $MSE = \frac{1}{m}\sum_{t=1}^{m}(y_t - \hat{y}t)^2$
  - *Mean Absolute Error* (MAE): $MAE = \frac{1}{m}\sum_{t=1}^{m}|y_t - \hat{y}_t|$
  - *Mean Absolute Percentage Error* (MAPE): $MAPE = \frac{1}{m}\sum_{t=1}^{m}\frac{|y_t - \hat{y_t}|}{y_t}$
- ✓ Per valutare **l'adattamento *out-of-sample*** si possono usare vari indicatori:
  - *Mean Error* (ME): $ME = \frac{1}{N-m}\sum_{t=m+1}^{N}(y_t - \hat{y}_t)$
  - *Mean Square Error* (MSE): $MSE = \frac{1}{N-m}\sum_{t=m+1}^{N}(y_t - \hat{y}_t)^2$
  - *Mean Absolute Error* (MAE): $MAE = \frac{1}{N-m}\sum_{t=m+1}^{N}|y_t - \hat{y}_t|$
  - *Mean Absolute Percentage Error* (MAPE): $MAPE = \frac{1}{N-m}\sum_{t=m+1}^{N}\frac{|y_t - \hat{y_t}|}{y_t}$

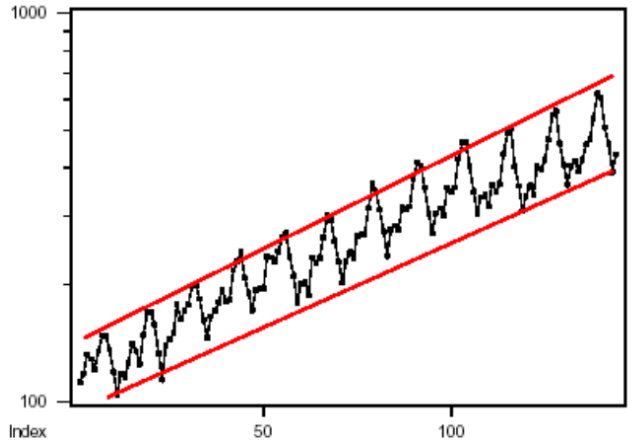
Modelli di scomposizione delle serie storiche

- ✓ L'approccio classico ipotizza che la serie storica sia generata come una funzione delle componenti già viste alla *slide 152*:

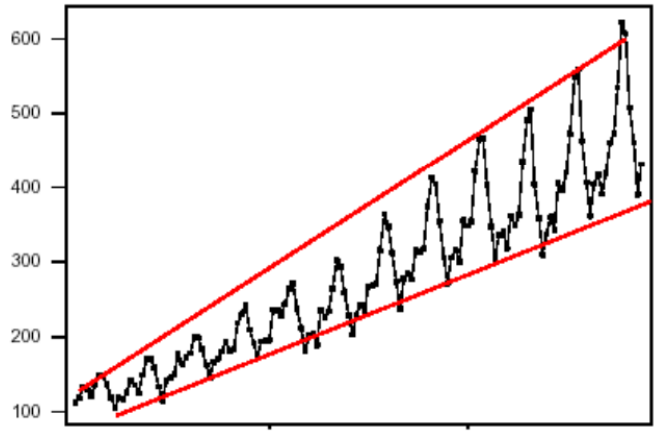
$$Y_t = f(T_t, C_t, S_t, e_t)$$

- ✓ La **parte deterministica** consiste nel **trend (*T*), ciclo (*C*) e stagionalità (*S*),** mentre ***e* è un disturbo aleatorio**
- ✓ Stimare la componente ciclica (*C*) è molto complesso e richiede moltissime osservazioni. Ci si accontenta di stimarla assieme al trend (*T*)
- ✓ La **componente deterministica** *f* può assumere diverse forme funzionali:

Additiva: $Y_t = T_t + S_t + e_t$



Moltiplicativa: $Y_t = T_t * S_t * e_t$



- ✓ Un modello **<u>moltiplicativo può sempre essere linearizzato</u>** con una **trasformazione logaritmica**

$$\ln(Y_t) = \ln(T_t) + \ln(S_t) + \ln(e_t)$$

Metodi di stima delle componenti

- ✓ **Le componenti trend (*T*) e stagionalità (*S*)** possono essere **stimate con metodi:**
  - **Empirici (perequativi):** consistono in un «lisciamento» che si adatta ai dati del campione permettendo di isolare una componente in-sample, ma **non permette di prevederla**
  - **Analitici (interpolativi):** consistono nella scelta di una **funzione analitica di cui stimare i parametri**, la quale **si può usare per prevedere** le singole componenti

Le medie mobili 1/2

- ✓ Le **medie mobili** sono una **metodologia** statistica per «**lisciare» le oscillazioni di una serie**, mettendone **in evidenza le componenti sistematiche**
- ✓ Sia *k* il numero di termini della media che andiamo a calcolare. **Se *k* è dispari** ciascuna media mobile si riferisce al tempo centrale rispetto a cui è calcolata. Vediamo un esempio per *k=3*

$$MM_3(y_t) = (y_{t-1} + y_t + y_{t+1}) / 3$$

- ✓ In sostanza è **una media di *k* termini centrati su y_t**
- ✓ Formalmente è:

$$MM_k(y_t) = \frac{\sum_{s=t-(k-1)/2}^{t+(k-1)/2} y_s}{k}$$

Si veda parte pratica in excel

Le medie mobili 2/2

✓ **Se *k* è pari** si calcola la media di due medie mobili contigue. Vediamo un esempio per *k=4*

- Si calcolano le seguenti medie mobili:

$$MM_4(y_{t-1,\,t}) = (y_{t-2} + y_{t-1} + y_t + y_{t+1}) / 4$$

$$MM_4(y_{t,\,t+1}) = (y_{t-1} + y_t + y_{t+1} + y_{t+2}) / 4$$

- Si fa la media aritmetica delle due medie:

$$MM_4(y_t) = (y_{t-2} + 2y_{t-1} + 2y_t + 2y_{t+1} + y_{t+2}) / 8$$

$$\boxed{MM_4(y_t) = (0{,}5y_{t-2} + y_{t-1} + y_t + y_{t+1} + 0{,}5y_{t+2}) / 4}$$

- In generale bisogna sommare i *k/2* termini prima del periodo *t* ed i *k/2* dopo il periodo *t*, pesando i termini estremi per *0,5* (dividendoli cioè per *2*) e dividere la somma per *k*

$$\boxed{MM_{12}(y_t) = (0{,}5y_{t-6} + y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2} + y_{t+3} + y_{t+4} + y_{t+5} + 0{,}5y_{t+6}) / 12}$$

✓ Il calcolo delle **medie mobili fa «perdere» *k/2* periodi** all'inizio e alla fine

Si veda parte pratica in excel

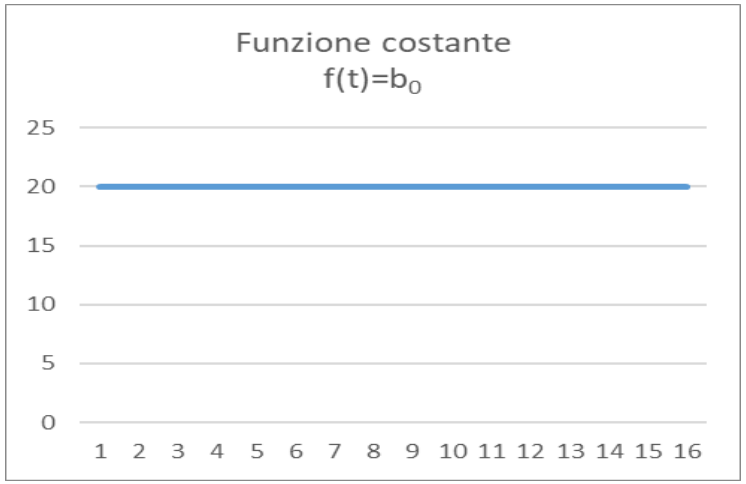
Stima del trend, della la stagionalità e della serie destagionalizzata con le medie mobili

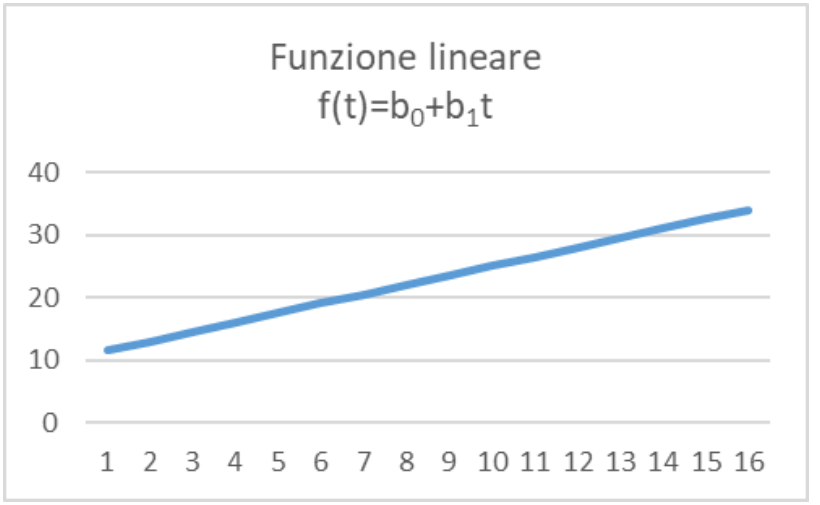
- ✓ Le **medie mobili dunque eliminano o riducono le oscillazioni** di periodo. Le useremo **per calcolare il trend (ciclo) della serie** che stiamo analizzando
- ✓ **L'intervallo (*k*)** su cui calcoliamo la media mobile **dipende dalla stagionalità** che stiamo osservando. Per dati mensili userò *k=12*, trimestrali *k=4*, ...
- ✓ **Per ottenere una stima *in-sample* di *T* ed *S*** si può procedere come segue (**modello additivo**):
  - **Calcolo delle medie mobili** come **prima stima** della componente di trend **(*T**)**
  - **Ricavo la componente stagionale ed accidentale** calcolando gli $y_t - T^*_t$
  - Ipotizzando una stagionalità costante e **per eliminare la componente accidentale**:
    - Si **calcola la media delle stagionalità grezze** dei vari anni per ogni periodo ottenendo ***k* coefficienti S^*_j**
    - Si **verifica che la media degli S^*_j sia zero** (principio della conservazione delle aree, altrimenti si centrano sottraendo la media)
    - Abbiamo finalmente **ottenuto i coefficienti netti di stagionalità**
- ✓ Per ottenere la **stima «definitiva» del trend-ciclo T^*_t**
  - Si deriva la **serie destagionalizzata $D_t = y_t - S^*_t$**
  - **Si stima il trend** eliminando le oscillazioni casuali con **un'ulteriore media mobile** di ampiezza opportuna

Si veda parte pratica in excel

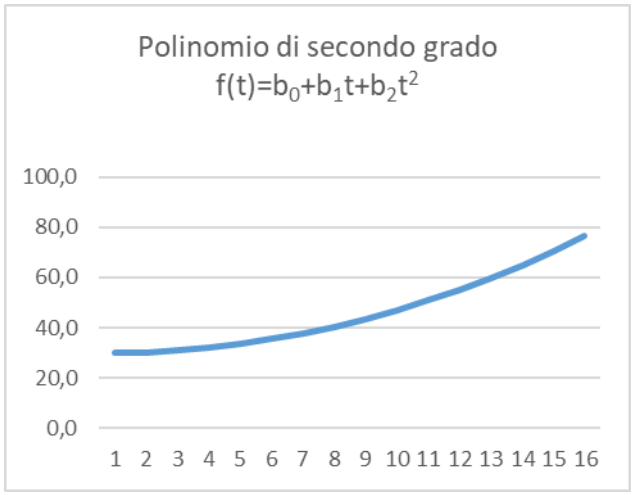
Stima analitica del trend 1/2

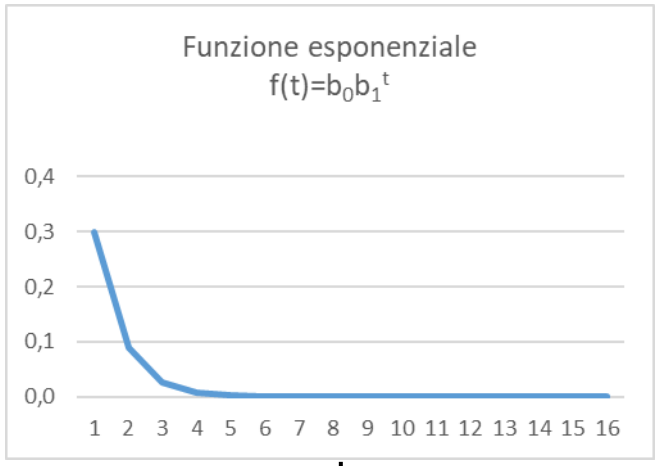
- ✓ Le medie mobili dunque sono una procedura di adattamento *in-sample,* utile ai fini interpretativi
- ✓ Fanno tuttavia perdere *k/2* periodo all'inizio e alla fine e sono inadatte alla previsione out-of-sample
  - ✓ Una **procedura alternativa è la stima del trend T^*_t con metodi analitici**, da cui **ottenere $S_t=y_t-T^*_t$.** Si **calcolano poi i coefficienti di stagionalità** e si può poi **proiettare la serie *out-of-sample***, sommando (se siamo nel caso additivo) **stagionalità e trend-ciclo**
- ✓ Cosa vuol dire stima del trend con **metodi analitici?** Significa **assegnare possibili forme funzionali al trend** stesso **in funzione del tempo**
  - ✓ Se stimiamo un trend con un modello OLS vanno applicate tutte le considerazioni viste nel capitolo 6, in termini di «bontà» della stima

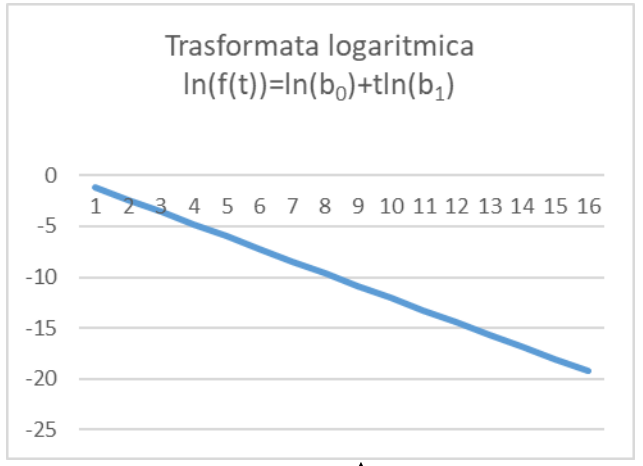




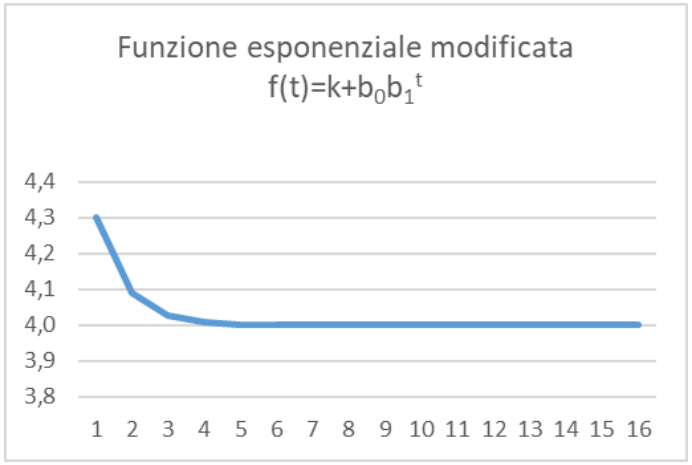
Stima analitica del trend 2/2

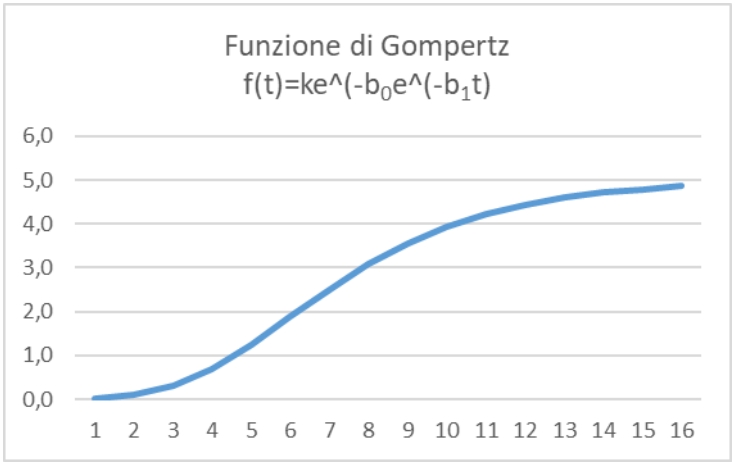






Linearizzata con i logaritmi





Si veda parte pratica in excel

Livellamento esponenziale 1/2

- ✓ Il **livellamento esponenziale** nasce nel 1957 come metodo pragmatico per la **previsione delle serie storiche basato sulle medie mobili**. In seguito è stato giustificato anche nel quadro della teoria moderna delle serie storiche come caso particolare dei modelli *ARMA/ARIMA*

- ✓ Si supponga di voler prevedere y_{t+1}, disponendo al tempo t, di una serie di osservazioni

$$y_{t-n}, y_{t-n+1}, .., y_{t-3}, y_{t-2}, y_{t-1}, y_t$$

- ✓ Si potrebbe pensare di ricorrere ad una media mobile «all'indietro» di alcuni termini

- ✓ **L'idea di base** del livellamento esponenziale è di **modificare l'approccio delle medie mobili attribuendo più importanza alle osservazioni più recenti** ed in particolare all'ultima (y_t)

- ✓ E' un metodo che si adatta bene a **previsioni di breve periodo**

Livellamento esponenziale 2/2

- ✓ Il **livellamento esponenziale costante o semplice** parte dall'ipotesi che la **serie sia stazionaria in media**
  - Dato che la **serie è stazionaria** in media, in prima approssimazione, si potrebbe prendere come **previsore del livello in *t+1* la media aritmetica della osservazioni**, dando lo stesso peso ad ogni osservazione (anche le più lontane nel tempo)

$$\hat{y}_{t+1} = \hat{M}_t = (y_t + y_{t-1} + \ldots + y_{t-n-2} + y_{t-n-1} + y_{t-n})/n$$

  - Il **livellamento esponenziale** semplice generalizza quanto sopra **assegnando ad ogni osservazione un peso w_j** la cui somma è pari a *1*, che decresce esponenzialmente a zero man mano che il dato diventa «più vecchio». Peso dunque di più i dati recenti, rispetto a quelli più lontani nel tempo
    - Definisco w_j **attraverso un parametro** α compreso tra zero e uno

$$w_j = \alpha (1-\alpha)^j \quad j=0, \ldots, n \text{ e } 0 \leq \alpha \leq 1$$

    - Ottengo una **previsione di *y*, come media pesata** delle osservazioni disponibili **pesando di più i dati recenti**, rispetto a quelli lontani nel tempo

$$\hat{y}_{t+1} = \hat{M}_t = \alpha y_t + \alpha (1-\alpha) y_{t-1} + \alpha (1-\alpha)^2 y_{t-2} + \alpha (1-\alpha)^3 y_{t-3} + \ldots + \alpha (1-\alpha)^n y_{t-n}$$

    - Sostituendo ricorsivamente il termine più lontano, otteniamo la **formula fondamentale del livello esponenziale costante**, dove ***α*** è detto **parametro di livellamento (*smoothing*)**

$$\boxed{\hat{y}_{t+1} = \alpha y_t + (1-\alpha) \hat{y}_t}$$

    - **Maggiore è *α*, maggiore è il peso** che diamo alle **osservazioni più recenti**

Livellamento esponenziale come correzione sequenziale degli errori di previsione

✓ Si vede come il **modello si fondi su una logica di aggiornamento sequenziale**:

- La previsione ad un periodo $\hat{y}_{t+1}$ è una media dell'ultimo periodo disponibile e di tutti i precedenti, questi ultimi sintetizzati nella previsione precedente $\hat{y}_t$

- Inoltre riscrivendo la formula come

$$\hat{y}_{t+1} = \alpha\, y_t + \hat{y}_t - \alpha\, \hat{y}_t = \hat{y}_t + \alpha\,(y_t - \hat{y}_t)$$

- Si nota come **la previsione corrente $\hat{y}_{t+1}$ sia uguale alla precedente $\hat{y}_t$ modificata per l'errore di previsione** $(y_t - \hat{y}_t)$ commesso al passo precedente, moltiplicato per il parametro di smussamento α. Si adotta pertanto una logica di correzione sequenziale degli errori di previsione

Si veda parte pratica in excel

Come scegliere il parametro α ?

✓ A questo punto rimane il problema di **come attribuire un valore ad *α***

- Il criterio di ottimalità per la sua stima dovrà essere basato sull'impiego pratico del modello
- Pertanto è naturale **cercare un $\hat{\alpha}$ tale da minimizzare gli errori di previsione**, per esempio sotto forma di somma di quadrati (la stima viene ottenuta con algoritmi numerici)

$$\text{Min}_{\alpha}\ SS\ (\hat{\alpha}) = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

- Un altro problema (minore) è come inizializzare la serie di valori previsti, ovvero cosa sostituire per $\hat{y}_1$: si può usare y_1 oppure una media dei primi valori

Si veda parte pratica in excel

Il metodo Holt e Winters 1/3

- ✓ Si tratta di **un'evoluzione dell'*exponential smoothing*** visto in precedenza, nel quale si aggiungono le componenti della serie legate alla **stagionalità ed al *trend***. Per questo si adattano a serie non stazionarie in media, che ammettono:
  - un *trend*
  - una componete stagionale
- ✓ Se la serie storica ammette una tendenza di fondo localmente rettilinea, un modo di adattare il livellamento esponenziale al caso in questione è di **scomporre il valore y_{t+1}** in:
  - un **livello medio** in *t*, m_t
  - **un trend** T_t tra il tempo *t* e *t+1* medio in *t*
- ✓ In generale, su un intervallo di lunghezza θ, il valore previsto della serie al tempo $t + \theta$ sarà esprimibile come:

$$\hat{y}_{t+\theta} = \hat{m}_t + \hat{T}_{t\,\theta}$$

Si veda parte pratica in excel

Il metodo Holt e Winters 2/3

✓ Anziché stimare congiuntamente le due componenti, **si scompone il procedimento utilizzando due modelli di livellamento esponenziale**:

- uno **per il livello medio**

$$\hat{m}_t = \delta_1 y_t + (1- \delta_1)(\hat{m}_{t-1} + \hat{T}_{t-1}) \qquad \text{con } 0 \leq \delta_1 \leq 1$$

- ed uno **per il trend**

$$\hat{T}_t = \delta_2 (\hat{m}_t - \hat{m}_{t-1}) + (1- \delta_2) \hat{T}_{t-1} \qquad \text{con } 0 \leq \delta_2 \leq 1$$

- dove il primo modello ricostruisce, secondo il solito processo ricorsivo/di correzione dell'errore, il livello medio, il secondo il trend (il cui valore osservato T_t è definito come differenza tra i livelli medi

Si veda parte pratica in excel

Il metodo Holt e Winters 3/3

✓ Nel caso **ci fosse anche una componente stagionale** (*s*), ovvero

$$\hat{Y}_{t+\theta} = \hat{m}_t + \hat{T}_t\theta + \hat{S}_{t+\theta-s}$$

- Si aggiungerebbe una **terza equazione**:

Livello medio ⟹ $\hat{m}_t = \delta_1 \overline{y}_t + (1-\delta_1)(\hat{m}_{t-1} + \hat{T}_{t-1})$ con $0 \leq \delta_1 \leq 1$

Trend ⟹ $\hat{T}_t = \delta_2 (\hat{m}_t - \hat{m}_{t-1}) + (1-\delta_2)\hat{T}_{t-1}$ con $0 \leq \delta_2 \leq 1$

Stagionalità ⟹ $\hat{S}_t = \delta_3 (\tilde{y}_t - \hat{m}_t) + (1-\delta_3)\hat{s}_{t-s}$ con $0 \leq \delta_3 \leq 1$

- dove nell'equazione del livello medio $\overline{y}_t$ è il valore destagionalizzato di y_t, e nell'equazione della stagionalità $\tilde{y}_t$ è il valore detrendizzato di y_t

Si veda parte pratica in excel

Statistica per l'impresa

8. Valutazione delle prestazioni economiche-finanziarie delle imprese

IL BILANCIO
- ✓ Le principali componenti e la sua riclassificazione
- ✓ Indici di bilancio

ANALISI STATISTICA DEGLI INDICI DI BILANCIO
- ✓ Valori medi
- ✓ Benchmarking

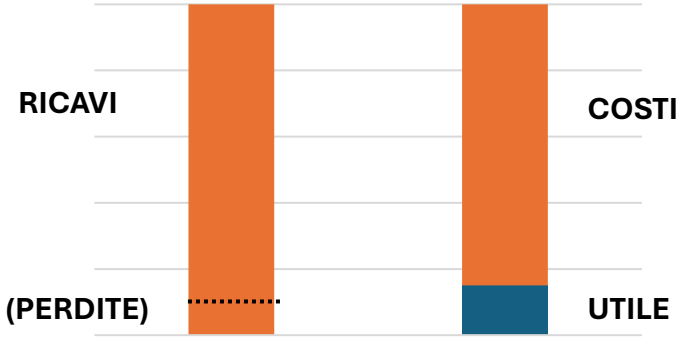
ANALISI MULTIVARIATA
- ✓ Analisi in componenti principali
- ✓ Analisi cluster

APPLICAZIONI
- ✓ Diagnosi precoce dell'insolvenza aziendale con l'analisi discriminante
- ✓ Modelli Logit

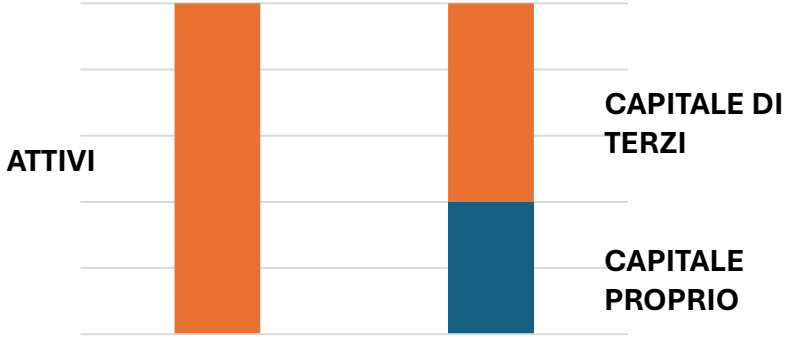
Il Bilancio e le aree gestionali

- ✓ Abbiamo visto all'inizio del corso che l'impresa dispone ed ha necessità di elaborare con metodi statistici più o meno complessi (dalla statistica descrittiva a quella inferenziale) diverse tipologie di dati di fonte interna ed esterna
- ✓ Tra questi dati particolare importanza hanno i dati di bilancio e gli indicatori che da essi si possono ricavare per giungere ad un giudizio complessivo sulla gestione dell'impresa (**analisi di bilancio**), attraverso l'indagine delle principali aree in cui si concretizza l'attività dell'impresa
- ✓ Il bilancio è costituito principalmente da:

CONTO ECONOMICO (flussi)



STATO PATRIMONIALE (stock)



- ✓ La normativa generale e quella di settore prescrivono poi ulteriori reportistiche, tra cui la **nota integrativa** (una sorta di «nota metodologica» del bilancio stesso) e la **relazione sulla gestione**

Indici di bilancio

- ✓ Tipicamente l'analista, **prima di fare un'analisi di bilancio, lo riclassifica,** ovvero scompone e ricompone i dati di bilancio (che sono redatti secondo specifici principi contabili), al fine di ottenere degli aggregati con una migliore valenza informativa per le proprie finalità
  - Per esempio il **Conto Economico** (profitti e perdite) vene **riclassificato** per separare la **Gestione Operativa** (gestione caratteristica), ovvero quella specifica dell'impresa in questione, da quella **Non Operativa** (per esempio quella finanziaria), da quella **Straordinaria**
  - Lo **Stato Patrimoniale** viene in genere riclassificato per mettere in evidenza per esempio la **natura degli attivi** (immateriali vs materiali), delle **scorte**, dei **crediti** dal lato degli attivi e dei **debiti** dal versante dei passivi
- ✓ Ogni «industria» ha una propria struttura di riclassificazione tipo. Dopo la riclassificazione l'analista calcola una serie di indicatori di bilancio che ne fotografano «l'equilibrio»

Analisi di bilancio per indici

Struttura dell'attivo
Attivi immobilizzati/Tot. Attivi
.....

Struttura finanziaria
Patrimonio netto / Totale passivo
Debiti / Totale passivo
.....

Situazione finanziaria
Debiti / Patrimonio netto (β)
Oneri finanziari / Debiti (RoD)

Redditività
RoE (Utile / Patrimonio netto)
RoS (Risultato operativo / Ricavi)
RoI (Risultato operativo / Tot. Attivi

- ✓ Si tratta alla fine dei rapporti visti nel capitolo 3 del libro, specie rapporti di composizione, coesistenza, derivazione e variazione

Valori medi e il bilancio somma 1/2

✓ **Obiettivo dell'analisi statistica** dei dati di bilancio è:

- **Confrontare gli indici dell'impresa con valori teorici o standard di settore** per cogliere l'equilibrio finanziario dell'impresa ed eventuali punti di forza e di debolezza (*gap analysis*). Per esempio per un determinato settore ed in un particolare momento storico l'indicatore di indebitamento non dovrebbe superare una certa soglia. Oppure gli attivi immateriali dovrebbero essere un parte non maggioritaria del totale dell'attivo
- **Confrontare gli indici dell'impresa con valori medi di posizione** ricavati da analisi statistiche di archivi di bilanci aziendali. In questo caso sono necessarie banche dati aziendali (Cerved, Bloomberg, banche dati di settore, …) ⟶ **BENCHMARKING**

✓ Analisi e comparazioni dei dati di bilancio passano attraverso la **qualità dei dati** stessi, che implica

- **Comparabilità**
- **Correttezza**
- **Rilevanza**
- **Completezza**

Valori medi e il bilancio somma 2/2

✓ Per calcolare i valori medi e le somme di un certo indicatore di bilancio x_i *(ratio)* per *n* imprese con cui confrontare la generica impresa *i* si possono calcolare

- **Media semplice del ratio**

$$x^* = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Media ponderata del ratio**

$$x^*_p = \frac{\sum_{i=1}^{n} x_i A_i}{\sum_{i=1}^{n} A_i}$$

⟶ Che corrisponde al rapporto tra le somme dei valori delle singole imprese

	U1	U2	U3	U4	Totale
Totale Attivo (€)	2.156.968	1.457.429	2.017.202	2.145.505	7.777.104
Reddito Operativo (€)	76.808	70.655	86.863	69.862	304.188
RoA	3,6%	4,8%	4,3%	3,3%	3,9%
			Media pesata		***3,9%***
			Media semplice		***4,0%***

	U1	U2	U3	U4	Totale
Totale Attivo (€)	8.803.610	2.156.968	2.017.202	2.145.505	15.123.285
Reddito Operativo (€)	963.130	76.808	86.863	69.862	1.196.663
RoA	10,9%	3,6%	4,3%	3,3%	7,9%
			Media pesata		***7,9%***
			Media semplice		***5,5%***

Si veda parte pratica in excel

Particolarità empiriche dei dati di bilancio

✓ Spesso nell'analisi empirica sui dati di bilancio si incontrano delle difficoltà

- **Legate alle caratteristiche delle distribuzioni**
  - **Popolazione eterogenea caratterizzata da sottogruppi** (comparabilità): a volte i dati delle società, ancorché fatti su dati di bilancio, riflettono un diverso *mix* di prodotti, di canali distributivi, di modelli di business, eccetera che rendono difficile il confronto o perlomeno non univoche le conclusioni. Le differenze si possono gestire creando dei raggruppamenti sulla base delle conoscenze del mercato o con regressioni multivariate, creando opportune variabili strumentali (*dummy*)
  - **Presenza di numerosi *outliers*** (per esempio imprese con valori molto piccoli e con indicatori non significativi
  - **Asimmetria**: le distribuzioni degli indicatori per impresa sono generalmente asimmetriche, dunque medie semplici/mediane o addirittura ipotesi di distribuzioni di tipo normale sono poco significative
- **Poste che possono assumere valori negativi e positivi**
  - Per esempio un'impresa con risultato negativo (perdita) e capitale negativo (praticamente fallita), avrebbe un *RoE* positivo. In questi casi il *ratio* non è significativo

Il Benchmarking[1] 1/2

- ✓ Abbiamo citato il Benchmarking con riferimento al confronto tra indicatori di differenti società
- ✓ Il **Benchmarking**, in ambito aziendale, è un **processo volto ad identificare i migliori standard** in termini di prodotti offerti, efficienza organizzativa, profittabilità, eccetera e **compiere i miglioramenti necessari per raggiungere questi *standard***
- ✓ Esistono differenti tipi di benchmarking, ognuno dei quali richiede in genere un trattamento statistico dei dati
- ✓ **Rispetto alla «popolazione» di riferimento** con cui ci vogliamo confrontare possiamo avere
  - **Benchmarking Interno**: l'impresa confronta al suo interno differenti società/uffici per cogliere le best practice e applicarle in maniera generalizzata. Ha il vantaggio che i dati sono facilmente disponibili e confrontabili, ma naturalmente risente di possibili eterogeneità dei contesti competitivi in cui le filiali operano
  - **Competitors Benchmarking**: l'impresa si confronta con i cosiddetti «*Peers*» di settore per individuare le aree di miglioramento. A volte non è semplice reperire le informazioni «pertinenti», ma le conclusioni sono molto utili
  - **Functional Benchmarking**: l'impresa si confronta con i concorrenti in particolari aree di attività/tecnologie (es. customer care)
  - **Generic Benchmarking**: l'impresa si confronta con operatori di altri settori (per esempio un'industria manifatturiera studia l'organizzazione di un'impresa di logistica per ottimizzare la gestione del magazzino

[1] Le slides del benchmarking presentate fanno riferimento *a Khurru S. Bhutta and Faizul Huq, «Benchmarking – best pratices: an integrated approach», 1999 University of Texas at Arlington*

Il Benchmarking 2/2

- ✓ **Rispetto alla cosa si va a misurare:**
  - **Performance benchmarking**: è per esempio il caso dei confronti con indicatori di bilancio
  - **Process Benchmarking**: si valuta i processi attuati dai concorrenti per valutare/spiegare differenti performance
  - **Strategic Benchmarking**: si valuta le differenti «traiettorie» strategiche dei concorrenti

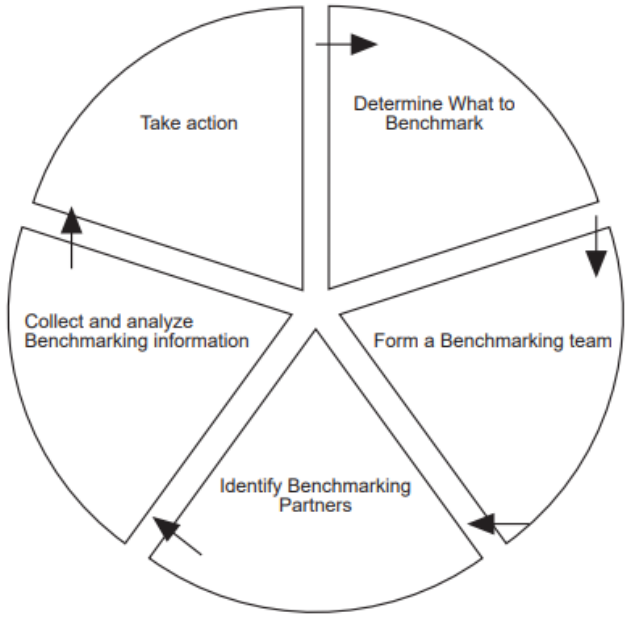
BENCHMARKING MATRIX

	Internal Benchmarking	Competitor Benchmarking	Functional Benchmarking	Generic Benchmarking
Performance Benchmarking				
Process Benchmarking				
Strategic Benchmarking				

Maggiore è l'intensità del colore, maggiore è la rilevanza

Le fasi del Benchmarking

- ✓ Il **Benchmarking** non è solo un confronto numerico volto a mettere in evidenza punti di forza e di debolezza di un'impresa
- ✓ Si tratta di **un processo** che deve interessare l'impresa nel continuo e che ha l'obiettivo di «industrializzare» il miglioramento
- ✓ Il processo iterativo è ben rappresentato dalla cosiddetta **Benchmarking Wheel**



Source: Adapted from Camp (1989)

Analisi multivariata degli indici di bilancio

- ✓ I metodi statistici multivariati servono per l'analisi descrittiva simultanea di più variabili e più unità
- ✓ Bell'analisi di bilancio possiamo trovarci di fronte a:
  - vari indici di bilancio ….
  - …. di varie e numerose imprese
- ✓ Essendo generalmente metodi «computazionalmente intensivi» hanno visto uno sviluppo considerevole in anni recenti, nell'ambito del fenomeno detto *data mining* (processo di estrazione di informazioni utili da grandi moli di dati). Vedremo due metodi:
  - **L'analisi delle componenti principali**, che è una tecnica di **riduzione del numero delle variabili**
  - **L'analisi** dei gruppi, o analisi **dei *cluster***, che permette di **classificare in gruppi omogenei** le unità osservate **sulla base delle variabili considerate**

Analisi delle componenti principali

- ✓ **Obiettivo**: sintetizzare e **ridurre i molteplici indici di bilancio** disponibili ad un **numero «ragionevole» di componenti** incorrelate e capaci di spiegare la maggior parte della varianza complessiva

- ✓ Date p variabili $x_1, x_2, \ldots, x_p$, osservate su n unità. Costruiamo $k<p$ variabili dette componenti principali tali che siano:
  - una combinazione lineare delle *p* variabili originali
  - incorrelate tra loro

- ✓ Con *p* variabili si possono ricavare al massimo *p* componenti principali (CP). La generica CP è così determinata:

$$C_h = a_{h1}\, x_1 + a_{h2}\, x_2 + \ldots + a_{hp}\, x_p$$

  - dove a_{hj} è il coefficiente della componente principale della variabile (indice di bilancio) *j*

- ✓ Il punteggio (*score*) dell'*i-esima* unità (la singola impresa) sulla *h-esima* componente è:

$$C_{ih} = \sum_{j=1}^{p} a_{hj}\, x_i$$

Determinare i coefficienti delle Componenti Principali 1/2

- ✓ I coefficienti a_{hj} dipendono dai dati che abbiamo a disposizione. Essi vengono determinati in modo che le Componenti Principali siano:
  - tra loro incorrelate
  - ordinate gerachicamente secondo la varianza

$$\text{Var}(C_1) \geq \text{Var}(C_2) \geq \ldots\ldots \geq \text{Var}(C_p)$$

  - In grado di riprodurre la varianza totale delle variabili originarie

$$\sum_{j=1}^{p} \text{var}(x_j) = \sum_{j=1}^{p} \text{var}(C_h)$$

Determinare i coefficienti delle Componenti Principali 2/2

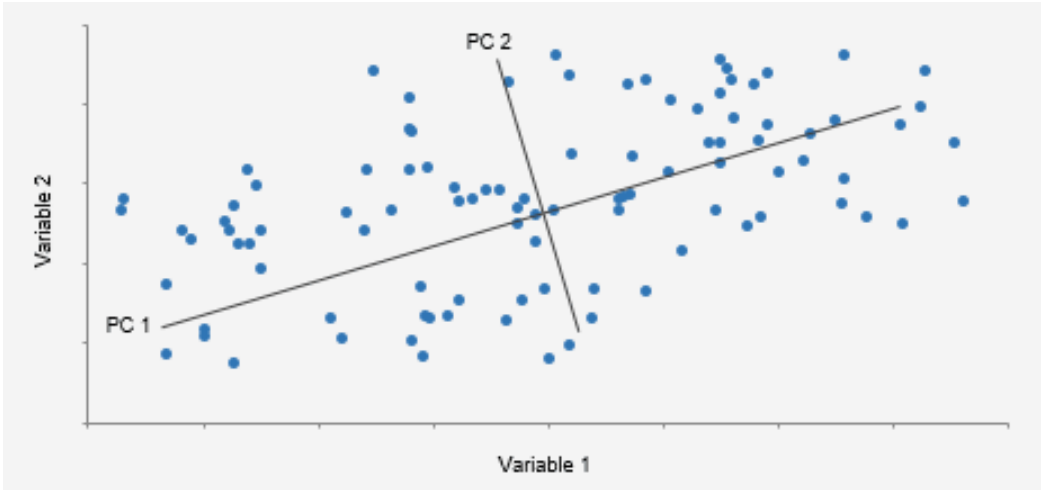
- ✓ I coefficienti a_{hj} dipendono dai dati che abbiamo a disposizione. Essi vengono determinati in modo che le **Componenti Principali** siano:
  - tra loro **incorrelate**
  - **ordinate** gerachicamente **secondo la varianza**

$$\text{Var}(C_1) \geq \text{Var}(C_2) \geq \ldots\ldots \geq \text{Var}(C_p)$$

  - In grado di **riprodurre la varianza totale** delle variabili originarie

$$\sum_{j=1}^{p} \text{var}(x_j) = \sum_{j=1}^{p} \text{var}(C_h)$$

- ✓ I **coefficienti a_{hj}** vengono concretamente **calcolati** in modo da individuare via via **le componenti principali con la varianza più grande** (intuitivamente dati che si concentrano tutti attorno ad un valore spiegano poco)
- ✓ Le **componenti** rappresentano una **combinazione lineare** degli **indicatori di bilancio** di partenza
- ✓ **Ogni impresa** avrà uno **score per ciascuna componente** principali, che rappresenta la sua «dimensione» nella stessa



Efficacia ed obiettivi delle Componenti Principali

✓ I **principali obiettivi** dell'analisi delle componenti principali sono dunque:

- **Individuare la posizione** di un'unità di osservazione (**impresa**) **rispetto alle altre**
- Comunicare in modo **compatto i risultati dell'analisi statistica**

✓ La **capacità delle CP** di ridurre le *p* variabili originarie dipende da:

- la **varianza delle variabili originarie**: le x_j con la massima varianza assumeranno peso maggiore nelle varie CP
- la **correlazione tra le variabili originarie**: tanto più le p variabili originarie sono correlate, tanto più la loro varianza potrà essere riprodotta da un numero limitato di CP

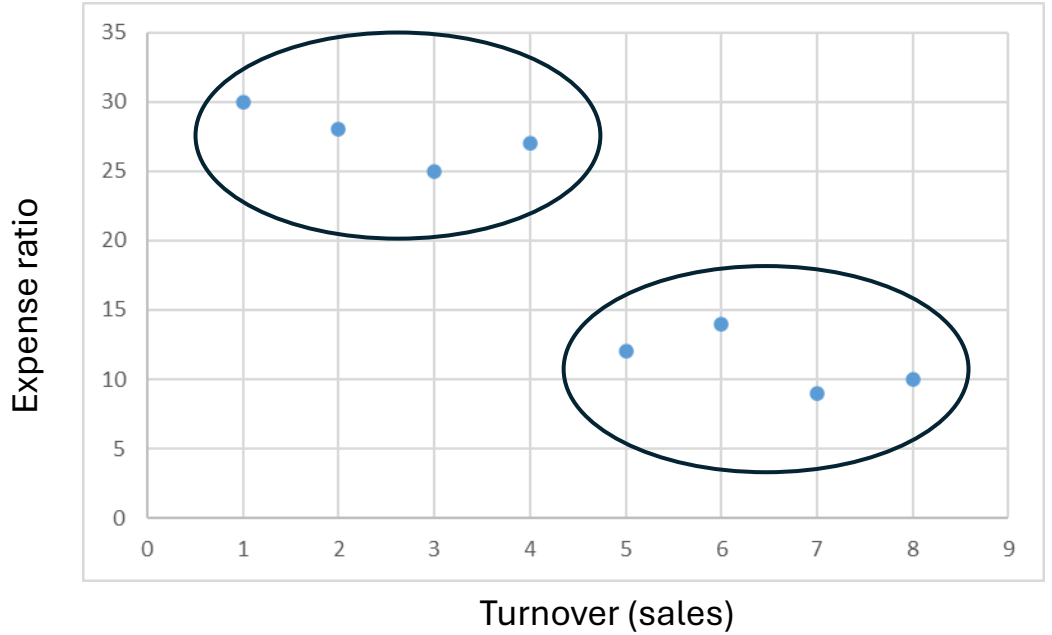
Le fasi dell'analisi in Componenti principali

✓ Le **fasi dell'analisi** in componenti principali

- **selezione delle unità** (imprese che vogliamo comparare con la nostra)
- **scelta delle variabili** ed eventuale standardizzazione (nel caso le diverse unità di misura distorcano le relazioni tra variabili)
- **calcolo delle CP** (= del coefficienti a_{hj})
- **scelta delle componenti** da considerare: si prendono le prime k componenti principali in modo che queste:
  - spieghino complessivamente almeno il **70% della varianza totale**
  - spieghino ciascuna una **varianza superiore alla media**
- **interpretazione** delle componenti principali e conseguente loro utilizzo nell'analisi del posizionamento dell'impresa

Analisi cluster 1/2

- ✓ **Obiettivo** è **suddividere le unità di analisi** (classificare) **rispetto ad alcune caratteristiche** in modo che le unità di **un certo gruppo (le imprese) siano tra loro «vicine»**.
- ✓ Per esempio vogliamo fare un benchmarking tra le imprese operanti sul mercato rispetto al rapporto spese su fatturato (*expense ratio*) e vogliamo suddividere le imprese in gruppi con un valore del rapporto simile per poi approfondire le analisi: facciamo un diagramma di dispersione con *x=fatturato e y=expense ratio*
  - Possiamo intuire che esistono due gruppi di imprese rispetto all'expense ratio. Le «piccole» hanno un ratio più alto, le seconde più basso (verosimilmente per la presenza di economie di scala)



- Quando abbiamo molti dati e più di tre dimensioni abbiamo bisogno di un sistema di classificazione «automatico»

Analisi cluster 2/2

- ✓ **Per identificare gruppi omogenei** di unità rispetto a p variabili di classificazione (nel nostro caso le imprese rispetto ad un insieme di indici di bilancio) si distinguono **metodi**:
  - **gerarchici**: i gruppi provengono dall'aggregazione progressiva di altri gruppi
  - **non gerarchici**: i gruppi provengono direttamente dall'aggregazione delle unità elementari
- ✓ **Metodi gerarchici**:
  - **divisivi**: partendo da un unico gruppo di *n* unità si suddivide lo stesso in gruppi fino ad ottenere gruppi ciascuno di *1* unità
  - **agglomerativi**: partendo da *0* gruppi e *n* unità, aggrego le unità in sottogruppi fino ad arrivare ad un solo gruppo
  - Il **numero ottimale dei gruppi** *g* è poi il **risultato della procedura agglomerativa** (intuitivamente ci si ferma a quel numero di gruppi che ci consente di bilanciare unità tra loro «simili» all'interno dei gruppi stessi e «lontani» tra gruppi diversi)

Analisi cluster gerarchica

- ✓ Si parte da una **matrice di misure di dissimilarità** (per variabili quantitative) a coppie. Dato un insieme di *p* variabili osservate sulle *n* unità, tali che $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]$ prese due unità U_i *e* U_j, indichiamo **la distanza** con:

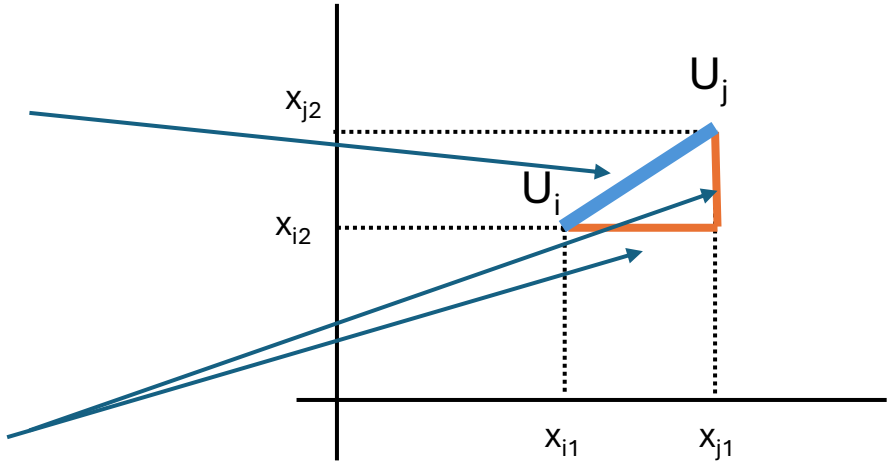
$$\mathbf{d\ (U_i, U_j) = d(x_i, x_j)}$$

- ✓ Tale distanza deve essere tale da **soddisfare le seguenti proprietà**:
  - **Identità**: $d\ (U_i, U_j) = 0$ se e solo se $x_i = x_j$
  - **Non negatività**: $d\ (U_i, U_j) \geq 0$ per ogni *i* e *j*
  - **Simmetria**: $d\ (U_i, U_j) = d\ (U_j, U_i)$
  - **Triangolarità**: $d\ (U_i, U_j) \leq d\ (U_j, U_w) + d\ (U_w, U_j)$ per ogni *i, j e w*

Alcune definizioni di distanza

✓ **Distanza euclidea:**

$$D(U_i, U_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$



✓ Distanza **city-block** (Manhattan):

$$D(U_i, U_j) = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$$

✓ **Distanza euclidea al quadrato** (non soddisfa la triangolarità ma ha alcuni vantaggi pratici):

$$D(U_i, U_j) = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$$

Fasi dell’analisi cluster

✓ Le fasi fondamentali dell’analisi cluster

- scelta delle unità (le variabili devono essere espresse nella stessa unità di misura. Nel caso bisogna standardizzare)
- scelta delle variabili di classificazione
- scelta dell’indice di distanza
- scelta del criterio gerarchico agglomerativo
- interpretazione dei risultati

La logica del raggruppamento gerarchico

- ✓ Con *n* unità, l'algoritmo opera in *n-1* passi
  - Si calcola la matrice delle distanze
  - Si uniscono le due unità meno distanti ottenendo il gruppo G_1
  - Si ricalcola la matrice di distanza tra le unità rimaste e il nuovo gruppo
  - Si uniscono le due unità o gruppi meno distanti ottenendo il Gruppo G_2
  - .....
  - Al passo *n-1* si uniscono i due gruppi, o il gruppo e l'unità rimasti nell'unico gruppo *G*

Il calcolo delle distanze tra gruppi

✓ L'algoritmo agglomerativo visto alla slide precedente presuppone di calcolare la **distanza tra gruppi** o tra gruppi e singoli elementi. Come facciamo?

- **Metodo del legame singolo**: la distanza tra $d(G_i, G_j)$ è uguale alla **minima distanza** tra le unità appartenenti al primo al primo e al secondo gruppo
- **Metodo del legame completo**: la distanza tra $d(G_i, G_j)$ è uguale alla **massima distanza** tra le unità appartenenti al primo al primo e al secondo gruppo
- **Metodo della distanza** media: prende la **media** di tutte le distanze tra le possibili coppie di unità appartenenti una al primo ed una al secondo gruppo
- **Metodo di Ward**: aggrega la coppia che comporta il **minimo incremento di devianza** all'interno dei Gruppi

Un esempio 1/5

- ✓ Date 6 imprese (*U1, …, U6*) e due indicatori di bilancio: *Beta* (Debiti/Patrimonio netto) e *Current Ratio* (Attivo Circolante/Passività Correnti)

Imprese	Beta (x1)	CR (x2)
U1	1,502	1,617
U2	2,953	0,673
U3	0,957	1,693
U4	1,631	0,584
U5	2,205	0,586
U6	2,016	0,508

- ✓ Sono numeri puri dunque posso fare calcolare direttamente la matrice delle differenze (usiamo il quadrato della distanza euclidea)

PASSO 1	U1	U2	U3	U4	U5	U6
U1	0,00					
U2	3,00	0,00				
U3	0,30	5,02	0,00			
U4	1,08	1,76	1,68	0,00		
U5	1,56	0,57	2,78	0,33	0,00	
U6	1,49	0,91	2,53	0,15	**0,04**	0,00

- ✓ La minima distanza è *0,04* tra *U5* e *U6*, pertanto creo il primo gruppo *G(U5, U6)* e costruisco una nuova matrice

Si veda parte pratica in excel

Un esempio 2/5

- ✓ Attenzione: come calcolo la differenza tra il gruppo *G1(U5, U6)* e le altre unità. Uso il criterio del legame completo (massima distanza tra componenti dei gruppi). Per esempio tra *U1* è distante 1,56 da *U5* e 1,49 da *U6* (vedi matrice precedente): dunque prendo 1,56 e così via

PASSO 2	G1(U5,U6)	U1	U2	U3	U4
G1(U5,U6)	0,00				
U1	1,56	0,00			
U2	0,91	3,00	0,00		
U3	2,78	**0,30**	5,02	0,00	
U4	0,33	1,08	1,76	1,68	0,00

- ✓ Vedo adesso che la distanza minima (0,3) si trova tra *U1* e *U3*, che aggrego dunque in un nuovo gruppo *G(U1, U3)*. Rifaccio la matrice:

PASSO 3	G1(U5,U6)	G2(U1,U3)	U2	U4
G1(U5,U6)	0,00			
G2(U1,U3)	2,78	0,00		
U2	0,91	5,02	0,00	
U4	**0,33**	1,68	1,76	0,00

Si veda parte pratica in excel

Un esempio 3/5

✓ Procedo nella stessa maniera al *passo 4:*

PASSO 4	G1(U5,U6, U4)	G2(U1,U3)	U2
G1(U5,U6, U4)	0,00		
G2(U1,U3)	2,78	0,00	
U2	**1,76**	5,02	0,00

✓ … e al *passo 5:*

PASSO 5	G1(U5,U6, U4, U2)	G2(U1,U3)
G1(U5,U6, U4, U2)	0,00	
G2(U1,U3)	**5,02**	0,00

✓ A questo punto mi fermo perché ho ottenuto un unico gruppo con tutti gli elementi

✓ Ma come faccio a capire quanto gruppi vale la pensa considerare (la fase di interpretazione dei risultati)?

Si veda parte pratica in excel

Un esempio 4/5

✓ Intanto sintetizziamo quanto abbiamo trovato (***iter* di raggruppamento**):

PASSO	UNITA'	GRUPPI	DISTANZA
0	U1, U2, U3, U4, U5, U6	6	0,00
1	G(U5, U6), U1, U2, U3, U4	5	0,04
2	G(U5, U6, U4), U1, U2, U3	4	0,30
3	G(U5, U6, U4, U2), U1, U3	3	0,33
4	G(U5, U6, U4, U2), G(U1, U3)	2	1,76
5	G (U1, U2, U3, U4, U5, U6)	1	5,02

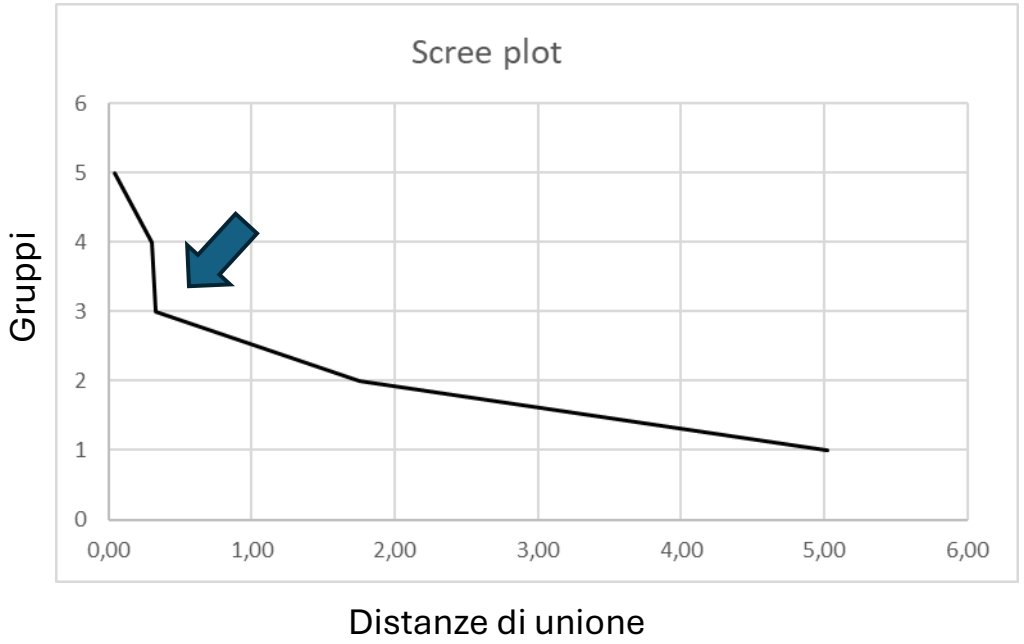
✓ Ad ogni passo poniamo il numero dei gruppi e la distanza relativa (si parte dal massimo numero dei gruppo perché usiamo un metodo agglomerativo). In questo caso si parte da una distanza *0*. La distanza aumenta man mano che aumentiamo il numero dei gruppi. Vediamo che aumenta molto dopo aver fissato un numero gruppi pari a tre

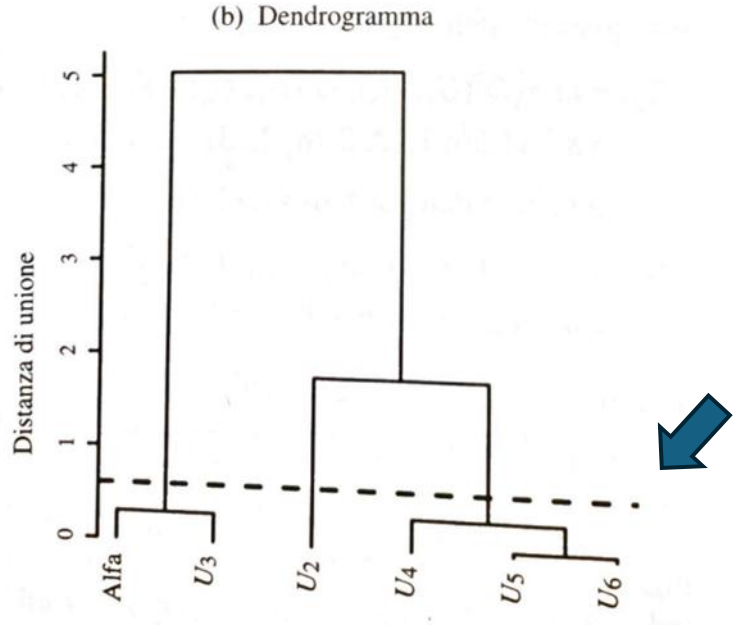
✓ La tabella ci fa intuire qualche cosa. Possiamo però farci aiutare da **due rappresentazioni grafiche per prendere la decisione finale**

Si veda parte pratica in excel

Un esempio 5/5

- ✓ Sintetizziamo quanto abbiamo trovato con due rappresentazioni grafiche:
  - Lo **scree plot** mette nelle x le distanze e nelle y il numero dei gruppi: l'idea è che ci fermiamo in corrispondenza del numero dei gruppi «che fa un gomito»: dunque ci fermiamo a 3
  - Il **dendrogramma** descrive graficamente l'aggregazione dei gruppi e le relative distanze come nella tabella della *slide* precedente. Ci fermiamo quando: ad aumenti significativi di distanze non si creano gruppi





Si veda parte pratica in excel

Diagnosi precoce dell'insolvenza aziendale: l'analisi discriminante

- ✓ A partire da un esempio concreto, parliamo di **analisi discriminante** che si presta ad essere applicata in una molteplicità di casi
- ✓ Il problema: **vogliamo valutare il rischio di insolvenza di un'impresa** a partire **da un *set* di informazioni di bilancio** (un numero *p* di indicatori)
- ✓ L'analisi discriminante parte **dall'individuazione** di un insieme di **indici di bilancio premonitori di crisi** aziendali. Ciò implica che dobbiamo disporre di un campione di dati che contenga sia imprese insolventi, sia imprese sane
- ✓ A partire da questi dati, l'analisi discriminante consente di attribuire le imprese, tramite uno score o una probabilità, nel gruppo delle imprese insolventi o in quello delle sane. In particolare:
  - il **comportamento passato di un gruppo di imprese** con certe caratteristiche viene **utilizzato per creare una sorta di modello di comportamento** genere che ci consente di prevedere quello futuro di altre imprese con caratteristiche simili
  - a partire dai dati passati infatti, l'analisi discriminante consente di **sintetizzare in un solo indice il profilo dell'imprese espresso da *p* indici**
  - **tale punteggio discrimina** se un'impresa appartenga **ad un gruppo o all'altro**

Fasi dell'analisi discriminante 1/4

La predisposizione di un modello per la previsione delle insolvenze aziendali richiede:

- ✓ **individuazione della popolazione** di riferimento e bipartizione nei due gruppi (sane/insolventi)
  - si raccolgono informazioni su un insieme di imprese omogenee *i* al tempo *t* e *t-d*
  - si individua un criterio non ambiguo per identificare le imprese insolventi
- ✓ **formazione di due campioni**
  - G_0**: le imprese insolventi**. Sono in genere assai meno numerosi delle sane, pertanto in genere si considerano tutte quelle disponibili
  - G_1**: le imprese sane** più numerose pertanto si selezione un campione, magari stratificato per mantenere le caratteristiche della popolazione obiettivo
- ✓ **selezione delle variabili**: generalmente si usano tre tipologie di variabili
  - indici di bilancio
  - variabili derivate dagli indici di bilancio (medie/varianze)
  - variabili macro

Fasi dell'analisi discriminante 2/4

✓ Scelta della **regola classificatoria** e del **valore critico**

- Si ipotizza che le osservazioni delle *p* variabili (indicatori di bilancio) sulle singoli imprese x_i sia una **variabile casuale X_i** che assume **diverse distribuzioni per le imprese del gruppo G_0 e $G1$** (le distribuzioni condizionate)

$$f(x_i|G_1) \qquad f(x_i|G_0)$$

- **Attribuiamo ogni impresa ai due gruppi in base a quanto «assomigliano» alle distribuzioni** stimate sopra. In termini propri si dice che usiamo la regola classificatoria della massima verosimiglianza:
  - se $f(x_i|G_1) \geq f(x_i|G_0)$ (ovvero se è più probabile trovare caratteristiche simili all'impresa nel gruppo G_1), allora l'impresa sarà attribuita al gruppo G_1
  - altrimenti l'impresa è attribuita al gruppo G_0
- Lo stesso concetto, per motivi tecnici di costruzione dell'analisi, può essere espresso dalla **funzione discriminante** $h(x_i) = ln\, f(x_i\ G_1) - ln\, f(x_i\ G_0)$
  - se $h(x_i) > 0$ allora l'impresa sarà attribuita al gruppo G_1
  - altrimenti l'impresa è attribuita al gruppo G_0
  - *il valore x* tale che* $h(x^*) = 0$ corrisponde al valore critico (*cut-off point*)

Fasi dell'analisi discriminante (un esempio) 3/4

- ✓ Nel caso normale univariato (*p=1*, ovvero lavoriamo con un solo indice di bilancio) e con varianza uguale tra i due gruppi, la funzione discriminante diviene la variabile stessa *($h(x_i) = x_i$)* e la regola di massima verosimglianza diviene:
  - Prendo una generica impresa $_i$ ed il suo indicatore di bilancio x_i
  - Se $x_i > x^*$ dove x^* è il valore di *cut-off*, allora la *i-esima* impresa è attribuita al gruppo G_1
  - Altrimenti la *i-esima* impresa è attribuita al gruppo G_0
  - In questo caso specifico si può dimostrare che il *cut-off* sia

$$x^* = (\mu_{G_0} + \mu_{G_1}) / 2$$

  - In altri termini: se x_i è meno distante dalla media di G_1 che dalla media di G_0, allora è attribuito al gruppo G_1

Fasi dell'analisi discriminante 4/4

- ✓ **Validazione della regola stimata**: una volta stimata la regola classificatoria bisogna valutarne la qualità. Si può distinguere tra:
  - **Analisi interna**: riclassifica le unità del campione, valutando a quale dei due gruppi il modello assegnerebbe l'impresa e lo confronta con la realtà. Avremo così:
    - la proporzione di unità (imprese) correttamente riclassificate
    - la proporzione di falsi positivi (unità di G_0 classificate erroneamente in G_1)
    - la proporzione di falsi negativi (unità di G_1 classificate erroneamente in G_0)
  - **Analisi esterna**: prevede di usare due campioni
    - uno che viene usato per stimare il modello (*training set*)
    - uno di cui sono noti gli esiti, ma che non è stato usato nella stima
    - tale analisi è la più indicata per valutare la capacità predittiva del modello

Il modello Logit 1/2

- ✓ Si tratta di un **modello di classificazione binaria** delle unità osservate spesso alla base del *machine learning*
- ✓ La **classificazione delle imprese** (*y*) potrà cioè assumere **solo due valori: *0* e *1*** (imprese che pagano dividendo vs imprese che non li pagano per esempio)
- ✓ Quello che **stimiamo è la probabilità che data una determinazione di x_{ji}, y_i sia uguale a *0* o a *1***, dove *j (j=1,.., k)* è il *j-esimo* indicatore dell'impresa *i*. La somma di queste probabilità deve fare *1,* ovvero:

$$P(y_i = 1 | x_{ij}) \quad \text{e} \quad P(y_i = 0 | x_{ij}) = 1 - P(y_i = 1 | x_{ij})$$

- ✓ Le x_{ji}, vengono poi usate per **stimare uno *score*:**

$$T_{x_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \ldots\ldots + \beta_k x_{ki}$$

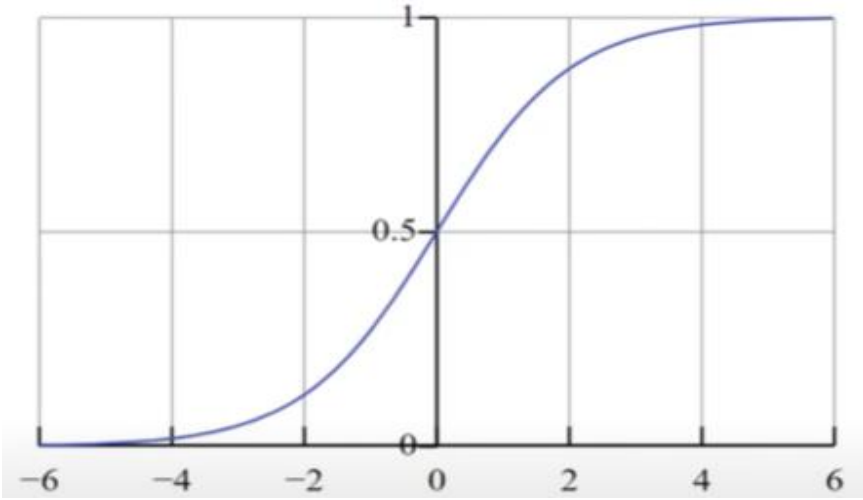
- ✓ I ***β*** vengono **stimati** con il metodo della **massima verosimiglianza**
- ✓ **Più alto è T_{x_i} maggiore è la probabilità che $P(y_i = 1 | x_{ij}) = 1$**

Il modello Logit 2/2

- ✓ T_{x_i} è un numero che può variare da +/- ∞, **non è** dunque **ancora una probabilità** (che varia tra zero e uno)
- ✓ Si applica allora una **trasformazione logistica:**

$$P(y_i = 1 \mid x_{ij}) = \frac{1}{1+\exp(-Txi)}$$

- ✓ Che presenta la **seguente distribuzione**:
  - funzione **crescente**
  - **compresa tra *0* e *1***
  - converge a ***0*** quando lo ***score* tende a - ∞**
  - converge a ***1*** quando lo ***score* tende a + ∞**



- ✓ Per valutare la **bontà di adattamento** si può usare la **percentuale di caso previsti correttamente** o altri indicatori più complessi