RICHIAMI METODOLOGICI

In questo documento sono riportati gli approfondimenti metodologici e applicativi richiamati nei diversi documenti, ordinati per argomento.

Si riporta per facilitare l'individuazione degli argomenti, una tabella contenente l'indicazione delle principali tecniche di analisi univariata e bivariata (descrittiva e inferenziale).

Gli approfondimenti metodologici relativi ad ogni argomento sono evidenziati in **giallo**. Per visualizzare la pagina contenente un approfondimento, selezionate la  nella barra del menu di Adobe:

e cliccate con il mouse.

Indice

Sintetizzare la distribuzione di un carattere (dispersione)	
Quantitativo	Campo di variazione Range Interquartile Varianza Scarto quadratico medio **Coefficiente di variazione e scarto quadratico medio**
Stimare le misure di sintesi di un carattere a partire da un campione	
Stimare la media Stimare la varianza Stimare una percentuale	Stima puntuale Stima per intervallo
Verificare l'ipotesi di distribuzione normale per un carattere	
Quantitativo	Test di normalità

Analisi bivariata

Rappresentare la distribuzione congiunta di	
Caratteri con poche modalità	Tabelle a doppia entrata Diagrammi a barre Diagrammi a bolle
Caratteri di cui almeno uno con molte modalità	Diagramma di dispersione
Valutare l'associazione tra due caratteri.	
Entrambi qualitativi nominali (analisi della connessione)	Chi-quadrato Lambda Indice di incertezza **Le misure di associazione** **Cautele nella valutazione delle misure di associazione** **La collassabilità di una tabella** **Associazione e causalità**
Entrambi (almeno) qualitativi ordinali (analisi della concordanza)	Tau di Kendall Coefficiente di Spearman

Entrambi quantitativi (analisi della relazione lineare)	Coefficiente di correlazione **Valori estremi e valori anomali (caso bivariato)** **Non robustezza del coefficiente di correlazione lineare** **Cautele nella valutazione del coefficiente di correlazione lineare** **Associazione e causalità**
Valutare la significatività dell'associazione	
Per tutti gli indici	Test sull'assenza di associazione
Valutare la dipendenza di un carattere (*Y*) da un altro (*X*)	
X e *Y* entrambi qualitativi nominali	Lambda asimmetrico Indice di incertezza asimm.
X e *Y* entrambi (almeno) qualitativi ordinali	Indici di concordanza)Tau di Kendall, Coefficiente di Spearman)
X e *Y* entrambi quantitativi (relazione lineare)	Coefficiente di correlazione Coefficiente di determinazione
Y quantitativa e *X* con poche modalità	Indice Eta
Analisi (funzionale) della dipendenza di un carattere (*Y*) da un altro (*X*)	
X e *Y* entrambi quantitativi (interpolazione)	Retta dei minimi quadrati Interpolanti diverse dalla retta **Interpolazione di una serie di dati: quale funzione scegliere?** **L'impatto della variabile esplicativa: caso lineare e non lineare** **Media mobile come perequazione di una serie storica di dati**
Analisi della distribuzione di un carattere *Y* dipendente da un carattere *X*	
Analisi stratificata o condizionata	Distribuzioni condizionate Misure di sintesi condizionate **Analisi stratificata: introduzione** **Analisi stratificata: le distribuzioni di frequenza** **Analisi stratificata: gli istogrammi** **Analisi stratificata: i box plot affiancati** **Analisi stratificata: le misure di sintesi** **Differenze tra medie (rappresentazione grafica)** **Associazione e causalità**

Modelli della dipendenza (previsivi) di un carattere quantitativo	
Dipendenza in media (var. esplicativa nominale o con poche modalità)	Anova Anova non parametrica **Analisi della varianza a una via** **Test post-hoc** **Analisi della varianza: caso di varianze non omogenee** **Analisi della varianza: caso non normale (test non parametrici)** **Analisi della varianza a due vie (cenni)**
(var. esplicativa con due modalità)	Test sull'uguaglianza tra due medie **Uguaglianza tra due medie (o popolazioni): Test T** **Test T non parametrico**
Dipendenza lineare (var. esplicativa quantitativa)	Analisi inferenziale della retta di regressione

Problemi preliminari

La censura delle osservazioni

Uno dei problemi comuni nella raccolta di dati tramite indagine riguarda il fatto che alcune osservazioni (o tutte) risultano censurate. In altre parole, si riesce a raccogliere informazioni solo fino al momento dell'intervista, ma ovviamente nulla si sa sul "dopo".
Per chiarire meglio di cosa si tratta, possiamo considerare ad esempio il caso di una società di gestione di fondi di investimento, nata nel 1996, nel cui database sono raccolte informazioni sui clienti della società. Nel grafico qui di seguito l'asse orizzontale rappresenta il tempo, mentre quello verticale rappresenta la durata del rapporto con la società. Dato che si sono raccolte informazioni sino all'inizio del 2000, è chiaro che a seconda dell'anno di entrata la lunghezza del periodo di attività dei clienti è diversa.
Ad esempio, il cliente A, entrato a metà del 1996, è stato osservato per un periodo di 3 anni e mezzo, mentre il cliente B, che è entrato a contatto con la rete all'inizio del 1999 ha "cumulato" un solo anno di attività.

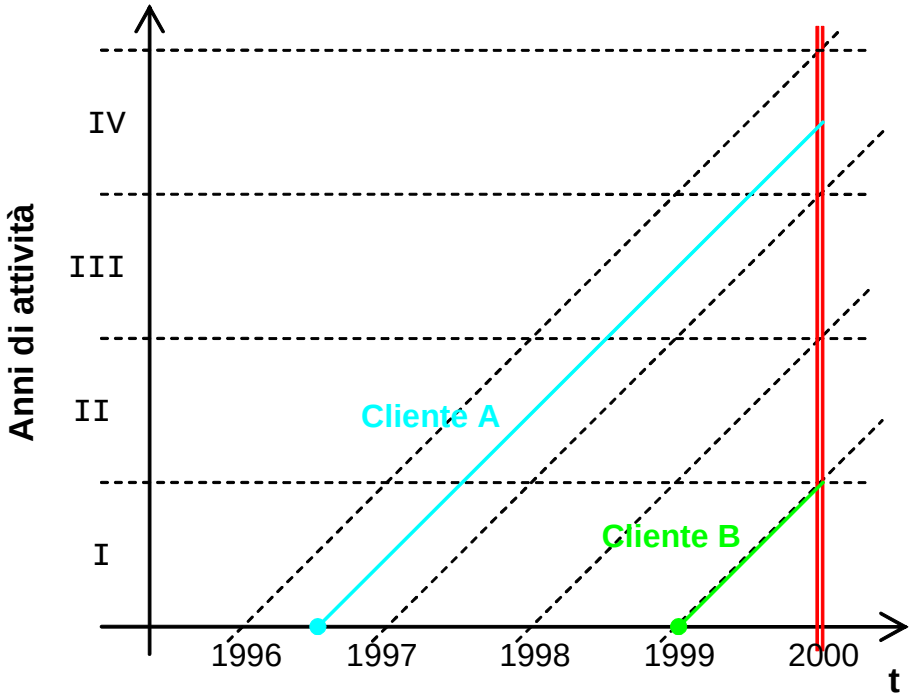
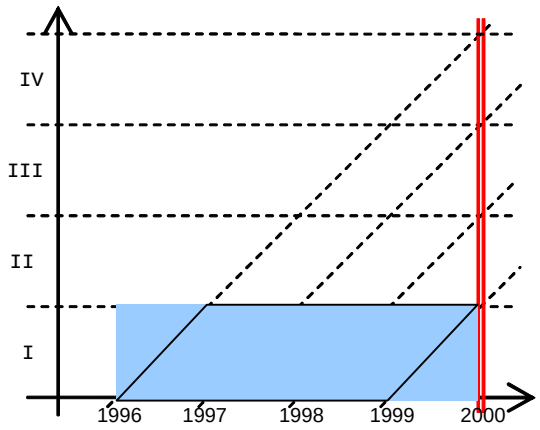
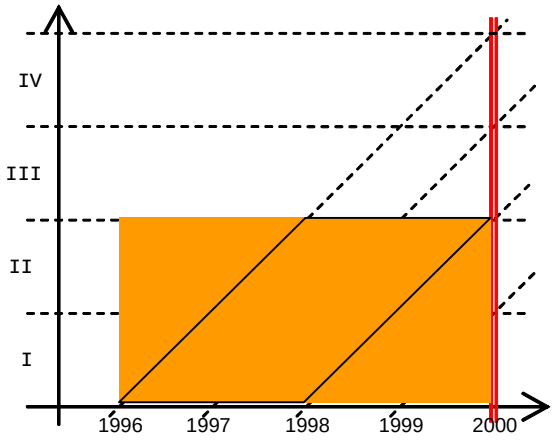


Diagramma per rappresentare i clienti secondo l'anno di entrata e il numero di anni di attività

Ovviamente, se consideriamo il totale dei soldi investiti da ogni cliente nel periodo di attività, risentiamo del fatto che la durata di attività è differente. A questo punto conviene quindi scorporare i clienti in gruppi omogenei per numero di anni di attività, e considerare il capitale investito durante quel fissato periodo di attività. Nei grafici seguenti, ecco ad esempio i clienti raggruppati per: almeno un anno di attività, almeno due anni di attività. Il capitale da considerare nelle analisi sarà perciò quello speso nella zona colorata corrispondente. Come è chiaro dalle figure, inoltre, questo tipo di selezione implica che vengano considerati man mano gruppi di numerosità diversa (i clienti che hanno avuto almeno un anno di attività sono più di coloro che hanno avuto almeno due anni di attività e così via).

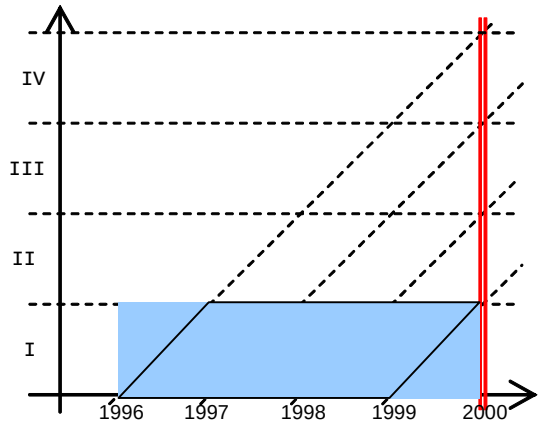


Clienti con almeno 1 anno di attività

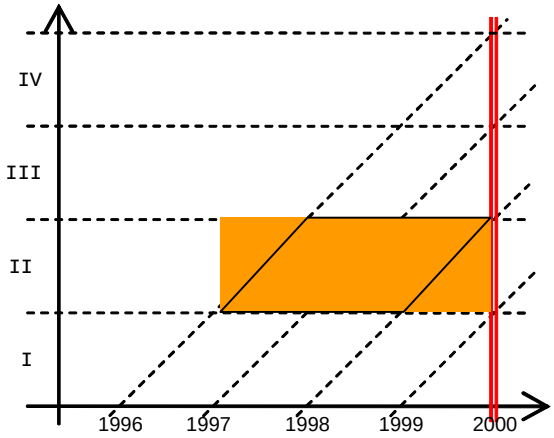


Clienti con almeno 2 anni di attività

In alternativa, possiamo anche considerare l'investimento complessivo disaggregato per anno di attività (primo anno di attività, secondo anno di attività, …). Ancora una volta le numerosità relative a diversi anni saranno diverse.



1° anno di attività



2° anno di attività

Dati mancanti

In alcune indagini può capitare che per certe unità statistiche non sia stato possibile rilevare il valore assunto da un particolare carattere di interesse.
Dobbiamo innanzitutto distinguere il problema del dato mancante da quello delle mancate risposte.
Nel secondo caso, il dato manca in quanto l'unità statistica si è "opposta" alla misurazione del dato. Ad esempio, questo può capitare quando la rilevazione dei caratteri di interesse avviene tramite questionario o tramite interviste telefoniche e l'intervistato non risponde ad una delle domande. Oppure può succedere quando l'indagine statistica prevede che l'unità statistica acconsenta alla rilevazione del dato in diverse occasioni e da un certo punto in poi alcuni soggetti non si presentano più all'appuntamento (pensiamo ad esempio ai test clinici, che prevedono cicli regolari di esami clinici: il soggetto in esame può decidere, da un certo punto in poi, di rinunciare al trattamento).

Non sempre però il dato mancante è assimilabile ad una mancata risposta.
Nelle indagini su diversi paesi, ad esempio, può capitare che gli uffici nazionali di statistica di alcuni stati abbiano deciso di non rilevare un certo carattere, che non risulta quindi disponibile. Oppure in alcuni stati potrebbero non esserci gli strumenti idonei alla rilevazione del carattere di interesse.

Nel caso della statistica descrittiva, la presenza di dati mancanti è sicuramente meno grave rispetto a quanto accade nell'ambito della statistica inferenziale.
In questo caso, è necessario infatti valutare con attenzione se l'assenza del dato è in qualche modo legata alle ***caratteristiche che si stanno studiando*.**
Consideriamo ad esempio un'analisi descrittiva relativa a diversi stati.
Se siamo interessati a studiare un certo carattere, ad esempio il tasso di istruzione, e mancano le informazioni relative agli stati più poveri, questo costituisce un problema in quanto possiamo ragionevolmente aspettarci che tali stati abbiano caratteristiche diverse rispetto agli altri. Procedere quindi nell'analisi trascurando l'esistenza dei dati mancanti può portare a distorsioni.
Per verificarlo analiticamente, consideriamo ad esempio la media di un certo carattere:

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} x_i$$

cioè la somma delle osservazioni rilevate sul carattere (per tutte le N unità statistiche che fanno parte del collettivo considerato) divisa per il numero totale di osservazioni. Ora supponiamo che la popolazione sia suddivisa in due gruppi, un primo gruppo per cui i dati sono disponibili ed un secondo gruppo per cui i dati non sono disponibili. La media *generale* può essere riscritta come:

$$\mu_X = \frac{1}{N}\left[\sum_{i=1}^{N_1} x_i + \sum_{i=1}^{N_2} x_i\right] = \frac{1}{N}\left[N_1 \frac{1}{N_1}\sum_{i=1}^{N_1} x_i + N_2 \frac{1}{N_2}\sum_{i=1}^{N_2} x_i\right] = \frac{1}{N}\left[N_1\mu_{X|1} + N_2\mu_{X|2}\right]$$
$$= p_1\mu_{X|1} + p_2\mu_{X|2}$$

dove $\mu_{X|1}$ $\mu_{X|2}$ indicano le medie del carattere X calcolate sui due gruppi, N_1 e N_2 indicano il numero di osservazioni nel primo e nel secondo gruppo e $p_1 = (N_1/N)$ e $p_2 = (N_2/N)$ indicano la percentuale di unità statistiche che fanno parte del primo e del secondo gruppo rispettivamente.
Ora, se trascuriamo il fatto che i dati relativi al secondo gruppo di osservazioni sono mancanti, valuteremo la media del carattere calcolando la media solo sul primo gruppo di osservazioni, $\mu_{X|1}$.
Tale operazione è ovviamente sensata solo se ci aspettiamo, sulla base di informazioni a priori, che la media calcolata sulle osservazioni del secondo gruppo, $\mu_{X|2}$, sia molto simile a $\mu_{X|1}$. Se ciò non accade, non ha ovviamente senso procedere in questo senso.
Diciamo quindi, in generale, che i dati mancanti possono essere trascurati solamente nel caso in cui l'assenza del dato può essere considerata *casuale* di modo che ci possiamo aspettare in qualche modo che le osservazioni su cui sono disponibili i dati costituiscano una sorta di *campione rappresentativo dell'intero collettivo di interesse.*

Dati mancanti: mancate risposte

In alcune indagini può capitare che per certe unità statistiche non sia stato rilevato uno o più caratteri di interesse. Per distinguere le mancate risposte a singole domande dalle mancate rilevazioni si usa talvolta, per le prime, il termine "mancate risposte parziali" e per le seconde "mancate risposte totali".

Il problema delle mancate risposte ha diverse implicazioni. Una riguarda il fatto che, in ambito inferenziale, la diminuzione del numero di unità statistiche porta a stimatori con varianze più elevate (si pensi alla varianza della media campionaria che è legata da una relazione inversa con l'ampiezza campionaria). Un secondo problema, molto importante, riguarda il fatto che i soggetti rispondenti possono avere caratteristiche diverse da quelle dei soggetti che non rispondono, e questo può portare a distorsioni nel processo inferenziale. Per verificarlo analiticamente, consideriamo ad esempio la ***vera*** media di un carattere *X* per una popolazione costituita da *N* unità, μ:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

cioè la somma dei valori assunti dal carattere *X* in corrispondenza di tutti gli *N* individui che costituiscono la popolazione, divisa per *N*. Ora supponiamo che la popolazione sia suddivisa in due gruppi: un primo gruppo, *R*, per cui i dati sono disponibili ed un secondo gruppo, *NR*, per cui i dati non sono disponibili. Siano rispettivamente μ_R e μ_{NR} le medie del carattere rispettivamente nella popolazione dei rispondenti e dei non rispondenti. La media generale può essere scritta in funzione delle due medie nelle sottopopolazioni:

$$\mu = \frac{1}{N}\left[\sum_{i=1}^{N_R} X_i + \sum_{i=1}^{N_{NR}} X_i\right] = \frac{1}{N}\left[N_R \frac{1}{N_R}\sum_{i=1}^{N_R} X_i + N_{NR}\frac{1}{n_{NR}}\sum_{i=1}^{N_{NR}} X_i\right] = \frac{1}{N}\left[N_R\mu_R + N_{NR}\mu_{NR}\right]$$

Consideriamo ora un campione di ampiezza *n* e supponiamo che solo n_R unità statistiche abbiano risposto. Si potrebbe assumere che le *r* risposte siano un campione proveniente dalla sola popolazione dei rispondenti, con media μ_R. La media campionaria calcolata sulle sole risposte è quindi non distorta per μ_R ma non per la media generale, μ. In particolare:

$$E(\bar{x}_R) = \mu_R$$

Ricordando la relazione tra la media generale e le due medie sarà:

$$\mu_R - \mu = \mu_R - \frac{1}{N}\left[N_R\mu_R + N_{NR}\mu_{NR}\right] = \frac{N\mu_R - \left[N_R\mu_R + N_{NR}\mu_{NR}\right]}{N} = \frac{(N - N_R)\mu_R - N_{NR}\mu_{NR}}{N}$$

$$= \frac{N_{NR}(\mu_R - \mu_{NR})}{N}$$

La distorsione della media campionaria calcolata sulla base delle sole risposte sarà quindi tanto maggiore quanto più differiscono μ_R e μ_{NR} e quanto maggiore risulta il tasso di non risposta $(N - N_R)/N$ (si noti che esattamente lo stesso ragionamento può essere fatto nel caso in cui si sia interessati a stimare una percentuale).

Dal punto di vista inferenziale dobbiamo quindi distinguere tra alcuni casi. Nel caso in cui le osservazioni sono mancanti "in modo casuale", la non risposta non è legata alle caratteristiche oggetto di interesse. In questo caso le osservazioni sui rispondenti rappresentano un campione rappresentativo della popolazione in quanto possiamo aspettarci che le due medie μ_R e μ_{NR} siano simili tra loro e quindi la distorsione dello stimatore basato sulle sole risposte sarà minima.

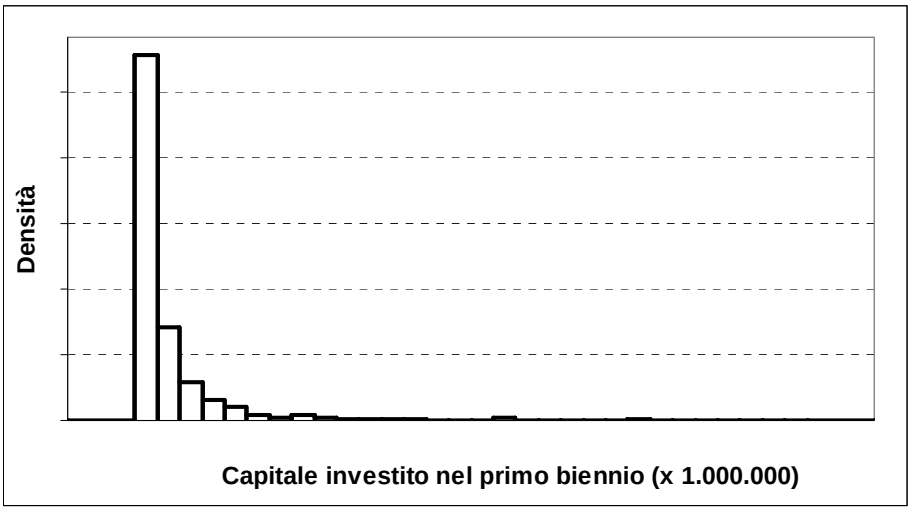
Al contrario, se siamo in presenza di dati mancanti non ignorabili, la scelta di non rispondere è legata al carattere oggetto di interesse e quindi la distorsione potrebbe essere non trascurabile (si pensi ad esempio ad un'indagine in cui si chiede quanti libri vengono letti in media all'anno; se i rispondenti sono per la maggior parte persone che non vogliono ammettere di non leggere, la stima della caratteristica di interesse sarà completamente distorta).

Analisi univariata

Istogramma: la scelta delle classi

Quando si rappresenta un carattere quantitativo continuo tramite istogramma, uno dei problemi da affrontare è la scelta delle classi da utilizzare per la rappresentazione. Oltre al numero, bisogna anche decidere gli estremi di tali classi, e infine anche se vale la pena rappresentarle tutte o meno. L'idea è di scegliere un numero di classi abbastanza elevato da permettere di comprendere la forma della distribuzione, ma, allo stesso tempo, abbastanza contenuto da consentire una rappresentazione sintetica.
Nel grafico che segue, a scopo esemplificativo, si è deciso di utilizzare per un primo tentativo 30 classi di uguale ampiezza per rappresentare un carattere continuo.

Estr. inf. delle classi	Estr. sup. delle classi	Freq. assolute	Freq. relative	Densità
0.0014	58.74	653	0.651697	0.011095
58.74	117.48	167	0.166667	0.002837
117.48	176.22	69	0.068862	0.001172
176.22	234.96	37	0.036926	0.000629
234.96	293.69	25	0.024950	0.000425
293.69	352.43	9	0.008982	0.000153
352.43	411.17	5	0.004990	0.000085
411.17	469.91	11	0.010978	0.000187
469.91	528.65	5	0.004990	0.000085
528.65	587.39	3	0.002994	0.000051
587.39	646.12	3	0.002994	0.000051
646.12	704.86	2	0.001996	0.000034
704.86	763.60	3	0.002994	0.000051
763.60	822.34	0	0	0
822.34	881.08	0	0	0
881.08	939.82	0	0	0
939.82	998.55	5	0.004990	0.000085
998.55	1057.29	0	0	0
1057.29	1116.03	0	0	0
1116.03	1174.77	0	0	0
1174.77	1233.51	0	0	0
1233.51	1292.25	0	0	0
1292.25	1350.98	2	0.001996	0.000034
1350.98	1409.72	1	0.000998	0.000017
1409.72	1468.46	1	0.000998	0.000017
1468.46	1527.20	0	0	0
1527.20	1585.94	0	0	0
1585.94	1644.68	0	0	0
1644.68	1703.42	0	0	0
1703.42	1762.15	1	0.000998	0.000017



Notiamo che la prima classe, modale, rappresenta circa il 65% della popolazione considerata. Quindi potrebbe essere utile procedere raggruppando i dati in classi di intervallo più sensate, riunendo la lunga coda destra in un'unica classe, senza perdita di informazioni, e disaggregando la prima classe.
Una scelta quindi più sensata potrebbe essere la seguente, dove l'ultima classe è però stata esclusa dalla rappresentazione.

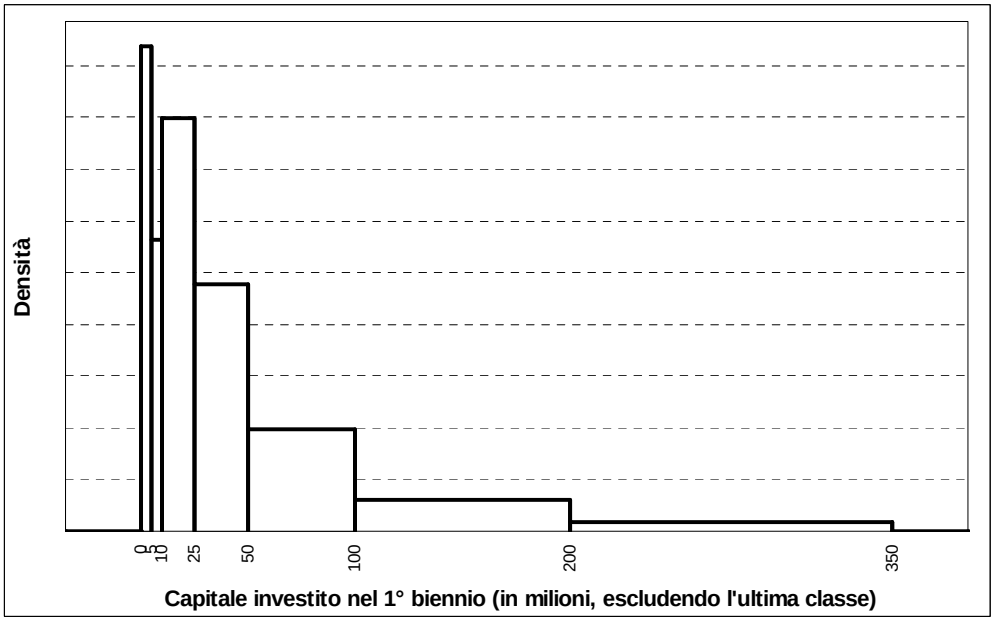


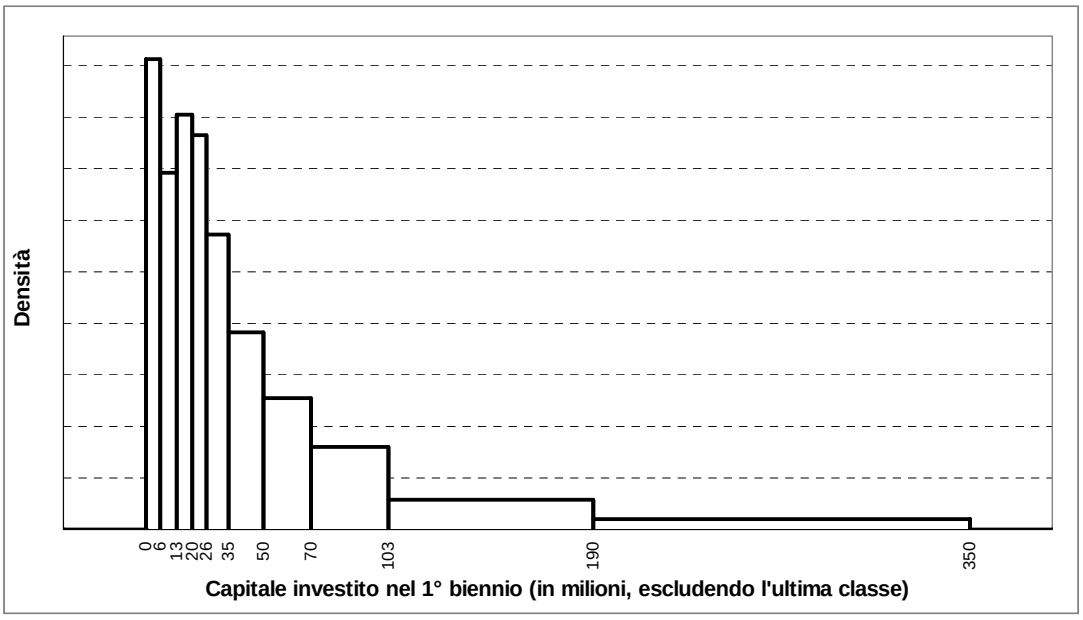
Tabella delle frequenze relativa all'istogramma (classi pre-assegnate)

Estr. inferiori	*Estr. superiori*	*Fr. assolute*	*Fr. relative*	*Densità*
0	5	90	0.08982036	0.0179641
5	10	54	0.05389222	0.0107784
10	25	230	0.22954092	0.0153027
25	50	229	0.22854291	0.0091417
50	100	189	0.18862275	0.0037725
100	200	116	0.11576846	0.0011577
200	350	51	0.0508982	0.0003393
350	1763	43	0.04291417	3.037E-05

In alternativa, un modo semplice ma efficace per comprendere la forma della distribuzione consiste nel considerare, come estremi delle classi, i decili della distribuzione. In questo modo, già abbiamo un'idea della forma del carattere, e possiamo in seguito decidere di raggruppare o scomporre alcune classi di particolare interesse.
Qui di seguito, ecco l'istogramma risultante selezionando, come estremi delle classi, i decili della distribuzione, in cui però l'ultima classe è stata scorporata in due e la seconda parte non è stata rappresentata.

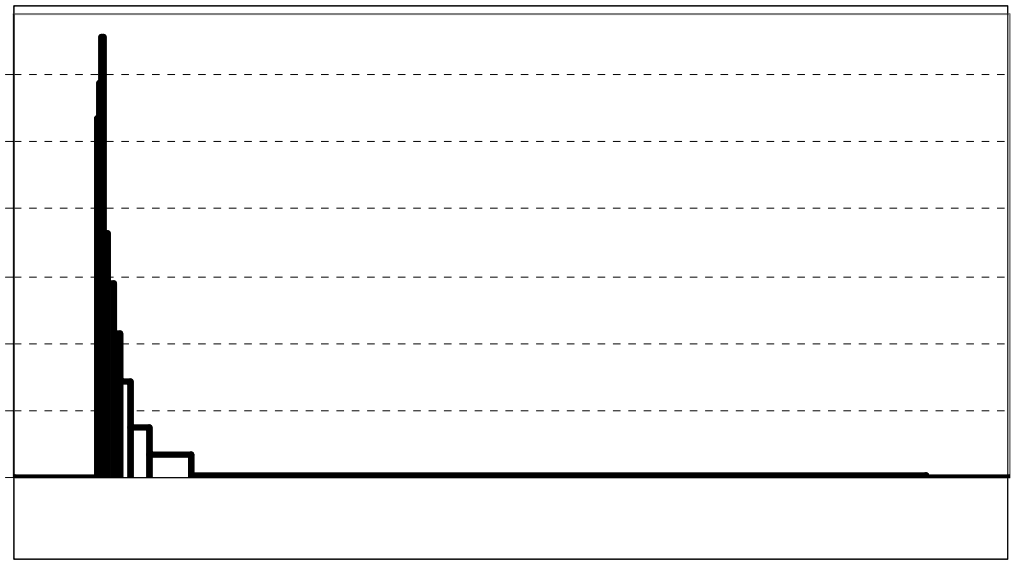
Tabella delle frequenze relativa all'istogramma (decili)

Estr. inferiori	*Estr. superiori*	*Fr. assolute*	*Fr. relative*	*Densità*
0	6	105	0.104790419	0.01746507
6	13	93	0.092814371	0.013259196
13	20	108	0.107784431	0.015397776
20	26	88	0.087824351	0.014637392
26	35	99	0.098802395	0.010978044
35	50	110	0.109780439	0.007318696
50	70	98	0.097804391	0.00489022
70	103	102	0.101796407	0.00308474
103	190	97	0.096806387	0.001112717
190	350	59	0.058882236	0.000368014
350	1763	43	0.042914172	3.0371E-05

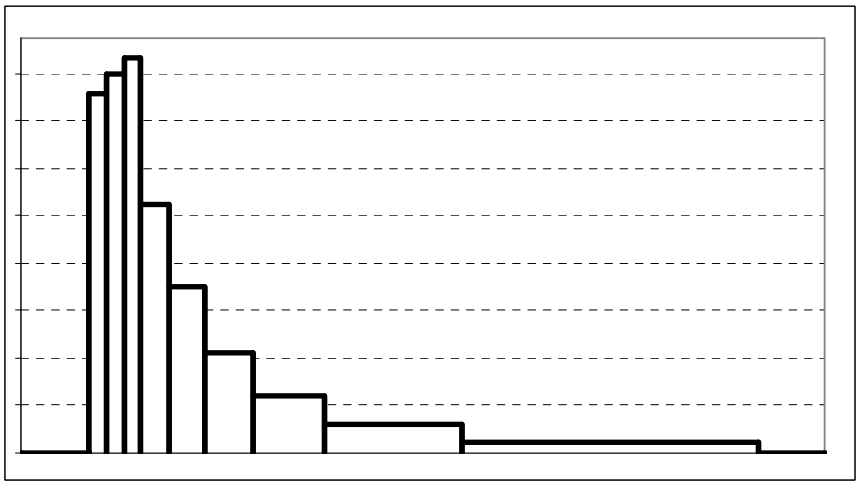


Istogramma: cautela nella scelta delle scale

Quando si rappresenta un carattere quantitativo continuo con una distribuzione fortemente *asimmetrica* può valere la pena di non rappresentare le classi estreme.
Per comprendere il motivo, consideriamo un carattere con campo di variazione molto ampio e con distribuzione fortemente asimmetrica. La presenza di asimmetria ci suggerisce di utilizzare una rappresentazioni con 10 classi di frequenza approssimativamente uguale.
L'istogramma ottenuto è il seguente:



Notiamo che tale rappresentazione grafica non consente di visualizzare proprio la porzione di asse reale più rilevante, cioè l'intervallo su cui si concentrano la maggior parte delle osservazioni. Per migliorare la visualizzazione dell'istogramma, possiamo decidere di eliminare l'ultima classe (cui compete una densità quasi nulla) dal grafico, riportando quindi le densità di frequenze solo per le prime nove classi di intervallo (attenzione: l'ultima classe non viene rappresentata ma viene comunque utilizzata per il calcolo delle densità di frequenza). L'istogramma ottenuto è il seguente:

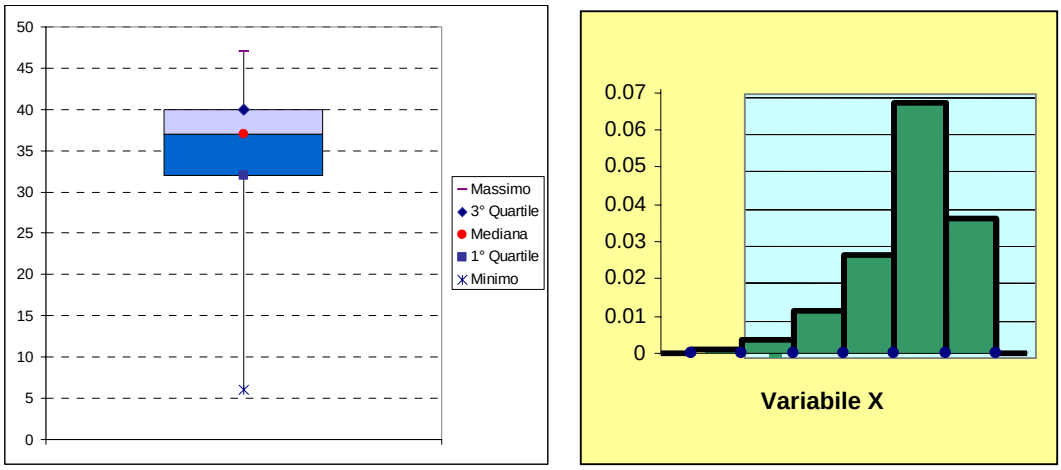


Come si nota, tale rappresentazione grafica risulta molto più efficace di quella considerata precedentemente.

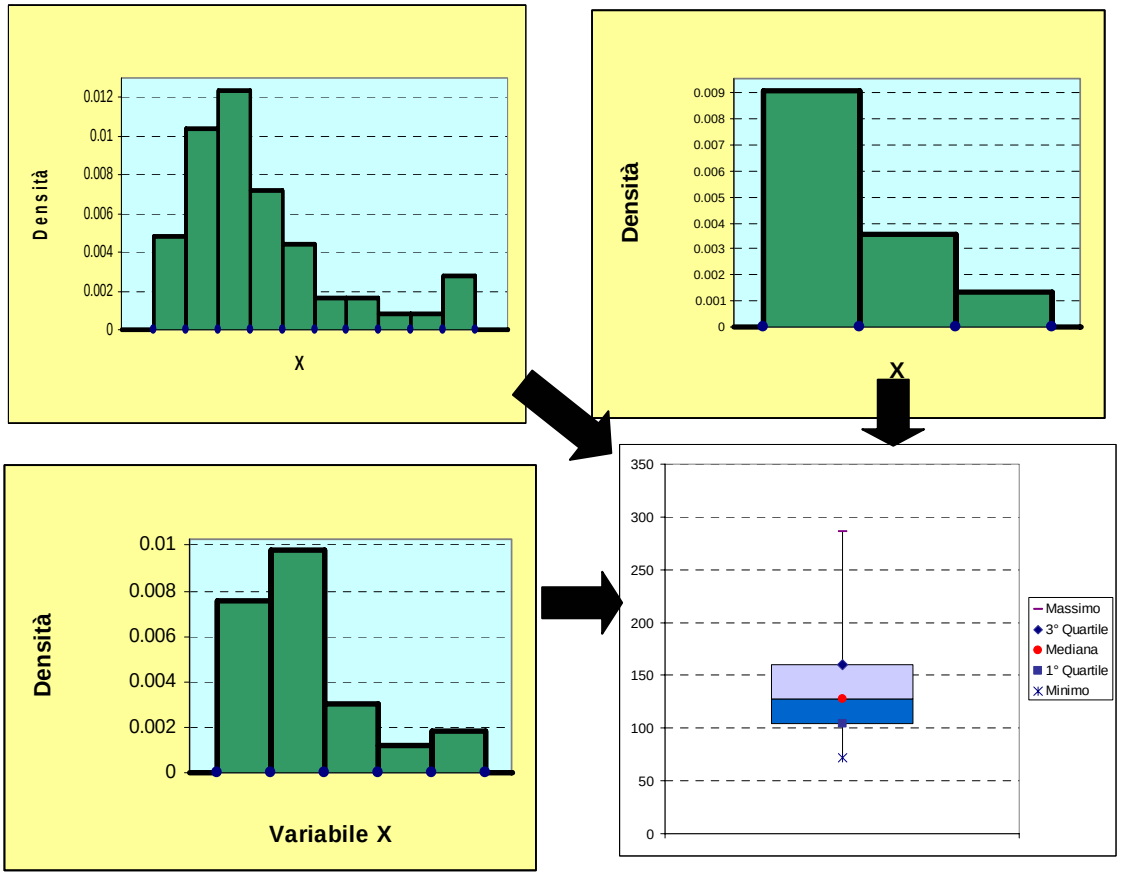
Box plot: una rappresentazione sintetica della distribuzione

Il box plot o diagramma a scatola e baffi, è un grafico, relativo a caratteri quantitativi - ottenuto a partire dai *5 numeri di sintesi* [minimo, 1° quartile (Q1), mediana, 3° quartile (Q3), massimo] - che descrive le caratteristiche salienti della distribuzione. Si ottiene riportando su un asse verticale (oppure orizzontale) i 5 numeri di sintesi. La scatola del box plot ha come estremi inferiore e superiore rispettivamente Q1 e Q3. La mediana divide la scatola in due parti. I baffi si ottengono congiungendo Q1 al minimo e Q3 al massimo. In alcuni grafici (ad esempio, quello ottenuto con SPSS) il baffo ha lunghezza pari a 1.5 volte l'altezza della scatola, data dalla distanza tra Q3 e Q1 – detto anche *range interquartile*; ovviamente è inferiore se il massimo valore osservato dista da Q3 meno di 1.5 volte il range interquartile.
Confrontando tra loro le lunghezze dei due baffi (che rappresentano le distanze tra Q1 e il minimo e tra Q3 e il massimo) e le altezze dei due rettangoli che costituiscono la scatola (che rappresentano le distanze tra Q1 e mediana e tra mediana e Q3) si ottengono informazioni sulla simmetria della distribuzione: questa è tanto più simmetrica quanto le lunghezze dei baffi risultano simili tra loro e le altezze dei due rettangoli risultano simili tra loro.
I baffi mettono inoltre in evidenza la presenza di eventuali outliers (osservazioni eccezionali) [*Valori estremi e outliers*].

Per rappresentare una distribuzione in modo sintetico, il box plot è un'ottima possibilità: con poche informazioni, si riesce a comprendere la sua forma, simmetrica o asimmetrica che sia.
Ad esempio, in questa figura notiamo che il box plot evidenzia efficacemente l'asimmetria della distribuzione del carattere.

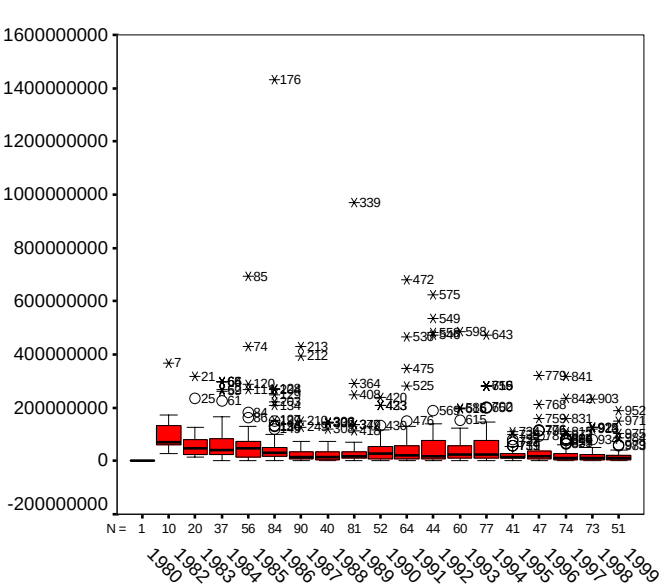


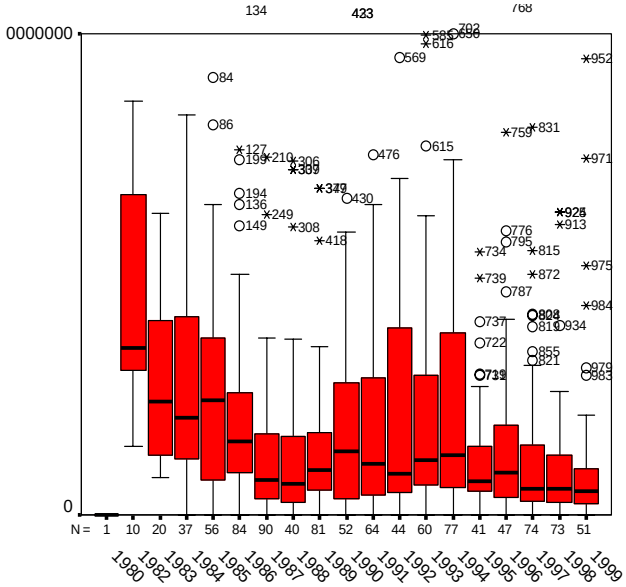
Da notare inoltre, che il box plot dà una rappresentazione ***univoca*** della distribuzione, ***a differenza dell'istogramma*** che può dare rappresentazioni diverse ***a seconda degli estremi delle classi scelte***.
Ad esempio, nella pagina di seguito sono riportati 3 istogrammi relativi ad uno stesso carattere, ma ottenuti scegliendo un numero differente di classi di diversa ampiezza.
Il box plot relativo alla distribuzione, però, non varia.


Densità
0.012
0.01
0.008
0.006
0.004
0.002
0
X
Densità
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
X
Densità
0.01
0.008
0.006
0.004
0.002
0
Variabile X
350
300
250
200
150
100
50
0
Massimo
3° Quartile
Mediana
1° Quartile
Minimo

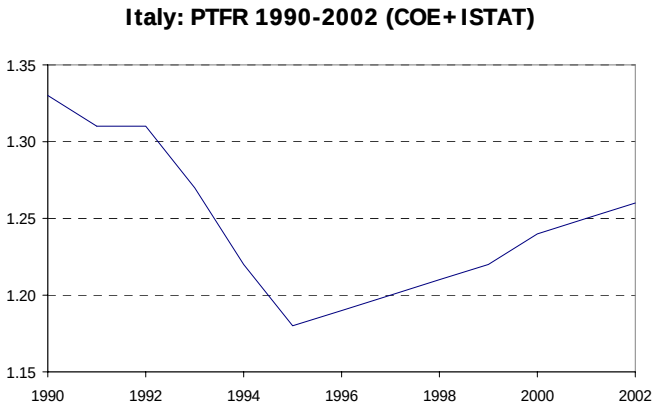
Box plot: cautele nella scelta delle scale

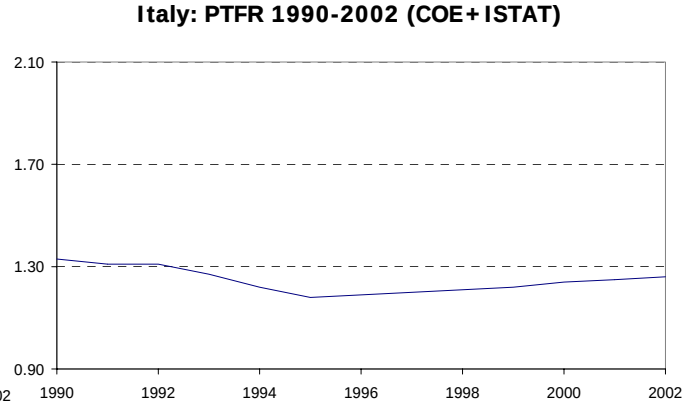
Nel rappresentare un grafico, la scelta degli estremi della scala può essere un elemento cruciale per mettere in luce alcuni elementi di particolare interesse.
Nel grafico che segue (Figura di sinistra), ad esempio, si nota come la presenza di osservazioni estreme non permetta una chiara visualizzazione delle differenze tra i diversi anni (variabile in ascissa).
Riscalando invece il grafico in modo da escludere tali osservazioni, le differenze tra gli anni sono molto più evidenti (grafico di destra).





Esempio analogo è quello riportato qui di seguito, dove però cambiamenti reali del tutto trascurabili del tasso di fecondità italiano sembrano invece essere notevoli a causa di una discutibile scelta degli estremi della scala. Nel secondo grafico gli estremi della scala sono stati scelti in modo sensato: mentre 0.9 rappresenta il valore più basso osservato per il tasso di fecondità totale, 2.10 è un valore soglia, poiché rappresenta il numero medio di figli che una donna dovrebbe avere per rimpiazzare nella generazione successiva se stessa ed il proprio partner.





***Valori estremi e valori anomali* (*caso univariato*)**

Si definiscono *valori estremi* i valori più grandi o più piccoli di una distribuzione. In senso più generale, l'espressione significa i valori prossimi alla coda di una distribuzione.
Con *valori anomali* (in inglese, outlier) ci si riferisce invece ai valori estremi di una distribuzione che si caratterizzano per essere estremamente elevati o estremamente bassi rispetto al resto della distribuzione e che rappresentano perciò casi isolati rispetto al resto della distribuzione.

In generale, per stabilire se un valore è estremo o anomalo, si fa riferimento alle misure di sintesi della posizione e di dispersione.

Distanza dalla media Tale criterio fa riferimento alla cosiddetta disuguaglianza di **Tchebycheff**. In base a tale disuguaglianza, per un carattere X con media μ e scarto quadratico medio σ si ha:

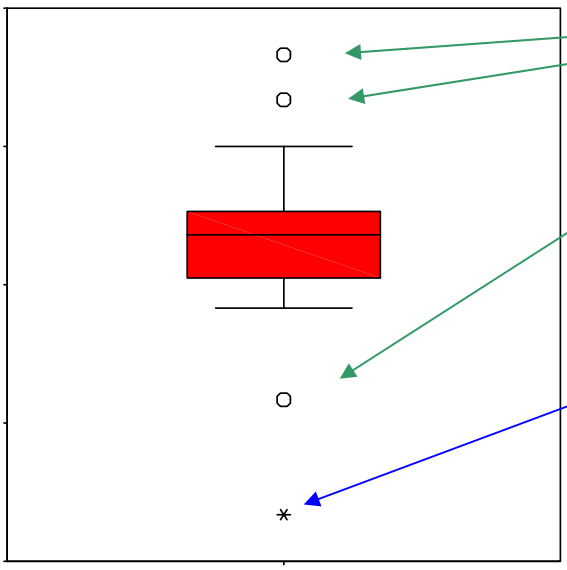
$$Freq\ (|X\ - \mu| < k\sigma) > 1 - \frac{1}{k^2}$$

In termini "pratici" si ha che, *qualunque sia la distribuzione di X*:

almeno il 75% delle osservazioni su X sono contenute nell'intervallo $\mu - 2\sigma$; $\mu + 2\sigma$
almeno l'89% delle osservazioni su X sono contenute nell'intervallo $\mu - 3\sigma$; $\mu + 3\sigma$

Utilizzando tale risultato, vengono considerati come *possibili* valori anomali quei valori che si discostano dalla media (aritmetica) per più di 3 volte lo scarto quadratico medio.
Ovviamente, in ogni caso è necessario considerare la distribuzione nella sua interezza, e vedere se i punteggi troppo alti o troppo bassi rappresentano casi isolati dal resto della distribuzione.

Distanza dai quartili. Un secondo criterio per stabilire quali sono i valori estremi fa riferimento al range interquartile, dato dalla differenza tra terzo e primo quartile, cioè l'ampiezza dell'intervallo entro cui cade il 50% delle osservazioni che occupano le *posizioni centrali* nella serie ordinata dei dati (quindi le osservazioni "meno anomale").



Viene considerato ***valore estremo*** un valore con scostamento positivo dal terzo quartile superiore a 1.5 volte il range interquartile o, simmetricamente, un valore con scostamento negativo dal primo quartile superiore (in valore assoluto) a 1.5 volte il range interquartile.
Viene invece considerato ***valore anomalo*** un valore con scostamento (positivo) dal terzo quartile o (negativo) dal primo quartile superiore a 3 volte il range interquartile.

In SPSS i valori estremi e anomali vengono evidenziati rispettivamente con un cerchio e con un asterisco come nella figura di fianco.

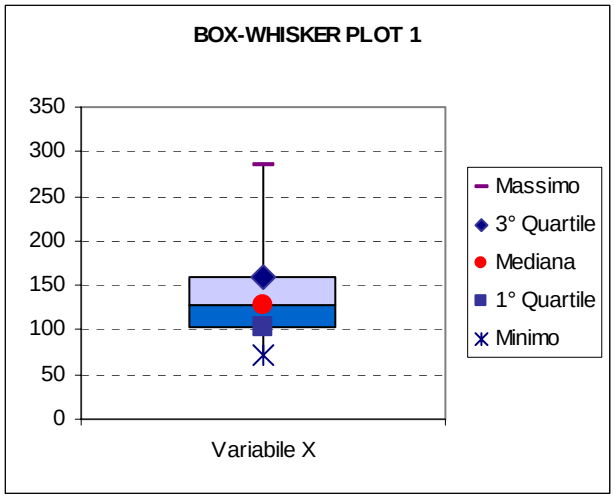
I valori anomali possono influenzare molti indicatori, come la media o la deviazione standard. Essi possono anche influenzare gli indici di associazione tra le variabili come il coefficiente di correlazione di Pearson.

In presenza di casi anomali che influenzano i risultati delle analisi è possibile utilizzare delle misure di sintesi che risultano meno influenzate dalla presenza di tali valori. Ad esempio, la mediana spesso può risultare più affidabile della media. Sono inoltre disponibili alcune misure di sintesi che risultano "robuste" alla presenza di tali valori, come ad esempio la media *troncata* che viene calcolata eliminando il 5% dei casi con punteggi più elevati e più bassi.
[*Media, mediana, media troncata per distribuzioni asimmetriche*]

In alcuni casi si può essere tentati di procedure alla *rimozione* di valori anomali che risultano *influenti*, ovvero che hanno un impatto eccessivo sulle misure di sintesi che si vogliono considerare (ad esempio la media, o il coefficiente di correlazione lineare). Tuttavia tale modo di procedere non è sempre prudente, a meno che non si siano identificati i motivi che rendono un'osservazione anomala e non si possa supporre che essa possa essere esclusa dall'analisi in quanto non coerente con il collettivo di riferimento. Negli altri casi, non è sensato procedere alla rimozione delle osservazioni anomale.
Supponiamo ad esempio che si stia studiando il reddito di una determinata classe lavorativa e che si osservi un reddito eccezionalmente elevato rispetto agli altri. Se, in base alle informazioni a nostra disposizione, siamo in grado di concludere che un reddito così elevato è spiegato dal fatto che l'individuo in questione lavora in un'azienda molto particolare, o ha una mansione molto particolare, di modo che non è omogeneo con il resto del collettivo, può essere sensato rimuovere l'osservazione. Se invece redditi elevati sono rari ma possibili nel collettivo considerato, rimuovendo l'osservazione anomala impoveriamo i risultati dell'indagine statistica, in quanto escludiamo dall'analisi un determinato segmento di popolazione.

Media, mediana, media troncata per distribuzioni asimmetriche

Ogni qual volta vogliamo avere delle informazioni sulla forma di una distribuzione, ovvero sulla sua simmetria o asimmetria, possiamo sfruttare strumenti diversi, sia grafici che non.
Il box plot, ad esempio, ci dà un'idea sia del *range* in cui è concentrato il 50% delle osservazioni che occupano i valori centrali della distribuzione, che delle "code" della distribuzione. Una coda particolarmente lunga a destra come nella figura sottostante ci farà pensare che la distribuzione è asimmetrica a destra. Informazioni analoghe si possono ottenere confrontando la mediana con la media interquartile (media tra il primo e il terzo quartile): la distribuzione si dice obliqua a destra se la mediana risulta più piccola rispetto alla media interquartile, e obliqua a sinistra se invece è la media interquartile a precedere la mediana.



Variabile X	Valore Indice
Minimo	72
1° Quartile	104
Mediana	128
3° Quartile	160
Massimo	286
Media interquartile	132

Quando vogliamo sintetizzare una distribuzione molto asimmetrica con una misura di tendenza centrale, dobbiamo tenere conto del fatto che la ***media***, la misura di sintesi più comunemente utilizzata è una misura *non robusta*. Ciò significa che la media è un indicatore sensibile ai valori estremi della distribuzione, e verrà quindi "attratta" da essi.
In questo caso, quindi, la media sarà influenzata dai valori che si trovano sulla "coda" (destra o sinistra) della distribuzione, quindi risulterà una sintesi poco efficace della massa di dati "più tipici".
Quando vogliamo utilizzare una misura di sintesi che descriva adeguatamente i dati più tipici, conviene allora utilizzarne una meno sensibile ai valori anomali. La più nota è la ***mediana***, il valore che occupa la posizione centrale nella serie ordinata dei dati (o, anche, che viene preceduta e seguita dallo stesso numero di osservazioni nella serie ordinata dei dati).
Un'altra possibilità è costituita dalla cosiddetta ***media troncata***. Questa è la media della distribuzione troncata ad una soglia fissata, di solito, al 5%, e non è altro se non la media calcolata sul 95% delle osservazioni che occupano i valori centrali della distribuzione. Di fatto, nel calcolo della media, si trascura una parte residuale della distribuzione, dove i valori sono più estremi.

Ovviamente, se la distribuzione è obliqua a destra, la mediana e la media troncata risulteranno inferiori alla media. Relazione opposta legherà le misure nel caso di distribuzione obliqua a sinistra. Naturalmente, le differenze tra le misure di sintesi vanno valutate utilizzando un termine di paragone. La stessa differenza tra media e la mediana, ad esempio, va valutata tenendo conto del campo di variazione del carattere. Se il campo di variazione è molto esteso, la differenza tra media e mediana può essere giudicata come "relativamente" piccola. Al contrario, quando il campo di variazione è contenuto, anche una piccola differenza tra media e mediana può essere giudicata "relativamente" grande.

Coefficiente di variazione e scarto quadratico medio

Per effettuare confronti di variabilità tra distribuzioni diverse conviene fare riferimento al coefficiente di variazione piuttosto che allo scarto quadratico medio (o deviazione standard), poiché il coefficiente di variazione tiene conto della media della distribuzione.
Ricordiamo che la varianza è la media delle differenze elevate al quadrato tra ciascuna delle osservazioni in un gruppo di dati e la media aritmetica dei dati stessi, μ.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Quindi rappresenta l'errore al quadrato che commettiamo, in media, sostituendo ad una generica osservazione, x_i, la media, μ. Lo scarto quadratico medio è la radice della varianza e quindi rappresenta la radice quadrata dell'errore (medio) al quadrato.
Consideriamo ora il seguente esempio. Supponiamo di aver rilevato in 10 negozi il prezzo di due beni A e B (ad esempio, il prezzo di una camicia di un certo filato e il prezzo di un completo da uomo) e di aver ottenuto i seguenti risultati:

Bene	Media	Campo di variazione	Varianza	Scarto quadratico medio
A	118,50	110	1550,25	39,37
B	1649,00	700	59409,00	243,74

Ora sembra abbastanza evidente che le fluttuazioni del prezzo del bene A sono superiori rispetto a quelle del bene B sebbene le misure di variabilità considerate siano tutte superiori per il secondo bene. Questa considerazione dipende dal fatto che la *media dei prezzi* per il bene B è più elevata di quanto non lo sia quella relativa al bene A.
Il coefficiente di variazione, definito a partire da media e scarto quadratico medio, è l'indice opportuno per confrontare la variabilità di due caratteri:

$$CV = \left(\frac{\sigma}{|\mu|}\right) \times 100\% \qquad \mu \neq 0$$

Il CV risulta particolarmente utile quando si vuole confrontare la dispersione di dati aventi differenti unità di misura o aventi range di variazione diversi.
Con riferimento all'esempio visto sopra, si ha:

Coefficiente di variazione:

Bene A	Bene B
0,33	0,15

Notiamo che contrariamente alle misure viste sopra, il coefficiente di variazione risulta più elevato per il bene A.

Analisi bivariata

Le misure di associazione

Le misure di associazione sono sintesi dei dati (o statistiche nel caso in cui vengano calcolate su un campione) che indicano la forza dell'associazione tra due caratteri.
Ovviamente, il *tipo di associazione investigato* dipende dalla natura dei caratteri considerati.

Misure di associazione *Nominale-Nominale*
Sono misure di associazione che vengono utilizzate per valutare la relazione tra due caratteri qualitativi. Vengono dette misure Nominale-Nominale perché possono essere calcolate per caratteri di qualunque natura (a differenza di altre misure che richiedono che le modalità siano almeno ordinabili o quantitative).
Cominciamo col dire che per associazione si intende in questo caso una generica tendenza di alcune modalità ad associarsi (cioè la presenza di un'attrazione tra le modalità dei due caratteri). Per valutare la forza di tale associazione (connessione) si fa sostanzialmente riferimento alle distribuzioni condizionate [*Analisi condizionata: le distribuzioni di frequenza*]. L'idea è che se i due caratteri non presentano alcun tipo di legame, e sono quindi detti ***indipendenti***, tutte le distribuzioni condizionate coincidono tra loro. Gli indici di associazione nominale-nominale misurano, in modo diverso, l'allontanamento dalla situazione di indipendenza.

Indici SIMMETRICI

CHI-QUADRATO DI PEARSON. Indicatore simmetrico di associazione tra caratteri di natura qualunque, basato sulle contingenze, ovvero le differenze tra frequenze assolute congiunte effettivamente rilevate e le frequenze assolute congiunte che si otterrebbero nel caso di indipendenza tra i caratteri.
Questo indice assume solo valori non negativi e vale 0 se e solo se i due caratteri sono indipendenti.
Il valore dell'indice chi-quadrato dipende dalla numerosità della popolazione e dal numero di modalità assunte dei due caratteri: aumenta con esse a parità di livello di associazione. Per questo motivo vengono introdotti indicatori ottenuti normalizzando l'indice chi-quadrato, cioè rendendolo indipendente dall'ampiezza della popolazione e dalle dimensioni della tabella.
Pur potendo, formalmente, essere calcolato per coppie di caratteri di natura arbitraria, questo indice *è poco significativo per caratteri quantitativi con molte modalità distinte* (ad esempio caratteri "continui") ed è utile in particolare per coppie di caratteri qualitativi nominali, in quanto rileva il grado di associazione non strutturata, ovvero informa su quanto i due caratteri sono genericamente legati senza indicazioni su come lo siano.

CONTINGENZA QUADRATICA MEDIA (PHI-QUADRATO DI PEARSON). Indicatore simmetrico di associazione non strutturata, si ottiene dividendo per la numerosità *n* della popolazione l'indice chi-quadrato di Pearson; in questo modo, il livello di associazione non dipende più da tale quantità. Ovviamente, conserva tutte le caratteristiche e le proprietà del chi-quadrato, per cui vale 0 se e solo se vi è indipendenza tra i caratteri e assume il valore massimo nel caso di massima connessione, cioè quando almeno uno dei due caratteri dipende perfettamente dall'altro, ovvero le distribuzioni subordinate di uno dei due caratteri rispetto alle modalità dell'altro sono tutte degeneri, cioè concentrate su un'unica modalità.
Il valore dell'indice dipende ancora dalle dimensioni della tabella: in particolare il suo valore massimo (assunto nel caso di massima connessione) è pari al minimo tra *r*-1 e *c*-1, dove *r* e *c* sono il numero di modalità distinte del primo e, rispettivamente, del secondo carattere, quando vi è massima connessione tra essi.

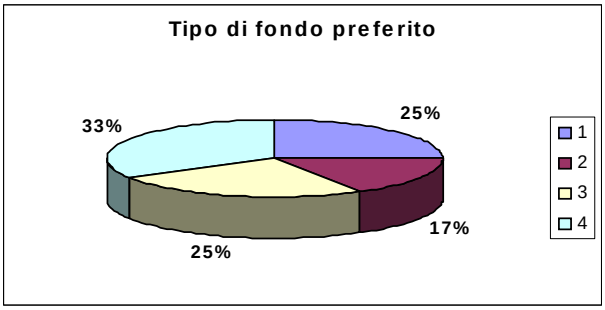
V DI CRAMER. Indice simmetrico di associazione, ottenuto come radice quadrata della contingenza quadratica media divisa per il suo valore massimo, dato dal minimo tra r-1 e c-1, dove r e c sono il numero di modalità distinte del primo e, rispettivamente, del secondo carattere, quando

vi è massima connessione tra essi. Per sua stessa costruzione, quindi, assume valori compresi tra 0 e 1, estremi inclusi. Vale 0 se e solo se vi è indipendenza tra i caratteri, vale 1 se e solo se vi è perfetta connessione, ovvero almeno uno dei due caratteri dipende perfettamente dall'altro (se il numero delle modalità del primo carattere e quello del secondo sono uguali, cioè se la tabella relativa alla distribuzione congiunta è quadrata, necessariamente la dipendenza perfetta di uno dei due caratteri dall'altro implica anche quella inversa). Come l'indice chi-quadrato, da cui è derivato, il V di Cramer fornisce informazioni non significative se riferito a caratteri "continui"; il suo obiettivo è quello di fornire indicazioni sul livello di associazione non strutturata tra caratteri, particolarmente qualitativi nominali.

Indici ASIMMETRICI

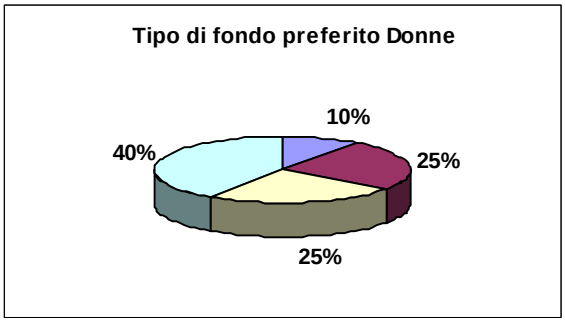
Gli indici basati sul chi-quadrato sono indici ***simmetrici***, nel senso che si è interessati a valutare l'esistenza di una relazione senza curarsi della *direzione* dell'associazione stessa. In alcuni casi, se i caratteri non sono indipendenti, si può pensare di utilizzare uno dei due caratteri per prevedere l'altro. Gli indici che misurano la forza dell'associazione facendo riferimento alla capacità previsiva del carattere indipendente si dicono ***asimmetrici***.

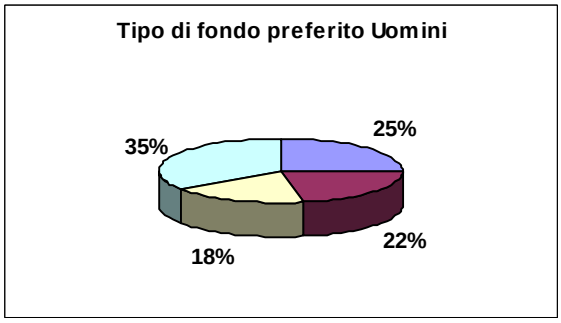
LAMBDA DI GOODMAN E KRUSKAL. Indice asimmetrico di associazione, volto a rilevare l'entità della dipendenza di un carattere *Y* da un carattere *X*, misurando il miglioramento nella previsione di *Y* che si ottiene conoscendo *X*, rispetto alla previsione ottenuta esclusivamente sulla base della distribuzione marginale di *Y*. Si basa sull'idea che, nel caso in cui non si possiedano informazioni sul carattere esplicativo, *X*, il carattere dipendente verrà previsto con la ***moda marginale***. Se invece si tiene conto di *X* si potrà prevedere il carattere con le ***mode condizionate***.
Calcolabile, in teoria, per caratteri di natura arbitraria (essendo basato sulle frequenze), è utilizzato soprattutto con riferimento a caratteri qualitativi nominali. Assume valori compresi tra 0 e 1, estremi inclusi; in particolare, vale 0 se e solo se la previsione di *Y* è la stessa, che si utilizzi o no la conoscenza di *X*, mentre vale 1 se e solo se ogni riga della tabella corrispondente alla distribuzione congiunta dei caratteri ha un'unica cella con frequenza positiva.
Il fatto che l'indice Lambda assuma valore nullo non implica che i due caratteri siano indipendenti: distribuzioni condizionate diverse possono infatti essere caratterizzate dalla stessa moda.



Supponiamo ad esempio che si sia interessati a studiare qual è tipo di fondo (tra 4 possibili fondi) preferito da un generico investitore.
La distribuzione (univariata) del carattere è rappresentata dal diagramma a torta di fianco.
Notiamo che la ***moda*** del carattere è **Fondo 4** (preferito dal 33% dei clienti)

Consideriamo ora le distribuzioni del tipo di fondo preferito condizionate al sesso del cliente riportate nei diagrammi di seguito: le due distribuzioni condizionate sono diverse tra loro (i due caratteri non sono indipendenti), ma le differenze non si riflettono nelle mode, entrambe **Fondo4**.





E' la "versione nominale" dell'Indice Eta quadro. Esiste anche una versione ***simmetrica dell'indice*** ottenuta a partire dalle due misure asimmetriche.

COEFFICIENTE DI INCERTEZZA (UNCERTAINTY)
Indice asimmetrico di associazione, volto a rilevare l'entità della dipendenza di un carattere *Y* da un carattere *X*. Per farlo, si misura l'eterogeneità della variabile dipendente utilizzando l'entropia (una "versione nominale" della varianza). L'entropia viene valutata nell'intera popolazione e nelle sottopopolazioni indotte dalle modalità della variabile esplicativa (si considera una media di tali entropie). Si calcola quindi la percentuale di entropia della variabile dipendente spiegata dalla variabile esplicativa (più precisamente si considera la differenza tra l'entropia di *Y* nell'intera popolazione e l'entropia media nelle sotto-popolazioni indotte da *X;* tale differenza viene divisa per l'entropia di *Y* nell'intera popolazione). E' l'analogo qualitativo dell'indice di determinazione, R^2.
Anche in questo caso, un indice nullo non implica che i due caratteri siano indipendenti.
Esiste anche una versione ***simmetrica dell'indice*** ottenuta a partire dalle due misure asimmetriche.

Misure di associazione *Ordinale-Ordinale*
Sono misure di associazione che vengono utilizzate per valutare la relazione tra due caratteri le cui modalità siano almeno ordinabili (quindi i caratteri devono essere almeno qualitativi ordinali).
In questo caso si indaga su un particolare tipo di associazione: consideriamo cioè un'associazione strutturata, ovvero valutiamo se a modalità elevate di un carattere tendono ad essere associate modalità elevate dell'altro.

TAU DI KENDALL. Indice simmetrico di associazione strutturata, relativo a coppie di caratteri qualitativi ordinali oppure quantitativi. Informa su quanto concordanti (o discordanti) siano due caratteri, quindi rileva, oltre al grado dell'associazione, anche il suo verso; in altri termini, misura l'entità della tendenza dei due caratteri ad associarsi in modo che a modalità di ordine elevato di un carattere corrispondano, con frequenze maggiori, modalità di ordine elevato dell'altro carattere, o viceversa. Questo indice è basato sul conteggio del numero di coppie di unità ordinate nello stesso modo su entrambi i caratteri e del numero di coppie di unità ordinate in modo opposto. Assume valori compresi tra -1 e 1; vale 1 se e solo se le unità del collettivo sono ordinate esattamente allo stesso modo rispetto ai due caratteri, vale -1 se e solo se le graduatorie sono esattamente invertite. Il valore 0 indica sostanziale indifferenza tra i due caratteri rispetto alla concordanza o discordanza; ciò non implica la indipendenza, che è condizione più forte. Tale indice è particolarmente utilizzato con riferimento a dati qualitativi ordinali, oppure a dati quantitativi ma con un numero ridotto di modalità distinte.

RHO DI SPEARMAN. Indice simmetrico di associazione, utilizzato per rilevare il grado di concordanza o discordanza tra caratteri con modalità ordinabili; è basato sui ranghi, ovvero sulle graduatorie di n unità rispetto a due criteri. Assume valori compresi tra -1 (che corrisponde a graduatorie perfettamente opposte) e 1 (che corrisponde a graduatorie, rispetto ai due criteri, perfettamente identiche). Il valore 0 indica assenza di concordanza/discordanza, che non necessariamente significa assenza di legame (cioè indipendenza). Quando riferito a caratteri quantitativi, è particolarmente robusto rispetto a valori estremi (outliers), in quanto basato sui ranghi e non sulle modalità effettive; questo fatto lo rende più significativo, in certi casi, di altri indici di associazione per caratteri quantitativi (come il coefficiente di correlazione lineare), che invece risentono della presenza di valori anomali.

GAMMA DI GOODMAN E KRUSKAL. Indice simmetrico di associazione strutturata, relativo a coppie di caratteri qualitativi ordinali oppure quantitativi. Questo indice informa su quanto

concordanti (o discordanti) siano due caratteri, misurando la riduzione dell'errore che si commette nel prevedere come una coppia di unità si ordina rispetto alle modalità di un carattere, quando si conosce il loro ordinamento rispetto all'altro carattere. Assume valori compresi tra -1 e 1; vale 1 se e solo se le unità del collettivo sono ordinate esattamente allo stesso modo rispetto ai due caratteri, vale -1 se e solo se le graduatorie sono esattamente invertite. Il valore 0 indica sostanziale indifferenza tra i due caratteri rispetto alla concordanza o discordanza; ciò non implica l'indipendenza, che è condizione più forte. Tale indice è particolarmente utilizzato con riferimento a dati qualitativi ordinali, oppure a dati quantitativi ma con un numero ridotto di modalità distinte.

COEFFICIENTE DI CORRELAZIONE LINEARE. Indicatore (relativo) del livello e del verso dell'***associazione lineare*** tra due caratteri quantitativi. Può assumere valori compresi tra –1 e 1 (estremi inclusi); valori positivi indicano associazione positiva (cioè, tendenza delle modalità dei due caratteri ad associarsi in modo concordante), valori negativi associazione ti tipo discordante. Quanto più il coefficiente di correlazione lineare, in valore assoluto, è vicino a 1, tanto più è elevato il grado di associazione lineare; nel caso in cui il suo valore è 1 (–1), l'associazione lineare è perfetta e i punti del grafico di dispersione corrispondente sono tutti allineati su una retta con coefficiente angolare positivo (rispettivamente negativo). Si noti che un valore del coefficiente di correlazione lineare pari a 0 non indica mancanza di associazione tra i due caratteri, ma mancanza di associazione lineare; potrebbero cioè essere presenti altri tipi di legame. E' un indicatore da utilizzare però con cautela nel caso in cui siano presenti, nella distribuzione congiunta, degli outliers, facilmente rilevabili nel corrispondente grafico di dispersione. E' sempre utile, quindi, valutare l'affidabilità del coefficiente di correlazione lineare attraverso un'analisi del corrispondente grafico di dispersione, per non incorrere in errori di interpretazione anche molto significativi [*Cautele nella valutazione del coefficiente di correlazione lineare e di determinazione*].

Misure di associazione *Quantitativo-Nominale*

RAPPORTO DI CORRELAZIONE ETA QUADRO

Indicatore volto a misurare la dipendenza di un carattere quantitativo *Y* da un carattere qualitativo (o anche quantitativo, ma con poche modalità distinte) *X*. E' ovviamente un indice asimmetrico, ottenuto dal confronto tra le distribuzioni subordinate di *Y* rispetto alle modalità di *X*, basato sul miglioramento della previsione del carattere *Y* utilizzando le informazioni su *X*. Assume valori compresi tra 0 ed 1; vale 0 se e solo se la previsione ottimale di *Y* non tiene conto di *X*, vale 1 se e solo se la previsione di *Y* basata su *X* è perfetta, ovvero l'errore di previsione è nullo. Il valore 0, comunque, non implica indipendenza tra i due caratteri, mentre in caso di indipendenza, al contrario, questo indicatore assume valore 0. [*Analisi stratificata: le misure di sintesi*]

Cautele nella valutazione delle misure di associazione

Quando si analizzano tabelle a doppia entrata e si sintetizzano le informazioni sull'associazione tra due caratteri per mezzo di indici bisogna prestare attenzione ai cosiddetti ***zeri strutturali***. Consideriamo la distribuzione congiunta di due caratteri e supponiamo che sia possibile rappresentarla per mezzo di una tabella a doppia entrata, in quanto entrambi i caratteri presentano un numero limitato di modalità. Supponiamo che una o più celle della tabella siano caratterizzate da frequenza nulla, cioè che nessuno degli individui considerati è caratterizzato da una certa coppia di modalità. A livello campionario dobbiamo distinguere due situazioni.

1) La frequenza congiunta nulla è dovuta al fatto che è bassa la percentuale di soggetti che presentano la coppia di modalità in esame nella popolazione. A livello campionario può quindi accadere che nessun individuo presenti la coppia di modalità *rara*. Diremo in questo caso che lo zero è uno zero ***campionario***.

2) La frequenza congiunta nulla è dovuta al fatto che è *impossibile* osservare una unità statistica caratterizzata dalla coppia di modalità in esame. Ad esempio, se stiamo rilevando per un collettivo di individui i due caratteri "Età" e "Numero di figli maggiorenni" è evidente che i soggetti con età inferiore ai 18 non potranno avere figli maggiorenni. Allo stesso modo se stiamo rilevando per un collettivo di clienti di una certa società il "Numero di anni trascorsi dal momento del primo contatto con la società" e il "Numero di anni durante il quale il cliente è stato attivo (cioè ha stipulato almeno un contratto con la società") è evidente che un soggetto che è cliente da soli due anni non potrà essere stato attivo per più di due anni: non potremo quindi osservare clienti con 3 anni o più di attività tra questi clienti.

Per comprendere come mai la presenza di zeri strutturali può portare a indici di associazione non attendibili, consideriamo la seguente tabella a doppia entrata, relativa a due caratteri *X* e *Y*.

X * Y Crosstabulation

			Y					
			1.00	2.00	3.00	4.00	5.00	Total
X	1.00	Count	60	90	60	75	15	300
		% within X	20.0%	30.0%	20.0%	25.0%	5.0%	100.0%
	2.00	Count	45	78	52	67	8	250
		% within X	18.0%	31.2%	20.8%	26.8%	3.2%	100.0%
	3.00	Count	88	116	76	92	28	400
		% within X	22.0%	29.0%	19.0%	23.0%	7.0%	100.0%
Total		Count	193	284	188	234	51	950
		% within X	20.3%	29.9%	19.8%	24.6%	5.4%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,921	8	,545
Likelihood Ratio	7,129	8	,523
Linear-by-Linear Association	,015	1	,903
N of Valid Cases	950		

Notiamo che le distribuzioni di *Y* condizionate ai diversi valori di *Y* sono piuttosto simili tra loro. In effetti, la statistica Chi-quadrato porta ad accettare l'ipotesi di indipendenza tra i due caratteri (ipotesi sotto la quale le distribuzioni condizionate sono tutte uguali tra di loro).

Consideriamo ora una seconda tabella ottenuta aggiungendo alla prima una riga, relativa ad una quarta modalità di *X*:

X *Y Crosstabulation

Count

		Y					
		1,00	2,00	3,00	4,00	5,00	Total
X	1,00	60	90	60	75	15	300
	2,00	45	78	52	67	8	250
	3,00	88	116	76	92	28	400
	4,00	99	10				109
Total		292	294	188	234	51	1059

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	252,937	12	,000
Likelihood Ratio	247,909	12	,000
Linear-by-Linear Association	53,521	1	,000
N of Valid Cases	1059		

Notiamo che l'ultima riga contiene alcune celle vuote. Il test Chi-quadrato in questo caso a rifiutare l'ipotesi di indipendenza. Ciò è dovuto al fatto che la distribuzione di Y condizionata a $X = 4$ (l'ultima riga) risulta piuttosto diversa dalle altre distribuzioni condizionate.
Ora è evidente che se le celle vuote nella tabella sono zeri strutturali, cioè se era impossibile osservare coppie con $X = 4$ e Y maggiore di 2, l'ipotesi di indipendenza viene rifiutata solo per "motivi tecnici", cioè perché abbiamo considerato nell'analisi anche una riga che ha necessariamente una struttura diversa da quella che caratterizza le altre righe. I valori assunti dagli indici di associazione in questo caso, andrebbero analizzati con una certa attenzione e le conclusioni andrebbero tratte con una certa cautela.

La collassabilità di una tabella

Nel cercare di identificare dei profili significativi di clienti, possiamo cominciare partendo da una tabella a molte vie, così da visualizzare le numerosità a seconda di ciascun incrocio delle molte variabili in esame.
Ovviamente, l'analisi di tabelle a più vie è piuttosto complicata dal punto di vista descrittivo. Un metodo per semplificare l'analisi di una tabella a più vie è quello di ***collassare*** la tabella con riferimento ad un certo carattere. L'idea è quella di trascurare uno dei caratteri considerati. Consideriamo ad esempio la Tabella 1 contenente le frequenze congiunte relative a tre caratteri rilevati su un collettivo di clienti di una società: Sesso, Classe di Età (codificata in due classi) e Redditività del cliente per l'azienda (Bassa, Alta).

Tab. 1. Tabella a 3 vie di (Sesso, Età, Redditività) e Misure di associazione

				Redditività		
Sesso				Alta	Bassa	Total
Femmina	Età	Alta	Count	77	34	111
			% within Età	69,4%	30,6%	100,0%
			% of Total	15,4%	6,8%	22,2%
		Bassa	Count	192	196	388
			% within Età	49,5%	50,5%	100,0%
			% of Total	38,5%	39,3%	77,8%
Maschio	Età	Alta	Count	91	40	131
			% within Età	69,5%	30,5%	100,0%
			% of Total	16,0%	7,0%	23,0%
		Bassa	Count	220	218	438
			% within Età	50,2%	49,8%	100,0%
			% of Total	38,7%	38,3%	77,0%

Chi-Square Tests: Associazione tra Età e Redditività

Sesso		Value	df	Asymp. Sig. (2-sided)
Femmina	Pearson Chi-Square	13,735	1	,000
	Likelihood Ratio	14,092	1	,000
Maschio	Pearson Chi-Square	15,058	1	,000
	Likelihood Ratio	15,455	1	,000

Symmetric Measures: Associazione tra Età e Redditività

Sesso			Value	Approx. Sig.
Femmina	Nominal by Nominal	Phi	,166	,000
		Cramer's V	,166	,000
		Contingency Coefficient	,164	,000
Maschio	Nominal by Nominal	Phi	,163	,000
		Cramer's V	,163	,000
		Contingency Coefficient	,161	,000

La prima considerazione interessante da fare è che la distribuzione delle frequenze (relative) congiunte di Età e Redditività risulta abbastanza simile nella sotto-popolazione dei maschi e delle femmine. Esattamente come nel caso in cui valutiamo se una distribuzione va studiata a livello marginale o condizionando ai valori di un'altra variabile [*Analisi condizionata: le distribuzioni di frequenza*] così in questo caso ci stiamo chiedendo se per studiare la relazione tra Età e Redditività è opportuno considerare anche il sesso di un cliente.
Notiamo che, essendo le distribuzioni congiunte molto simili nelle sottopopolazioni dei maschi e delle femmine, risultano molto simili anche le distribuzioni della redditività condizionata alla fascia di età. Sono più redditizi i clienti con età Alta (il 70% dei clienti con età Alta risulta caratterizzato da un'elevata redditività, mentre solo il 50% dei clienti con età Bassa risulta altamente redditizio).

La similitudine tra le tabelle condizionate al sesso si riflette anche nelle misure di associazione, che sono molto simili nelle due sotto-popolazioni dei maschi e delle femmine.
Le considerazioni appena fatte suggeriscono che l'analisi della relazione tra Età e Redditività può essere condotta trascurando il sesso. Possiamo quindi pensare di collassare la tabella rispetto al sesso, unendo in un'unica popolazione maschi e femmine. La tabella collassata si ottiene ovviamente sommando le frequenze congiunte relative ad una certa coppia di modalità (relativa a Redditività ed Età) per i maschi e per le femmine.

Tab. 2. Tabella a 2 vie di (Età, Redditività) ottenuta collassando la Tab. 1 rispetto al Sesso

			Redditività		
			Alta	Bassa	Total
Età	Alta	Count	168	74	242
		% within Età	69,4%	30,6%	100,0%
		% of Total	15,7%	6,9%	22,7%
	Bassa	Count	412	414	826
		% within Età	49,9%	50,1%	100,0%
		% of Total	38,6%	38,8%	77,3%
Total		Count	580	488	1068
		% within Età	54,3%	45,7%	100,0%
		% of Total	54,3%	45,7%	100,0%

Chi-Square Tests: Associazione tra Età e Redditività

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	28,806	1	,000
Likelihood Ratio	29,561	1	,000

Symmetric Measures: Associazione tra Età e Redditività

		Value	Approx. Sig.
Nominal by Nominal	Phi	,164	,000
	Cramer's V	,164	,000
	Contingency Coefficient	,162	,000

La tabella collassata è molto simile a quelle che caratterizzavano la distribuzione congiunta anche condizionatamente al sesso. Allo stesso modo risultano piuttosto simili a quelle ottenute precedentemente le misure di associazione (non il chi-quadro in quanto il suo valore dipende anche dall'ampiezza campionaria – si noti tuttavia che, anche in questo caso, l'indice chi-quadrato è altamente significativo, come nei due casi precedenti).

Le tabelle a più vie vanno collassate solo dopo un'accurata analisi delle interrelazioni tra I caratteri coinvolti. Questo perché in alcuni casi collassare una tabella può portare a distorsioni nei risultati. Uno di questi "effetti collaterali" è noto come paradosso di Simpson. Illustriamo con un esempio un caso in cui la tabella a più vie non va assolutamente collassata. Consideriamo la Tabella 3 contenente ancora le frequenze congiunte relative ai tre caratteri Sesso, Classe di Età e Redditività rilevati su un certo numero di clienti.
Notiamo che in questo caso le due distribuzioni congiunte di Redditività e Sesso appaiono diverse quando si controlla l'età del cliente (quindi, quando si procede ad un'analisi stratificata rispetto alla classe di età). Analizzando le distribuzioni della Redditività subordinate al sesso si osserva come ad essere più redditizi, condizionatamente al sesso, sono i clienti maschi: le percentuali di maschi con elevata redditività sono più alte delle percentuali di femmine con elevata redditività sia nella popolazione dei clienti più giovani che in quella dei clienti con età più elevata.

Tab. 3. Tabella a 3 vie di (Età, Sesso, Redditività)

Età				Redditività		Total
				Alta	Bassa	
Alta	Sesso	Femmina	Count	280	159	439
			% within Sesso	63,8%	36,2%	100,0%
			% of Total	49,1%	27,9%	77,0%
		Maschio	Count	91	40	131
			% within Sesso	69,5%	30,5%	100,0%
			% of Total	16,0%	7,0%	23,0%
Bassa	Sesso	Femmina	Count	90	100	190
			% within Sesso	47,4%	52,6%	100,0%
			% of Total	18,0%	20,0%	38,0%
		Maschio	Count	160	150	310
			% within Sesso	51,6%	48,4%	100,0%
			% of Total	32,0%	30,0%	62,0%

Collassando la tabella otteniamo la Tabella 4: abbiamo un'informazione molto diversa da quella ottenuta considerando anche l'età del cliente: a risultare più redditizie sono le donne (il 58% delle donne ha alta redditività, mentre solo il 51% degli uomini ha alta redditività).

Tab. 4. Tabella a 2 vie di (Sesso, Redditività) ottenuta collassando la Tab. 3 rispetto alla classe di Età

			Redditività		Total
			Alta	Bassa	
Sesso	Femmina	Count	370	259	629
		% within Sesso	58,8%	41,2%	100,0%
		% of Total	34,6%	24,2%	58,8%
	Maschio	Count	251	190	441
		% within Sesso	56,9%	43,1%	100,0%
		% of Total	23,5%	17,8%	41,2%
Total		Count	621	449	1070
		% within Sesso	58,0%	42,0%	100,0%
		% of Total	58,0%	42,0%	100,0%

Per comprendere come mai si ha questa inversione nei risultati dobbiamo tenere presente il fatto che tra i clienti con età elevata il 69.5% dei clienti maschi e il 63.8% dei clienti femmine hanno redditività alta. Tra i clienti con età più bassa, le percentuali di clienti con elevata redditività passano rispettivamente a 51.6% e 46.4% per i maschi e per le femmine. Notiamo ora nella Tabella 5 che i maschi con età bassa sono molto più numerosi dei maschi con età elevata. Quindi, quando la tabella è collassata rispetto all'età, l'elevata percentuale di maschi con redditività alta viene fortemente diminuita perché "mediata" con la più bassa percentuale che caratterizza i maschi con età bassa altamente redditizi. Essendo questo secondo gruppo più numeroso, "pesa" di più nella determinazione della percentuale media che risulta quindi inferiore rispetto a quella delle femmine. Per le femmine infatti, si verifica l'esatto contrario: i clienti femmina con elevata età (più redditizi) sono di più rispetto a quelli con età più bassa (meno redditizi).

Il paradosso di Simpson evidenzia che anche se collassare le tabelle a più vie può risultare un'operazione notevolmente semplificatrice, tuttavia è necessario utilizzare una grande cautela per evitare che si verifichino casi come quello descritto.

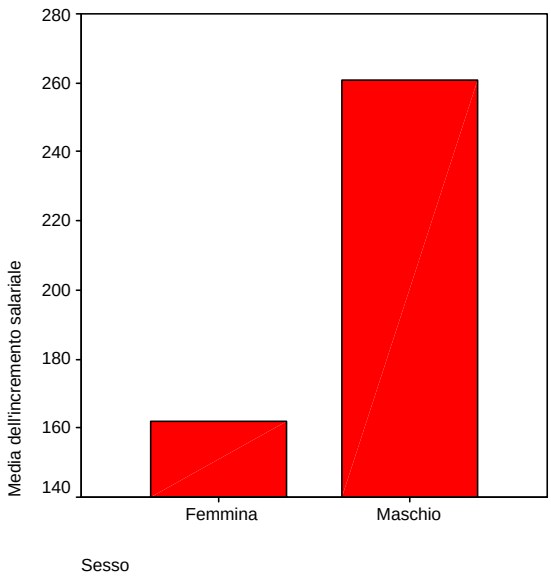
Associazione e causalità

Anche quando due caratteri risultano caratterizzati da una associazione, anche di forte entità, non si può mai "forzare" l'interpretazione di un nesso *causale* tra le variabili stesse.

Per comprenderne il motivo, consideriamo alcuni esempi.

Connessione: Consideriamo un collettivo di fumatori che stanno cercando di smettere di fumare e accettano di partecipare ad una "terapia di gruppo". Dopo un certo periodo di trattamento, si rileva per ogni fumatore se ha smesso o no di fumare; sono inoltre disponibili alcune informazioni sulle caratteristiche socio-demografiche dei soggetti, quali sesso, titolo di studio, composizione del nucleo familiare e così via.
Supponiamo di osservare che il carattere "Smette di fumare" è fortemente associato con il carattere "Titolo di studio": in particolare, l'analisi delle distribuzioni condizionate evidenzia che coloro con elevati titoli di studio risultano meno propensi a smettere di fumare.
Se possiamo dire che risulta evidentemente più complicato smettere di fumare per coloro che hanno un elevato titolo di studio, non possiamo sicuramente affermare che *un elevato titolo di studio indebolisce la forza di volontà nello smettere di fumare*. Per fare considerazioni di questo tipo, dovremo prendere in considerazione altri aspetti del problema trascurati nell'analisi effettuata, che ci portino a comprendere come mai le persone con elevato titolo di studio hanno meno successo nello smettere di fumare. Nell'analisi fatta abbiamo probabilmente trascurato alcuni caratteri rilevanti (uno dei più semplici che ci può venire in mente è, ad esempio, lo stress cui un soggetto è sottoposto durante il lavoro).

Dipendenza in media Supponiamo, come secondo esempio, di rilevare per un gruppo di impiegati presso una certa azienda il sesso e l'incremento salariale medio annuale. Le due medie sono riportate nel diagramma che segue.

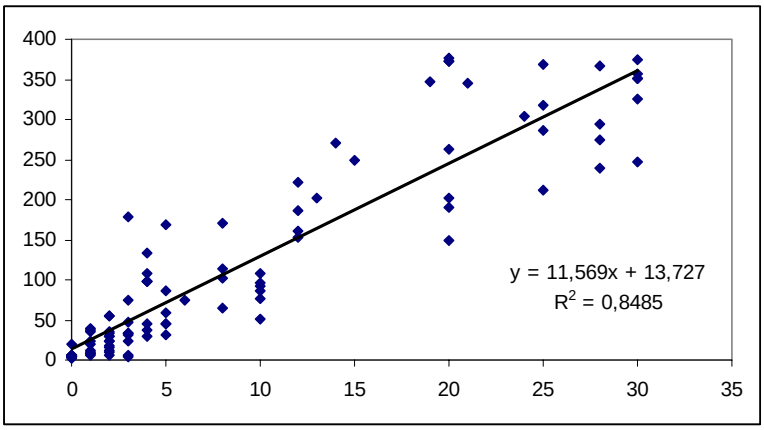


Notiamo che le medie dei due caratteri risultano molto diverse l'una dall'altra. Possiamo quindi concludere che l'incremento salariale dipende in media dal sesso del soggetto considerato. Tuttavia, il termine *dipendenza in media* indica solo che a partire dal sesso è possibile prevedere l'incremento salariale meglio di quanto non si possa fare considerando la media marginale.
Tuttavia, prima di concludere che nell'azienda considerata il fatto di essere una donna *comporta* un incremento salariale inferiore, dovremo valutare attentamente il motivo di questo risultato. Se è innegabile che tra i due caratteri esista un'associazione non è possibile asserire che questa associazione si traduca in un nesso di *causalità*.

Il risultato ottenuto potrebbe essere legato alle mansioni lavorative cui sono preposti maschi e femmine nell'azienda considerata. I maschi potrebbero infatti occupare posizioni legate a salari più elevati (con, quindi, incrementi più sostanziali).

Dipendenza lineare (o correlativa). Come ultimo esempio, consideriamo il diagramma di dispersione riportato di seguito, relativo al Fatturato e al Numero di occupati in un collettivo di aziende. Notiamo che esiste una forte associazione lineare tra i due caratteri: l'indice R^2 è infatti

molto elevato, e la retta di regressione del fatturato sul numero di occupati spiega circa l'84% della varianza del fatturato.

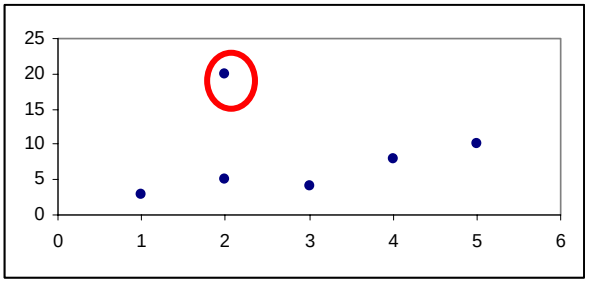


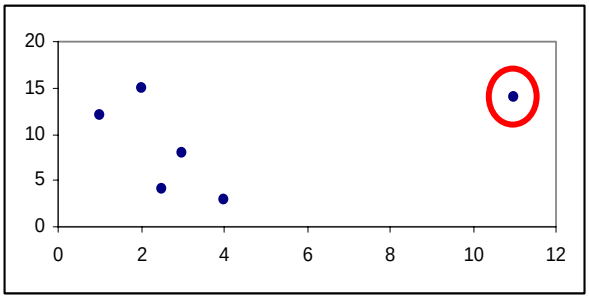
La forza dell'associazione lineare potrebbe quindi essere usata per prevedere il fatturato di un'azienda a partire dall'informazione sul suo numero di occupati (ad esempio, per motivi fiscali, per valutare se il fatturato di un'azienda è plausibile o meno).

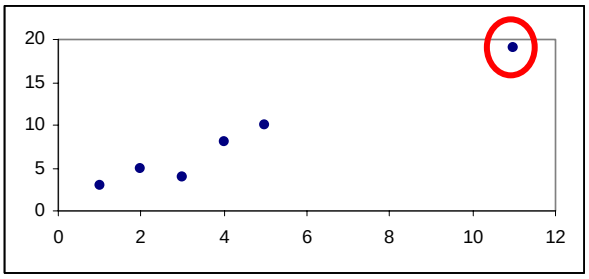
Tuttavia, evidentemente, nulla può farci concludere che il fatturato *dipende* dal numero di occupati e che quindi se un'azienda assume nuovi lavoratori vedrà aumentare il proprio fatturato. Diremo che aziende di grandi dimensioni (e quindi con un numero molto elevato di dipendenti) hanno anche elevati fatturati, e l'esistenza di tale relazione può essere utilizzata per fare previsioni sul fatturato.

Valori estremi e valori anomali (*caso bivariato*)

Ricordiamo che nel caso univariata si definiscono *valori estremi* i valori più grandi o più piccoli di una distribuzione, cioè quelli che si "collocano" in una posizione anomala rispetto agli altri sulla retta dei numeri reali. [*Valori anomali: caso univariato*]. Nel caso bivariato una coppia di modalità è considerata *anomala* quando si colloca in una posizione anomala rispetto alle altre coppie di modalità nel piano cartesiano.







Le osservazioni (bivariate) anomale rispetto ad una coppia di caratteri (*X*, *Y*) non sono necessariamente anomale dal punto di vista univariato, come mostrano i diagrammi di dispersione di fianco. Ad esempio, nel primo diagramma la "coppia anomala" (individuata da un cerchio rosso) è anomala rispetto alla variabile in ordinata ma non rispetto a quella in ascissa; la situazione è opposta per l'osservazione cerchiata nel secondo diagramma, anomala solo rispetto alla variabile in ascissa. Nell'ultimo diagramma di dispersione, invece, l'osservazione risulta anomala con riferimento ad entrambi i caratteri.

Le coppie anomale ***possono*** distorcere il coefficiente di correlazione, ρ [*Non robustezza di* ρ] [*Cautele nella valutazione di* ρ] e attirare a sé la retta di regressione.

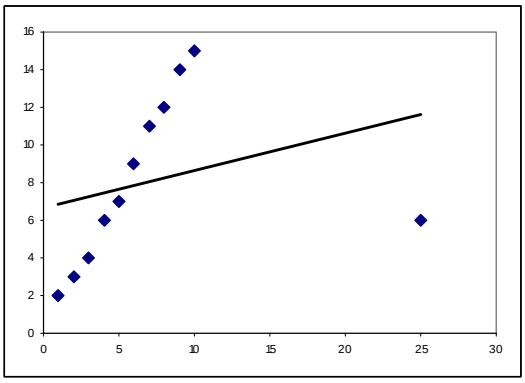
Non robustezza del coefficiente di correlazione lineare

Il *coefficiente di correlazione lineare*, ρ, misura la *forza* della relazione lineare che lega tra loro due caratteri ***quantitativi***.
Come noto, la media è una misura di sintesi non robusta, in quanto fortemente condizionata da valori estremi o anomali [*Valori estremi e outliers*]. Il coefficiente di correlazione lineare tra due generici caratteri, X e Y , è definito a partire da medie:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{M(XY) - M(X)M(Y)}{\sqrt{M(X^2) - [M(X)]^2}\sqrt{M(Y^2) - [M(Y)]^2}}$$

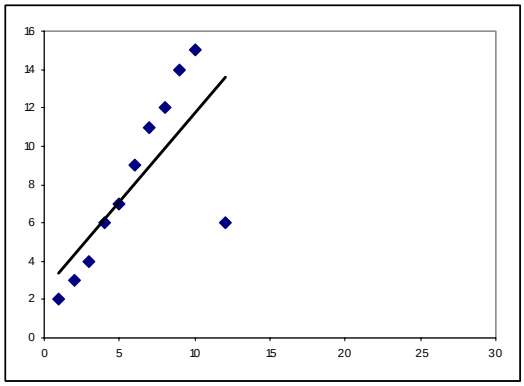
X_1	X_2	X_3	Y
1	1	1	2
2	2	2	3
3	3	3	4
4	4	4	6
5	5	5	7
6	6	6	9
7	7	7	11
8	8	8	12
9	9	9	14
10	10	10	15
25	12	8	6

Conseguentemente, ρ è una misura di sintesi *non robusta* ed è quindi attratto dalla presenza di valori anomali. Per rendersene conto, è sufficiente considerare il seguente esempio. Nella tabella a seguire sono riportate 11 osservazioni sul carattere dipendente Y e sui tre caratteri esplicativi X_1, X_2, e X_3. Si noti che i tre caratteri esplicativi sono caratterizzati dagli stessi valori eccetto che per l'ultima osservazione.
Di seguito i tre diagrammi di dispersione relativi a (Y, X_1), (Y, X_2) e (Y, X_3), con le rette di regressione.

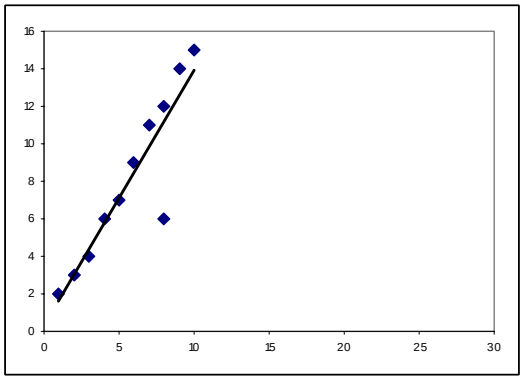


$\rho(Y, X_1) = 0.292$

Notiamo dai diagrammi che modificando il valore assunto dalla variabile esplicativa in corrispondenza dell'ultima osservazione, il coefficiente di correlazione lineare vede modificarsi in modo sostanziale il suo valore, passando da 0.292 a 0.725 a 0.913.
Sostanzialmente, questo esempio evidenzia che basta *una sola osservazione* a modificare drasticamente il valore assunto dal coefficiente di correlazione.



$\rho(Y, X_2) = 0.725$



$\rho(Y, X_3) = 0.913$

Con riferimento al problema della robustezza, possiamo considerare due misure alternative di *concordanza* che risultano meno sensibili alla presenza di valori anomali. Tali misure sono l'indice

di Kendall e l’indice di Spearman.
Per le tre coppie di valori considerati otteniamo i seguenti risultati:

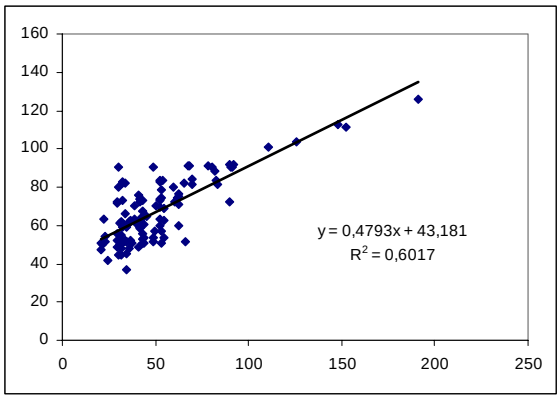
Indice	(Y, X_1)	(Y, X_2)	(Y, X_3)
Pearson Correlation	.292	.725	.913
Kendall's tau_b	.771	.771	.870
Spearman's rho	.779	.779	.911

Notiamo come l’indice tau di Kendall e l’indice di Spearman risultino molto meno sensibili alla variazione del valore della variabile esplicativa per l’ultima osservazione. Tali misure danno meglio conto della “forma” dell’intera nuvola di punti, smorzando il peso assunto dall’osservazione anomala.

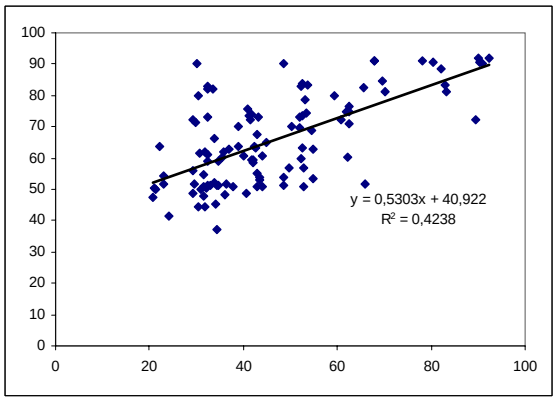
Cautele nella valutazione del coefficiente di correlazione lineare e di determinazione

Il *coefficiente di correlazione lineare*, ρ, misura la *forza* della relazione lineare che lega tra loro due caratteri ***quantitativi***. ρ è una misura di sintesi *non robusta* ed è quindi attratto dalla presenza di valori anomali. [*Non robustezza del coefficiente di correlazione lineare*] In alcuni casi, la presenza di tali valori rende ρ *non affidabile*. Diremo che il valore assunto da ρ è ***affidabile*** quando sintetizza efficacemente la relazione lineare tra i due caratteri descrivendo quindi in modo soddisfacente la "forma" della nuvola dei punti. Stesse considerazioni valgono per il coefficiente di determinazione, $R^2 = \rho^2$, che misura la bontà di adattamento della retta di regressione alla nuvola dei punti. Illustriamo il problema con alcuni esempi.

Consideriamo il diagramma di dispersione a sinistra: si ha ρ = 0.776 (e R^2 = 0.6017). Tali valori suggeriscono una relazione lineare diretta di forte entità tra i due caratteri. Analizzando il diagramma di dispersione, notiamo però alcuni valori anomali che attirano a sé la retta di regressione. Eliminando tali valori risulta ρ = 0.651 e R^2 scende a 0.4238. In questo caso, il valore iniziale di ρ, 0.776, non è affidabile, nel senso che non descrive adeguatamente la relazione tra i due caratteri.

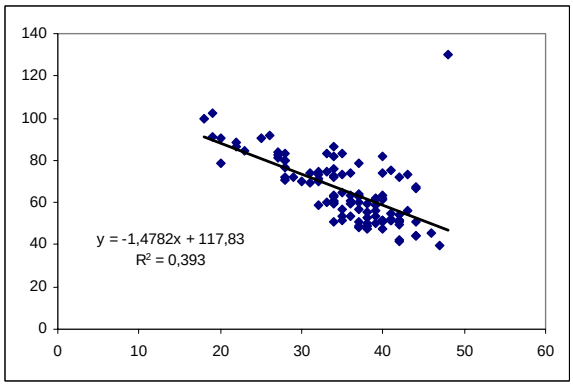


Coefficienti non affidabili

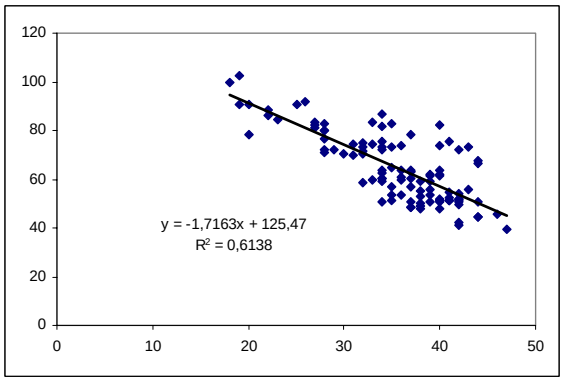


Coefficienti affidabili

Ovviamente, non è sempre detto che osservazioni anomale inflazionino in positivo il valore del coefficiente di correlazione lineare. Nei due diagrammi che seguono troviamo un caso opposto a quello precedente: nel diagramma a sinistra risulta ρ = -0.6269, e R^2 = 0.393. Notiamo tuttavia che la retta di regressione è spostata verso l'alto per la presenza di un'osservazione anomala, che la attira a sé. In questo senso, i coefficienti non sono affidabili in quanto descrivono più una singola osservazione che la totalità dei dati. Rimuovendo tale osservazione ρ passa a -0.7835, indicando una relazione inversa di forte entità tra i due caratteri.



Coefficienti non affidabili



Coefficienti affidabili

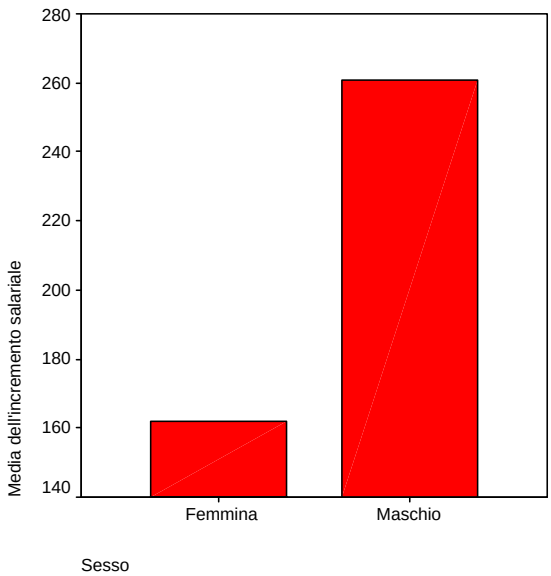
Associazione e causalità

Anche quando due caratteri risultano caratterizzati da una associazione, anche di forte entità, non si può mai "forzare" l'interpretazione di un nesso *causale* tra le variabili stesse.

Per comprenderne il motivo, consideriamo alcuni esempi.

Connessione: Consideriamo un collettivo di fumatori che stanno cercando di smettere di fumare e accettano di partecipare ad una "terapia di gruppo". Dopo un certo periodo di trattamento, si rileva per ogni fumatore se ha smesso o no di fumare; sono inoltre disponibili alcune informazioni sulle caratteristiche socio-demografiche dei soggetti, quali sesso, titolo di studio, composizione del nucleo familiare e così via.
Supponiamo di osservare che il carattere "Smette di fumare" è fortemente associato con il carattere "Titolo di studio": in particolare, l'analisi delle distribuzioni condizionate evidenzia che coloro con elevati titoli di studio risultano meno propensi a smettere di fumare.
Se possiamo dire che risulta evidentemente più complicato smettere di fumare per coloro che hanno un elevato titolo di studio, non possiamo sicuramente affermare che *un elevato titolo di studio indebolisce la forza di volontà nello smettere di fumare*. Per fare considerazioni di questo tipo, dovremo prendere in considerazione altri aspetti del problema trascurati nell'analisi effettuata, che ci portino a comprendere come mai le persone con elevato titolo di studio hanno meno successo nello smettere di fumare. Nell'analisi fatta abbiamo probabilmente trascurato alcuni caratteri rilevanti (uno dei più semplici che ci può venire in mente è, ad esempio, lo stress cui un soggetto è sottoposto durante il lavoro).

Dipendenza in media Supponiamo, come secondo esempio, di rilevare per un gruppo di impiegati presso una certa azienda il sesso e l'incremento salariale medio annuale. Le due medie sono riportate nel diagramma che segue.

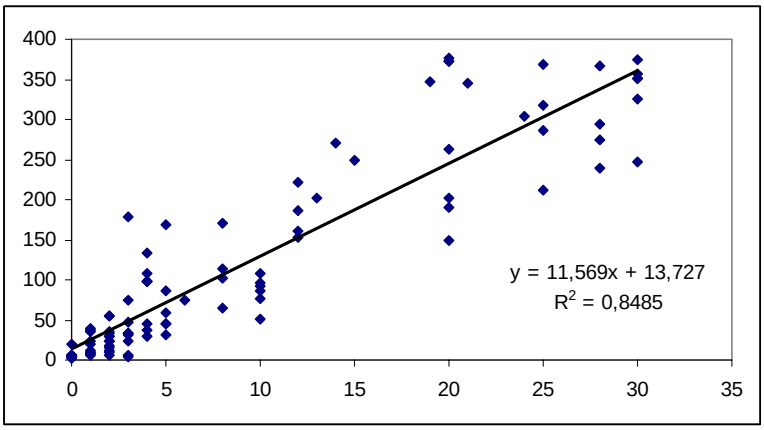


Notiamo che le medie dei due caratteri risultano molto diverse l'una dall'altra. Possiamo quindi concludere che l'incremento salariale dipende in media dal sesso del soggetto considerato. Tuttavia, il termine *dipendenza in media* indica solo che a partire dal sesso è possibile prevedere l'incremento salariale meglio di quanto non si possa fare considerando la media marginale.
Tuttavia, prima di concludere che nell'azienda considerata il fatto di essere una donna *comporta* un incremento salariale inferiore, dovremo valutare attentamente il motivo di questo risultato. Se è innegabile che tra i due caratteri esista un'associazione non è possibile asserire che questa associazione si traduca in un nesso di *causalità*.

Il risultato ottenuto potrebbe essere legato alle mansioni lavorative cui sono preposti maschi e femmine nell'azienda considerata. I maschi potrebbero infatti occupare posizioni legate a salari più elevati (con, quindi, incrementi più sostanziali).

Dipendenza lineare (o correlativa). Come ultimo esempio, consideriamo il diagramma di dispersione riportato di seguito, relativo al Fatturato e al Numero di occupati in un collettivo di aziende. Notiamo che esiste una forte associazione lineare tra i due caratteri: l'indice R^2 è infatti

molto elevato, e la retta di regressione del fatturato sul numero di occupati spiega circa l'84% della varianza del fatturato.

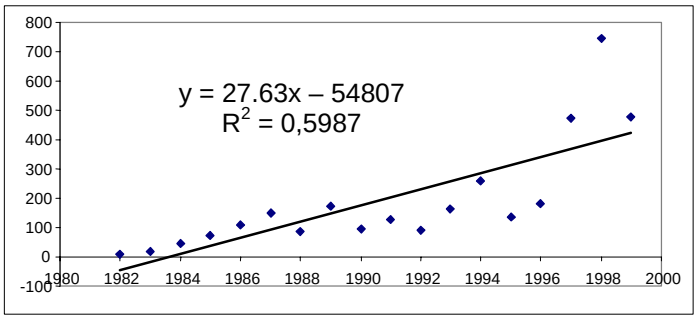


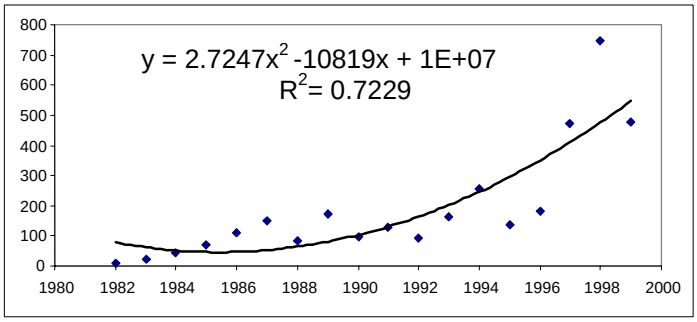
La forza dell'associazione lineare potrebbe quindi essere usata per prevedere il fatturato di un'azienda a partire dall'informazione sul suo numero di occupati (ad esempio, per motivi fiscali, per valutare se il fatturato di un'azienda è plausibile o meno).

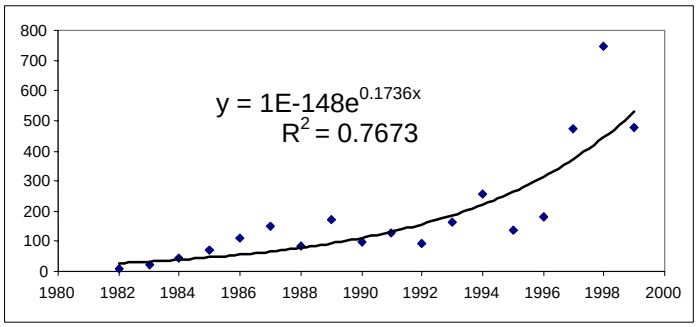
Tuttavia, evidentemente, nulla può farci concludere che il fatturato *dipende* dal numero di occupati e che quindi se un'azienda assume nuovi lavoratori vedrà aumentare il proprio fatturato. Diremo che aziende di grandi dimensioni (e quindi con un numero molto elevato di dipendenti) hanno anche elevati fatturati, e l'esistenza di tale relazione può essere utilizzata per fare previsioni sul fatturato.

Interpolazione di una serie di dati: quale funzione scegliere?

Consideriamo una coppia di caratteri quantitativi (*X*, *Y*) rilevati su un certo collettivo di unità statistiche. Le coppie di modalità osservate possono essere riportate su un diagramma di dispersione, che consente di visualizzare la distribuzione congiunta dei due caratteri cioè, in questo caso, il "modo" in cui le coppie si dispongono nel piano cartesiano e la forma assunta dalla nuvola di punti delle modalità osservate. Il diagramma di dispersione consente di fare alcune considerazioni sul tipo di relazione che lega uno dei due caratteri, ad esempio *Y*, quello le cui modalità sono – per convenzione – riportate sull'asse delle ordinate, all'altro, *X*. Possiamo perciò valutare se tra i due caratteri esiste una relazione diretta (i valori di *Y* crescono all'aumentare dei valori di *X*) o inversa (i valori di *Y* diminuiscono all'aumentare dei valori di *X*).
Se si vuole utilizzare *X* per prevedere *Y*, si cercherà di individuare una opportuna *funzione analitica di X*, che *interpoli* in modo soddisfacente la nuvola di punti osservata, cioè che "passi" vicino a tutti i punti nel diagramma di dispersione.
L'idea è quella di trovare una funzione che approssimi bene la relazione tra *Y* e *X*, di cui le osservazioni registrate sono un set di possibili valori. Allo stesso tempo, però, l'interesse è anche quello di non inserire troppi parametri nell'equazione. Sostanzialmente, si cercherà di trovare la miglior interpolazione delle coppie di modalità osservate ottenibile con un'equazione con un numero di parametri limitato.
Una misura della bontà dell'adattamento dell'interpolante è data dal coefficiente di determinazione, R^2, che misura la percentuale di varianza di *Y* spiegata dalla funzione di *X* prescelta: tanto più elevato è R^2, tanto più la funzione scelta può essere utilizzata per prevedere *Y* sulla base di *X*.







Anche se siamo abituati a cercare di individuare relazioni di tipo lineare tra coppie di caratteri, non è detto che non ci siano altri tipi di funzioni interpolanti che meglio si prestano a descrivere il tipo di relazione tra i due caratteri, come ad esempio, funzioni polinomiali, logaritmiche o esponenziali di *X*.

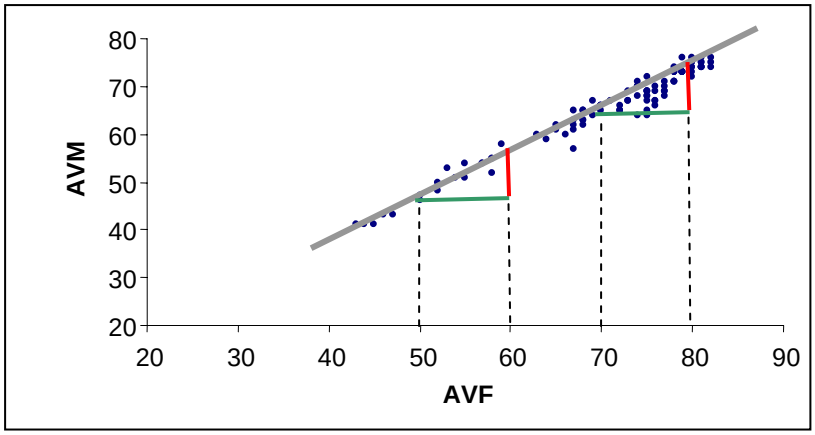
Nei diagrammi di fianco, sono state utilizzate 3 diverse funzioni interpolanti: quella lineare, quella polinomiale di secondo grado, e quella esponenziale.
Visivamente ci rendiamo conto che le ultime due funzioni approssimano meglio la relazione tra i due caratteri.

Questo è confermato anche dai valori di R^2 che risultano rispettivamente, 0,60; 0,72 e 0,77.
La miglior interpolante (in termini di capacità esplicativa) è quindi l'interpolante esponenziale, che consente anche una descrizione più parsimoniosa (vengono utilizzati solo 2 parametri) della relazione tra *X* e *Y* rispetto a quella fornita dallo sviluppo polinomiale.

L'impatto della variabile esplicativa: caso lineare e non lineare

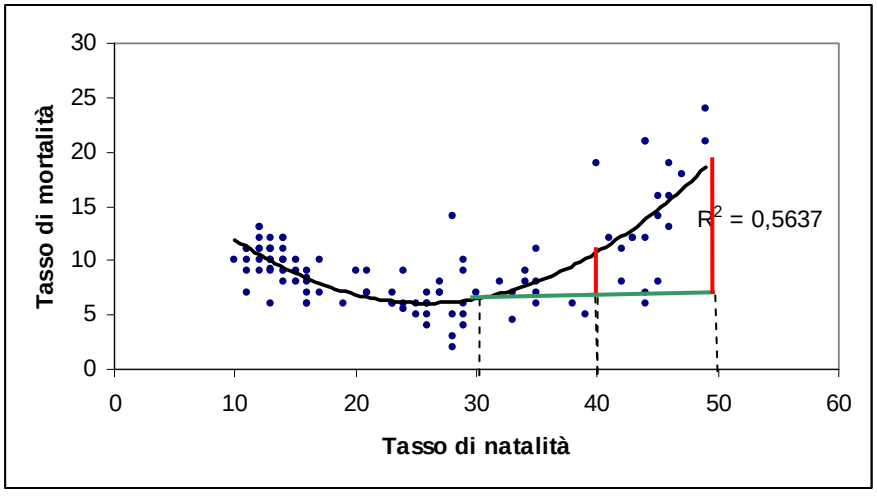
Quando cerchiamo di descrivere il legame tra due variabili tramite una forma funzionale, lo scopo è quello di descrivere come variano i valori di un carattere (considerato come dipendente) al diminuire o all'aumentare dell'altro (detto variabile indipendente). Scegliere una forma funzionale anziché un'altra implica però descrivere in modo diverso il rapporto tra le due variabili. Qui di seguito, ecco 3 esempi chiarificatori.

Il primo esempio riguarda il rapporto che sussiste tra la speranza di vita alla nascita dei maschi e delle femmine.



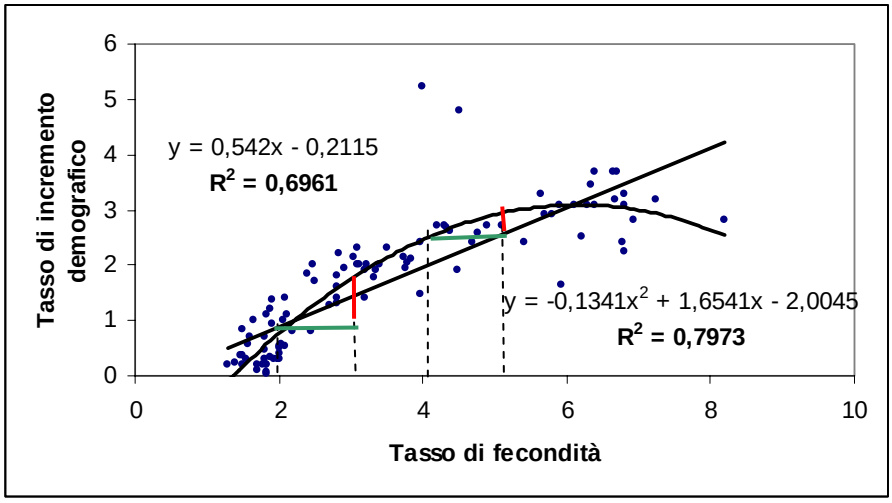
Ipotizzando un legame lineare imponiamo che una variazione di un anno nella speranza di vita delle donne porta ad una certa variazione nella speranza di vita degli uomini, e che tale variazione rimane invariata considerando qualsiasi valore della speranza di vita femminile. Non è importante quindi quale sia il valore iniziale assunto dalla speranza di vita femminile, ma importa solo la variazione ipotizzata.
Sostanzialmente, ci aspettiamo che i guadagni nella speranza di vita maschile di un paese che passa da una speranza di vita femminile di 50 anni a 51 anni siano gli stessi di un altro paese, che passa da una speranza di vita femminile di 79 anni a 80 anni.

Caso diverso è invece quello di un legame non lineare. Consideriamo ad esempio il diagramma di seguito dove sono rappresentati i tassi di natalità e quelli di mortalità per diversi paesi. La relazione più opportuna in questo caso è di tipo quadratico. Si osservi come l'***impatto*** di una certa variazione del tasso di natalità sul tasso di mortalità cambi a seconda del "punto di partenza".



Ad esempio, nel tratto ascendente della curva, a seguito di un aumento del tasso di natalità da 30 a 40 si osserva una variazione prevista sul tasso di mortalità molto inferiore rispetto a quella conseguente ad una variazione, sempre pari a 10, da 40 a 50.
Per paesi con tassi di natalità superiori a 30 all'aumentare del tasso di natalità si osserva un aumento del tasso di mortalità. A differenza di quanto non accada nel caso lineare, l'entità dell'aumento ***aumenta*** man mano che ci si allontana da 30.
Ovviamente, nel tratto decrescente della curva si nota come un aumento del tasso di natalità è associato ad una diminuzione del tasso di mortalità e che la variazione del tasso di natalità ha un impatto superiore man mano che ci si allontana da 30.

Come ultimo esempio, è interessante valutare che cosa significa quando due caratteri sono legati da una relazione quadratica convessa. Nel diagramma di dispersione viene riportata la distribuzione congiunta del tasso di fecondità e del tasso di incremento demografico. In questo caso notiamo una situazione ancora diversa. Confrontiamo due variazioni della stessa entità del tasso di fecondità, da 2 a 3 e da 4 a 5 nel tratto crescente della curva.



Si osservi come l'impatto sul tasso di incremento demografico risulti decisamente più marcato quando il punto di partenza è il valore più basso, 2. Ciò indica che un aumento del tasso di fecondità in paesi con tassi di fecondità bassi ha un impatto molto più marcato di quanto non accada nei paesi con tasso di fecondità più elevato.

La curva polinomiale da un certo punto in poi addirittura decresce ad indicare non più un affievolimento dell'impatto ma anzi una diminuzione del tasso di incremento demografico all'aumentare del tasso di fecondità. L'interpretazione di questo risultato dal punto di vista "demografico" ed economico è abbastanza immediata: nei paesi più ricchi (quelli con tasso di fecondità più basso) un aumento del tasso di fecondità si traduce in un aumento del tasso di incremento demografico. Nei paesi più poveri, il tasso di mortalità è più elevato e quindi non tutti i nuovi nati sopravvivono, e il tasso di incremento demografico non "reagisce" quindi allo stesso modo a variazioni del tasso di natalità.

Concludendo: quando scegliamo una funzione interpolante ci basiamo solitamente sulle misure di adattamento, ad esempio l'indice di determinazione R^2.
Se invece dell'interpolante lineare scegliamo altri tipi di curve, più adatte, è importante tener conto che tali curve ***contribuiscono*** a migliorare l'analisi sulla relazione che esiste tra i due caratteri. L'analisi della relazione tra due caratteri viene arricchita considerando se e come cambia rispetto al caso lineare la variazione nella variabile dipendente associata ad una certa variazione della variabile esplicativa.

Media mobile come perequazione di una serie di dati

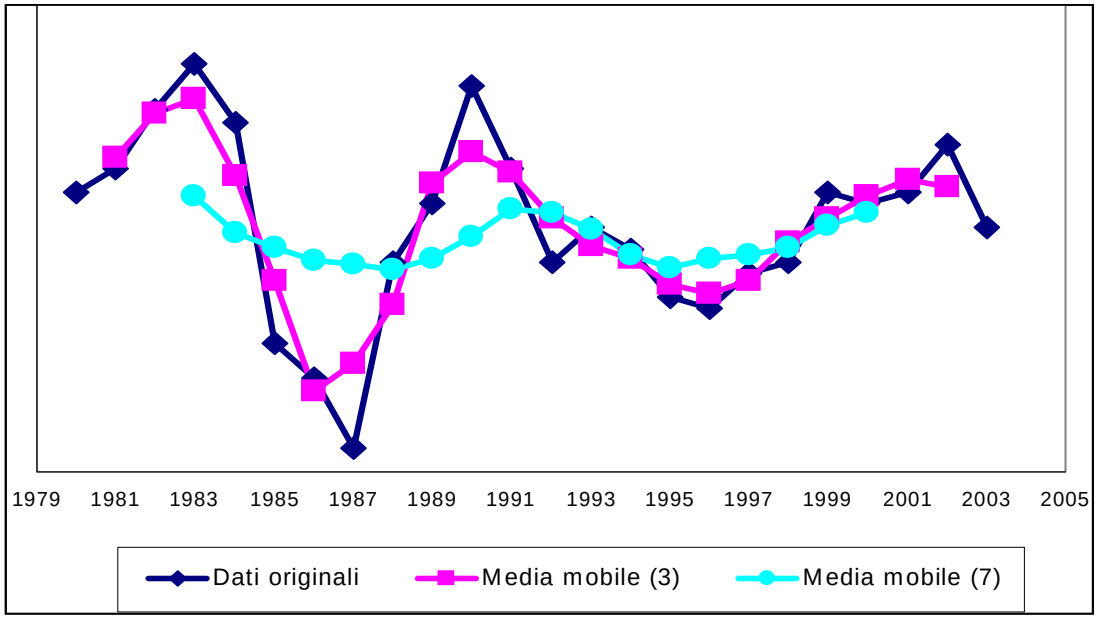
A partire dalla serie di dati originali, si possono costruire tramite le medie mobili di ordine *k* delle serie perequate di dati. Sostanzialmente, se X_t è il valore osservato al tempo *t* questo viene *filtrato* (o *perequato*) con la media dei valori osservati nei periodi adiacenti (passati e futuri). In questo modo, valori anomali vengono "smussati", e serie particolarmente irregolari vengono rese più regolari.
In generale, la **media mobile di ordine (2*m* + 1)** è calcolata come:

$$X_t^* = \sum_{s=-m}^{m} \frac{X_{t+s}}{2m+1} = \frac{\overbrace{X_{t-m} + X_{t-m+1} + \ldots + X_{t-1}\ldots}^{m} + X_t + \overbrace{X_{t+1} + \ldots X_{t+m-1} + X_{t+m}}^{m}.}{2m+1}$$

Si noti che nella serie perequata si ***perdono*** *m* osservazioni all'inizio e *m* osservazioni alla fine.

La media mobile di ordine 3, ad esempio, sostituisce ogni elemento (dove possibile) con la media dell'elemento precedente, elemento stesso e elemento successivo. Ovviamente non sarà possibile filtrare con la media mobile il primo valore (non ha alcuna osservazione prima di sé) e l'ultimo valore (non ha alcuna osservazione dopo di sé).

Nel grafico seguente, sono state riportate la serie di dati originali e la serie di medie mobili di ordine 3 e 7. Medie mobili di ordine superiori smussano (o *lisciano*) di più la serie, fornendo grafici più regolari che permettono di osservare l'andamento di medio/lungo periodo del fenomeno oggetto di studio. Medie mobili di ordine inferiore "seguono invece la serie più da vicino" e consentono di visualizzare l'andamento di breve periodo del fenomeno, attenuando le oscillazioni di brevissimo periodo.



Analisi stratificata

Nell'analisi univariata si studia la distribuzione di un singolo carattere facendo riferimento all'intero collettivo di riferimento.
In alcuni casi, la popolazione può essere naturalmente suddivisa in *strati* (ad esempio, zona geografica di residenza, livello di istruzione, e così via). Molto spesso possiamo aspettarci che la distribuzione del carattere considerato *vari* da strato a strato. In questi casi, può essere opportuno confrontare le distribuzioni nei diversi strati in modo da comprendere se

a) è necessario tener conto dello strato, in quanto la distribuzione del carattere di interesse ***cambia*** da strato a strato;

b) non è necessario tener conto dello strato in quanto le distribuzioni del carattere sono abbastanza ***simili*** (se non addirittura identiche) nei diversi strati.

Quando si studiano congiuntamente due caratteri, gli strati presi in considerazione possono essere quelli indotti dalle modalità di uno dei due caratteri. In questo caso, si parla anche di *analisi condizionata.*

In generale, l'analisi stratificata viene condotta quando il *numero degli strati è piuttosto limitato* (di modo che sono in numero ridotto le sottopopolazioni da prendere in considerazione).

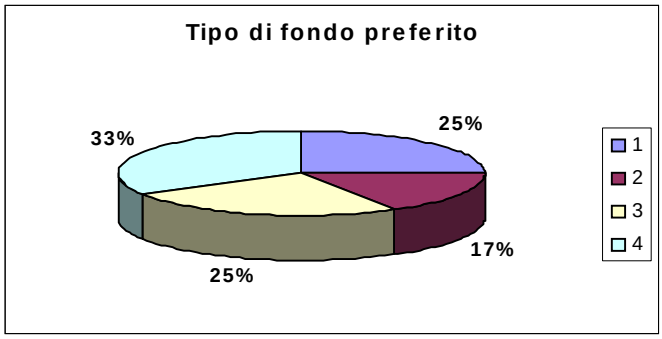
Ovviamente l'analisi stratificata può riguardare caratteri *dipendenti* di diversa natura (qualitativi, nominali o ordinali; quantitativi, discreti o continui). L'analisi stratificata dovrà essere condotta utilizzando strumenti di rappresentazione grafica e sintesi del carattere secondo le stesse modalità utilizzate in ambito di analisi univariata.

Analisi stratificata o condizionata: le distribuzioni di frequenza

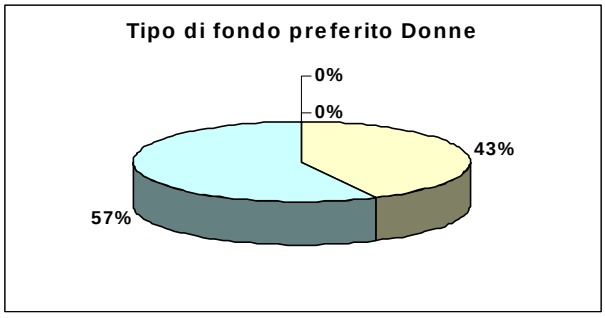
Ricordiamo che si parla di analisi condizionata quando si vuole studiare la distribuzione di un carattere (detto nel seguito carattere *dipendente*) non a livello *marginale* (analisi univariata) cioè considerando l'intera popolazione di interesse ma nelle sottopopolazioni indotte dalle modalità di una seconda variabile (detta nel seguito *variabile esplicativa*).
Molto spesso possiamo aspettarci che la distribuzione del carattere considerato *sia diversa* nelle diverse sottopopolazioni. In questi casi, può essere quindi opportuno confrontare le distribuzioni nelle diverse sottopopolazioni in modo da comprendere se
a) è necessario tener conto della variabile esplicativa, in quanto la distribuzione varia da sottopopolazione a sottopopolazione
b) non è necessario tener conto della variabile esplicativa in quanto le distribuzioni del carattere sono abbastanza simili nelle diverse sottopopolazioni.

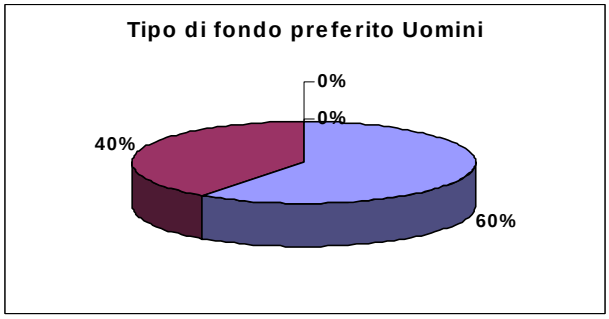
Quando il carattere dipendente assume un numero contenuto di modalità, le distribuzioni condizionate possono essere confrontate *direttamente*. In generale, un'analisi condizionata che procede confrontando direttamente le distribuzioni condizionate viene condotta quando il *numero delle modalità della variabile esplicativa è piuttosto limitato* (di modo che sono in numero ridotto le sottopopolazioni da prendere in considerazione).

Supponiamo ad esempio che si sia interessati a studiare qual è tipo di fondo (tra 4 possibili fondi) preferito da un generico investitore. La distribuzione (univariata) del carattere è rappresentata dal seguente diagramma a torta:

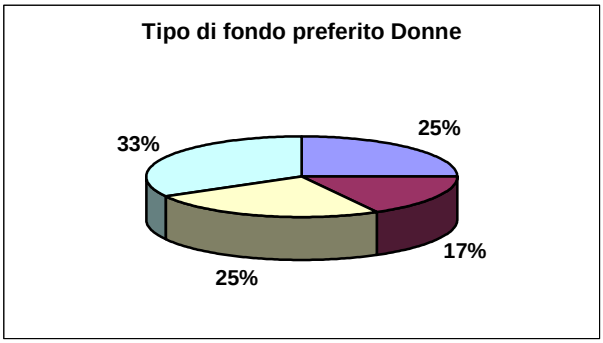


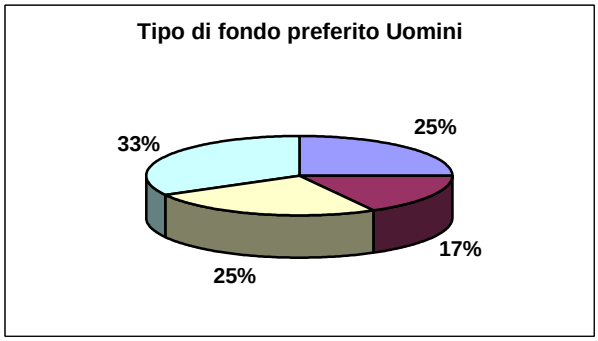
Ora supponiamo di *prendere in considerazione il sesso* e di studiare come si distribuisce il carattere nelle due sotto-popolazioni delle donne e degli uomini. Uno dei possibili risultati potrebbero essere i seguenti.





E’ evidente che in questo caso, l’analisi univariata del carattere sarebbe incompleta, in quanto è opportuno *tener conto* del sesso di un cliente se si vuole comprendere meglio il fenomeno dipendente.
Se invece osservassimo due distribuzioni come quelle delle figure che seguono, identiche tra di loro e identiche alla distribuzione marginale, concluderemmo che non è necessario tener conto del *sesso del cliente* in quanto questo appare *ininfluente*. Diciamo in questo caso che i due caratteri sono ***statisticamente indipendenti***.





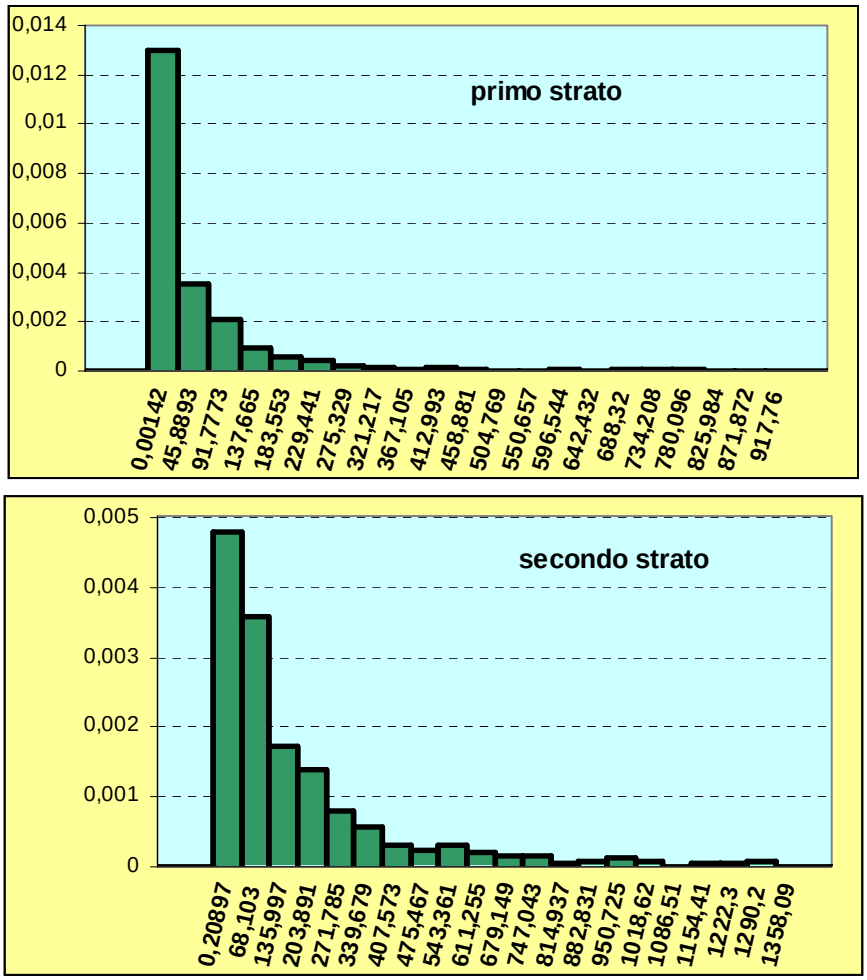
Se due caratteri non sono statisticamente indipendenti diremo che sono **connessi**.
Ovviamente, una volta stabilito che le distribuzioni condizionate differiscono l’una dall’altra, e che i caratteri sono quindi connessi, è importante valutare *quanto* tali distribuzioni si allontanano dalla situazione di indipendenza statistica (o di mancanza di associazione).
A tale scopo si utilizzano le misure di associazione. In particolare, la misura di associazione più comunemente utilizzata per valutare la forza della connessione è il chi-quadro, eventualmente normalizzato. [*Misure di associazione*]

Analisi stratificata: gli istogrammi

Ricordiamo che si parla di analisi stratificata quando si vuole studiare la distribuzione di un carattere (detto nel seguito carattere *dipendente*) non a livello *marginale* (analisi univariata) cioè considerando l'intera popolazione di interesse ma nelle sottopopolazioni indotte dalle modalità di una seconda variabile (detta nel seguito *variabile esplicativa*).

Quando il carattere dipendente è *continuo* o assume un numero di modalità tale da rendere poco agevole un confronto tra distribuzioni di frequenza, il confronto tra le distribuzioni avviene per mezzo di istogrammi.

Per effettuare confronti sensati tra i diversi istogrammi bisogna tenere conto di alcuni aspetti. Ne evidenziamo uno con un esempio.
Rappresentiamo graficamente le distribuzioni di un carattere in due strati utilizzando 20 classi di uguale ampiezza.

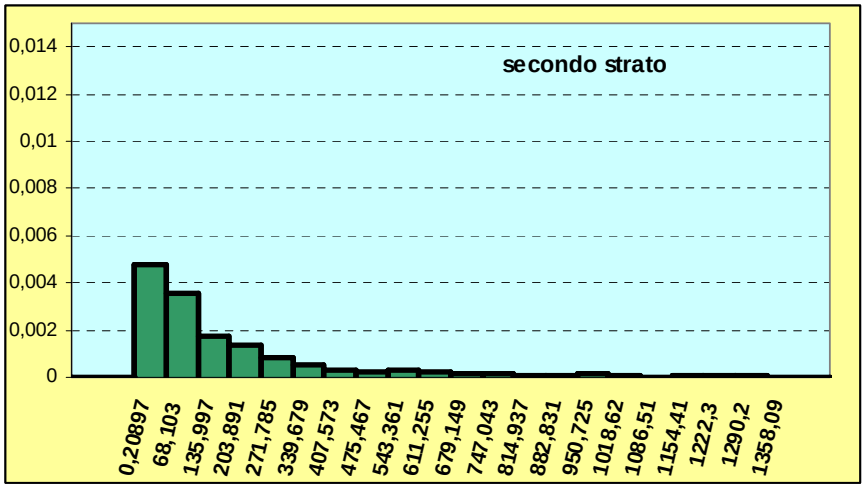


Fare considerazioni a partire da questi due grafici non è sensato in quanto:

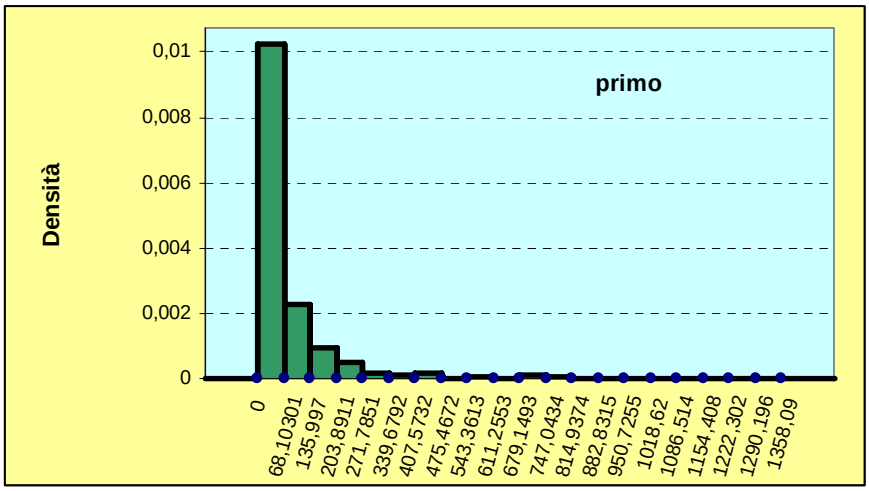
1) l'asse verticale non ha la stessa scala per entrambi i diagrammi
2) l'asse orizzontale non ha la stessa scala per entrambi i caratteri e, inoltre l'ampiezza delle classi che visivamente sembra uguale per entrambi gli istogrammi è diversa.

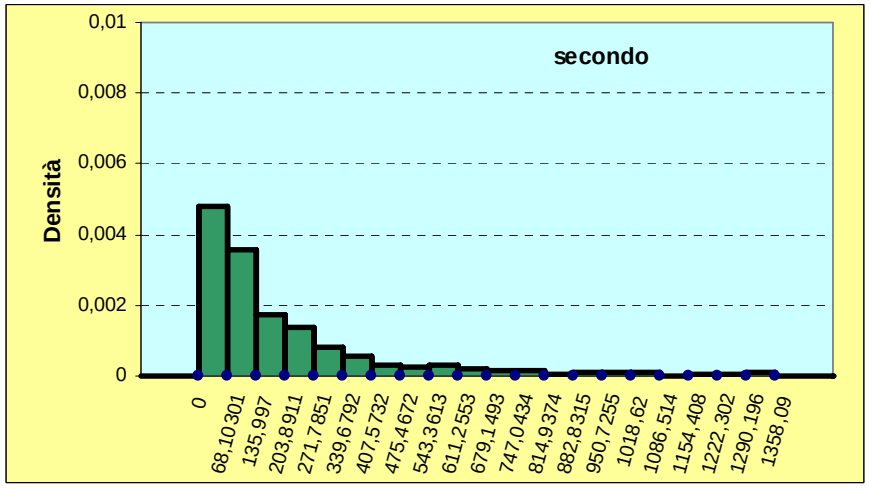
Questo si verifica perché un computer produce sempre l'output nel modo migliore possibile dal punto di vista visivo, che non necessariamente è il migliore anche dal punto di vista dell'affidabilità del risultato.

Già modificando la scala per l'asse verticale del secondo istogramma notiamo una distribuzione molto differente.



La rappresentazione *migliore per il confronto* è la seguente: le classi scelte e la scala dell'asse verticale sono le medesime.

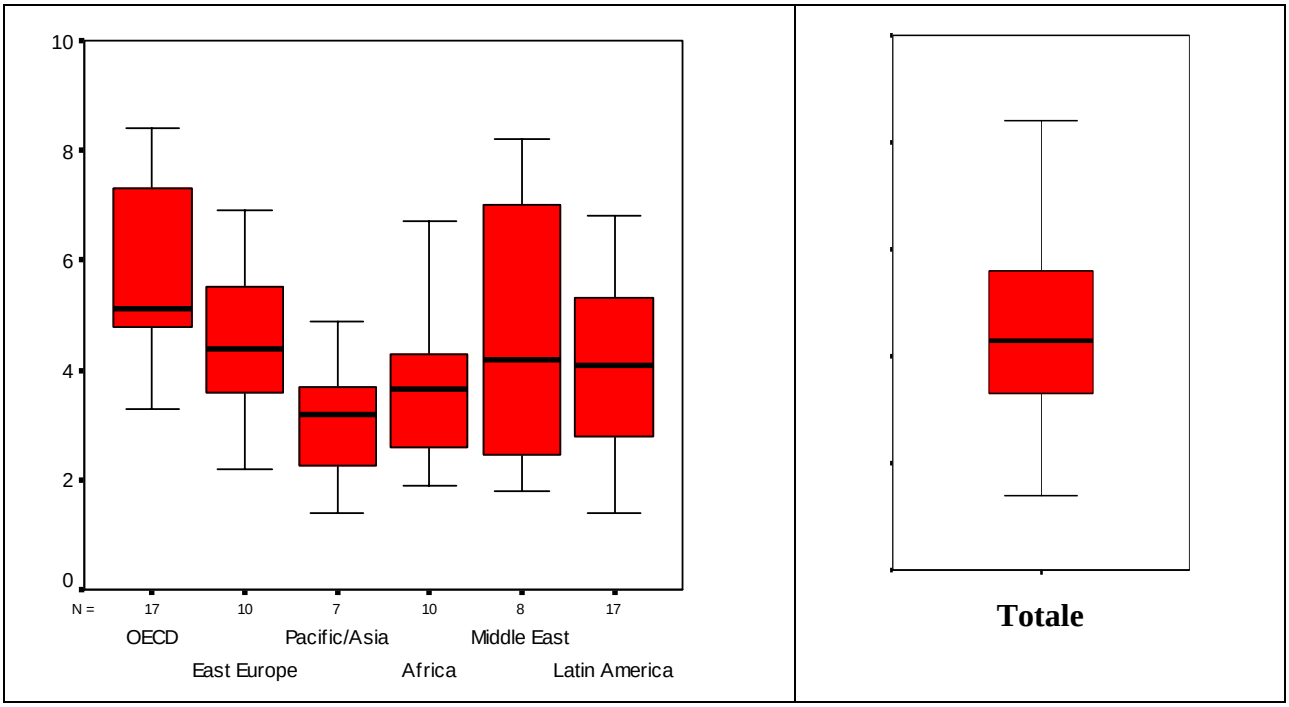




Analisi stratificata: box plot affiancati

Quando si procede all'analisi stratificata di un carattere quantitativo continuo, le distribuzioni nei diversi strati possono essere rappresentate graficamente o per mezzo di istogrammi o per mezzo di box plot. La prima scelta è piuttosto laboriosa, e richiede la scelta preliminare delle classi da utilizzare e l'utilizzo di alcune cautele. [*Analisi stratificata: gli istogrammi*].

Una scelta più semplice è quella di ricorrere ai box plot *affiancati* [*Box plot: una rappresentazione sintetica della distribuzione*]. I box-plot che sintetizzano la distribuzione del carattere in ogni strato vengono riportati nello stesso grafico, in modo che siano più evidenti eventuali differenze nelle distribuzioni e sia più agevole un confronto tra le stesse.
Consideriamo ad esempio un carattere rilevato su un insieme di stati. Potremmo considerare la distribuzione del carattere trattando l'intera popolazione come omogenea (il box-plot totale o *marginale* riportato sulla destra) oppure valutare se la distribuzione del carattere *varia* a seconda della macro area geografica di appartenenza dello stato. A tale scopo consideriamo i box plot affiancati.



Si nota in questo caso l'utilità dell'analisi stratificata: sulla destra è riportato il box-plot relativo all'intera popolazione, che suggerisce una sostanziale simmetria nella distribuzione del carattere. L'analisi dei box plot affiancati evidenzia invece come tale asimmetria "nasconda" sottopopolazioni in cui la forma della distribuzione è molto diversa: la distribuzione del carattere varia a seconda dello strato preso in considerazione. In situazioni come questa è *insensato* procedere ad un'analisi univariata del carattere ed è necessario *tener conto dello strato*.

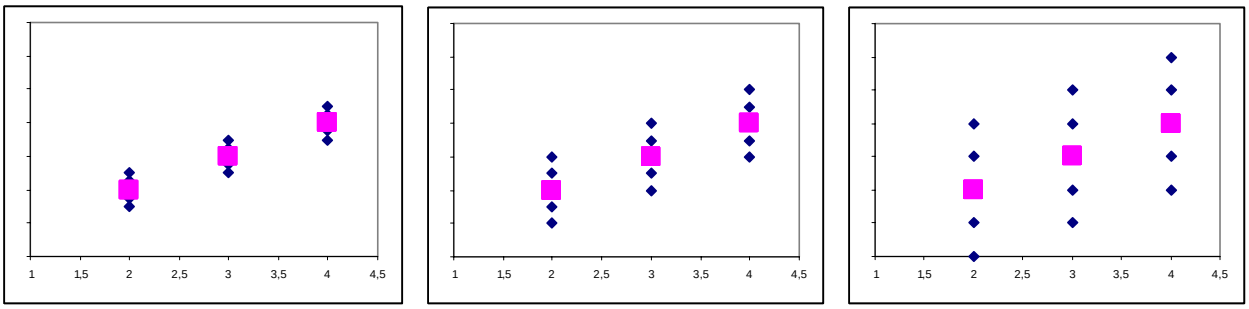
Ovviamente, l'analisi stratificata sarebbe del tutto inutile nel caso in cui *tutti i box plot* avessero, almeno approssimativamente, la stessa forma del box plot relativo alla distribuzione del carattere nell'intera popolazione. In questo caso, è inutile appesantire l'analisi con un'analisi stratificata e l'analisi univariata è sufficiente.

Analisi stratificata: le misure di sintesi

Ricordiamo che si parla di analisi stratificata quando si vuole studiare la distribuzione di un carattere (detto nel seguito carattere *dipendente*) non a livello *marginale* ovvero considerando l'intera popolazione di interesse (analisi univariata) ma nelle sottopopolazioni indotte dalle modalità di una seconda variabile (detta nel seguito *variabile esplicativa*).

Quando il carattere dipendente è quantitativo ed assume molte modalità, le distribuzioni condizionate non possono essere confrontate direttamente, in quanto saranno presumibilmente diverse l'una dall'altra. Ciò può essere dovuto solo al fatto che il carattere assume molte modalità e non invece al fatto che ci sono forti differenze tra le distribuzioni del carattere dipendente condizionate ai valori della variabile esplicativa (che induce gli strati). Una possibilità consiste perciò nel confrontare tra loro gli istogrammi o i box plot delle distribuzioni condizionati.
Per rendere più semplice il confronto tra le distribuzioni stratificate (condizionate) queste vengono sintetizzate con una opportuna ***misura della posizione***, cioè con un valore che "rappresenti" l'insieme delle modalità osservate.
All'analisi delle distribuzioni condizionate viene quindi spesso affiancata un'analisi e un confronto tra misure di sintesi *condizionate*, ad esempio la mediana o la media, a seconda della forma della distribuzione.
Ovviamente, se le distribuzioni condizionate (quindi gli istogrammi e/o i box-plot) sono *molto diverse tra loro* possiamo aspettarci che questa differenza si rifletta anche nelle misure di sintesi, che risulteranno quindi anch'esse molto diverse tra loro.

Una delle misure di sintesi condizionate più comunemente utilizzate è la ***media condizionata***. Tanto più ***diverse*** tra loro sono le medie condizionate, tanto più il carattere *dipendente* è *spiegato -in media-* dal carattere esplicativo. Ovviamente, osservare medie diverse porta a concludere che c'è dipendenza in media, ma non consente di trarre conclusioni sulla ***forza della dipendenza in media***.
Per comprendere il motivo consideriamo i tre diagrammi di dispersione riportati di seguito. Le medie del carattere in ordinata condizionate a quello in ascissa sono individuate da un quadratino.



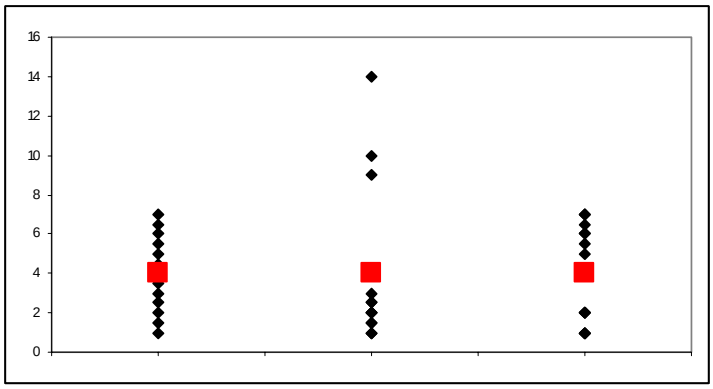
Notiamo che le tre medie condizionate sono le stesse nei tre diagrammi considerati. Tuttavia, le medie condizionate nel primo diagramma risultano molto più rappresentative delle distribuzioni condizionate di quanto non lo siano le medie condizionate degli altri diagrammi di dispersione. Quindi, oltre a valutare quanto le medie sono ***diverse*** tra di loro sarà necessario considerare anche quanto sono ***rappresentative*** delle distribuzioni condizionate.
A questo tipo di obiettivo assolve l'indice **Eta**, che misura la ***forza della dipendenza in media di un carattere dipendente Y da un carattere esplicativo X.*** L'indice Eta varia tra 0 e 1 ed è tanto più vicino ad 1 quanto più le medie condizionate possono essere efficacemente utilizzate per prevedere *Y* in funzione di *X* (utilizzando solo le medie condizionate).

Nel caso in cui le medie condizionate risultino tutte molto simili tra loro e simili alla marginale, diremo che il carattere *dipendente* è ***indipendente in media*** dal carattere esplicativo. Ciò significa

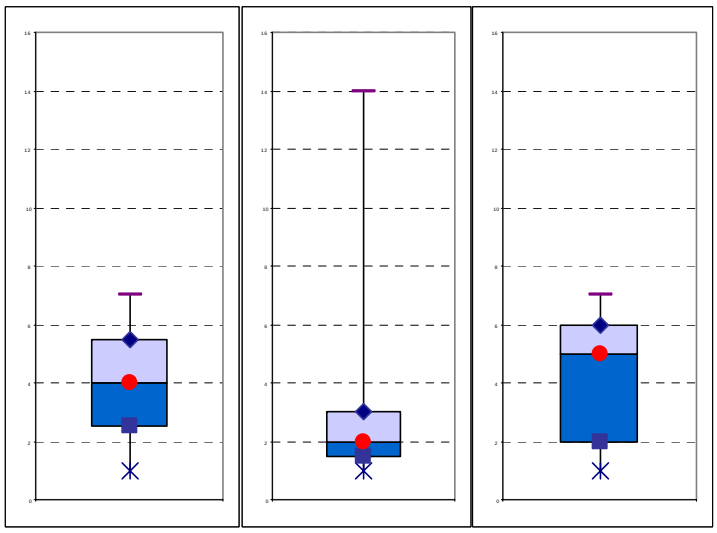
che non è sensato utilizzare le medie condizionate per effettuare previsioni su *Y* o, meglio, che l'analisi delle medie condizionate non arricchisce in alcun modo l'analisi marginale (in quanto le medie condizionate risultano simili alla marginale). In questo caso, l'indice **Eta** risulterà prossimo a zero.

Ovviamente, se il carattere *condizionante* (la variabile esplicativa, quella che definisce gli strati) non influenza in alcun modo il carattere *dipendente*, le distribuzioni condizionate saranno molto simili (al limite identiche) tra di loro. In questo caso i due caratteri si diranno ***statisticamente indipendenti***. Le medie condizionate risulteranno naturalmente molto simili, se non identiche, tra di loro. In termini statistici diremo che ***l'indipendenza statistica*** **implica** ***l'indipendenza in media***.

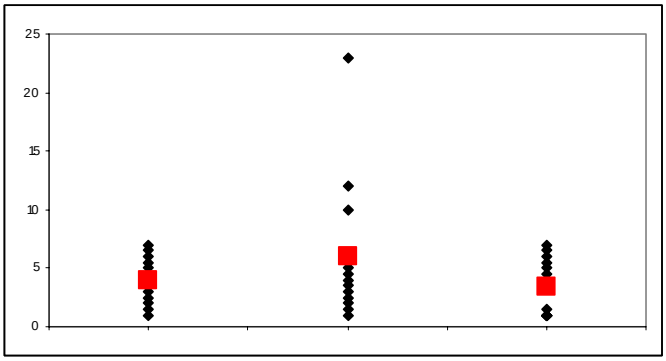
Bisogna però enfatizzare con riferimento a questo punto che *medie condizionate uguali* (o molto simili) tra *loro non sempre sono indicatrici di un legame debole tra i due caratteri*: se le medie condizionate sono ***uguali*** tra loro questo non significa che le distribuzioni condizionate stesse siano tra loro uguali e uguali alla distribuzione marginale.
Può infatti accadere che le misure di sintesi delle distribuzioni condizionate non riflettano la differenza tra queste.



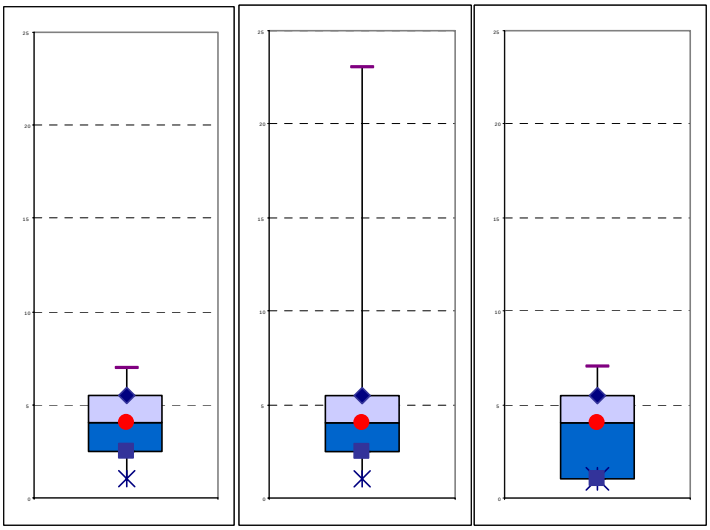
Ad esempio, nel diagramma di dispersione di fianco le distribuzioni condizionate sono diverse tra loro, ma tale differenza non si riflette nelle medie condizionate che risultano uguali tra di loro. Nella figura sottostante sono riportati i tre box plot condizionati evidentemente diversi tra loro.
Quindi, se le medie condizionate dovessero coincidere tra di loro questo non dovrebbe farci dedurre che tra il carattere dipendente e il carattere esplicativo non ci sia una relazione. Ancora, in termini statistici si dice che **l'*indipendenza in media* non implica l'*indipendenza statistica***.



Notiamo che nella figura sopra le differenze tra le distribuzioni condizionate, caratterizzate da medie identiche, si riflettono invece nella mediana. La considerazione appena fatta dipende dal fatto che le medie sono misure *non robuste*, cioè sensibili alla presenza di valori estremi o anomali. Quindi potrebbe verificarsi che medie condizionate risultino identiche o molto simili tra loro a causa della presenza di valori anomali.



Consideriamo ad esempio il diagramma di dispersione e i tre box plot riportati di seguito. Notiamo che in questo caso le medie sono diverse tra di loro, ma in realtà le tre distribuzioni condizionate sono piuttosto simili tra di loro (a parte per valori anomali) e le differenze osservate tra le tre medie sono riconducibili alla presenza di valori anomali.
Le differenze riscontrate tra le medie non si riflettono, come evidente dai box plot sotto riportati, in differenze tra le mediane.



Il messaggio è chiaro: ***quando la distribuzione del carattere è molto asimmetrica non ha molto senso o può essere fuorviante sintetizzare le distribuzioni condizionate con le medie condizionate.*** La sensibilità delle medie ai valori anomali può infatti far risultare simili medie che sintetizzano distribuzioni molto diverse oppure portare a medie diverse che sintetizzano distribuzioni in realtà molti simili.

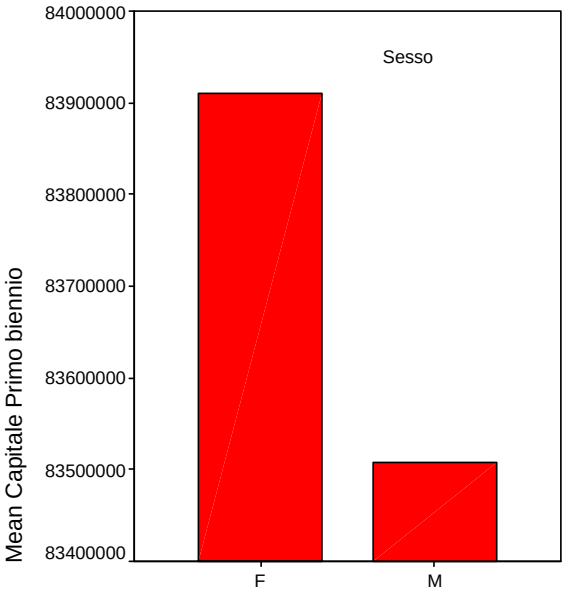
In questi casi è più opportuno considerare le ***mediane condizionate***.

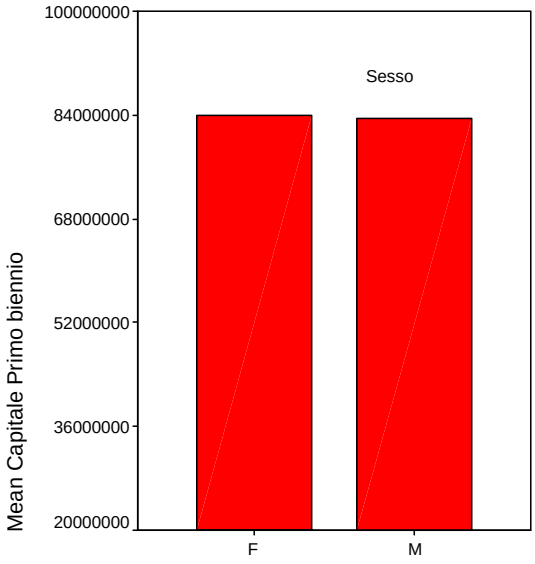
Prima di concludere, dobbiamo evidenziare che differenze eventualmente riscontrate tra le medie o le mediane di distribuzioni condizionate valutate su osservazioni ***campionarie*** non possono far trarre conclusioni sulla forza della dipendenza in media del carattere dipendente da quello esplicativo. Per fare inferenza su tale dipendenza è necessario procedere ad una valutazione inferenziale del risultato osservato.

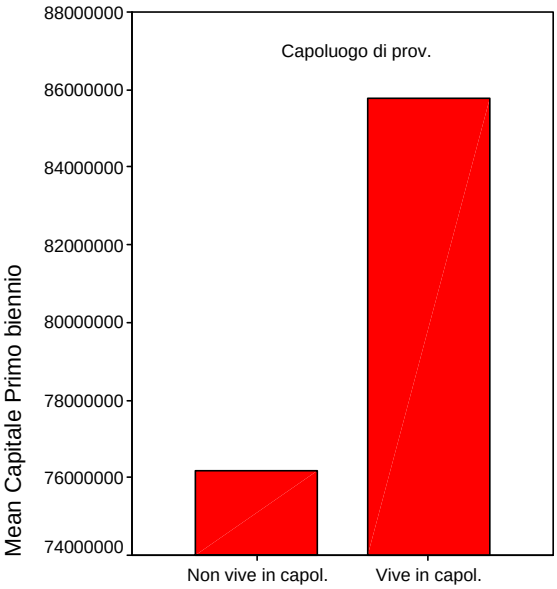
A questo compito assolve l'***analisi della varianza*** il cui scopo è quello di valutare se la differenza osservata a livello descrittivo tra le medie condizionate può essere considerata significativa anche al livello dell'intera popolazione. [*Analisi della varianza*]
Quando la distribuzione del carattere dipendente è fortemente asimmetrica, di modo che non ha senso o affidabilità confrontare tra loro le medie condizionate, la valutazione della dipendenza (in sintesi) di un carattere dipendente da quello esplicativo avviene considerando la cosiddetta ***analisi della varianza non parametrica***, [*Analisi della varianza non parametrica*] basata sul confronto tra mediane o tra i ranghi delle osservazioni (ricordiamo che il rango di un'osservazione è la sua posizione nella lista ordinata delle osservazioni. La mediana è l'osservazione che occupa il rango centrale).

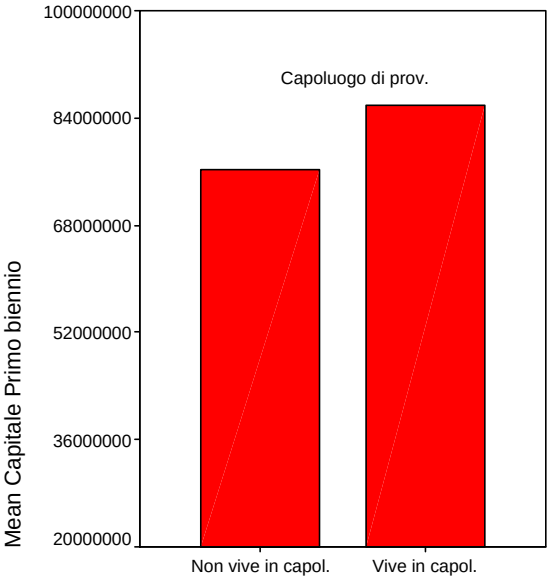
Differenza tra medie. Rappresentazione grafica

Per valutare la dipendenza di un carattere quantitativo da un carattere qualitativo, si deve fare riferimento alle medie condizionate: un carattere quantitativo *Y* dipende (in media) dal carattere qualitativo *X* se le medie di *Y* condizionate alle diverse modalità di *X* risultano diverse tra di loro.
[*Analisi stratificata: le misure di sintesi*]
Un primo strumento che potremmo essere tentati di utilizzare è lo strumento grafico, che ci dà un'idea della presenza/assenza di diversità tra le medie. In realtà, però, come sappiamo, a seconda della scala di misura che utilizziamo le differenze possono essere più o meno marcate. Ad esempio, nei grafici che seguono si può vedere come scelte diverse della scala di rappresentazione possono far apparire le differenze esistenti come più o meno marcate.









Tra l'altro, comunque, lo strumento grafico ci può dare solo un'impressione della differenza tra le medie, ma l'unico strumento che possiamo usare per valutare se le differenze eventualmente osservate tra le medie sono o meno significative, è quello fornito dall'analisi della varianza.

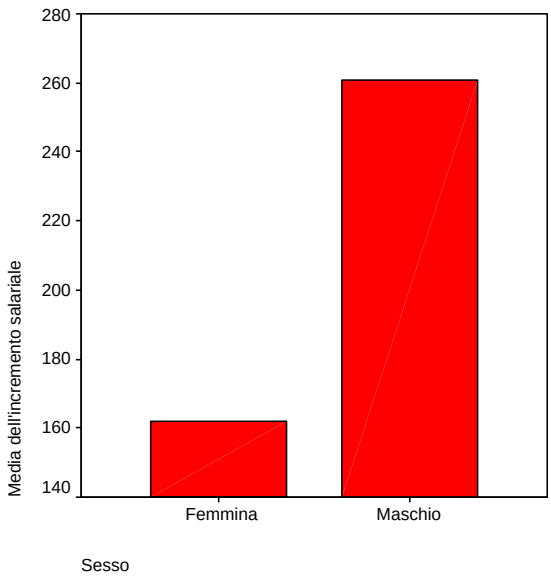
Associazione e causalità

Anche quando due caratteri risultano caratterizzati da una associazione, anche di forte entità, non si può mai "forzare" l'interpretazione di un nesso *causale* tra le variabili stesse.

Per comprenderne il motivo, consideriamo alcuni esempi.

Connessione: Consideriamo un collettivo di fumatori che stanno cercando di smettere di fumare e accettano di partecipare ad una "terapia di gruppo". Dopo un certo periodo di trattamento, si rileva per ogni fumatore se ha smesso o no di fumare; sono inoltre disponibili alcune informazioni sulle caratteristiche socio-demografiche dei soggetti, quali sesso, titolo di studio, composizione del nucleo familiare e così via.
Supponiamo di osservare che il carattere "Smette di fumare" è fortemente associato con il carattere "Titolo di studio": in particolare, l'analisi delle distribuzioni condizionate evidenzia che coloro con elevati titoli di studio risultano meno propensi a smettere di fumare.
Se possiamo dire che risulta evidentemente più complicato smettere di fumare per coloro che hanno un elevato titolo di studio, non possiamo sicuramente affermare che *un elevato titolo di studio indebolisce la forza di volontà nello smettere di fumare*. Per fare considerazioni di questo tipo, dovremo prendere in considerazione altri aspetti del problema trascurati nell'analisi effettuata, che ci portino a comprendere come mai le persone con elevato titolo di studio hanno meno successo nello smettere di fumare. Nell'analisi fatta abbiamo probabilmente trascurato alcuni caratteri rilevanti (uno dei più semplici che ci può venire in mente è, ad esempio, lo stress cui un soggetto è sottoposto durante il lavoro).

Dipendenza in media Supponiamo, come secondo esempio, di rilevare per un gruppo di impiegati presso una certa azienda il sesso e l'incremento salariale medio annuale. Le due medie sono riportate nel diagramma che segue.

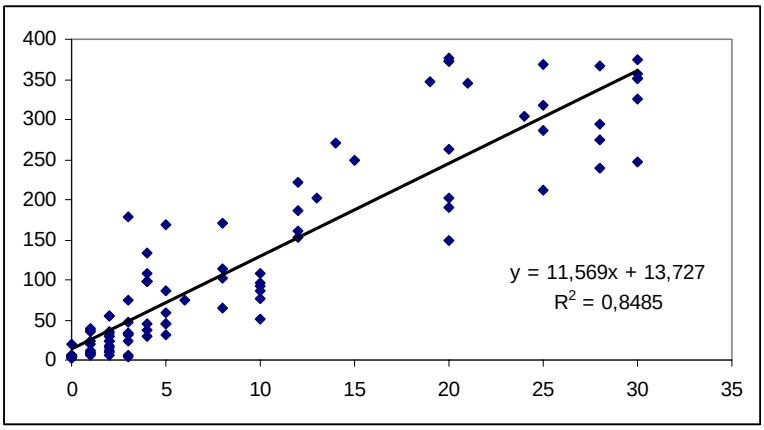


Notiamo che le medie dei due caratteri risultano molto diverse l'una dall'altra. Possiamo quindi concludere che l'incremento salariale dipende in media dal sesso del soggetto considerato. Tuttavia, il termine *dipendenza in media* indica solo che a partire dal sesso è possibile prevedere l'incremento salariale meglio di quanto non si possa fare considerando la media marginale.
Tuttavia, prima di concludere che nell'azienda considerata il fatto di essere una donna *comporta* un incremento salariale inferiore, dovremo valutare attentamente il motivo di questo risultato. Se è innegabile che tra i due caratteri esista un'associazione non è possibile asserire che questa associazione si traduca in un nesso di *causalità*.

Il risultato ottenuto potrebbe essere legato alle mansioni lavorative cui sono preposti maschi e femmine nell'azienda considerata. I maschi potrebbero infatti occupare posizioni legate a salari più elevati (con, quindi, incrementi più sostanziali).

Dipendenza lineare (o correlativa). Come ultimo esempio, consideriamo il diagramma di dispersione riportato di seguito, relativo al Fatturato e al Numero di occupati in un collettivo di aziende. Notiamo che esiste una forte associazione lineare tra i due caratteri: l'indice R^2 è infatti

molto elevato, e la retta di regressione del fatturato sul numero di occupati spiega circa l'84% della varianza del fatturato.



La forza dell'associazione lineare potrebbe quindi essere usata per prevedere il fatturato di un'azienda a partire dall'informazione sul suo numero di occupati (ad esempio, per motivi fiscali, per valutare se il fatturato di un'azienda è plausibile o meno).

Tuttavia, evidentemente, nulla può farci concludere che il fatturato *dipende* dal numero di occupati e che quindi se un'azienda assume nuovi lavoratori vedrà aumentare il proprio fatturato. Diremo che aziende di grandi dimensioni (e quindi con un numero molto elevato di dipendenti) hanno anche elevati fatturati, e l'esistenza di tale relazione può essere utilizzata per fare previsioni sul fatturato.

Analisi della varianza (a una via)

Ricordiamo che si parla di analisi stratificata (o condizionata) quando si vuole studiare la distribuzione di un carattere (detto nel seguito carattere *dipendente*) non a livello *marginale* ovvero considerando l'intera popolazione di interesse (analisi univariata) ma nelle sottopopolazioni indotte dalle modalità di una seconda variabile (detta nel seguito *esplicativa*). [*Analisi stratificata*]

Quando il carattere è quantitativo ed assume molte modalità, un confronto diretto tra le distribuzioni non è sensato. Se è vero che possono essere confrontati tra loro gli istogrammi e/o i box plot delle distribuzioni condizionate [*Analisi stratificata: gli istogrammi*] [*Analisi stratificata: i box plot*] è anche vero che spesso per rendere più semplice il confronto tra le distribuzioni stratificate (condizionate) queste vengono sintetizzate con una opportuna ***misura della posizione***, cioè con un valore che "rappresenti" l'insieme delle modalità osservate.
Una delle misure di sintesi condizionate più comunemente utilizzate è la ***media condizionata***. Tanto più ***diverse*** tra loro sono le medie condizionate, tanto più il carattere *dipendente* è *spiegato - in media -* dal carattere esplicativo.

Quando si ha a che fare con un campione di osservazioni, le differenze eventualmente osservate tra le medie condizionate (***campionarie***) devono essere analizzate da un punto di vista inferenziale. Si è quindi interessati a valutare se esse sono ***significative***, cioè se riflettono reali differenze anche nella popolazione oppure se sono dovute ***al caso*** (cioè sono legate in qualche modo al fatto che stiamo considerando campioni e non popolazioni). L'analisi della varianza (o *ANOVA*) risponde a questa domanda nel caso in cui gli strati siano più di due. Nel caso in cui si considerino solo due sotto-popolazioni, allo stesso scopo risponde il test T per l'uguaglianza tra due medie, di cui l'ANOVA è un'estensione [*Test T per verificare l'uguaglianza tra due medie*].

Le assunzioni alla base dell'ANOVA sono le seguenti:

1) Le osservazioni devono essere tra loro *indipendenti*
2) La variabile dipendente deve avere distribuzione *normale*
3) Le varianze all'interno degli strati devono essere *omogenee* (cioè simili tra loro).

Molto semplicemente, possiamo dire che l'analisi della varianza verifica l'ipotesi ***nulla***

$$\mathbf{H_0}: \textbf{\textit{tutte le medie sono uguali tra di loro}}$$

che con riferimento alle medie condizionate può essere tradotta in

$$\mathbf{H_0}: \mu_1 = \mu_2 = \ldots = \mu_K = \mu$$

L'ipotesi ***alternativa*** è:

$$\mathbf{H_1}: \textbf{\textit{almeno una media è diversa dalle altre}}$$

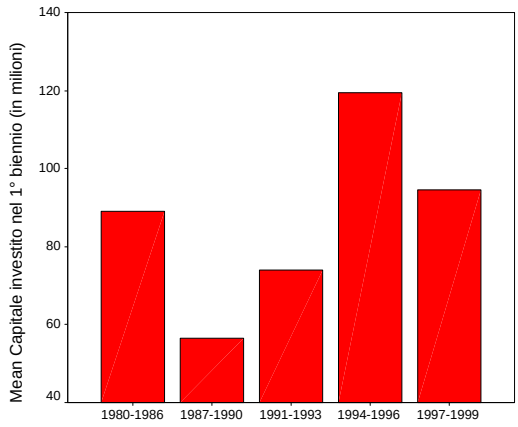
che con riferimento alle medie condizionate può essere tradotta in

$$\mathbf{H_1}: \textbf{esiste almeno uno strato } \boldsymbol{k} \textbf{ per cui } \mu_k \neq \mu$$

Prima di procedere, è utile notare che l'ipotesi nulla viene rifiutata se almeno due strati hanno medie significativamente diverse tra loro. Non è quindi detto che, se l'ipotesi nulla viene rifiutata, si possa concludere che tutte le medie sono diverse tra loro (ovvero, non è detto che l'effetto dell'appartenenza agli strati sia sempre significativo; è sufficiente un solo strato con una media significativamente diversa da quella di un altro perché l'ipotesi nulla sia rifiutata).

Non vogliamo addentrarci nei dettagli "tecnici". Diciamo solo che l'ipotesi nulla viene verificata facendo riferimento ad una statistica test, ***F***, che, sotto l'ipotesi nulla ha distribuzione nota. Ipotizzando che l'ipotesi nulla sia vera (e che quindi le medie siano tutte diverse tra loro) la statistica *F* dovrebbe assumere valori "piccoli". Valori elevati della statistica *F* sono quindi

"anomali" sotto l'ipotesi nulla e "compatibili" con quella alternativa. La verifica di ipotesi si basa, come di consueto, sulla determinazione del p-value che misura la probabilità di estrarre campioni caratterizzati da un valore della statistica *F* più elevati di quello osservato per il campione in esame. Valori molto bassi del p-value (o comunque inferiori al livello di significatività prescelto) indicano quindi che sotto H_0 il risultato campionario osservato è molto anomalo e deve quindi farci propendere per la decisione di rifiutare H_0.



Consideriamo ad esempio un campione di clienti di una società. Per ognuno rileviamo il *Capitale investito nel 1° biennio* e l'*anno in cui il soggetto è diventato cliente della società*.
Le medie condizionate risultano diverse tra loro almeno dal punto di vista descrittivo. Si è interessati a verificare se le differenze riscontrate tra le medie campionarie a livello descrittivo debbano essere considerate significative oppure se possano essere ritenute effetto di oscillazioni casuali (dovute al fatto che stiamo considerando dei campioni).

L'analisi della varianza del capitale investito sull'anno di ingresso fornisce i seguenti risultati.

ANOVA

Capitale investito nel 1° biennio (in milioni)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	450436.238	4	112609.060	4.743	.001
Within Groups	23671055.148	997	23742.282		
Total	24121491.386	1001			

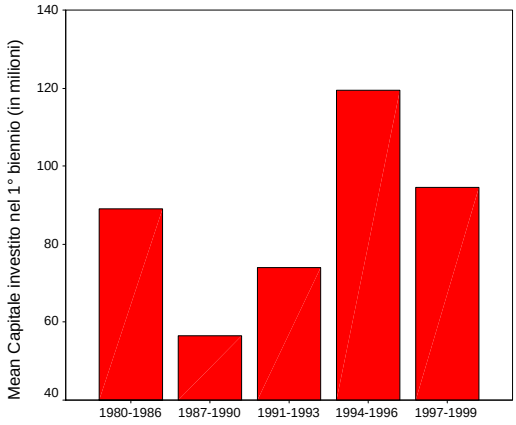
Il p-value è molto basso. Indipendentemente dal valore scelto per la probabilità di commettere un errore di primo tipo, rifiuteremo l'ipotesi nulla, secondo la quale ***tutte le medie sono uguali,*** cioè, con riferimento all'esempio, rifiutiamo l'ipotesi che il capitale medio investito nel 1° biennio di attività non vari in media al variare dell'anno.
Se il valore del p-value ci fa propendere per il rifiuto dell'ipotesi nulla ciò, come già evidenziato, non significa che le medie sono tutte significativamente diverse l'una dall'altra ma, piuttosto, che c'è almeno una coppia di medie che risultano statisticamente diverse tra loro.
E' quindi necessario individuare quali sono le medie diverse tra loro, procedendo quindi a verificare l'uguaglianza tra tutte le possibili coppie di medie attraverso opportuni test, detti ***test post-hoc***.
[*I test post-hoc per l'individuazione delle differenze significative in un'ANOVA*]

Ricordiamo che affinché i risultati ottenuti con il modello ANOVA siano affidabili è necessario che siano soddisfatte le due ipotesi che 1) le varianze all'interno degli strati devono essere *omogenee* (cioè simili tra loro) 2) la variabile dipendente deve avere distribuzione *normale*. Entrambe le ipotesi vanno verificate prima di trarre conclusioni. Nel caso in cui non risultassero soddisfatte è necessario/opportuno ricorrere a tecniche alternative a quelle standard.

Nell'esempio appena visto, nessuna delle due ipotesi è soddisfatta, ed è quindi necessario procedere utilizzando altri metodi.
[*Caduta delle ipotesi dell'ANOVA: varianze non omogenee*]
[*Caduta delle ipotesi dell'ANOVA: non normalità*]

I test post-hoc per l'individuazione delle differenze significative nell'ANOVA

Ricordiamo che nell'Analisi della Varianza (ANOVA) si considerano le medie di una variabile, dipendente, negli strati indotti dalle modalità di una seconda variabile, detta esplicativa (o fattore). L'intento è quello di verificare l'ipotesi nulla che tutte le medie siano uguali tra di loro contro l'ipotesi alternativa che almeno una coppia di medie presenti una differenza statisticamente significativa. [*Analisi della varianza*]



Consideriamo ad esempio un campione di clienti di una società, per ognuno dei quali viene rilevato il *Capitale investito nel 1° biennio* e il periodo *in cui il soggetto è diventato cliente della società.*
Si è interessati a verificare se le differenze riscontrate tra le medie campionarie a livello descrittivo debbano essere considerate significative oppure se possano essere ritenute effetto di oscillazioni casuali (dovute al fatto che stiamo considerando dei campioni).

L'analisi della varianza del capitale investito sull'anno di ingresso fornisce i seguenti risultati.

ANOVA

Capitale investito nel 1° biennio (in milioni)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	450436.238	4	112609.060	4.743	.001
Within Groups	23671055.148	997	23742.282		
Total	24121491.386	1001			

Il p-value è molto basso e ci porta a rifiutare l'ipotesi nulla, secondo la quale ***tutte le medie sono uguali***. Ciò non significa che le medie sono tutte significativamente diverse l'una dall'altra ma, piuttosto, che c'è almeno una coppia di medie la cui differenza è statisticamente significativa.
E' quindi necessario individuare quali sono le medie diverse tra loro, procedendo quindi a verificare l'uguaglianza tra tutte le possibili coppie di medie attraverso opportuni test.
In tali test, detti ***test post-hoc***, per ogni coppia di medie l'ipotesi nulla è che la differenza tra queste sia pari a zero, mentre l'alternativa è che le due medie differiscano significativamente tra loro.
Il test più semplice per effettuare tale confronto consiste nel verificare per ogni coppia di medie μ_j e μ_w l'ipotesi nulla H_0: $\mu_j = \mu_w$ contro l'ipotesi alternativa H_1: $\mu_j \neq \mu_w$ ad un livello di significatività prefissato, α. Tale modo di procedere (procedura ***LSD***) tiene sotto controllo l'errore di primo tipo relativo ad ognuno dei singoli confronti (detto anche Comparisonwise Error). Se si vuole tener conto dell'errore *complessivo* di primo tipo, cioè l'errore di primo tipo relativo a tutti i confronti considerati (detto anche Experimentwise Error) si possono applicare procedure più raffinate. Un primo tipo di procedure consiste nel *correggere* il livello di significatività α tenendo conto del fatto che si stanno effettuando confronti multipli (ad esempio, la correzione di ***Bonferroni*** o di ***Sidak***). Tali test sono più conservativi rispetto a quelli che tengono sotto controllo l'errore relativo ad ogni singolo controllo (ossia possono non evidenziare differenze significative anche se le differenze sono presenti). Esistono altri test meno conservativi e basati su altro tipo di considerazioni per l'analisi di confronti multipli (per esempio, il metodo di ***Tukey***, di ***Duncan***, di ***Student-Newmann-Keuls***).
Con riferimento all'esempio citato, ecco i risultati ottenuti con i test menzionati.

Multiple Comparisons

Dependent Variable: Capitale investito nel 1° biennio (in milioni)

	(I) Anno di ingresso (classi)	(J) Anno di ingresso (classi)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1980-1986	1987-1990	32,5682	14,29757	,023	4,5114	60,6249
		1991-1993	15,2936	15,98339	,339	-16,0714	46,6585
		1994-1996	-30,3643	16,06357	,059	-61,8866	1,1580
		1997-1999	-5,3482	15,29889	,727	-35,3699	24,6735
	1987-1990	1980-1986	-32,5682	14,29757	,023	-60,6249	-4,5114
		1991-1993	-17,2746	15,21834	,257	-47,1383	12,5891
		1994-1996	-62,9325	15,30253	,000	-92,9613	-32,9036
		1997-1999	-37,9163	14,49777	,009	-66,3660	-9,4667
	1991-1993	1980-1986	-15,2936	15,98339	,339	-46,6585	16,0714
		1987-1990	17,2746	15,21834	,257	-12,5891	47,1383
		1994-1996	-45,6579	16,88833	,007	-78,7986	-12,5171
		1997-1999	-20,6417	16,16272	,202	-52,3586	11,0751
	1994-1996	1980-1986	30,3643	16,06357	,059	-1,1580	61,8866
		1987-1990	62,9325	15,30253	,000	32,9036	92,9613
		1991-1993	45,6579	16,88833	,007	12,5171	78,7986
		1997-1999	25,0161	16,24202	,124	-6,8563	56,8886
	1997-1999	1980-1986	5,3482	15,29889	,727	-24,6735	35,3699
		1987-1990	37,9163	14,49777	,009	9,4667	66,3660
		1991-1993	20,6417	16,16272	,202	-11,0751	52,3586
		1994-1996	-25,0161	16,24202	,124	-56,8886	6,8563
Bonferroni	1980-1986	1987-1990	32,5682	14,29757	,229	-7,6551	72,7915
		1991-1993	15,2936	15,98339	1,000	-29,6725	60,2596
		1994-1996	-30,3643	16,06357	,590	-75,5559	14,8273
		1997-1999	-5,3482	15,29889	1,000	-48,3885	37,6921
	1987-1990	1980-1986	-32,5682	14,29757	,229	-72,7915	7,6551
		1991-1993	-17,2746	15,21834	1,000	-60,0883	25,5391
		1994-1996	-62,9325	15,30253	,000	-105,9830	-19,8819
		1997-1999	-37,9163	14,49777	,090	-78,7029	2,8702
	1991-1993	1980-1986	-15,2936	15,98339	1,000	-60,2596	29,6725
		1987-1990	17,2746	15,21834	1,000	-25,5391	60,0883
		1994-1996	-45,6579	16,88833	,070	-93,1698	1,8540
		1997-1999	-20,6417	16,16272	1,000	-66,1123	24,8288
	1994-1996	1980-1986	30,3643	16,06357	,590	-14,8273	75,5559
		1987-1990	62,9325	15,30253	,000	19,8819	105,9830
		1991-1993	45,6579	16,88833	,070	-1,8540	93,1698
		1997-1999	25,0161	16,24202	1,000	-20,6775	70,7097
	1997-1999	1980-1986	5,3482	15,29889	1,000	-37,6921	48,3885
		1987-1990	37,9163	14,49777	,090	-2,8702	78,7029
		1991-1993	20,6417	16,16272	1,000	-24,8288	66,1123
		1994-1996	-25,0161	16,24202	1,000	-70,7097	20,6775
Tukey HSD	1980-1986	1987-1990	32,5682	14,29757	,153	-6,5039	71,6402
		1991-1993	15,2936	15,98339	,874	-28,3855	58,9726
		1994-1996	-30,3643	16,06357	,323	-74,2624	13,5338
		1997-1999	-5,3482	15,29889	,997	-47,1566	36,4603
	1987-1990	1980-1986	-32,5682	14,29757	,153	-71,6402	6,5039
		1991-1993	-17,2746	15,21834	,788	-58,8629	24,3137
		1994-1996	-62,9325	15,30253	,000	-104,7509	-21,1141
		1997-1999	-37,9163	14,49777	,068	-77,5355	1,7028
	1991-1993	1980-1986	-15,2936	15,98339	,874	-58,9726	28,3855
		1987-1990	17,2746	15,21834	,788	-24,3137	58,8629
		1994-1996	-45,6579	16,88833	,054	-91,8099	,4941
		1997-1999	-20,6417	16,16272	,705	-64,8108	23,5274
	1994-1996	1980-1986	30,3643	16,06357	,323	-13,5338	74,2624
		1987-1990	62,9325	15,30253	,000	21,1141	104,7509
		1991-1993	45,6579	16,88833	,054	-,4941	91,8099
		1997-1999	25,0161	16,24202	,536	-19,3697	69,4019
	1997-1999	1980-1986	5,3482	15,29889	,997	-36,4603	47,1566
		1987-1990	37,9163	14,49777	,068	-1,7028	77,5355
		1991-1993	20,6417	16,16272	,705	-23,5274	64,8108
		1994-1996	-25,0161	16,24202	,536	-69,4019	19,3697

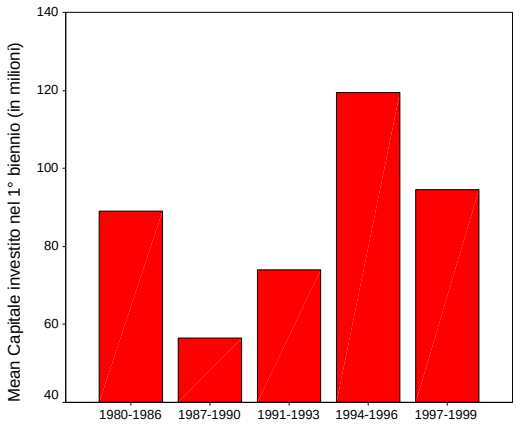
Sono state evidenziate in grigio le coppie di medie significativamente diverse (al livello 0.05)

Vediamo che il test LSD (Comparisonwise Error) individua una differenza significativa (al livello 0.05) tra il capitale medio investito dai clienti entrati nel periodo 1987-1990 (che presentano la media *più bassa*) e quelli entrati nel periodo 1980-1986, 1994-1996 e 1997-1999 e una differenza significativa tra i clienti entrati nel periodo 1991-1993 e quelli entrati nel periodo 1994-1996.
Si osservi come differenze più marcate a livello descrittivo non sempre si traducono in differenze più significative dal punto di vista inferenziale (ad esempio, la differenza tra la media del 1991-1993 e la media del 1994-1996 è significativa; la differenza tra la media del 1991-1993 e la media del 1997-1999, che sembra di entità piuttosto simile a quella precedente, non risulta invece significativa).
Consideriamo ora i risultati ottenuti con (alcuni) test per confronti multipli (Experimentwise Error). Notiamo che questi test sono più conservativi (tendono quindi rifiutare meno facilmente l'ipotesi nulla, secondo cui le coppie di medie sono uguali tra di loro). In particolare, il test di Bonferroni individua una sola differenza significativa (al livello 0.05): quella tra il capitale medio investito dai clienti entrati nel periodo 1987-1990 (che presentano la media *più bassa*) e quelli entrati 1994-1996 (che presentano la media *più bassa*). Allo stesso risultato si perviene utilizzando il test di Tukey (HDS).

Bisogna evidenziare che i test post hoc considerati si basano tutti sull'ipotesi che le varianze all'interno degli strati considerati siano uguali tra di loro (ipotesi di *omogeneità delle varianze*).
Nel caso in cui tale ipotesi non risultasse soddisfatta è necessario/opportuno ricorrere a test alternativi a quelli descritti.
[*Caduta delle ipotesi dell'ANOVA: varianze non omogenee*]

Caduta delle ipotesi alla base del modello ANOVA: varianze non omogenee

Ricordiamo che nell'Analisi della Varianza (ANOVA) si considerano le medie di una variabile, dipendente, negli strati indotti dalle modalità di una seconda variabile, detta esplicativa (o fattore). L'intento è quello di verificare l'ipotesi nulla che tutte le medie siano uguali tra di loro contro l'ipotesi alternativa che almeno una coppia di medie presenti una differenza statisticamente significativa. [*Analisi della varianza*]
Una delle ipotesi alla base del modello ANOVA è che le varianze all'interno degli strati siano uguali tra loro. E' quindi necessario verificare l'ipotesi nulla di omogeneità delle varianze: se tale ipotesi viene rifiutata, la procedura standard è inadeguata.



Consideriamo ad esempio un campione di clienti di una società, per ognuno dei quali viene rilevato il *Capitale investito nel 1° biennio* e il periodo *in cui il soggetto è diventato cliente della società.*
Si è interessati a verificare se le differenze riscontrate tra le medie campionarie a livello descrittivo debbano essere considerate significative oppure se possano essere ritenute effetto di oscillazioni casuali (dovute al fatto che stiamo considerando dei campioni).

Test per la verifica dell'omogeneità delle varianze. Uno dei più famosi test per questa ipotesi nulla è quello di Bartlett. Tale test è basato sull'ipotesi che la distribuzione del carattere dipendente sia normale, ed è poco robusto a deviazioni da tale ipotesi. Per ovviare a tale problema si preferisce quindi di solito ricorrere a test che siano affidabili anche nel caso non normale, come ad esempio il test di **Levene** (utilizzato in SPSS).

Con riferimento all'esempio considerato, il test di Levene fornisce i risultati riportati di fianco.
Il p-value è molto basso, e l'ipotesi nulla (secondo la quale le varianze sono uguali all'interno degli strati) viene quindi rifiutata.

Test of Homogeneity of Variances

Capitale investito nel 1° biennio (in milioni)

Levene Statistic	df1	df2	Sig.
8.200	4	997	.000

ANOVA nel caso di varianze non omogenee. Nel caso in cui cada l'ipotesi di omogeneità delle varianze, è opportuno verificare l'ipotesi nulla dell'anova (le medie negli strati sono tutte uguali) ricorrendo a test robusti all'assunzione di omogeneità delle varianze. Tra questi, quelli più comunemente implementati nei pacchetti statistici sono il test di Welch e il test di Brown-Forsythe.
Con riferimento all'esempio precedente si ha:

Robust Tests of Equality of Means

Capitale investito nel 1° biennio (in milioni)

	Statistic[a]	df1	df2	Sig.
Welch	5.203	4	446.649	.000
Brown-Forsythe	4.391	4	658.041	.002

a. Asymptotically F distributed.

Entrambi i test portano a rifiutare l'ipotesi che le medie siano uguali tra loro. Ciò non implica che le medie siano tutte significativamente diverse l'una dall'altra ma, piuttosto, che c'è almeno una

coppia di medie la cui differenza risulta significativa. E' quindi necessario individuare quali sono le medie diverse tra loro, procedendo quindi a verificare l'uguaglianza tra tutte le possibili coppie di medie attraverso i cosiddetti ***test post-hoc***.
[*I test post-hoc "standard" per l'individuazione delle differenze significative nell'ANOVA*]

Quando le varianze non sono omogenee i test post hoc standard non possono essere utilizzati. Procedure adeguate in questo caso sono ad esempio i test di ***Tamhane*** o di ***Games-Howell*** o il test di ***Duncan***. Nella tabella di seguito sono riportati i risultati ottenuti con i primi due test (il primo test è più conservativo).

Multiple Comparisons

Dependent Variable: Capitale investito nel 1° biennio (in milioni)

	(I) Anno di ingresso (classi)	(J) Anno di ingresso (classi)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tamhane	1980-1986	1987-1990	32.5682	12.42617	.088	-2.4694	67.6057
		1991-1993	15.2936	14.23431	.964	-24.7966	55.3837
		1994-1996	-30.3643	20.45615	.776	-88.0747	27.3461
		1997-1999	-5.3482	16.60928	1.000	-52.1082	41.4118
	1987-1990	1980-1986	-32.5682	12.42617	.088	-67.6057	2.4694
		1991-1993	-17.2746	10.55218	.662	-47.0390	12.4898
		1994-1996	-62.9325*	18.08832	.006	-114.1397	-11.7252
		1997-1999	-37.9163	13.58680	.055	-76.2591	.4264
	1991-1993	1980-1986	-15.2936	14.23431	.964	-55.3837	24.7966
		1987-1990	17.2746	10.55218	.662	-12.4898	47.0390
		1994-1996	-45.6579	19.37507	.176	-100.3905	9.0747
		1997-1999	-20.6417	15.25802	.857	-63.6335	22.3500
	1994-1996	1980-1986	30.3643	20.45615	.776	-27.3461	88.0747
		1987-1990	62.9325*	18.08832	.006	11.7252	114.1397
		1991-1993	45.6579	19.37507	.176	-9.0747	100.3905
		1997-1999	25.0161	21.18126	.934	-34.7105	84.7428
	1997-1999	1980-1986	5.3482	16.60928	1.000	-41.4118	52.1082
		1987-1990	37.9163	13.58680	.055	-.4264	76.2591
		1991-1993	20.6417	15.25802	.857	-22.3500	63.6335
		1994-1996	-25.0161	21.18126	.934	-84.7428	34.7105
Games-Howell	1980-1986	1987-1990	32.5682	12.42617	.069	-1.5275	66.6638
		1991-1993	15.2936	14.23431	.820	-23.7267	54.3139
		1994-1996	-30.3643	20.45615	.573	-86.5180	25.7893
		1997-1999	-5.3482	16.60928	.998	-50.8636	40.1672
	1987-1990	1980-1986	-32.5682	12.42617	.069	-66.6638	1.5275
		1991-1993	-17.2746	10.55218	.475	-46.2370	11.6878
		1994-1996	-62.9325*	18.08832	.006	-112.7276	-13.1373
		1997-1999	-37.9163*	13.58680	.044	-75.2227	-.6100
	1991-1993	1980-1986	-15.2936	14.23431	.820	-54.3139	23.7267
		1987-1990	17.2746	10.55218	.475	-11.6878	46.2370
		1994-1996	-45.6579	19.37507	.131	-98.9015	7.5858
		1997-1999	-20.6417	15.25802	.658	-62.4830	21.1995
	1994-1996	1980-1986	30.3643	20.45615	.573	-25.7893	86.5180
		1987-1990	62.9325*	18.08832	.006	13.1373	112.7276
		1991-1993	45.6579	19.37507	.131	-7.5858	98.9015
		1997-1999	25.0161	21.18126	.762	-33.1044	83.1367
	1997-1999	1980-1986	5.3482	16.60928	.998	-40.1672	50.8636
		1987-1990	37.9163*	13.58680	.044	.6100	75.2227
		1991-1993	20.6417	15.25802	.658	-21.1995	62.4830
		1994-1996	-25.0161	21.18126	.762	-83.1367	33.1044

*. The mean difference is significant at the .05 level.

Entrambi i test portano a concludere che il capitale medio investito dai clienti entrati nel periodo 1987-1990 (che risulta il minimo) è significativamente diverso dal capitale medio massimo, che è quello investito dai clienti entrati nel 1994-96. Il test meno conservativo (il secondo) suggerisce una differenza significativa tra la media nel 1987-90 e quella dell'ultimo periodo, 1997-99 (il secondo valore della media in ordine di grandezza – quindi la media più bassa è significativamente diversa dalle due più elevate).

Caduta delle ipotesi alla base del modello ANOVA: non normalità

Ricordiamo che nell'Analisi della Varianza (ANOVA) si considerano le medie di una variabile, dipendente, negli strati indotti dalle modalità di una seconda variabile, detta esplicativa (o fattore). L'intento è quello di verificare l'ipotesi nulla che tutte le medie siano uguali tra di loro contro l'ipotesi alternativa che almeno una coppia di medie presenti una differenza statisticamente significativa. [*Analisi della varianza*]

Una delle ipotesi alla base del modello ANOVA è che la variabile dipendente abbia una distribuzione normale. La procedura ANOVA "standard" è robusta a deviazioni dall'ipotesi di normalità. Tuttavia, se la distribuzione del carattere dipendente è fortemente asimmetrica i risultati dell'ANOVA possono non essere affidabili.
Ciò è dovuto al fatto che le medie sono misure di sintesi molto sensibili alla presenza di valori estremi. Quindi, se si vuole verificare se la variabile esplicativa (o fattore) ha un effetto sulla variabile dipendente, può essere più opportuno fare riferimento a sintesi più robuste, che garantiscono quindi confronti più significativi. [*Analisi stratificata: le misure di sintesi*].

Prima di trarre conclusioni sui risultati ottenuti utilizzando l'approccio "standard" è quindi necessario verificare l'ipotesi che la variabile dipendente abbia una distribuzione normale: se questa viene rifiutata la procedura standard è inadeguata ed è necessario ricorrere a procedure diverse.

Test per la verifica dell'ipotesi di normalità. Uno dei più famosi test per questa ipotesi nulla è quello di test di Kolmogorov-Smirnov. Riportiamo di fianco un esempio di tale test. Notiamo che il p-value è molto basso, e l'ipotesi nulla (secondo la quale la distribuzione è normale) viene quindi rifiutata.

One-Sample Kolmogorov-Smirnov Test

		Capitale investito nel 1° biennio (in milioni)
N		1002
Normal Parameters [b]	Mean	84,0965
	Std. Deviation	155,23335
Most Extreme Differences	Absolute	,294
	Positive	,254
	Negative	-,294
Kolmogorov-Smirnov Z		9,306
Asymp. Sig. (2-tailed)		,000

b. Calculated from data.

ANOVA non parametrica. Nel caso in cui cada l'ipotesi di normalità della distribuzione e nel caso in cui la distribuzione del carattere dipendente sia molto asimmetrica, l'ANOVA standard va sostituita con l'ANOVA non parametrica, i cui risultati non dipendono dall'assunzione di normalità. I due test non parametrici più importanti sono i test di Kruskal Wallis e il test delle Mediane.
Non entriamo nei dettagli ma descriviamo solo sinteticamente l'idea su cui si basano tali test, sottolineando come primo aspetto di rilievo il fatto che tali test si basano sui ranghi e sulle mediane e quindi sulla posizione delle osservazioni nella sequenze ordinata dei dati (e non sui valori) nel primo caso e su una misura di sintesi robusta – la mediana appunto – nel secondo.
Per quanto riguarda il test di Kruskal-Wallis, l'idea è quella di considerare i ranghi associati ad ogni osservazione. Consideriamo per semplicità 10 osservazioni su *Y*, variabile dipendente e *X* variabile esplicativa (categorica).

Y	2	5	7	8	10	6	9	11	13	15
X	1	1	1	1	1	2	2	2	2	2

Associamo ad ogni osservazione su *Y* il suo rango, cioè la posizione che occupa nella sequenza ordinata dei dati.

Y (ordinata)	2	5	6	7	8	9	10	11	13	15
X	1	1	2	1	1	2	1	2	2	2
Rango	1	2	3	4	5	6	7	8	9	10

Determiniamo ora la somma dei ranghi relativi alle osservazioni nel primo strato ($X = 1$) e la somma dei ranghi relativi alle osservazioni del secondo strato ($X = 2$).

Somma dei ranghi ($X = 1$) = (1 + 2 + 4 + 5 + 7) = **19**
Somma dei ranghi ($X = 2$) = (3 + 6 + 8 + 9 + 10) = **36**.

L'idea è che se il fattore non fosse significativo, le distribuzioni condizionate dovrebbero risultate "confuse" e quindi le somme dei ranghi (o delle opportune medie per tener conto che il numero di osservazioni può variare da strato a strato) dovrebbero risultare vicine tra loro. Se invece le distribuzioni fossero molto diverse (ad esempio, se per $X = 1$ i valori assunti da Y sono molto più bassi dei valori assunti nel caso in cui $X = 2$), tali differenze dovrebbero riflettersi nei ranghi.
A partire da tali quantità viene costruita una statistica test per verificare l'ipotesi nulla che ***le popolazioni condizionate siano identiche tra loro***. Se gli strati sono ***due*** il test costruito a partire dai ranghi si chiama test di **Wilcoxon-Mann-Whitney**; se gli strati sono ***più di due*** il test si chiama test di **Kruskal-Wallis**.
Per quanto riguarda il **test delle mediane**, questo è basato sull'idea che la differenza tra le distribuzioni condizionate (o negli strati) dovrebbe riflettersi in una differenza tra le mediane.
Considerando ancora i dati utilizzati precedentemente, si osserva che la mediana è 8.5 (la media dei due valori che occupano le due posizioni centrali – la 5 e la 6). Per ognuno dei due gruppi generati dai diversi valori di X si conta quanti sono i valori che superano la mediana.

Y (ordinata)	2	5	6	7	8	9	10	11	13	15
X	1	1	2	1	1	2	1	2	2	2

Quindi

Numero di osservazioni superiori alla mediana ($X = 1$) = **1**
Numero di osservazioni superiori alla mediana ($X = 2$) = **4**.

Anche in questo caso, tali quantità vengono utilizzate per verificare l'ipotesi nulla che le due popolazioni condizionate siano identiche tra loro.

Qualora il p-value che caratterizza le statistiche test alla base dei procedimenti inferenziali appena descritti risulti tanto basso da portare al rifiuto dell'ipotesi nulla, secondo la quale ***tutte le popolazioni condizionate sono uguali,*** ciò non significa tutte le popolazioni sono significativamente diverse l'una dall'altra ma, piuttosto, che c'è almeno una coppia di popolazioni che risultano statisticamente diverse tra loro. E' quindi necessario individuare, attraverso opportuni test, quali sono le coppie di popolazioni diverse tra loro.
In tali test, detti ***test post-hoc***, per ogni coppia di popolazioni l'ipotesi nulla è che queste siano uguali tra loro, mentre l'alternativa è che differiscano significativamente tra loro.
Ricordando i test post-hoc nel caso standard, [*Test post-hoc nell'ANOVA standard*] tali confronti possono essere effettuati singolarmente per ogni coppia di popolazioni (tale modo di procedere tiene sotto controllo l'errore di primo tipo relativo ad ognuno dei singoli confronti - detto anche Comparisonwise Error). Se si vuole tener conto dell'errore *complessivo* di primo tipo, cioè l'errore di primo tipo relativo a tutti i confronti considerati - Experimentwise Error – nel caso standard vengono applicate procedure più raffinate. Non sono disponibili test non parametrici del secondo tipo; dovremo quindi procedere a confrontare tutte le differenze separatamente.
[*Test non parametrici per verificare l'uguaglianza tra due popolazioni*]

Analisi della varianza a più vie

Ricordiamo che nell'Analisi della Varianza (ANOVA) a una via si considerano le medie di una variabile, dipendente, negli strati indotti dalle modalità di una seconda variabile, detta esplicativa (o fattore). L'intento è quello di verificare l'ipotesi nulla che tutte le medie siano uguali tra di loro contro l'ipotesi alternativa che almeno una coppia di medie presenti una differenza statisticamente significativa. [*Analisi della varianza a una via*]

Ciò significa che stiamo valutando se la variabile esplicativa (o fattore) considerata ha un *effetto significativo* sulla media, nel senso che le medie nelle sotto-popolazioni indotte dalle modalità della variabile esplicativa non sono tutte uguali tra di loro. Nell'ANOVA a due (o più vie) si prende in considerazione l'effetto congiunto di due (o più) fattori sulla media della variabile dipendente. In particolare, se consideriamo due fattori, le coppie di modalità osservate inducono un numero di sotto-popolazioni più elevato (invece di considerare le sotto-popolazioni indotte dalle modalità di un carattere consideriamo quelle indotte dalle coppie di modalità di due caratteri). L'obiettivo in questo caso è quello di comprendere se e come variano le medie condizionate al variare della sotto-popolazione. [*Analisi stratificata: le misure di sintesi*]

Si noti che potremmo a questo punto domandarci perché sia necessario procedere con un'ANOVA a due vie e non semplicemente procedere con due ANOVA a una via, ciascuna volta a verificare la significatività dei due effetti separatamente.
Per comprenderlo consideriamo un esempio. Viene effettuata un'indagine campionaria sui dipendenti di una multinazionale. Si è interessati a studiare il carattere "l'incremento salariale (mensile)". In particolare si vuole valutare se la media del carattere risulta o meno differente nelle due sotto-popolazioni dei soggetti che appartengono ad una minoranza e dei soggetti che non appartengono ad una minoranza. Di seguito riportiamo i risultati del Test T per verificare l'uguaglianza tra due medie. [*Test T per verificare l'uguaglianza tra due medie*]

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Aumento salariale mensile	Equal variances assumed	16,727	,000	3,562	472	,000	51,9748	14,59185	23,30173	80,64778
	Equal variances not assumed			4,122	211,21	,000	51,9748	12,60769	27,12173	76,82778

Il test porta a concludere che c'è una differenza (molto) significativa tra le medie delle due sotto-popolazioni. Dovremmo concludere che la politica salariale dell'azienda è discriminatoria? Se pensiamo al problema in esame, ci rendiamo conto che possono esserci alcuni fattori legati alla variabile esplicativa (tipo di mansione lavorativa, livello di istruzione, anni di studio) che potrebbero spiegare il risultato osservato. Quando ci sono altri caratteri, oltre a quello considerato, che sono legati a quest'ultimo e che possono influenzare il carattere dipendente, dovremmo preliminarmente eliminare il loro effetto sulla variabile dipendente prima di fare confronti tra le medie del carattere esplicativo considerato.
La domanda corretta in questo caso non è quindi "Ci sono differenze tra l'incremento salariale delle minoranze e quello delle non minoranze?" - che si può tradurre in "Considerati due soggetti, uno appartenente ad una minoranza e l'altro no, il loro incremento salariale è significativamente diverso?" - bensì "Considerati due soggetti con identiche caratteristiche per quanto riguarda livello

di istruzione e mansione lavorativa, l'incremento salariale cambia significativamente a seconda che uno dei due appartenga ad una minoranza?".
L'ANOVA a due (o più vie) fornisce quindi indicazioni sull'impatto di un certo effetto sulla media dl carattere dipendente *tenendo sotto controllo altri fattori che possono essere significativi*. Consideriamo ad esempio il risultato dell'ANOVA a due vie dell'incremento salariale su "Appartenenza ad una minoranza" e "Categoria lavorativa".

Tests of Between-Subjects Effects

Dependent Variable: INCREASE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4214233,145[a]	3	1404744,382	158,567	,000
Intercept	10280629,306	1	10280629,31	1160,470	,000
Minoranza	8803,080	1	8803,080	,994	,319
Categoria lavorativa	3994931,655	2	1997465,827	225,473	,000
Error	4163740,100	470	8859,021		
Total	30436945,302	474			
Corrected Total	8377973,245	473			

a. R Squared = ,503 (Adjusted R Squared = ,500)

Notiamo che l'appartenenza ad una minoranza, che prima risultava un fattore significativo non lo è più: se teniamo conto (o teniamo *sotto controllo*) la categoria lavorativa, l'incremento salariale mensile medio non differisce significativamente nelle due sotto-popolazioni di soggetti che appartengono ad una minoranza e di quelli che non vi appartengono. Ciò è dovuto al fatto che i soggetti che appartengono ad una minoranza hanno probabilmente mansioni lavorative che vengono meno premiate dalla politica di incremento salariale. [*Associazione e causalità*]

Ovviamente l'ANOVA a due vie ha spesso anche lo scopo di studiare la dipendenza di un carattere da due caratteri contemporaneamente (non necessariamente uno dei due è un carattere "di disturbo"). Ad esempio, di seguito riportiamo i risultati dell'anova dell'incremento salariale su Categoria lavorativa e Livello di istruzione. Notiamo che entrambi i fattori sono significativi (quindi l'incremento medio mensile varia sia al variare del livello di istruzione che al variare della categoria lavorativa). Notiamo che è stato inserito l'effetto ***interazione*** che serve a valutare se anche l'incrocio tra le modalità dei due caratteri considerati ha un effetto sulla media del carattere dipendente. Anche tale effetto è significativo.
Ovviamente, dire che un fattore (effetto principale o interazione) è significativo, non può portare a concludere che tutte le medie nelle sotto-popolazioni considerate sono diverse tra di loro ma che esiste almeno una coppia di medie diverse. [*I test post-hoc*]

Tests of Between-Subjects Effects

Dependent Variable: INCREASE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4646021,453[a]	9	516224,606	64,183	,000
Intercept	2209171,263	1	2209171,263	274,670	,000
Categoria lavorativa	186362,309	2	93181,155	11,585	,000
Livello di istruzione	209287,940	3	69762,647	8,674	,000
Interazione	82329,532	4	20582,383	2,559	,038
Error	3731951,792	464	8043,000		
Total	30436945,302	474			
Corrected Total	8377973,245	473			

a. R Squared = ,555 (Adjusted R Squared = ,546)

Test T per verificare l'uguaglianza tra due medie

Ricordiamo che si parla di analisi stratificata (o condizionata) quando si vuole studiare la distribuzione di un carattere (detto nel seguito carattere *dipendente*) non a livello *marginale* ovvero considerando l'intera popolazione di interesse (analisi univariata) ma nelle sottopopolazioni indotte dalle modalità di una seconda variabile (detta nel seguito *variabile esplicativa*).
[*Analisi stratificata*]

Quando il carattere è quantitativo ed assume molte modalità, un confronto diretto tra le distribuzioni non è sensato. Se è vero che possono essere confrontati tra loro gli istogrammi e/o i box plot delle distribuzioni condizionate [*Analisi stratificata: gli istogrammi*] [*Analisi stratificata: i box plot*] è anche vero che spesso per rendere più semplice il confronto tra le distribuzioni stratificate (condizionate) queste vengono sintetizzate con una opportuna ***misura della posizione***, cioè con un valore che "rappresenti" l'insieme delle modalità osservate.
Una delle misure di sintesi condizionate più comunemente utilizzate è la ***media condizionata***. Tanto più ***diverse*** tra loro sono le medie condizionate, tanto più il carattere *dipendente* è *spiegato - in media -* dal carattere esplicativo.

Quando si ha a che fare con un campione di osservazioni, le differenze eventualmente osservate tra le medie condizionate (***campionarie***) devono essere analizzate da un punto di vista inferenziale. Si è quindi interessati a valutare se esse sono ***significative***, cioè se riflettono reali differenze anche nella popolazione oppure se sono dovute ***al caso*** (cioè sono legate in qualche modo al fatto che stiamo considerando campioni e non popolazioni).
Il test T risponde a questa domanda nel caso in cui le medie poste a confronto siano **due** (qualora si fosse interessati a confrontare più di due medie si dovrebbe ricorrere all'analisi della varianza [*Analisi della varianza*]).
Le assunzioni alla base del test T sono le seguenti:

1) Le osservazioni devono essere tra loro *indipendenti*
2) La variabile dipendente deve avere distribuzione *normale*
3) Le varianze all'interno degli strati devono essere *uguali*.

Molto semplicemente, possiamo dire che date le due distribuzioni (condizionate) in esame, caratterizzate da medie μ_1 e μ_2 il test T verifica l'ipotesi ***nulla*:**

$$\mathbf{H_0: \mu_1 = \mu_2}$$

contro l'ipotesi ***alternativa***

$$\mathbf{H_1: \mu_1 \neq \mu_2}$$

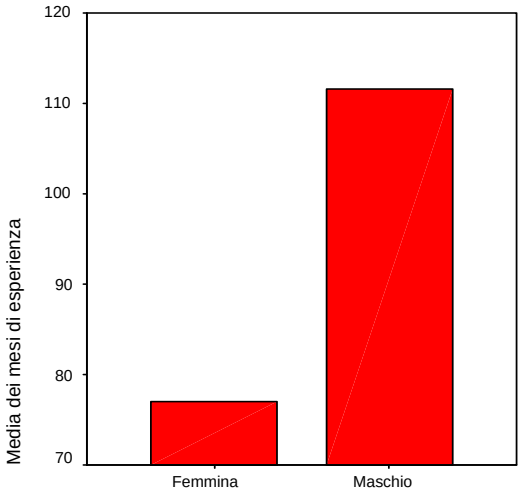
Non vogliamo addentrarci nei dettagli "tecnici". Diciamo solo che l'ipotesi nulla viene verificata facendo riferimento ad una statistica test, ***t***, costruita a partire dalle medie condizionate e di un'opportuna stima della varianza della variabile dipendente (supposta uguale nei due strati). Tale statistica sotto l'ipotesi nulla ha distribuzione nota, in particolare una distribuzione *T* di Student. Ipotizzando che l'ipotesi nulla sia vera (e che quindi le medie siano diverse tra loro) la statistica *t* dovrebbe assumere valori "piccoli". Valori elevati della statistica *t* sono quindi "anomali" sotto l'ipotesi nulla e "compatibili" con quella alternativa. La verifica di ipotesi si basa, come di consueto, sulla determinazione del p-value che misura la probabilità di estrarre campioni caratterizzati da un valore della statistica *t* più elevati di quello osservato per il campione in esame. Valori molto bassi del p-value (o comunque inferiori al livello di significatività prescelto) indicano quindi che sotto H_0 il risultato campionario osservato è molto anomalo e deve quindi farci propendere per la decisione di rifiutare H_0.

Prima di procedere illustrando un esempio, ricordiamo che una delle due ipotesi alla base del test T è che le varianze delle due sotto-popolazioni siano uguali tra loro. E' quindi necessario verificare l'ipotesi nulla di omogeneità delle varianze: se tale ipotesi viene rifiutata, la procedura standard è inadeguata.

Test per la verifica dell'omogeneità delle varianze. Uno dei più famosi test per questa ipotesi nulla è quello di Bartlett. Tale test è basato sull'ipotesi che la distribuzione del carattere dipendente sia normale, ed è poco robusto a deviazioni da tale ipotesi. Per ovviare a tale problema si preferisce quindi di solito ricorrere a test che siano affidabili anche nel caso non normale, come ad esempio il test di **Levene** (utilizzato in SPSS).

Consideriamo alcuni esempi.

Viene effettuata un'indagine campionaria sui dipendenti di una multinazionale. Si è interessati a studiare il carattere "Mesi di esperienza lavorativa precedenti all'assunzione". In particolare si vuole valutare se la media del carattere risulta o meno differente nelle due sotto-popolazioni dei maschi e delle femmine.
Le due medie campionarie risultano piuttosto differenti. Vogliamo verificare se la differenza a livello campionario è significativa, di modo che le conseguenti considerazioni possano essere estese all'intera popolazione.
Di seguito sono riportati i risultati del test T.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Mesi di esperienza	Equal variances assumed	2,582	,109	3,631	472	,000	34,58	9,524	15,869	53,297
	Equal variances not assumed			3,678	471,4	,000	34,58	9,404	16,105	53,062

Concentriamo innanzitutto la nostra attenzione sul test relativo all'omogeneità delle varianze. Il p-value è piuttosto elevato (in generale superiore ai valori standard di significatività). Si può quindi decidere di accettare l'ipotesi nulla e di concludere che le due varianze sono uguali tra di loro. Possiamo quindi considerare i risultati del test T "classico" (riga corrispondente a "Equal variances assumed"). Il p-value che caratterizza il valore della statistica *t* è molto basso, e quindi decidiamo di rifiutare l'ipotesi nulla e di concludere che le medie sono diverse. L'esperienza lavorativa (in mesi) precedente all'assunzione media delle impiegate femmine risulta inferiore a quella dei maschi. Notiamo che dire che il sesso influenza in media l'esperienza lavorativa precedente all'assunzione non implica assolutamente che il *sesso* influenzi l'esperienza lavorativa. L'individuazione di una dipendenza statistica non può e non deve necessariamente tradursi in un nesso di causa-effetto. [*Associazione e causalità*]

Consideriamo ora un secondo esempio. Supponiamo ora di essere interessati a valutare se l'incremento salariale (mensile) medio differisce o meno nelle due sotto-popolazioni dei maschi e delle femmine.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Aumento salariale	Equal variances assumed	69,411	,000	8,640	472	,000	98,6485	11,41737	76,21331	121,08363
	Equal variances not assumed			9,143	379,7	,000	98,6485	10,78998	77,43288	119,86406

In questo caso l'ipotesi che le due varianze siano uguali è rifiutata dal test di Levene. Dobbiamo quindi considerare i risultati del test T modificato a tenere conto della differenza delle varianze (riga corrispondente a "Equal variances not assumed"). Il p-value che caratterizza il valore della statistica *t* è molto basso, e quindi decidiamo di rifiutare l'ipotesi nulla e di concludere che le medie sono diverse. L'aumento salariale medio delle impiegate femmine risulta inferiore a quello dei maschi. Di nuovo questo non può portarci a concludere che l'azienda discrimini in base al sesso, in quanto possono esserci dei fattori che non sono stati presi in considerazione che motivano il risultato.
In particolare, tale analisi non prende in considerazione la mansione lavorativa, che risulta fortemente associata al sesso (la maggior parte dei manager dell'azienda sono maschi e la maggior parte degli impiegati sono femmine). [*Associazione e causalità*]

Ricordiamo che affinché i risultati ottenuti con il test T siano affidabili è necessario che la variabile dipendente abbia distribuzione *normale*. Nel caso in cui tale condizione non sia soddisfatta è necessario ricorrere a test non parametrici.
[*Test non parametrici per verificare l'uguaglianza tra due popolazioni*].

Test non parametrici per verificare l'uguaglianza tra due popolazioni

Ricordiamo che una delle ipotesi alla base del test T per verificare l'uguaglianza tra due medie condizionate (cioè le medie di un certo carattere dipendente in due gruppi indotti dalle modalità di un carattere detto esplicativo – o fattore) è che la variabile dipendente abbia una distribuzione normale. Ciò è dovuto al fatto che le medie sono misure di sintesi molto sensibili alla presenza di valori estremi. Quindi se si vuole verificare se la variabile esplicativa (o fattore) ha un effetto sulla variabile dipendente può essere più opportuno fare riferimento a sintesi più robuste, che garantiscono quindi confronti più significativi. [*Analisi stratificata: le misure di sintesi*].

Prima di trarre conclusioni sui risultati ottenuti utilizzando l'approccio "standard" è quindi necessario verificare l'ipotesi che la variabile dipendente abbia una distribuzione normale: se questa viene rifiutata la procedura standard è inadeguata ed è necessario ricorrere a procedure diverse.

Test per la verifica dell'ipotesi di normalità. Uno dei più famosi test per questa ipotesi nulla è quello di test di Kolmogorov-Smirnov. Riportiamo di fianco un esempio di tale test. Notiamo che il p-value è molto basso, e l'ipotesi nulla (secondo la quale la distribuzione è normale) viene quindi rifiutata.

One-Sample Kolmogorov-Smirnov Test

		Aumento
N		474
Normal Parameters[b]	Mean	215,7265
	Std. Deviation	133,08800
Most Extreme Differences	Absolute	,178
	Positive	,178
	Negative	-,142
Kolmogorov-Smirnov Z		3,884
Asymp. Sig. (2-tailed)		,000

b. Calculated from data.

Nel caso in cui cada l'ipotesi di normalità della distribuzione è opportuno verificare l'uguaglianza tra le medie ricorrendo a procedure non parametriche. Tali test sono sostanzialmente simili a quelli utilizzati nell'ANOVA non parametrica.
Non entriamo nei dettagli ma descriviamo solo sinteticamente l'idea su cui si basa il test più importante, detto test di **Wilcoxon-Mann-Whitney**. L'idea è quella di considerare i ranghi associati ad ogni osservazione. Consideriamo per semplicità 10 osservazioni su *Y*, variabile dipendente e *X* variabile esplicativa (categorica).

Y	2	5	7	8	10	6	9	11	13	15
X	1	1	1	1	1	2	2	2	2	2

Associamo ad ogni osservazione su *Y* il suo rango, cioè la posizione che occupa nella sequenze ordinata dei dati.

Y (ordinata)	2	5	6	7	8	9	10	11	13	15
X	1	1	2	1	1	2	1	2	2	2
Rango	1	2	3	4	5	6	7	8	9	10

Determiniamo ora la somma dei ranghi relativi alle osservazioni nel primo strato ($X = 1$) e la somma dei ranghi relativi alle osservazioni del secondo strato ($X = 2$).

Somma dei ranghi ($X = 1$) = (1 + 2 + 4 + 5 + 7) = **19**
Somma dei ranghi ($X = 2$) = (3 + 6 + 8 + 9 + 10) = **36**.

L'idea è che se il fattore non è significativo, le distribuzioni condizionate dovrebbero risultate "confuse" e quindi le somme dei ranghi (o delle opportune medie per tener conto che il numero di

osservazioni può variare da strato a strato) dovrebbero risultare prossime tra loro. Se invece le distribuzioni sono molto diverse (ad esempio, se $X = 1$ i valori assunti da Y sono molto più bassi dei valori assunti nel caso in cui $X = 2$), tali differenze dovrebbero riflettersi nei ranghi.
A partire da tali quantità viene costruita una statistica test per verificare l'ipotesi nulla che ***le due popolazioni condizionate sono identiche tra loro***.
Il test descritto non è l'unico, ma non è questa la sede per descrivere anche gli altri.