12 L'ANALISI MONOVARIATA

di Annalisa Di Benedetto

1. A cosa serve l'analisi monovariata?

L'analisi monovariata mira a descrivere la distribuzione di una variabile – cioè di un singolo carattere operativizzato, in un determinato insieme di dati – e le sue caratteristiche rilevanti. È il punto di partenza dell'analisi dei dati ed è estremamente informativa, per quanto tecnicamente semplice.

Le informazioni che l'analisi monovariata può fornire rispondono a domande elementari, ma non secondarie né scontate. La stessa rilevanza di un obiettivo cognitivo può dipendere da quanto emerge in questa fase, così come possono dipenderne la progettazione delle successive fasi di analisi e la considerazione dei relativi risultati.

Ad esempio, considerando come variabile rilevante il reddito, prima ancora di porsi domande su *chi ha di più o di meno* (i.e. quali sono le caratteristiche da cui il reddito dipende: dal genere, al titolo di studio, al tipo di attività economica) è necessario sapere *quanti hanno quanto* (qual è la *distribuzione* del reddito), *quanto* avrebbe ciascuno se tutti avessero lo stesso reddito (qual è la *media* della distribuzione), *quanto* sono diversi l'uno dall'altro i redditi rilevati (qual è la *variabilità* della distribuzione).

La descrizione della distribuzione e le caratteristiche che è possibile esaminare dipendono dal tipo di proprietà in esame e dalle modalità di operativizzazione adottate, cioè dal tipo di variabile (si veda il Cap. 3) e più precisamente dalle sue caratteristiche logico-matematiche: cioè alle possibilità di confrontarne le modalità in termini di uguale/diverso,

maggiore/minore, e/o di effettuare con pieno significato operazioni matematiche sui valori che assume.
Nel seguito si farà riferimento alla distinzione tra variabili categoriali nominali, categoriali ordinali e cardinali (e quasi-cardinali), sia per poter presentare chiaramente i valori caratteristici delle distribuzioni per ciascun tipo di variabile, sia per poter evidenziare le differenze tra le possibili modalità di rappresentazione grafica che, come si avrà modo di sottolineare, hanno una valenza analitica oltre che una valenza comunicativa.
Al di là degli scopi di descrizione e sintesi delle distribuzioni, le ulteriori funzioni dell'analisi monovariata sono (Marradi, 1993):

1. controllare la plausibilità dei valori;
2. segnalare squilibri nella distribuzione e opportunità di aggregazione;
3. consentire una valutazione critica del lavoro.

Prima ancora dell'analisi dei dati vera e propria, infatti, l'esame delle distribuzioni delle variabili permette di effettuare un controllo sulla qualità dei dati in matrice e di riflettere sull'eventuale necessità di rivedere alcune definizioni operative (ad esempio aggregando alcune delle modalità registrate) in vista delle analisi successive.
I controlli connessi alla plausibilità dei valori (già discussi nel Cap. 11) sfruttano gli strumenti dell'analisi monovariata; tuttavia, vanno condotti prima dell'analisi vera e propria, in modo tale che i risultati non risentano degli effetti di problematiche legate a codifiche errate (cioè errori di inserimento dei dati in matrice dovuti, ad esempio, all'immissione di codici non previsti nella classificazione originaria) e risposte non dovute.
Inoltre, l'analisi monovariata permette di riflettere sull'adeguatezza delle variabili che derivano dalle definizioni operative adottate rispetto agli obiettivi cognitivi, sia sul piano semantico che sul piano statistico. L'esame di ciascuna delle variabili può infatti evidenziare elementi che spingono a una riflessione sulle scelte adottate e portare a una revisione delle definizioni operative sul piano classificatorio, ad esempio prevedendo l'aggregazione di diverse modalità. Queste necessità possono emergere sia nei primi controlli a monte sia nel corso dell'analisi vera e propria, non solo in relazione all'analisi monovariata, ma anche in relazione all'analisi bi- e multivariata.

Infine, la terza funzione posta in evidenza da Marradi (1993) è relativa alla presentazione dei risultati. Potrebbe sembrare scontato, ma non è così: la rendicontazione della ricerca sociale non sempre tiene conto di questo invito alla completezza. È importante ricordare che includere i risultati dell'analisi monovariata nella presentazione degli esiti della ricerca permette di ancorare al piano descrittivo le successive conclusioni – descrittive e/o esplicative – e argomentare le scelte operative fatte. Ciò assicura non solo la pubblicità dei risultati, ma anche la loro riproducibilità e controllabilità.
È importante sottolineare che le analisi presentate nel seguito possono essere svolte anche su un foglio di calcolo (come Excel per Office), ma tutte le applicazioni per l'analisi statistica dei dati prevedono funzionalità specifiche per l'analisi monovariata: dalla produzione di tabelle di frequenza al calcolo dei valori caratteristici della distribuzione[1]. La logica dell'analisi è più rilevante degli strumenti che si utilizzano per realizzarla; per questo si farà riferimento alle formule statistiche che permettono sia di effettuare l'analisi su fogli di calcolo, sia di comprendere le operazioni che sottostanno ai risultati ottenuti utilizzando programmi di analisi statistica.

2. La distribuzione di frequenza

Il primo passo dell'analisi monovariata è l'analisi della distribuzione di frequenza, cioè in pratica dei conteggi di quanti casi/unità presentano ciascuna delle modalità che una certa variabile assume nell'insieme di dati in analisi.
Una distribuzione di frequenza, infatti, riporta quante volte una modalità – o un valore – si presenta nell'insieme di dati per la variabile in esame. È, dunque, una sintesi delle informazioni contenute in una singola colonna di una matrice classica casi per variabili (si veda il Cap. 11) e può

[1] Con riferimento ai programmi più diffusi i comandi per ottenere questi output sono semplici e abbastanza intuitivi. Ad esempio, per le distribuzioni di frequenza, si possono utilizzare i comandi: *frequencies* per SPSS, *tabulate* per STATA, *proc freq* per SAS, *table* per R.

essere presentata sia in forma tabellare che, come si vedrà, in forma grafica.
Dal punto di vista statistico la frequenza assoluta (n_i) corrisponde al numero di casi che presentano una specifica modalità (x_i) per una specifica variabile. La distribuzione delle frequenze assolute per una variabile X, con *k* modalità, su un insieme di dati con N casi è tale per cui la somma delle frequenze assolute è pari al numero totale dei casi nell'insieme:

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + [\dots] + n_k = N$$

Ad esempio, nella Tabella 1 si considera il genere dei rispondenti all'indagine ISTAT "Aspetti della vita quotidiana" del 2020[2]. In totale i rispondenti sono stati 42.810 (*N*) e la variabile "genere" presenta due modalità: "femmina" (k_1) e "maschio" (k_2). Il numero 22.292 è la *frequenza assoluta* della modalità "femmina" (n_1), il numero 20.518 è la *frequenza assoluta* della modalità "maschio" (n_2): la loro somma – essendo la variabile nota per tutti i casi – corrisponde al totale dei rispondenti (*N*).

Tab. 1 – Distribuzione di frequenza del genere dei rispondenti (AVQ 2020)

Genere	***n***	***p***	***%***
Femmina	22.292	0,521	52,1%
Maschio	20.518	0,489	47,9%
Totale	**42.810**	**1**	**100,0%**

Le distribuzioni di frequenza assolute (*n*) possono risultare poco agevoli da leggere e ancor meno immediate da confrontare, per questa ragione si accompagnano con le *frequenze relative* (p), che si ottengono rapportando il numero di casi che presentano una certa modalità al numero totale dei

[2] Quasi tutti gli esempi presentati nel capitolo sono elaborati a partire dal file dei microdati pubblicamente accessibili dell'indagine "Aspetti della vita quotidiana" del 2020 (https://www.istat.it/it/archivio/129956). Si tratta di un'indagine campionaria a cadenza annuale, realizzata con tecnica PAPI (*Paper And Pencil Interview*) e condotta su un campione di circa 20.000 famiglie e 50.000 individui. L'indagine è dedicata a una serie di aspetti fondamentali della vita quotidiana e ai relativi comportamenti, presenta quindi l'opportunità di individuare diversi aspetti interessanti. Per questi esempi la didascalia delle tabelle e delle figure riporta "AVQ 2020"; per gli esempi provenienti da altre fonti di dati la fonte è specificata puntualmente in una nota.

casi. La frequenza relativa di una modalità corrisponde alla quota di casi che presentano quella modalità nell'insieme di dati.
In termini statistici, la frequenza relativa della *i*-esima modalità di una variabile (p_i) è data dal rapporto della frequenza assoluta per l'*i*-esima modalità (n_i) sul numero totale dei casi nell'insieme (N):

$$p_i = \frac{n_i}{N}$$

Nell'esempio riportato (Tab. 1), rapportando il numero di rispondenti di sesso femminile al totale si ottiene la frequenza relativa 0,521. È però evidente che – per quanto l'informazione sia la stessa – si legge e comprende più semplicemente il valore percentuale, pari al 52,1%: oltre 52 rispondenti su 100 sono femmine. Le percentuali sono le frequenze relative più utilizzate. Indicate con il simbolo %, si calcolano rapportando a 100 il totale dei casi, ossia moltiplicando per 100 le frequenze relative:

$$\% = \frac{n_i}{N}(100)$$

In caso la variabile in esame sia ordinale o cardinale – nel caso cioè in cui sia possibile considerare le modalità della variabile in termini di maggiore/minore – è possibile presentare anche una *distribuzione di frequenza cumulata*, cioè una distribuzione che presenti in corrispondenza di ciascuna modalità la somma delle frequenze corrispondenti alla modalità e a tutte le modalità inferiori (o superiori; in questo caso ci si riferisce a *distribuzioni di frequenza retrocumulate*).
Ad esempio, per l'età, può essere utile rendere evidente quanti casi hanno 18 anni e qual è la loro quota sul totale, ma anche quanti casi in totale hanno *meno* di 18 anni e qual è la loro proporzione sul totale. La distribuzione cumulata può essere calcolata sia per le frequenze assolute che per le frequenze relative, ma soltanto per variabili che siano almeno ordinali (per cui sia quindi possibile determinare l'ordinamento rilevante).
Nell'esempio in Tabella 2 è presentata la distribuzione di frequenza del titolo di studio dei rispondenti all'indagine ISTAT "Aspetti della vita quotidiana" del 2020. La presenza delle frequenze cumulate (indicate in tabella con *p(cum)* e *%(cum)*) ci permette di osservare in modo intuitivo che i rispondenti che hanno un diploma o un titolo di studio inferiore sono l'84,6% (che corrisponde alla somma delle frequenze relative delle

modalità "Licenza di scuola elementare o nessun titolo", "Licenza di scuola media" e "Diploma").
La tabella ci dà modo di evidenziare un ulteriore elemento: l'eventuale presenza di dati mancanti *non dovuti* (cioè effettivamente mancanti, a seguito di una mancata risposta da parte di rispondenti cui la domanda era destinata) o *dovuti* (cioè mancanti perché non rilevanti o rilevabili; in genere queste mancate risposte sono *attese*, ad esempio, seguono una domanda filtro che comporta la non risposta da parte di un sottogruppo di rispondenti alla domanda successiva; si veda a questo proposito il Cap. 11).

Tab. 2 – Distribuzione di frequenza del titolo di studio dei rispondenti (AVQ 2020)

Titolo di studio	**n**	**p**	**%**	**p (val)**	**% (val)**	**n (cum)**	**p (cum)**	**% (cum)**
Licenza di scuola elementare o nessun titolo	8.682	0,203	14,4%	0,216	21,6%	8.682	0,216	21,6%
Licenza di scuola media	11.351	0,265	32,6%	0,283	28,3%	20.033	0,499	49,9%
Diploma	13.970	0,326	26,5%	0,348	34,8%	34.003	0,846	84,6%
Laurea o post laurea	6.177	0,144	20,3%	0,154	15,4%	40.180	1,000	100,0%
Totale validi	**40.180**			**1,000**	**100,0%**			
Mancanti non dovuti	*882*	0,022	*2,1%*					
Mancanti dovuti (età inferiore ai 5 anni)	*1. 748*	0,039	*4,1%*					
Totale	**42.810**	1,000	**100,0%**					

In tabella il numero dei mancanti *non dovuti* corrisponde al numero di rispondenti a cui il titolo di studio è stato richiesto, ma non lo hanno indicato; invece, il numero di mancanti *dovuti* corrisponde al numero di casi per cui il titolo di studio non era richiesto perché non rilevante: cioè ai rispondenti con età pari a 5 anni o inferiore (cioè in età pre-scolare).
Quando il valore della variabile in esame non è noto per tutti i casi è sempre opportuno riportare adeguatamente l'informazione (sui dati mancanti) nella rendicontazione dell'analisi. I dati mancanti possono essere presentati in tabella come nell'esempio, oppure riportati in una nota, a seconda dello stile di presentazione adottato e della rilevanza informativa nel contesto (si veda ad esempio la nota alla Tabella 3).
Nell'analisi monovariata la considerazione dei valori mancanti è necessaria per valutare la completezza dell'informazione e permette di calcolare le frequenze relative non solo sul totale dei casi ma anche sul totale dei *casi validi.* Entrambe le informazioni sono rilevanti. Ad esempio,

i diplomati sono il 26,5% del totale dei rispondenti, ma il 34,8% dei rispondenti per cui il titolo di studio è noto; questa seconda percentuale corrisponde alla percentuale *valida*, calcolata sul totale dei casi che hanno fornito una risposta, depurata quindi dai dati mancanti (Tabella 2).
Le frequenze assolute e relative possono essere calcolate indipendentemente dal tipo di variabile: è sempre possibile contare quanti casi presentino una certa modalità o un certo valore.
È importante, tuttavia, considerare che la presentazione tabellare delle distribuzioni di frequenza per le variabili cardinali è consigliabile soltanto se il numero di valori che la variabile assume è molto limitato.
Ad esempio, le risposte alla domanda: "Attualmente quanto si ritiene soddisfatto della sua vita nel complesso?", essendo articolate con una scala Cantril da 0 a 10 e dunque trattabili con qualche cautela come una variabile quasi-cardinale[3], possono essere senz'altro presentate adeguatamente da una distribuzione di frequenza (Tabella 3).

Tab. 3 - Distribuzione di frequenza delle risposte alla domanda: Attualmente quanto si ritiene soddisfatto della sua vita nel complesso? (n, AVQ 2020)*

Soddisfazione	n	p (val)	% (val)	p (cum)	% (cum)
0	222	0,006	0,6%	0,006	0,6%
1	115	0,003	0,3%	0,009	0,9%
2	201	0,005	0,5%	0,014	1,4%
3	409	0,011	1,1%	0,025	2,5%
4	764	0,020	2,0%	0,046	4,6%
5	2.836	0,076	7,6%	0,122	12,2%
6	5.795	0,155	15,5%	0,277	27,7%
7	9.674	0,259	25,9%	0,536	53,6%
8	11.002	0,295	29,5%	0,831	83,1%
9	4.011	0,107	10,7%	0,938	93,8%
10	2.317	0,062	6,2%	1,000	100,0%
Totale*	**37.346**	**1,000**	**100,0%**		

** Dati mancanti non dovuti: 675 (1,6%); dati mancanti dovuti (età inferiore ai 14 anni): 4.789 (11,2%)*

Una tabella che riporti, ad esempio, la distribuzione di frequenza del reddito potrebbe teoricamente richiedere tante righe quanti sono i casi in matrice (posto che nessuno dei casi abbia un reddito identico) e dunque non presentare alcun vantaggio dal punto di vista sintetico, informativo e descrittivo rispetto alla matrice stessa. In caso di variabili cardinali di

[3] Per la questione dell'effettiva quasi-cardinalità delle risposte a scale autoancoranti e sulle necessarie cautele si rimanda nuovamente al Capitolo 7.

questo genere, o in ogni caso in cui le modalità siano molto numerose, è preferibile optare per una rappresentazione grafica e anche in quel caso, per permettere una lettura e un'interpretazione più agevole, è possibile prevedere il raggruppamento dei valori in classi.
Nella Tabella 4, ad esempio, si presenta la distribuzione di frequenza del numero di libri letti negli ultimi 12 mesi dai rispondenti all'indagine.

Tab. 4 - Distribuzione di frequenza del numero di libri letti negli ultimi 12 mesi dai rispondenti (AVQ 2020)

Numero di libri letti negli ultimi 12 mesi	n	p	%	p (cum)	% (cum)
Nessuno	23081	0,571	57,1%	0,571	57,1%
1	1782	0,044	4,4%	0,615	61,5%
2	3101	0,077	7,7%	0,691	69,1%
3	2724	0,067	6,7%	0,759	75,9%
4	1724	0,043	4,3%	0,801	80,1%
5	1602	0,040	4,0%	0,841	84,1%
6	1101	0,027	2,7%	0,868	86,8%
7	399	0,010	1,0%	0,878	87,8%
8	572	0,014	1,4%	0,892	89,2%
9	119	0,003	0,3%	0,895	89,5%
10	1446	0,036	3,6%	0,931	93,1%
11	60	0,001	0,1%	0,932	93,2%
12	517	0,013	1,3%	0,945	94,5%
13	59	0,001	0,1%	0,947	94,7%
14	51	0,001	0,1%	0,948	94,8%
15	450	0,011	1,1%	0,959	95,9%
16	45	0,001	0,1%	0,960	96,0%
17	22	0,001	0,1%	0,961	96,1%
18	42	0,001	0,1%	0,962	96,2%
19	12	0,000	0,0%	0,962	96,2%
20	560	0,014	1,4%	0,976	97,6%
21-25	221	0,005	0,5%	0,981	98,1%
26-30	247	0,006	0,6%	0,987	98,7%
31-40	170	0,004	0,4%	0,992	99,2%
41-50	189	0,005	0,5%	0,996	99,6%
51 e più	153	0,004	0,4%	1,000	100,0%
Totale*	**40449**	**1,000**	**100,0%**		

** Dati mancanti non dovuti: 613(1,4%); dati mancanti dovuti (età inferiore ai 5 anni): 1748 (4,1%)*

La variabile è cardinale (trattandosi di numeri reali interi), ma in parte è ricondotta in classi: al di sopra dei venti libri letti i valori numerici sono stati aggregati nelle classi: 21-25, 26-30, 31-40, 41-50 e 51 e più. In questo

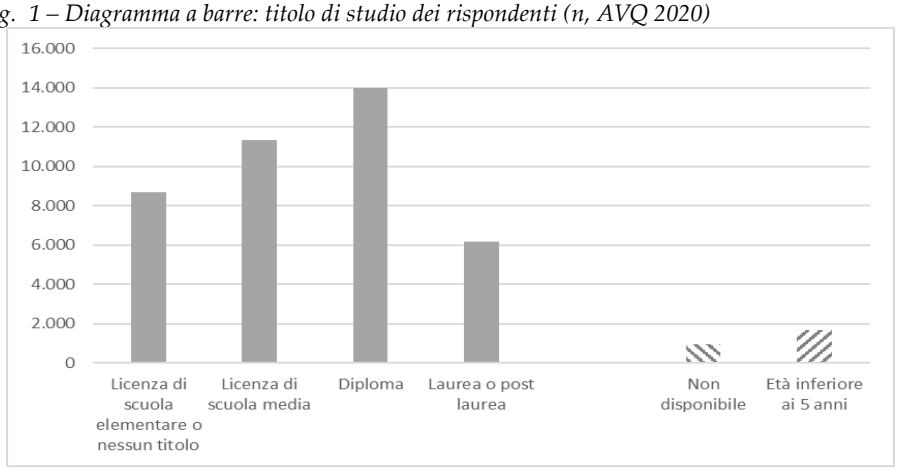
caso l'aggregazione è stata effettuata dall'ISTAT al rilascio del file di microdati a uso pubblico (nel quadro delle ricodifiche effettuate al fine di limitare il rischio di violazione della riservatezza), ma rappresenta comunque un buon esempio dell'opportunità della riconduzione in classi delle variabili cardinali al fine della presentazione della distribuzione di frequenza: la tabella si presenta anche così molto articolata e poco agevole da leggere, anche perché le frequenze sono fortemente eterogenee. La tabella potrebbe essere resa ulteriormente sintetica e chiara aggregando in classi anche gli altri valori; ad esempio, una classe di valori da 16 a 20 includerebbe 681 casi.
Sui possibili criteri che possono essere adottati nei casi in cui si ritiene opportuna una aggregazione si tornerà più avanti riprendendo anche questo esempio (si veda il paragrafo 6).

2.1. Le rappresentazioni grafiche delle distribuzioni di frequenza

Le rappresentazioni grafiche delle distribuzioni di frequenza possono essere più o meno adeguate in base al tipo di variabile in esame, ma tutti i tipi di variabile possono essere rappresentati utilizzando diagrammi a barre (per quanto nel caso delle variabili cardinali sia preferibile, come si vedrà, optare per altre soluzioni).
I diagrammi a barre (detti anche *ortogrammi*), infatti, riportano su un asse le modalità della variabile e sull'altro il valore della frequenza (assoluta o relativa)[4]. Nel caso dei diagrammi a barre è l'altezza della barra a dare conto della frequenza della modalità (la Figura 1 è riferita alla Tabella 2).
I dati mancanti possono essere presentati nel grafico come in Figura 1, oppure riportati in nota al grafico come per la Figura 4 più avanti, a seconda dello stile adottato e della rilevanza informativa nel contesto.

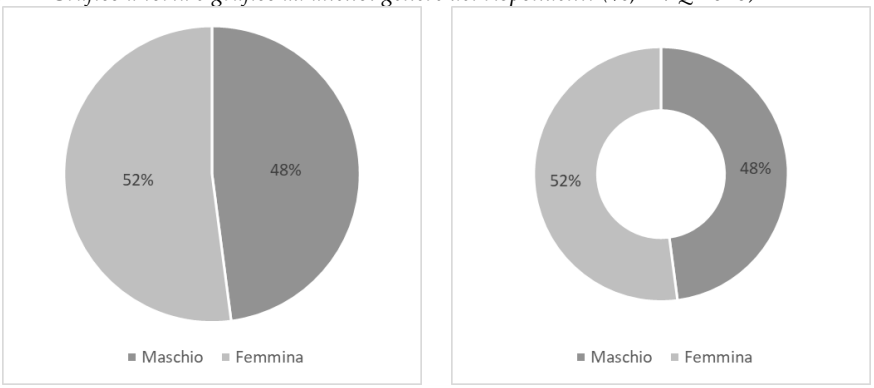
[4] Questo tipo di grafico è utile anche al confronto tra distribuzioni di frequenza, come si vedrà più avanti e poi per l'analisi bivariata (nel capitolo successivo).

Fig. 1 – Diagramma a barre: titolo di studio dei rispondenti (n, AVQ 2020)



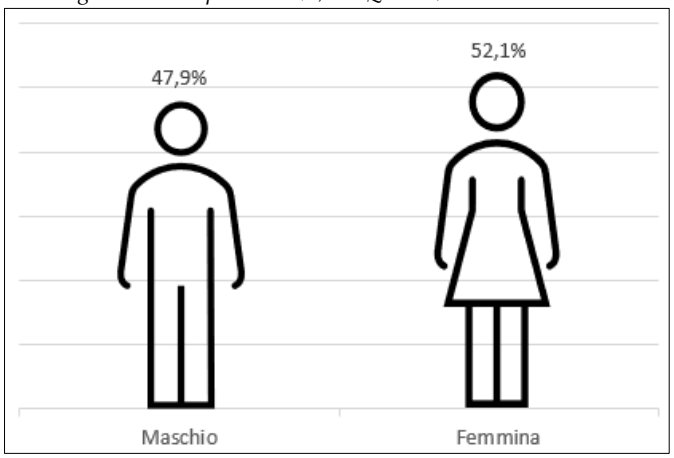
È anche possibile optare per diagrammi di composizione (*areogrammi*) in cui le frequenze sono rappresentate con aree proporzionali (come per i grafici a torta o ad anello, ma anche dei diagrammi a barre suddivise, utili nella rappresentazione congiunta di due variabili). I grafici a torta o ad anelli sono generalmente utilizzati per variabili categoriali nominali, in questo caso l'ordine delle variabili non ha infatti alcuna importanza, e risultano più efficaci nel caso vi siano poche modalità; la differenza tra i due tipi di grafico è puramente estetica (si veda la Figura 2, riferita all'esempio in Tabella 1). In caso di variabili ordinali è più utile scegliere una rappresentazione che permetta di considerare l'ordinamento delle modalità e dunque, come per la Figura 1, un diagramma a barre.

Fig. 2 – Grafico a torta e grafico ad anello: genere dei rispondenti (%, AVQ 2020)



Nell'insieme degli areogrammi rientrano, inoltre, i meno classici ma più accattivanti *ideogrammi*, in cui la grandezza delle figure utilizzate come rappresentazioni delle modalità è proporzionale alla loro frequenza (ad esempio, si veda la Figura 3).

Fig. 3 - Ideogramma: genere dei rispondenti (n, AVQ 2020)



Questo tipo di rappresentazione è utilizzato soprattutto a fini divulgativi, ad esempio nelle infografiche, ma presenta l'inconveniente di rendere meno semplici i confronti nel caso le figure utilizzate siano molto dissimili tra loro.

Le variabili cardinali permettono di utilizzare altre rappresentazioni grafiche. Gli *istogrammi*, ad esempio, sono utilizzati per rappresentare le classi di variabili cardinali e pur apparendo simili ai diagrammi a barre sono in effetti areogrammi: è l'area dei rettangoli rappresentati ad essere proporzionale alla frequenza della classe corrispondente.

Fig. 4 - Istogramma: statura dei rispondenti in centimetri (n, AVQ 2020)*

**Mancanti non dovuti: 341 (0,80%)*

Nell'esempio in Figura 4 si presenta la distribuzione della statura (in cm) per i rispondenti all'indagine: l'altezza di ciascuna colonna è proporzionale alla frequenza della classe e – in questo caso – ha la stessa base, dato che ciascuna classe fa riferimento a un insieme di valori di un'ampiezza di 5 cm, con l'eccezione delle categorie estreme (statura inferiore o uguale a 150 cm; statura superiore ai 200 cm).

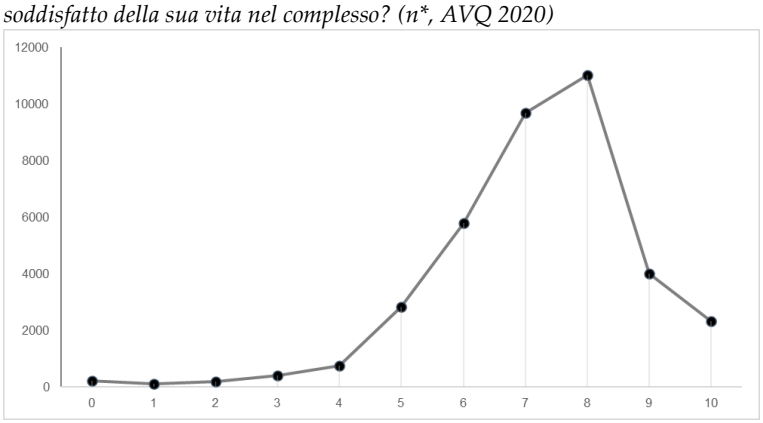
I poligoni di frequenza, sempre con riferimento alle variabili cardinali, sono costruiti con lo stesso principio dei diagrammi a barre, ma con un vero e proprio piano cartesiano[5] in cui sull'asse delle modalità è riportato in questo caso effettivamente il valore numerico della variabile. La frequenza di ciascun valore nell'insieme di dati è rappresentata come un punto. La distribuzione di frequenza può essere quindi raffigurata come una linea spezzata che unisce i punti. La Figura 5 presenta il poligono di

[5] Il piano cartesiano è un sistema di riferimento formato da due rette ortogonali (l'asse delle ascisse y e l'asse delle ordinate x), orientate e definite da una unità di misura, che si intersecano in punto (origine). Su questo piano ciascun punto è individuabile in base a due valori (coordinate cartesiane): uno sull'asse delle x l'altro sull'asse delle y.

frequenza relativo alla soddisfazione per la vita nel suo complesso (Tabella 3).

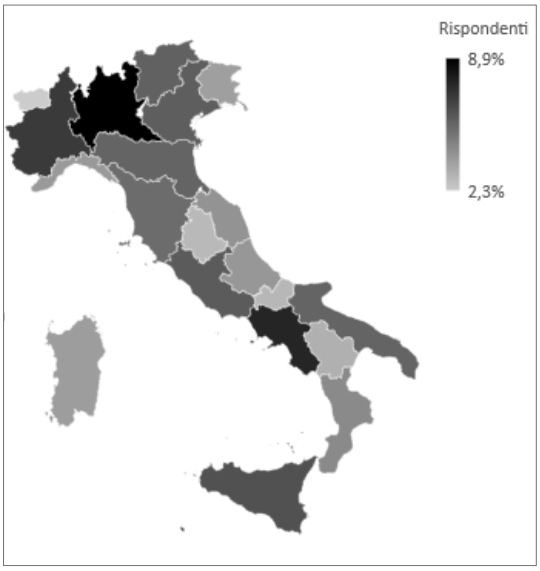
I poligoni di frequenza hanno una certa rilevanza nell'analisi monovariata, ma – come i grafici a barre – sono utili anche per la rappresentazione grafica congiunta di più variabili. Essendo costituiti – infatti – da semplici linee, permettono un confronto agevole tra diverse distribuzioni, pur mantenendo ciascuna distribuzione rappresentata indipendentemente dall'altra: l'ottica resta quindi di confronto e non ancora di analisi congiunta (su questo punto si tornerà più avanti).

Fig. 5 – Poligono di frequenza delle risposte alla domanda: Attualmente quanto si ritiene soddisfatto della sua vita nel complesso? (n, AVQ 2020)*



** Mancanti non dovuti: 675 (1,6%); dati mancanti dovuti (età inferiore ai 14 anni): 4.789 (11,2%)*

Fig. 6 – Grafico a mappa: numero di rispondenti per regione di residenza (n, AVQ, 2020)*



**Dati mancanti non dovuti: 42 (0,10%); Dati oscurati per impedire identificazioni: 160 (0,37%)*

Nel caso in cui la variabile in esame sia una variabile territoriale, come ad esempio lo stato di nascita, la regione di residenza, ecc., la distribuzione di frequenza può essere raffigurata anche con un grafico a mappa, cioè presentata tramite la rappresentazione cartografica del valore della frequenza per i territori corrispondenti alle modalità (in genere si utilizzano gradazioni di colore proporzionali alla frequenza, assoluta o relativa).
È il caso di sottolineare qui che le rappresentazioni cartografiche prevedono in genere almeno due variabili: quella riferita all'aspetto territoriale e quella in esame; la rappresentazione della distribuzione di frequenza della stessa variabile territoriale è l'unica eccezione.
Nell'esempio in Figura 4 è riprodotta la distribuzione di frequenza della regione di residenza dei rispondenti, sempre con riferimento all'indagine "Aspetti della vita quotidiana" del 2020.

3. I valori caratteristici della distribuzione

La distribuzione di frequenza permette di esaminare in modo completo e nel dettaglio il modo in cui la variabile è distribuita in un insieme di dati.

È utile però anche disporre di informazioni che ci permettano di sintetizzare queste informazioni. Ad esempio, una volta esaminata la distribuzione del titolo di studio, sarà possibile evidenziare quale sia il titolo più diffuso e quanto il livello di istruzione sia omogeneo o meno nella popolazione in esame. A questo fine è possibile calcolare e utilizzare i valori caratteristici della distribuzione: qual è la modalità o il valore che più di ogni altro sintetizza il contenuto informativo della variabile (la *tendenza centrale* della distribuzione), se le frequenze sono concentrate o disperse – cioè come si distribuiscono le altre modalità o valori (qual è la *dispersione* o *variabilità* della distribuzione). Questi valori sono diversi a seconda del tipo di variabile. Di qui in avanti, di conseguenza, si terrà sempre ben presente la distinzione tra variabili categoriali nominali, categoriali ordinali e cardinali (o quasi-cardinali).

3.1. Le misure di tendenza centrale

Le misure di tendenza centrale sono valori caratteristici della distribuzione che ne indicano il baricentro: la modalità o il valore che più di ogni altro sintetizza l'informazione relativa alla variabile in analisi.
Nel caso di variabili categoriali nominali (le cui modalità risultino non ordinabili), l'unica misura di tendenza centrale che è possibile individuare è la *moda*: la modalità prevalente, cioè la modalità di una variabile che si presenta con la maggiore frequenza in una data distribuzione.
Essendo la risposta alla semplice domanda "quale modalità è più frequente?" la moda può essere individuata per tutti i tipi di variabile: nella Tabella 1 è la modalità "femmina" a presentare la frequenza più alta e di conseguenza a rappresentare la moda della distribuzione della variabile "genere"; nella Tabella 2 è la modalità "diploma" la moda della variabile "titolo di studio"; nella Tabella 3 è la modalità "8" la moda della variabile relativa alla soddisfazione per la vita nel suo complesso.
La moda (*Mo*) è una *media lasca* perché non dipende da tutti i valori della distribuzione ma solo da alcuni di essi. Nel caso la frequenza più alta si presenti due volte, cioè per due diverse modalità, entrambe le modalità

rappresentano la moda della distribuzione e la distribuzione stessa è detta *bimodale* (se i valori modali sono più di due la distribuzione è *plurimodale*).
Nel caso di variabili categoriali ordinali, oltre alla moda è possibile individuare la *mediana* della distribuzione. La mediana è la modalità che corrisponde al centro della distribuzione ordinata; dunque, può caratterizzare esclusivamente distribuzioni di variabili che siano ordinabili (ordinali o cardinali).
In termini statistici, considerata la distribuzione di una variabile ordinale X con k modalità su N casi, la mediana è la i-esima modalità che nella distribuzione ordinata occupa il posto centrale. È importante tenere presente che se la distribuzione è dispari avrà un solo posto centrale, ma nel caso in cui la distribuzione sia pari, i posti centrali saranno due e dunque (a meno che ad entrambi corrisponda la stessa modalità) potrebbero essere due anche le mediane corrispondenti. La collocazione della mediana nella distribuzione è quindi calcolabile come:

N dispari	*N pari*
$Me = \frac{N+1}{2}$	$Me = \frac{N}{2}$ e $\frac{N}{2} + 1$

In una distribuzione ordinata che presenti le frequenze relative cumulate è molto semplice individuare la mediana: corrisponde alla modalità in cui cadono le frequenze relative 0,5 e 0,51 o le percentuali del 50 e 51%).
La mediana è più informativa della moda perché fornisce informazioni sulla tendenza centrale della variabile considerando l'ordinamento delle sue modalità. È sempre utile, tuttavia, sottolineare che moda e mediana non necessariamente corrispondono. Nell'esempio della Tabella 2 si è visto che la moda è la modalità "diploma", con una frequenza assoluta pari a 13.970. Il numero di casi validi è pari a 40.180; di conseguenza, la mediana è collocata nelle posizioni:

$$Me = \frac{40.180}{2} = 20.090 \quad \text{e} \quad \frac{40.180}{2} + 1 = 20.191$$

Dato che le frequenze assolute cumulate (*n(cum)*) indicano che le posizioni fino a 20.033 sono occupate dalla modalità "licenza di scuola media" e che la modalità "diploma" occupa le successive posizioni fino a quella numero 34.003, la mediana è unica e corrisponde proprio alla

modalità "diploma" (che occupa le posizioni 20.090 e 20.091 nella distribuzione ordinata).
È possibile anche utilizzare come riferimento le frequenze relative cumulate (*f(cum)*) dato che la mediana corrisponde alla modalità che nella distribuzione cumulata delle frequenze relative occupa la posizione corrispondente a 0,5 (o a 50% se si tratta di percentuali): ad esempio nel caso della Tabella 3 è il valore "6" a rappresentare la mediana, dato che occupa le posizioni che nella distribuzione ordinata vanno dal 27,7% al 53,6%. Mentre nell'esempio del titolo di studio la moda e la mediana corrispondono, nell'esempio della soddisfazione la moda è 8, ma la mediana è 6 (si tornerà più avanti sulla rilevanza di questa informazione).
Nel caso di variabili cardinali è possibile utilizzare la *media* come misura di tendenza centrale. Esistono diversi tipi di medie algebriche. La più conosciuta, comune e semplice è la media aritmetica[6] (quando ci si riferisce alla media senza ulteriori specificazioni vuol dire che si sta parlando di media aritmetica).
La *media aritmetica* è ottenuta dal rapporto tra la somma dei valori della variabile in esame e il numero di casi. Corrisponde al valore che la variabile avrebbe in caso di equidistribuzione, cioè se il totale fosse distribuito equamente (diviso in parti uguali) per tutti i casi. Come già accennato, essa può essere calcolata soltanto per variabili cardinali, dato che i valori della variabile devono avere un pieno significato numerico per poter essere sommati e rapportati, operazioni che non possono essere effettuate per le variabili categoriali, neppure nel caso siano ordinali.
In termini statistici la media aritmetica $\bar{X}$ è data dalla somma dei valori X_i per i casi i-esimi da 1 a N, rapportata al totale dei casi (N):

$$\bar{X} = X_1 + X_2 + X_{\dots} + X_N = \sum_{i=1}^{N} \frac{X_i}{N}$$

[6] Altri tipi di medie algebriche, non approfondite in questa sede, sono la media quadratica (utilizzata ad esempio per le superfici), la media geometrica (utilizzata ad esempio per i tassi di interesse o inflazione, essendo più sensibile ai piccoli valori rispetto alla media aritmetica), la media armonica (utile nel caso in cui sia necessario calcolare una media di rapporti, ad esempio la media della velocità o dei prezzi al consumo). Per una presentazione estesa si rimanda a manuali di statistica descrittiva, ad esempio Di Ciaccio e Borra (2003).

A partire da una distribuzione di frequenza la media è calcolabile come la somma dei valori X_i moltiplicati per la loro frequenza n_i per le k modalità della variabile X, rapportata al totale dei casi (N):

$$\bar{X} = \frac{X_1 n_1 + X_2 n_2 + X_{\dots} n_{\dots} + X_k n_k}{N} = \frac{\sum_{i=1}^{k} X_i n_i}{N}$$

Riprendendo l'esempio in Tabella 2, la variabile "soddisfazione" presenta 11 modalità (da 0 a 10), per calcolare la media si moltiplica ciascuno dei valori della soddisfazione per la propria frequenza (colonna *nx* nella Tabella 5), e si rapporta il totale al numero di casi, ottenendo una media pari a 7,2.

Tab. 5 – Calcolo della media per le risposte alla domanda: Attualmente quanto si ritiene soddisfatto della sua vita nel complesso? (n, AVQ 2020)*

Soddisfazione	**n**	***nx***
0	222	0
1	115	115
2	201	402
3	409	1.227
4	764	3.056
5	2.836	14.180
6	5.795	34.770
7	9.674	67.718
8	11.002	88.016
9	4.011	36.099
10	2.317	23.170
Totale*	**37.346**	**268.753**

$$\bar{X} = \frac{268.753}{37.346} = 7{,}2$$

** Dati mancanti non dovuti: 675 (1,6%); dati mancanti dovuti (età inferiore ai 14 anni): 4.789 (11,2%)*

Quando le tre misure di tendenza centrale coincidono, la distribuzione di frequenza è perfettamente *simmetrica* rispetto al valore centrale (si tornerà anche su questo più avanti), ma generalmente non è così.
Come è possibile osservare anche dalle frequenze relative in Tabella 3 e nel grafico in Figura 5, infatti, i valori superiori a 5 della soddisfazione per la vita nel suo complesso presentano frequenze più elevate dei valori inferiori a 5. In relazione a questa *asimmetria* la media della variabile è pari a 7,2 e la stessa distribuzione presenta mediana pari a 6 e moda pari a 8.
Se le distribuzioni di variabili cardinali danno la possibilità di individuare tutte e tre le misure di tendenza centrale (moda, mediana e media), in linea generale è preferibile utilizzare la misura più informativa: la media.

La media aritmetica, infatti, considera il *valore* della variabile su tutti i casi, mentre la mediana considera la *posizione* di tutti i casi ma solo il valore del caso centrale e la moda considera il valore con la maggiore *frequenza* ma non tutti gli altri. È però importante ricordare che anche le altre due misure forniscono elementi interessanti ed è opportuno tenere conto del possibile apporto informativo di ciascuna di esse.

Ad esempio, nel caso del reddito la media presenta uno svantaggio informativo: l'informazione che ci offre è quanto avrebbe ciascuno se tutti avessero lo stesso reddito, ma questa informazione potrebbe risultare poco adatta a rendere contro della situazione reale. In questo caso potrebbe essere più interessante sapere qual è il valore mediano: cioè sotto quale soglia di reddito si colloca la metà dei casi. Nel caso, infatti, la distribuzione del reddito risulti poco equilibrata un piccolo numero di casi con redditi molto elevati potrebbe influenzare in modo rilevante il valore del reddito medio, mentre la mediana resta una media di posizione che non risente della presenza di valori estremi.

Nella prospettiva dell'analisi dei dati è importante evidenziare fin d'ora che la media aritmetica ha diverse proprietà:

1. il valore della media aritmetica è interno al campo di variazione della variabile ($X_{min} < \bar{X} < X_{max}$);
2. la somma degli scarti dalla media (cioè la somma delle differenze tra i singoli valori presenti nella distribuzione e il valore della media) è nulla;
3. il quadrato della somma degli scarti dalla media (*devianza*) è sempre inferiore alla somma dei quadrati degli scarti da un qualsiasi altro valore della distribuzione (cioè, la media è il valore rispetto al quale la dispersione dei valori della distribuzione risulta minima; si riprenderà questo elemento nel prossimo paragrafo);
4. la media aritmetica è *associativa*, cioè la media per un dato insieme di dati (N) è uguale alla somma delle medie ($\bar{X}_s$) dei possibili S sottoinsiemi dei dati, ciascuna ponderata (cioè moltiplicata) per la numerosità dei casi del sottoinsieme cui è riferita (N_s);
5. la media aritmetica è *invariante per trasformazioni affini*, cioè se si trasformano tutti i valori per uno stesso parametro (cioè se si somma o sottrae, moltiplica o divide ciascuno dei valori della distribuzione per uno stesso numero), la media aritmetica subisce la stessa

trasformazione (in altri termini la media della nuova distribuzione sarà pari alla media aritmetica della distribuzione originale trasformata con la stessa operazione che è stata applicata ai singoli valori).

3.2. Le misure di dispersione e variabilità

Oltre alle misure di tendenza centrale i valori caratteristici della distribuzione includono misure di *dispersione* e *variabilità*. La domanda a cui questi valori rispondono è: "quanto e come variano i valori di un dato carattere all'interno di una popolazione?", "qual è la tendenza della variabile ad assumere modalità o valori differenti nell'insieme di casi in analisi?"
Nel caso delle variabili nominali, essendo possibile considerare soltanto l'uguaglianza o differenza tra gli stati sulla proprietà, è possibile utilizzare quali misure di dispersione gli *indici di omogeneità*[7]. Una distribuzione è *del tutto* omogenea quando *tutti* i casi presentano la stessa modalità, *del tutto* eterogenea quando *ciascuno* dei casi presenta una modalità diversa.
Le distribuzioni reali generalmente si collocano tra questi due estremi e risultano tanto più omogenee quanto più i casi si concentrano su una sola modalità (risultando quindi squilibrate), tanto più eterogenee quanto più i casi si disperdono tra le modalità (risultando quindi equilibrate).
Tanto più una distribuzione è omogenea tanto più alta sarà la probabilità che due unità scelte a caso appartengano alla stessa categoria.
La probabilità che due casi si collochino sulla stessa modalità i è pari al quadrato delle frequenze relative p; l'indice di omogeneità O è quindi calcolabile come la somma del quadrato delle frequenze relative per le modalità da 1 a k:

$$O = p_1^2 + p_2^2 + p_{...}^2 + p_k^2 = \sum_{i=1}^{k} p_i^2$$

[7] È possibile calcolare anche *indici di eterogeneità*, nel seguito non saranno approfonditi dato che sono essenzialmente calcolabili come complementi a 1 degli indici di omogeneità.

Riprendendo l'esempio della Tabella 1, l'indice di omogeneità è calcolabile come:

$$O = (0{,}521)^2 + (0{,}481)^2 = 0{,}271 + 0{,}239 = 0{,}510$$

L'indice dipende da un lato da quanto i casi sono concentrati o meno, dall'altro da quante modalità (*k*) presenta la variabile. Infatti, il massimo dell'omogeneità (tutti i casi su una singola variabile) dà un indice pari a 1, il massimo dell'eterogeneità (un numero uguale di casi su ciascuna modalità) dà un indice pari a $\frac{1}{k}$, cioè al reciproco del numero delle modalità.

Nell'esempio del genere il valore dell'indice è molto vicino al valore per la massima eterogeneità pari a 0,5 (la variabile ha due modalità, quindi con k=2 il massimo dell'eterogeneità è pari a $\frac{1}{2}$ = 0,5).

Dato che il valore dell'indice dipende anche da quante modalità presenta la variabile può essere utile normalizzare[8] il valore, in modo tale che sia comparabile per variabili che presentino un diverso numero di modalità, facendo quindi in modo che il minimo non sia pari a $\frac{1}{k}$ ma a 0.

La normalizzazione si effettua dividendo la differenza tra 0 e il suo valore minimo teorico per la differenza tra il massimo e il minimo teorico:

$$O_{rel} = \frac{(O - \frac{1}{k})}{(1 - \frac{1}{k})} = \frac{(k \cdot O - 1)}{(k - 1)}$$

Il risultato O_{rel} è detto *indice di omogeneità relativa* e assume valore 1 in caso completa omogeneità e valore 0 in caso di completa eterogeneità, indipendentemente dal numero delle modalità della variabile in esame.

È possibile calcolare l'indice di omogeneità anche per la distribuzione del titolo di studio riportata in Tabella 2:

$$O = (0{,}216)^2 + (0{,}283)^2 + (0{,}348)^2 + (0{,}154)^2 = \\ = 0{,}044 + 0{,}075 + 0{,}113 + 0{,}022 = 0{,}254$$

Nel caso del titolo di studio, che ha 4 modalità, alla massima eterogeneità corrisponderebbe un valore pari a 0,25 ($\frac{1}{4}$) anche in questo caso, quindi la distribuzione risulta molto eterogenea.

[8] Si rimanda al par. 5 di questo capitolo per una spiegazione estesa della normalizzazione.

Entrambe le variabili risultano eterogenee, ma presentano un numero diverso di modalità; quindi, per poterli confrontare è opportuno normalizzare entrambi gli indici:

$$\text{Genere:} \quad O_{rel} = \frac{(k \cdot O - 1)}{(k-1)} = \frac{(2 \cdot 0{,}510 - 1)}{(2-1)} = \frac{(1{,}02-1)}{(1)} = 0{,}02$$

$$\text{Titolo di studio:} \quad O_{rel} = \frac{(k \cdot O - 1)}{(k-1)} = \frac{(4 \cdot 0{,}254 - 1)}{(4-1)} = \frac{(1{,}017-1)}{(3)} = \frac{(0{,}017)}{(3)} = 0{,}006$$

Alla luce degli indici di omogeneità relativa è possibile affermare che la distribuzione del titolo di studio è più eterogenea di quella del genere.
Nel caso delle variabili ordinali (e naturalmente delle variabili cardinali) le modalità, oltre che in termini di uguaglianza/disuguaglianza, possono essere considerate in termini di maggiore/minore e dunque, come già constatato, è possibile ordinarle. Di conseguenza è possibile determinare non solo quanto la distribuzione è omogenea/eterogenea, ma anche quanto è concentrata o dispersa rispetto al valore centrale (cioè alla mediana).
La distribuzione ordinata delle modalità permette di individuare diversi *valori di posizione* caratteristici della distribuzione, che – come la mediana – derivano appunto dalla collocazione nella distribuzione ordinata e indicano quanti casi si collocano al di sopra o al di sotto di una certa modalità – nel caso della mediana il 50% dei casi è collocato su modalità pari o uguale e il 50% su modalità superiori o uguale.
Questi valori di posizione sono detti *quantili* e dividono i casi di una distribuzione ordinata di frequenza in gruppi di uguale numerosità (ad esempio i *decili* dividono la distribuzione in 10 parti, i *percentili* in 100 parti, ecc.). Quanto più la distribuzione risulta dispersa tra molte modalità o valori, tanto più i valori di posizione risultano distanti l'uno dall'altro.
Tra i valori di posizione più utilizzati ci sono i *quartili*, che dividono la distribuzione ordinata in quattro parti di uguale numerosità. Nella distribuzione ordinata il 25% dei casi è al di sotto e il 75% al di sopra del primo quartile (Q_1); il 50% dei casi è al di sotto e il 50% al di sopra del secondo quartile (Q_2, che corrisponde alla mediana); il 75% dei casi è al di sotto e il 25% al di sopra del terzo quartile (Q_3).
In base ai quartili è possibile calcolare un indice di variabilità della distribuzione, la *differenza interquartile* (Q), come la differenza tra il terzo e il primo quartile della distribuzione:

$$Q = Q_3 - Q_1$$

Se la differenza tra il primo e il terzo quartile risulta alta la distribuzione è dispersa. Essa indica infatti che il 50% centrale dei casi risulta collocato su più modalità o valori; se invece la differenza è modesta indica che il 50% centrale dei casi (cioè il 50% incluso tra il primo e il terzo quartile) è concentrato su un numero ristretto di valori o modalità.

Tornando all'esempio del titolo di studio: il primo quartile della distribuzione ordinata è collocato nella modalità licenza media (che occupa le posizioni dal 21,6% al 49,9%, la mediana come già visto corrisponde al "diploma", modalità in cui cade anche il terzo quartile della distribuzione (dato che occupa le posizioni dal 49,9% all'84,6%).

Nel caso delle variabili cardinali la differenza interquartile assume un pieno valore numerico e si calcola tra i valori della variabile corrispondenti ai quartili nella distribuzione. Nel caso di variabili ordinali l'indice assume una valenza informativa meno puntuale, si calcola infatti assegnando al quartile il valore numerico corrispondente alla posizione della modalità nella distribuzione ordinata delle modalità e indica semplicemente quante modalità separano i due quartili[9]. Con riferimento all'esempio del titolo di studio una sola modalità divide il primo ("Licenza di scuola media" =2) dal terzo quartile ("Diploma"=3):

$$Q = Q_3 - Q_1 = 3 - 2 = 1$$

La distribuzione è quindi abbastanza concentrata intorno al valore mediano, dato che tra il primo e il terzo quartile cambia una sola modalità.

Con riferimento all'esempio della soddisfazione il primo quartile è collocato in corrispondenza del valore 3 e il terzo quartile in corrispondenza del valore 8:

$$Q = Q_3 - Q_1 = 8 - 3 = 5$$

La differenza interquartile ha un significato più pieno in questo caso, dato che è possibile considerare il numero 5 non solo come numero di modalità ma anche come valore: cinque punti (idealmente[10] equidistanti) separano il primo e il terzo quartile.

Le caratteristiche delle variabili cardinali, infatti, fanno sì che nell'analisi della variabilità sia possibile utilizzare indici che sfruttano non solo la

[9] Indici di dispersione più completi e complessi per le variabili ordinali sono noti, anche se poco utilizzati. Si veda ad esempio: Marradi, 1993.

[10] La questione dell'effettiva quasi-cardinalità delle risposte a scale autoancoranti si rimanda nuovamente al Capitolo 7.

frequenza o la posizione nella distribuzione, ma anche il valore numerico assunto dalla variabile.
Una prima informazione sulla variabilità è il *campo di variazione*, cioè l'intervallo di valori incluso tra il valore minimo e il valore massimo che la variabile assume nell'insieme di dati: [X_{min} ; X_{max}]. Il valore è calcolabile come la differenza tra il massimo e il minimo:

$$C = X_{max} - X_{min}$$

La soddisfazione per la vita nel suo complesso, come già visto, è stata operativizzata con una scala che ha un campo di variazione [0 ; 10], la statura riportata in Figura 4 ha un campo di variazione [0 ; 203]. È quindi evidente che la seconda variabile può potenzialmente assumere un numero di valori molto più alto rispetto alla prima (204 valori contro 11). Questo primo indice risente fortemente della presenza di valori estremi, basandosi su due valori di posizione (il massimo e il minimo), ma ha in comune una caratteristica essenziale con gli altri indici di variabilità, cioè assume valore 0 se tutti i casi presentano lo stesso valore (in altri termini è nullo se la variabilità è nulla).
La variabilità è infatti proporzionale alla "diversità" dei valori nella distribuzione. Considerato che, nel caso di variabili cardinali, se tutti i casi presentassero lo stesso valore questo sarebbe pari alla media, la differenza dei valori dalla media rende conto della variabilità.
Si è detto, però, che la somma delle differenze dalla media è sempre pari a 0, il calcolo della variabilità su questa base rende quindi necessario trasformare le differenze in valori positivi.
Un modo per farlo è considerare il valore assoluto delle differenze (considerando quindi anche le differenze di segno negativo come differenze positive). La media dei valori assoluti delle differenze dalla media è detto *scostamento semplice medio*. In termini statistici lo scostamento semplice medio (*ssm*) per una variabile X è dato dalla somma delle differenze tra i valori X_i e la media $\bar{X}$ in valore assoluto, divisa per il numero di casi:

$$ssm = \frac{\sum[X_i - \bar{X}]}{N} = \frac{\sum_{i=1}^{n}[X_i - \bar{X}]}{N} = \frac{\sum_{k=1}^{n} n_k([X_k - \bar{X}])}{N}$$

Esso può essere scritto come la sommatoria delle differenze in valore assoluto tra i valori delle k modalità e la media (X_k - $\bar{X}$) per le loro

frequenze (n_k). L'esempio in Tabella 6 segue questa modalità di calcolo per quest'indice e i seguenti.

Tab. 6 - Calcolo degli indici di variabilità per le risposte alla domanda: Attualmente quanto si ritiene soddisfatto della sua vita nel complesso? (n, AVQ 2020)*

Soddisfazione	n_k	$X_k - \bar{X}$	$\lvert X_k - \bar{X} \rvert$	$n_k(\lvert X_k - \bar{X} \rvert)$	$(X_k - \bar{X})^2$	$(X_k - \bar{X})^2 n_k$
0	222	-7,2	7,20	1.598,4	51,84	11.508,5
1	115	-6,2	6,20	713,0	38,44	4.420,6
2	201	-5,2	5,20	1.045,2	27,04	5.435,0
3	409	-4,2	4,20	1.717,8	17,64	7.214,8
4	764	-3,2	3,20	2.444,8	10,24	7.823,4
5	2.836	-2,2	2,20	6.239,2	4,84	13.726,2
6	5.795	-1,2	1,20	6.954,0	1,44	8.344,8
7	9.674	-0,2	0,20	1.934,8	0,04	387,0
8	11.002	0,8	0,80	8.801,6	0,64	7.041,3
9	4.011	1,8	1,80	7.219,8	3,24	12.995,6
10	2.317	2,8	2,80	6.487,6	7,84	18.165,3
Totale*	**37.346**			**45.156,2**	**163,2**	**97.062,4**

Media	**7,200**	
Scostamento semplice medio	**1,209**	**= 45.156,2 / 37.346**
Devianza	**163,240**	
Varianza	**2,599**	**= 97.062,4 / 37.346**
Deviazione standard	**1,612**	**= $\sqrt{2,599}$**

** Dati mancanti non dovuti: 675 (1,6%); dati mancanti dovuti (età inferiore ai 14 anni): 4.789 (11,2%)*

Lo *scostamento semplice medio* è intuitivo: si tratta della differenza media dalla media, ma presenta dei limiti dato che – essendo una media aritmetica – risente nella stessa misura di differenze di diversa entità.
Un indice di variabilità più sensibile alle differenze maggiori è lo *scarto quadratico medio* (o *deviazione standard o scarto-tipo*). In questo caso, anziché considerare il valore assoluto delle differenze della media si considera il loro valore elevato al quadrato. Così facendo, infatti, non solo i segni negativi si annullano, ma gli scarti maggiori assumono un peso maggiore. Nel calcolo della deviazione standard rientrano alcuni concetti fondamentali, come la somma del quadrato delle differenze dalla media, che è detta *devianza*:

$$\text{Dev} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{k=1}^{n}(X_k - \bar{X})^2\, n_k$$

Inoltre, la media dei quadrati delle differenze dalla media è una misura fondamentale di variabilità, detta *varianza* (σ^2):

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N} = \frac{\sum_{k=1}^{n}(X_k - \bar{X})^2\, n_k}{N}$$

Il calcolo di entrambi i valori è presentato nella Tabella 6, con riferimento all'esempio della soddisfazione per la vita nel suo complesso.
È importante evidenziare questi passaggi del calcolo poiché nei prossimi capitoli il concetto di varianza sarà richiamato più volte: l'obiettivo dell'analisi dei dati è quello di individuare le ragioni della variabilità tra i casi e le caratteristiche matematiche della varianza ne fanno una delle caratteristiche della distribuzione più utilizzate nell'analisi delle relazioni tra variabili.
Nell'analisi monovariata, tuttavia, l'indice di riferimento più comune è la *deviazione standard* (σ), che si ottiene dalla radice quadrata della varianza, cioè dalla radice quadrata del rapporto tra la somma del quadrato delle differenze dalla media e il numero di casi:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N}} = \sqrt{\frac{\sum_{k=1}^{n}(X_k - \bar{X})^2 n_k}{N}}$$

La radice quadrata ha lo scopo di riportare la deviazione standard nello stesso ordine di grandezza dei valori della variabile e della media, rendendo la deviazione standard di più semplice lettura rispetto alla varianza. È questa caratteristica a farne l'indice di variabilità di riferimento per gli obiettivi nell'analisi monovariata, nonostante alcuni limiti.
La *deviazione standard* risente – inevitabilmente – della grandezza della media della variabile. Si ipotizzano – a fine esemplificativo – qui di seguito due distribuzioni alternative della soddisfazione: una con una media molto alta (A), l'altra con una media che si assesta su valori centrali (B) come in Tabella 7.
La maggior parte dei valori nel caso della distribuzione A è inferiore alla media. Di conseguenza tutti gli scarti positivi hanno un'entità ridotta, mentre nel caso della distribuzione B gli scarti positivi e negativi possono avere la stessa entità.

Tab. 7 – Distribuzioni esemplificative A e B: calcolo della media e degli indici di variabilità

	Distribuzione A							Distribuzione B						
Soddisfazione	n_k	$n_k X_k$	$X_k - \bar{X}$	$\lvert X_k - \bar{X} \rvert$	$n_k(\lvert X_k - \bar{X} \rvert)$	$(X_k - \bar{X})^2$	$(X_k - \bar{X})^2 n_k$	n_k	$n_k X_k$	$X_k - \bar{X}$	$\lvert X_k - \bar{X} \rvert$	$n_k(\lvert X_k - \bar{X} \rvert)$	$(X_k - \bar{X})^2$	$(X_k - \bar{X})^2 n_k$
0	0	0	-7,2	7,20	0,0	51,84	0,0	5	0	-5,0	5,00	25,0	25	125,0
1	1	1	-6,2	6,20	6,2	38,44	38,4	5	5	-4,0	4,00	20,0	16	80,0
2	1	2	-5,2	5,20	5,2	27,04	27,0	10	20	-3,0	3,00	30,0	9	90,0
3	2	6	-4,2	4,20	8,4	17,64	35,3	10	30	-2,0	2,00	20,0	4	40,0
4	3	12	-3,2	3,20	9,6	10,24	30,7	10	40	-1,0	1,00	10,0	1	10,0
5	4	20	-2,2	2,20	8,8	4,84	19,4	20	100	0,0	0,00	0,0	0	0,0
6	4	24	-1,2	1,20	4,8	1,44	5,8	10	60	1,0	1,00	10,0	1	10,0
7	3	21	-0,2	0,20	0,6	0,04	0,1	10	70	2,0	2,00	20,0	4	40,0
8	7	56	0,8	0,80	5,6	0,64	4,5	10	80	3,0	3,00	30,0	9	90,0
9	35	315	1,8	1,80	63,0	3,24	113,4	5	45	4,0	4,00	20,0	16	80,0
10	40	400	2,8	2,80	112,0	7,84	313,6	5	50	5,0	5,00	25,0	25	125,0
Totale*	**100**	**857**			**224,2**	**163,2**	**588,2**	**100**	**500**			**210,0**	**110,0**	**690,0**

	A	B
Media	8,720	5,000
Scostamento semplice medio	2,392	2,100
Devianza	657,200	690,000
Varianza	6,572	6,900
Deviazione standard	2,564	2,627
Coefficiente	0,294	0,525

Se due variabili presentano medie molto diverse, pur avendo lo stesso campo di variazione, è dunque necessario confrontarle utilizzando un indice di variabilità che tenga conto anche della grandezza della media. Il *coefficiente di variazione*[11] rapporta il valore della deviazione standard a quello della media:

$$C_v = \frac{\sigma}{\bar{X}}$$

L'indice ha un minimo teorico uguale a 0 (quando tutti i casi presentano lo stesso valore), mentre il massimo dipende dalle caratteristiche della distribuzione[12]. Spesso, tuttavia, il suo valore è espresso in percentuale: moltiplicandolo per 100 è infatti leggibile come la percentuale della media a cui corrisponde la deviazione standard.

Nell'esempio riportato sopra la distribuzione A ha un coefficiente di variazione di 0,294, quindi la sua deviazione standard corrisponde a meno di un terzo della media (il 29,4%), il coefficiente per la distribuzione B è pari a 0,525: oltre la metà della media (il 52,5%). È evidente che la deviazione standard rendeva conto di una variabilità simile, pur in presenza di medie e distribuzioni molto diverse, mentre il coefficiente di variazione permette di tenere conto dell'entità della media. Si vedranno qui di seguito le possibili rappresentazioni grafiche per i valori caratteristici della distribuzione appena mostrati, per poi presentare ulteriori caratteristiche della distribuzione che hanno rilevanza nell'analisi monovariata e che danno conto di aspetti più specifici.

3.3. Rappresentazioni grafiche per i valori caratteristici delle distribuzioni

La rappresentazione grafica dei valori caratteristici delle distribuzioni può essere effettuata diversamente a seconda del tipo di variabile e della rappresentazione scelta.

[11] L'utilizzo del coefficiente di variazione può essere problematico per distribuzioni che presentino valori sia positivi che negativi, dato che l'ordine di grandezza della media non è effettivo. Potrebbe essere preferibile applicare una trasformazione alla distribuzione.

[12] È possibile normalizzare i coefficienti di variazione rapportandoli al loro massimo teorico, ma questi indici – detti *indici relativi di variabilità* – sono poco utilizzati, essendo strettamente legati alle ipotesi necessarie per determinare il valore massimo teorico del coefficiente di variazione (Di Ciaccio e Borra, 2003).

Le misure di tendenza centrale possono essere poste in evidenza con espedienti grafici per tutti i tipi di variabile, sfruttando il colore delle barre o delle aree (come nell'esempio in Figura 7 dove la mediana presenta una tonalità più scura) o inserendo elementi separati come punti e aree (come in Figura 8, dove le medie lasche (moda e mediana) sono rappresentate da punti di colore diverso lungo il poligono di frequenza, la media da una linea tratteggiata e la deviazione standard da un'area colorata nello sfondo, dal punto $\bar{X} - \sigma$ al punto $\bar{X} + \sigma$ sull'asse delle ascisse).

Fig. 7 – Diagramma a barre: titolo di studio dei rispondenti (n, AVQ 2020)

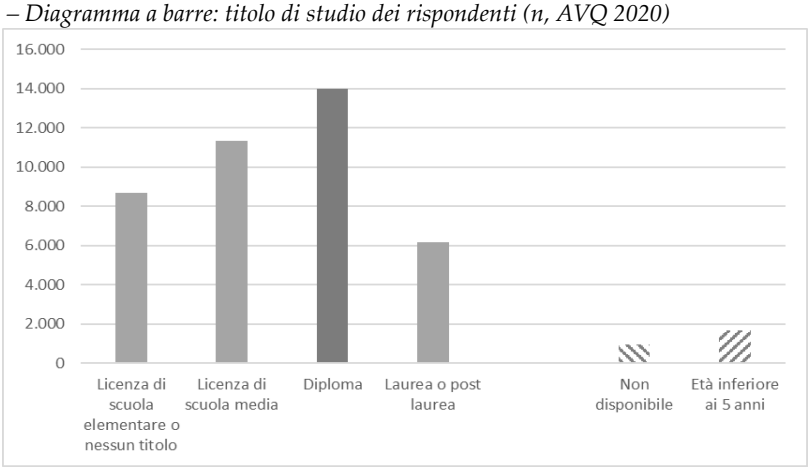
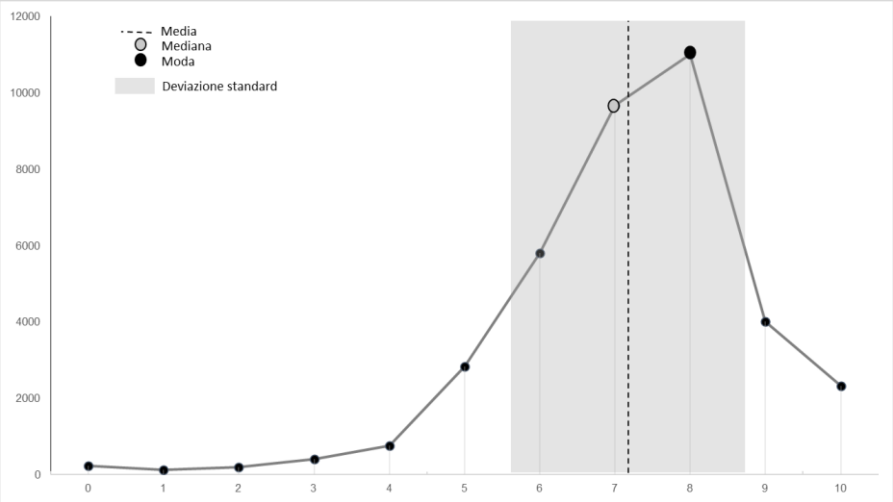


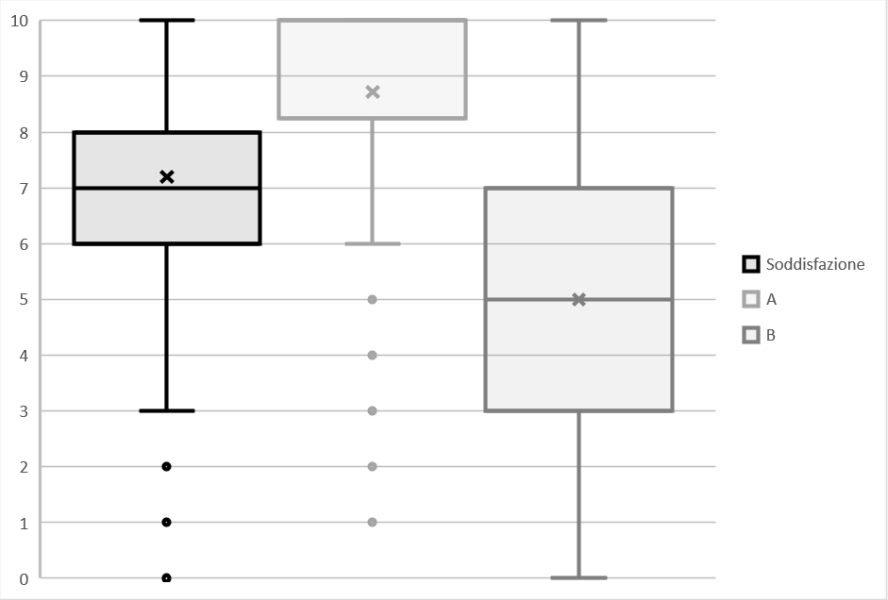
Fig. 8 – Poligono di frequenza delle risposte alla domanda: Attualmente quanto si ritiene soddisfatto della sua vita nel complesso? (n, misure di tendenza centrale e deviazione standard; AVQ 2020)*



Al di là degli espedienti grafici che di caso in caso possono essere scelti per rendere più chiara la presentazione, nel caso delle variabili cardinali è possibile optare per grafici sintetici che presentino diversi valori caratteristici della distribuzione: le scatole a baffi o – più comunemente – *box plot*.

I *box plot* presentano lungo l'asse delle ascisse i valori della variabile, e la distribuzione è rappresentata da un rettangolo (scatola) che parte dal primo quartile e arriva al terzo, presentando una linea in corrispondenza della mediana; la media è rappresentata con un punto (o una croce come in Figura 9), e sotto e sopra la scatola sono presenti delle linee (baffi) che rappresentano la distribuzione. I punti singoli esterni ai baffi rappresentano i valori con frequenza molto bassa (nell'esempio in Figura 9 si tratta dei valori con una frequenza relativa inferiore all'1%).

Fig. 9 – Box plot: Soddisfazione nella vita nel suo complesso (AVQ 2020 e distribuzioni esemplificative A e B in Tabella 7)



I vantaggi di queste rappresentazioni grafiche sono la sintesi e l'immediatezza: la Figura 9 ci permette di visualizzare diverse caratteristiche delle tre distribuzioni già esaminate, semplificandone il confronto.
È ad esempio evidente dalla figura che le tre distribuzioni presentano medie diverse e che soltanto nel caso della distribuzione B media e mediana coincidono.

4. La concentrazione

La variabilità è riferibile a qualsiasi carattere cardinale, ma è possibile considerare anche un ulteriore valore caratteristico della distribuzione, la concentrazione, che è propriamente riferibile solo ad alcune di esse.
Le variabili cardinali, infatti, possono essere riferite a caratteristiche di vario genere. Tra queste ci sono i *caratteri quantitativi trasferibili*, cioè riferiti a proprietà che è possibile immaginare siano trasferibili da un caso all'altro. Non è il caso dell'età per gli individui o del settore di attività per

le aziende. È però il caso del reddito o del fatturato, del numero di dispositivi digitali posseduti o del numero di dipendenti.
È con riferimento a questo genere di variabili che è possibile chiedersi non solo quanto variano nell'insieme di dati, ma anche quanto sono *concentrate*. Si è già osservato che la media corrisponde al valore che la variabile assumerebbe se ciascuna unità ne possedesse la stessa quantità, questa situazione corrisponderebbe a un'*equidistribuzione*.
Al contrario, si avrebbe la massima concentrazione se l'ammontare totale della variabile fosse posseduto da un solo caso.
L'analisi della concentrazione è un tipo specifico di analisi della variabilità.
Il più noto e diffuso indice di concentrazione è il *rapporto di concentrazione di Gini*, che si basa sul confronto tra la distribuzione riscontrata e la massima concentrazione teorica.
A partire dalla distribuzione ordinata della variabile in esame, l'indice richiede il calcolo delle proporzioni dei casi (p_i) e delle relative quantità sulla variabile (q_i) della variabile e le relative proporzioni cumulate (che saranno indicate con p_{ci} e q_{ci}):

$$p_i = \frac{n_i}{N} \qquad\qquad q_i = \frac{n_i x_i}{N\bar{X}}$$

Nell'esempio in Tabella 8, nel caso del reddito si calcolano le proporzioni cumulate della popolazione e le proporzioni cumulate del reddito; la distribuzione esemplificativa fa riferimento un reddito mensile complessivamente pari a 26.500 euro (X) distribuito su 15 casi (N).
Se il reddito fosse equidistribuito, p_{ci} e q_{ci} sarebbero uguali. Vale a dire che il reddito sarebbe suddiviso esattamente come i casi in termini di proporzioni cumulate. Quanto più invece la distribuzione è concentrata tanto più le proporzioni p_{ci} e q_{ci} sono diverse (quindi tanto maggiore risulterà la differenza tra loro).

Tab. 8 – Distribuzione esemplificativa per il calcolo dell'indice di Gini: valore del reddito mensile (X), relativo numero di casi (N) e passaggi del calcolo

x_i	n_i	$n_i x_i$	p_i	q_i	p_c	q_c	$p_c - q_c$
1.000	5	5.000	0,333	0,189	0,333	0,189	0,145
1.200	5	6.000	0,333	0,226	0,667	0,415	0,252
1.500	3	4.500	0,200	0,170	0,867	0,585	0,282
3.000	1	3.000	0,067	0,113	0,933	0,698	0,235
8.000	1	8.000	0,067	0,302	1,000	1,000	0,000
Totale	**15**	**26.500**	**1,000**	**1,000**			

L'indice di concentrazione di Gini (R) è dunque calcolabile come la somma delle differenze ($p_{ci} - q_{ci}$), normalizzata sulla somma delle proporzioni (p_{ci}):

$$R = \frac{\sum_{i=1}^{n-1}(p_{ci} - q_{ci})}{\sum_{i=1}^{n-1}(p_{ci})}$$

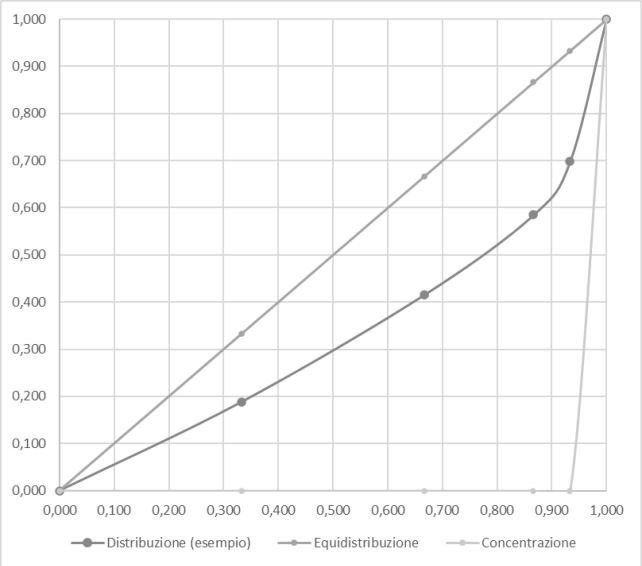
Il risultato è un numero puro (non dipende cioè dall'unità della variabile) e assume un valore incluso tra 0 (equidistribuzione) e 1 (massima concentrazione). Nell'esempio riportato in Tabella 8 l'indice è pari a:

$$R = \frac{0{,}145+0{,}252+0{,}282-0{,}235}{0{,}033+0{,}667+0{,}867+0{,}933} = \frac{0{,}913}{2{,}800} = 0{,}326$$

Per comprendere meglio questo passaggio logico, è più semplice fare riferimento a una rappresentazione grafica, attraverso un piano cartesiano che riporti sull'asse delle ordinate (y) le proporzioni cumulate per il valore (q_{ci}) e sull'asse delle ascisse (x) le proporzioni cumulate dei casi (p_{ci}).
La Figura 10 permette di visualizzare che in caso di equidistribuzione i punti corrispondenti alle i coppie (p_{ci}, q_{ci}) si collocherebbero tutti lungo la bisettrice, in caso di massima concentrazione tutte le i coppie (p_{ci}, q_{ci}) cadrebbero sull'asse delle ascisse tranne una, essendo tutte le proporzioni cumulate dei valori pari a 0 tranne una (quella corrispondente al caso che deterrebbe il totale della quantità in analisi).
La spezzata corrispondente alla distribuzione di X è detta *spezzata di concentrazione* (nel caso di variabili continue detta *curva di Lorenz*, che la propose come rappresentazione proprio in base delle definizioni del coefficiente di Gini), che è sempre compresa tra la bisettrice e l'asse delle ascisse. Quanto più la distribuzione è concentrata, tanto maggiore

risulterà l'area compresa tra la bisettrice e la spezzata di concentrazione, detta *area di concentrazione*.
Su questa base – se il carattere è continuo – è possibile definire l'indice di Gini come il rapporto tra l'area di concentrazione di una distribuzione e il suo massimo (si veda in proposito Di Ciaccio e Borra, 2003).

Fig. 10 – Spezzata di concentrazione della distribuzione in Tab. 8



La spezzata di concentrazione offre maggiori informazioni rispetto al solo valore del rapporto di concentrazione di Gini: dà modo di osservare l'intera distribuzione della quantità in analisi e di evidenziare in corrispondenza di quali proporzioni la distribuzione si allontana maggiormente dall'equidistribuzione. Il rapporto di concentrazione resta tuttavia utile sia in termini informativi che comparativi, grazie alla sua sinteticità.
L'uso più comune del rapporto di concentrazione di Gini è riferibile proprio all'analisi delle disuguaglianze nella distribuzione della ricchezza, ma può trovare utilità anche con riferimento ad altri ambiti in cui la concentrazione risulta di particolare rilevanza: dagli studi sull'integrazione (ad esempio con riferimento alla collocazione su un certo territorio dei residenti stranieri o dei rifugiati, ecc.) a quelli sulla digitalizzazione (ad esempio con riferimento al possesso di dispositivi digitali o all'accesso alla banda larga).

Si riporta, come esempio su dati reali, il calcolo dell'indice sugli introiti di musei, monumenti ed aree archeologiche statali per regione 2020 (Tabella 9). Oltre l'80% degli introiti è concentrato in tre sole regioni: Lazio, Toscana e Campania. L'indice di concentrazione di Gini risulta pari a:

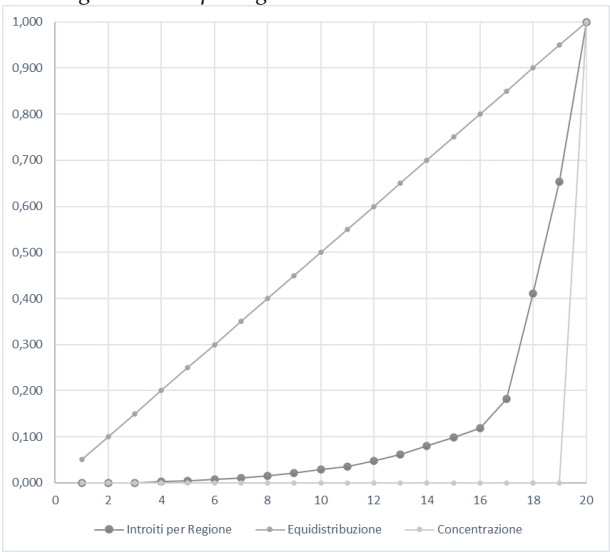
$$R = \frac{\sum_{i=1}^{n-1}(p_{ci} - q_{ci})}{\sum_{i=1}^{n-1}(p_{ci})} = \frac{7{,}719}{10{,}500} = 0{,}735$$

Tab. 9 - Introiti netti di musei, monumenti ed aree archeologiche statali per regione 2020 (introiti in euro)

Regione	**Introiti netti (Euro)**	**n**	$\mathbf{p}_i$	$\mathbf{q}_i$	$\mathbf{p}_{ic}$	$\mathbf{q}_{ic}$	$\mathbf{p}_{ic} - \mathbf{q}_{ic}$
Valle d'Aosta		1	0,050	0,000	0,050	0,000	0,050
Trentino-Alto Adige		1	0,050	0,000	0,100	0,000	0,100
Sicilia		1	0,050	0,000	0,150	0,000	0,150
Molise	103.935	1	0,050	0,002	0,200	0,002	0,198
Liguria	107.485	1	0,050	0,003	0,250	0,005	0,245
Abruzzo	113.900	1	0,050	0,003	0,300	0,008	0,292
Basilicata	126.465	1	0,050	0,003	0,350	0,011	0,339
Sardegna	181.632	1	0,050	0,004	0,400	0,015	0,385
Puglia	283.487	1	0,050	0,007	0,450	0,022	0,428
Umbria	290.933	1	0,050	0,007	0,500	0,029	0,471
Calabria	296.991	1	0,050	0,007	0,550	0,036	0,514
Friuli-Venezia Giulia	499.699	1	0,050	0,012	0,600	0,048	0,552
Piemonte	585.521	1	0,050	0,014	0,650	0,062	0,588
Marche	737.108	1	0,050	0,018	0,700	0,079	0,621
Veneto	807.206	1	0,050	0,019	0,750	0,098	0,652
Emilia-Romagna	864.843	1	0,050	0,021	0,800	0,119	0,681
Lombardia	2.659.346	1	0,050	0,063	0,850	0,182	0,668
Campania	9.600.236	1	0,050	0,229	0,900	0,411	0,489
Toscana	10.197.385	1	0,050	0,243	0,950	0,654	0,296
Lazio	14.535.757	1	0,050	0,346	1,000	1,000	0,000
Totale	**41.991.929**	**20**			***(10,5)***		***(7,719)***

Fonte: ISTAT - Statistiche culturali 2020 (https://www.istat.it/it/archivio/264586)

Fig. 11 - Spezzata di concentrazione della distribuzione degli Introiti netti di musei, monumenti ed aree archeologiche statali per regione 2020



Fonte: ISTAT - Statistiche culturali 2020 (https://www.istat.it/it/archivio/264586)

È interessante notare che in questo caso la spezzata di distribuzione è molto più vicina all'asse delle x, e dunque alla massima concentrazione, di quanto non lo fosse la spezzata dell'esempio precedente, a conferma del fatto che questa rappresentazione grafica, grazie alla forma della spezzata, rende conto della concentrazione della distribuzione (Figura 11).

5. Classificare e confrontare

Tra gli obiettivi principali dell'analisi monovariata c'è quello di creare i presupposti per poter confrontare tra loro le distribuzioni di diverse variabili. Questa funzione che risulta connessa con quella evidenziata da Marradi: la riflessione sulle definizioni operative e la loro eventuale revisione sul piano classificatorio o sul piano statistico.

La revisione di una definizione operativa sul piano classificatorio, a livello monovariato, è sostanzialmente riducibile alla possibilità di aggregare tra loro diverse modalità di una variabile.

La scelta di rivedere la classificazione è legata a considerazioni semantiche, sulla base del significato e della similarità delle modalità da

aggregare, e statistiche, sulla base della loro frequenza e considerando la distribuzione di frequenza complessiva.
Ad esempio, nella Tabella 10 si presenta la distribuzione di frequenza delle risposte alla domanda: "Consideri tutte le attività sportive praticate negli ultimi 12 mesi. Con che frequenza le ha praticate nell'anno?" (AVQ 2020).
La frequenza nell'anno delle attività sportive praticate è una variabile ordinale e la sua distribuzione presenta almeno due modalità con frequenze molto basse rispetto alle altre: "Una volta al mese" (7,6%) e "Qualche volta durante l'anno" (1,8%). Inoltre, la modalità "Due volte a settimana" raccoglie oltre un terzo (il 35% circa) delle risposte.

Tab. 10 – Distribuzione di frequenza della frequenza, nell'anno, delle attività sportive praticate (n, AVQ 2020)*

Frequenza nell'anno delle attività sportive praticate	n	p	%	p (cum)	% (cum)
a) Cinque o più volta a settimana	1321	0,08	8,5%	0,08	8,5%
b) Tre o quattro volte a settimana	4338	0,28	27,8%	0,36	36,3%
c) Due volte a settimana	5507	0,35	35,3%	0,72	71,6%
d) Una volta a settimana	2371	0,15	15,2%	0,87	86,8%
e) Due o tre volte al mese	1184	0,08	7,6%	0,94	94,4%
f) Una volta al mese	286	0,02	1,8%	0,96	96,2%
g) Qualche volta durante l'anno	585	0,04	3,8%	1,00	100,0%
Totale	**15592**	**1,00**	**100,0%**		

** Dati mancanti non dovuti: 844 (1,9%); dati mancanti dovuti (rispondenti che non praticano sport, neppure saltuariamente): 26.374 (61,6%)*

È possibile aggregare tra loro le modalità seguendo un criterio semantico: ad esempio se ci interessasse soprattutto distinguere chi pratica sport almeno una volta a settimana dagli altri potremmo aggregare tra loro le modalità *a*, *b*, *c*, *d* e le modalità *e*, *f* e *g* (come nell'opzione A in Tabella 11). Il risultato darebbe però una distribuzione molto sbilanciata, con solo il 13,2% dei rispondenti che praticano attività sportive meno di una volta a settimana.
Se invece volessimo soprattutto ottenere una distribuzione bilanciata (ad esempio in vista delle analisi successive, come si vedrà dal prossimo capitolo) dal punto di vista delle frequenze sarebbe più opportuno adottare un criterio più statistico e aggregare le modalità in modo tale da

ottenere classi con frequenze quanto più simili possibile tra loro: potremmo aggregare tra loro le categorie *a* e *b*, lasciare com'è la categoria *c* e aggregare le tre categorie *e*, *e* ed *f* (come nell'opzione B).

Tab. 11 – Opzioni di aggregazione per le modalità della variabile: frequenza, nell'anno, delle attività sportive praticate (n, AVQ 2020)*

Opzione A	**n**	**%**
Almeno una volta a settimana	13537	86,8%
Meno di frequente	2055	13,2%
Totale	**15592**	**100,0%**

Opzione B	**n**	**%**
Più di due volte a settimana	5659	36,3%
Due volte a settimana	5507	35,3%
Una volta a settimana o meno di frequente	4426	28,4%
Totale	**15592**	**100,0%**

Opzione C	**n**	**%**
Più di due volte a settimana	5659	36,3%
Due volte a settimana	5507	35,3%
Più volte al mese	3555	22,8%
Una volta al mese o meno di frequente	871	5,6%
Totale	**15592**	**100,0%**

Nella maggior parte dei casi, ad ogni modo, le considerazioni di ordine semantico e quelle di ordine statistico vanno considerate congiuntamente: l'obiettivo di ottenere una distribuzione non troppo sbilanciata non deve lasciare in secondo piano la capacità di discriminare efficacemente i soggetti in base alla proprietà rilevata. Riprendendo nuovamente l'esempio, si potrebbe scegliere di tenere distinti i rispondenti che praticano sport più di due volte a settimana (aggregando le modalità *a* e *b*), due volte a settimana (modalità c) o più di una volta al mese (aggregando le modalità *d* ed *e*) da quelli che lo praticano meno di frequente (aggregando le modalità *f* e *g*), si otterrebbe così una distribuzione non bilanciata ma in grado tenere traccia delle differenze rilevanti tra le abitudini sportive dei rispondenti (nella Tabella 11, l'opzione C).

È importante evidenziare che nel caso di variabili cardinali l'aggregazione delle modalità corrisponde all'identificazione di classi, esempio tipico è quello dell'età.

È possibile utilizzare i percentili per la costruzione di classi (è comune, ad esempio, l'uso dei quartili come soglie), o considerare altri valori caratteristici della distribuzione come riferimenti, come la media e la deviazione standard (ottenendo ad esempio due classi alte al di sopra

della media– una entro e una al di là della deviazione standard – e due classi basse al di sotto della media – sempre una entro e una al di là della deviazione standard). Raramente, tuttavia, queste soluzioni risultano completamente adeguate rispetto agli obiettivi cognitivi. Le classi, come le categorie, devono essere costruite tenendo insieme l'aspetto semantico e quello statistico: avere classi di uguale ampiezza con frequenze molto diverse può essere inutile rispetto agli obiettivi cognitivi quanto l'avere classi con la stessa frequenza ma ampiezze completamente eterogenee.
Per semplicità si riprende l'esempio relativo al numero di libri letti negli ultimi 12 mesi già presentato in Tabella 4 (qui riportato per le sole frequenze assolute e percentuali in Tabella 12). L'ISTAT nel file di microdati – per le ragioni già richiamate legate alla necessità di oscuramento dei dati – presenta per questa variabile i dati dettagliati fino al valore 20, poi raggruppa i valori in cinque classi, due di ampiezza 5 (*21-25* e *26-30*), due di ampiezza 10 (*31-40* e *41-50*) e una classe residuale per i valori oltre una certa soglia (*51 e più*). Si tratta quindi di una distribuzione solo parzialmente ricondotta in classi, che non presentano la stessa ampiezza. Le frequenze dei valori e delle classi sono molto eterogenee e la distribuzione risulta poco agevole da presentare (poiché molto articolata) e interpretare (sia per le dimensioni della tabella che per la distribuzione delle frequenze).
Sempre nella Tabella 12 sono presentate anche due opzioni di aggregazione: A e B. La prima presenta classi di uguale ampiezza (ad eccezione della classe aperta "*51 e più*") e permette di evidenziare molto efficacemente le frequenze decrescenti delle classi (cioè il fatto che tanti più libri letti prevede la classe tanto meno rispondenti vi ricadono), ma presentando ben tre classi con frequenza inferiore all'1% sostanzialmente riferite ai lettori forti, cioè ai rispondenti che leggono più di 30 libri l'anno. La seconda opzione non presenta classi di uguale ampiezza, ma bilancia le considerazioni sul significato del valore della variabile e quelle attinenti alla relativa frequenza. Si ottiene così una variabile con 5 classi di diversa ampiezza e diversa frequenza, ma che permette di evidenziare efficacemente l'incidenza maggiore dei lettori non abituali (*meno di 5 libri l'anno*) rispetto ai lettori abituali (*tra 6 e 10*) e costanti (*tra 11 e 20*), oltre che dai lettori forti (*più di 21 libri l'anno*).

Tab. 12 - Distribuzione di frequenza del numero di libri letti negli ultimi 12 mesi dai rispondenti (n, AVQ 2020) e opzioni di aggregazione*

AVQ	n	%
Nessuno	23081	57,1%
1	1782	4,4%
2	3101	7,7%
3	2724	6,7%
4	1724	4,3%
5	1602	4,0%
6	1101	2,7%
7	399	1,0%
8	572	1,4%
9	119	0,3%
10	1446	3,6%
11	60	0,1%
12	517	1,3%
13	59	0,1%
14	51	0,1%
15	450	1,1%
16	45	0,1%
17	22	0,1%
18	42	0,1%
19	12	0,0%
20	560	1,4%
21-25	221	0,5%
26-30	247	0,6%
31-40	170	0,4%
41-50	189	0,5%
51 e più	153	0,4%
Totale*	**40449**	**100,0%**

Opzione A	n	%
Nessuno	23081	57,1%
1-10	14570	36,0%
11-20	3264	8,1%
21-30	468	1,2%
31-40	170	0,4%
41-50	189	0,5%
51 e più	153	0,4%
Totale*	**40449**	**100,0%**

Opzione B	n	%
Nessuno	23081	57,1%
Meno di 5	9331	23,1%
Tra 6 e 10	5239	13,0%
Tra 11 e 20	1818	4,5%
21 e più	1540	3,8%
Totale*	**40449**	**100,0%**

** Dati mancanti non dovuti: 613(1,4%); dati mancanti dovuti (età inferiore ai 5 anni): 1748 (4,1%)*

L'aggregazione delle categorie o la costruzione di classi di valori sono utili anche al fine del confronto tra diverse distribuzioni di frequenza, anche con dati secondari, nel caso le definizioni operative prevedano articolazioni diverse ma siano riconducibili a categorie comuni. Ad esempio, rispetto alle abitudini di lettura l'ISTAT, nei propri report[13], evidenzia soprattutto due categorie: quella complessiva dei lettori, che hanno letto almeno un libro negli ultimi 12 mesi, e quella dei "lettori forti" che hanno letto più di 12 libri negli ultimi 12 mesi. Immaginando di voler confrontare dati primari sulle abitudini di lettura con le serie storiche ISTAT sarebbe quindi utile adottare una definizione operativa che

[13] Produzione e lettura di libri – 2020; reperibile da: https://www.istat.it/it/files//2022/02/REPORT_PRODUZIONE_E_LETTURA_LIBRI_2020.pdf. I dati provengono dall'indagine Aspetti della vita quotidiana.

preveda – o permetta la costruzione ex post, come evidenziato fin qui – di queste categorie.
Il confronto tra variabili può dunque essere semplificato dalla riconduzione a categorie comuni o che prevedano la stessa articolazione.
Nel caso di variabili cardinali, tuttavia, la riconduzione in classi potrebbe comportare una eccessiva perdita di informazioni e può essere opportuno prevedere, invece, una *trasformazione*.
Il caso più evidente in cui si rende necessaria una trasformazione è relativo al confronto tra distribuzioni di variabili con grandezze differenti, ad esempio i voti ottenuti all'esame di maturità e quelli di laurea.
Intuitivamente per rendere comparabili le due distribuzioni è necessario ricondurle a uno stesso campo di variazione: trasformarle cioè in modo tale che presentino entrambe lo stesso valore minimo e lo stesso valore massimo. Una procedura comune prevede di sottrarre a ciascuno dei valori della distribuzione il valore minimo teorico e di rapportare la differenza ottenuta alla differenza tra il massimo e il minimo teorico:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Questa operazione è detta *normalizzazione*, permette di ottenere un numero puro che varia tra 0 e 1 e non risente dell'unità di misura originale della variabile.
Si è già vista la sua utilità – sia in relazione all'indice di omogeneità che in relazione al coefficiente di variazione – per confrontare parametri relativi a distribuzioni diverse, qui la normalizzazione si applica a tutti i valori della distribuzione, trasformando ciascuno di essi per ricondurlo al campo di variazione [0 ; 1].
Un'altra trasformazione utile al confronto di variabili cardinali è la *standardizzazione*, che mira a eliminare le differenze di scala e di dispersione tra le variabili.
Il calcolo della variabile standardizzata prevede che si calcoli per ciascun valore lo scarto dalla media ($\bar{X}$), trasformando quindi la distribuzione in una distribuzione delle differenze dalla media. Il valore della differenza dalla media si rapporta poi alla deviazione standard, in modo tale che la

distribuzione della variabile standardizzata (generalmente indicata con z) abbia media 0 e deviazione standard 1:

$$z_i = \frac{(x_i - \bar{X})}{\sigma}$$

Né la normalizzazione né la standardizzazione modificano in alcun modo le frequenze nella distribuzione: hanno effetto sul solo valore della variabile.

Si vedrà di seguito che queste trasformazioni sono utili nell'analisi bi e multivariata, ancora più che nel confronto tra distribuzioni monovariate.