Lezione 2
L’unità statistica e le variabili statistiche

PROF. ROBERTO COSTA

SCIENZE DELL’EDUCAZIONE - STATISTICA SOCIALE (305SF)

Dove eravamo rimasti

Nella scorsa lezione abbiamo parlato di cos'è la statistica e perché è importante conoscerla.

Abbiamo visto le fasi di una ricerca quantitativa (progettazione, rilevazione, analisi, diffusione).

Abbiamo approfondito il concetto di dato.

C'era anche un piccolo lavoro da fare, ci siete riusciti? Avete trovato i metadati?

Dove eravamo rimasti

Ecco le definizioni:

OCCUPATI: Persone tra 15 e 89 anni che nella settimana di riferimento:

- hanno svolto almeno un'ora di lavoro a fini di retribuzione o di profitto, compresi i coadiuvanti familiari non retribuiti;
- sono temporaneamente assenti dal lavoro perché in ferie, con orario flessibile (part time verticale, recupero ore, etc.), in malattia, in maternità/paternità obbligatoria, in formazione professionale retribuita dal datore di lavoro;
- sono in congedo parentale e ricevono e/o hanno diritto a un reddito o a prestazioni legate al lavoro, indipendentemente dalla durata dell'assenza;
- sono assenti in quanto lavoratori stagionali ma continuano a svolgere regolarmente mansioni e compiti necessari al proseguimento dell'attività (da tali mansioni e compiti va escluso l'adempimento di obblighi legali o amministrativi);
- sono temporaneamente assenti per altri motivi e la durata prevista dell'assenza è pari o inferiore a tre mesi.

DISOCCUPATI: Persone non occupate tra i 15 e i 74 anni che:

- hanno effettuato almeno un'azione attiva di ricerca di lavoro nelle quattro settimane che precedono la settimana a cui le informazioni sono riferite e sono disponibili a lavorare (o ad avviare un'attività autonoma) entro le due settimane successive;

oppure

- inizieranno un lavoro entro tre mesi dalla settimana a cui le informazioni sono riferite e sarebbero disponibili a lavorare (o ad avviare un'attività autonoma) entro le due settimane successive, qualora fosse possibile anticipare l'inizio del lavoro.

Dove eravamo rimasti

INATTIVI: Persone che non fanno parte delle forze di lavoro, cioè quelle non classificate come occupate o in cerca di occupazione (disoccupate). Rientrano nella categoria coloro che:

- non hanno cercato lavoro nelle ultime quattro settimane e non sono disponibili a lavorare entro due settimane dall'intervista;
- pur non avendo cercato un lavoro nelle ultime quattro settimane si sono dichiarati disponibili a iniziare un lavoro entro due settimane dall'intervista;
- hanno cercato un lavoro nelle ultime quattro settimane, ma che non sono disponibili a iniziare un lavoro entro due settimane dall'intervista.

***Script QB01** (obbligatorio da leggere)*
Le prossime domande riguardano l'attività lavorativa svolta "LA SCORSA SETTIMANA", cioè la settimana che va "DA LUNEDI' … A DOMENICA…". Consideri qualsiasi attività lavorativa retribuita, in proprio o alle dipendenze, con o senza contratto.

QB01 "LA SCORSA SETTIMANA" [NOME] ha svolto almeno un'ora di lavoro da cui ha ricavato o ricaverà un guadagno?
Se necessario leggere: Deve escludere le ore di lavoro svolte al fine di percepire il reddito di cittadinanza o la frequenza di un corso di dottorato di ricerca nel caso siano le uniche ore di lavoro svolte.
ATTENZIONE: NON va considerato il lavoro NON RETRIBUITO svolto presso la ditta di un familiare.

Sì 1|_|
No 2|_|

QB02 Sempre nella settimana che va "DA LUNEDI' … A DOMENICA…", [NOME] ha lavorato almeno un'ora, senza essere pagata/o, presso la ditta di un familiare?

Sì 1 |_|
No 2 |_|

QB03 Sempre in quella settimana, [NOME] aveva comunque un lavoro che non ha svolto, ad esempio per un periodo limitato di ridotta attività, per malattia, per vacanza, per un corso di formazione legato al suo lavoro, per cassa integrazione guadagni, etc.?
Se necessario leggere: consideri il lavoro da cui ha ricavato o ricaverà un guadagno o il lavoro non pagato solo se effettuato abitualmente presso la ditta di un familiare.

Sì 1 |_|
No 2 |_|

QB04 E invece, nelle quattro settimane che vanno "DA LUNEDI' 4 settimane precedenti … A DOMENICA…", "NOME", ha svolto lavori occasionali o lavoretti guadagnando qualcosa, ad esempio come baby-sitter, cameriere, ripetizioni, consegne, etc.?
Se necessario leggere: Deve escludere le ore di lavoro svolte al fine di percepire il reddito di cittadinanza nel caso siano le uniche ore di lavoro svolte.

Sì (escluse eventuali ore di lavoro per percepire il reddito di cittadinanza) 1 |_|
No, ho lavorato solo poche ore per percepire il reddito di cittadinanza 2 |_|
No, nessuna ora di lavoro 3 |_|

QB05 [NOME] ha svolto almeno un'ora di questi lavori nella settimana che va "DA LUNEDI' … A DOMENICA…"?

QB06 Qual è il motivo principale per cui non ha lavorato in quella settimana?

Ferie e festività 1|_|
Regimi di orario flessibile (incluso part time verticale) o compensazione delle ore di lavoro straordinario 2|_|
Malattia, problemi di salute, infortunio 3|_|
Congedo di maternità obbligatorio o congedo di paternità 4|_|
Congedo parentale, ossia assenza facoltativa fino al dodicesimo anno del bambino 5|_|
Altri motivi familiari (esclusa maternità obbligatoria e congedo parentale) 6|_|
Formazione professionale direttamente collegata al lavoro oppure retribuita dal datore di lavoro(incluso il dottorato retribuito) 7|_|
Formazione professionale non direttamente collegata al lavoro oppure non retribuita dal datore di lavoro 8|_|
Cassa Integrazione Guadagni (CIG ordinaria o straordinaria) 9|_|
Ridotta attività/mancanza lavoro (esclusa CIG) 10|_|
Lavoro stagionale (ad es. bagnino, raccoglitore di frutta, cameriere in montagna d'inverno etc.) 11|_|
Ha un lavoro che non ha ancora iniziato 12|_|
Fa un lavoro occasionale 994|_|
*Altro motivo (specificare) (**QB06_a**)_____* 996|_|

QB07 In quella settimana era assente o proprio non aveva impegni di lavoro?

Aveva un lavoro ma era assente 1|_|
Non aveva un lavoro 2|_|

QB08 Durante il periodo di assenza continua a percepire una remunerazione (congedo parentale retribuito)?
Se necessario leggere: consideri anche il solo versamento dei contributi previdenziali.

Sì il 50% o più della retribuzione 1 |_|
Sì meno del 50% della retribuzione 2 |_|
No, non retribuito, solo contributi 3 |_|
No, non retribuito 4 |_|

QB09 Questo periodo di assenza dal lavoro durerà meno o più di tre mesi, da quando è iniziato a quando terminerà?
Consideri solo il periodo di assenza facoltativa escludendo l'assenza obbligatoria per la nascita del bambino (se l'assenza facoltativa avviene subito dopo il periodo di assenza obbligatoria)

Meno di tre mesi 1|_|
Tre mesi o più 2|_|

Dove eravamo rimasti

Confesso che la scorsa lezione ho fatto un piccolo «esperimento» con voi per testare la reazione a un dato palesemente anomalo. E' plausibile che in un anno in Italia muoia un milione di individui in più rispetto all'anno precedente?

In Italia muoiono circa 650mila persone all'anno, pertanto sarebbe davvero anomalo che, durante la pandemia fosse deceduto un milione di persone in più rispetto alla media del periodo 2015-2019.

Vediamo i dati:

Decessi per anno e mese													
	gennaio	febbraio	marzo	aprile	maggio	giugno	luglio	agosto	settembre	ottobre	novembre	dicembre	TOTALE
media 2015-2019	68.324	57.416	58.267	51.801	50.724	48.501	51.811	51.041	46.548	51.590	51.462	58.133	645.620
2020	62.019	56.070	86.501	72.809	52.440	48.589	51.422	53.744	49.326	59.861	78.470	74.895	746.146
2021	74.550	59.389	68.507	63.434	54.802	52.201	53.668	56.594	51.456	54.463	54.870	65.101	709.035
Fonte: Istat, Tavola decessi totali regionali mensili per la media degli anni 2015-2019, per gli anni 2020-2021 e per i mesi di gennaio-agosto 2022													

L’unità statistica

L’unità statistica è l’unità elementare su cui vengono rilevate le variabili oggetto delle analisi statistiche.

Nella ricerca sociale l’unità statistica è solitamente l’essere umano, del quale voglio studiare i comportamenti, le abitudini culturali, le relazioni interpersonali, lo status socio economico, ecc.

L’individuo non è l’unica unità statistica, possono costituire un’unità statistica le famiglie, le classi scolastiche, le aziende, le organizzazioni non profit, ecc.

Dobbiamo distinguere tra:

Unità di raccolta (o di rilevamento): presso le quali vengono raccolte le informazioni.

Unità di analisi (o di riferimento): a cui si riferiscono le informazioni raccolte.

Ad es. intervisto il responsabile del personale (unità di raccolta) per acquisire informazioni sul benessere degli addetti di un’impresa (unità di analisi).

La popolazione

La popolazione (o collettivo) è un insieme di unità omogenee rispetto a una o più caratteristiche.

Nella ricerca sociale le popolazioni devono essere contestualizzate nello spazio e nel tempo.

A questo proposito si distinguono:

Popolazioni o collettivi di stato: definibili un unico istante di tempo (ad es. residenti in un comune al 1.1.2022);

Popolazioni o collettivi di movimento: definibili in un intervallo di tempo (nati in un comune nel 2022, i laureati dell'Università di Trieste nell'anno accademico 2022-2023).

La popolazione

Si distingue inoltre tra:

La **popolazione empirica**: se tutte le unità che la compongono possono effettivamente fare parte della ricerca (immatricolati al primo anno del corso di studi di scienze dell'educazione nell'a. a. 2022-23).

La **popolazione teorica**: se alcune unità non possono essere osservate (persone entrate clandestinamente in Italia nel 2021).

Questo impatta ad esempio sulla costruzione del campione casuale rappresentativo della popolazione.

Tipo di rilevazione

Una volta definita la popolazione e le sue unità, il ricercatore decide se la raccolta dei dati debba riguardare tutta la popolazione o solo una parte di essa.

Si parla nel primo caso di **rilevazione totale o censuaria** e nel secondo caso di **rilevazione parziale o campionaria**.

Il contenimento dei tempi e dei costi di una ricerca sono tra i motivi principali per optare per una rilevazione campionaria.

Ci sono anche altri motivi, come ad esempio nei test di qualità di certi prodotti, laddove il test determina l'inutilizzabilità del prodotto (ad es. analisi della durata di una lampadina).

La numerosità campionaria dev'essere congrua alle analisi sui dati che si intende fare.

Ad esempio devo garantire una adeguata numerosità a eventuali sottogruppi della popolazione che voglio analizzare (per genere, territorio, età, ecc.).

La rilevazione parziale o campionaria

Il campione deve riprodurre in scala le caratteristiche della popolazione da cui viene estratto.

Le unità che comporranno il campione devono essere estratte in modo casuale, ovvero nel rispetto delle leggi della probabilità (**campione probabilistico**).

In questo caso potremo generalizzare i risultati della rilevazione all'intera popolazione di riferimento.

Come acquisire i dati

Definiti gli obiettivi e la popolazione di riferimento, bisogna decidere come si vogliono raccogliere i dati.

Nella ricerca sociale quantitativa lo strumento più utilizzato è il **questionario**, formato da diverse domande predefinite, che seguono un preciso ordine.

In base ai contenuti del questionario si distinguono questionari **strutturati**, **semistrutturati** e **non strutturati**.

Abbiamo già visto che una possibile alternativa è il ricorso a dati secondari, ovvero già raccolti, seppur con finalità diverse, in altre circostanze (ad es. dati di fonte amministrativa, big data, ecc.).

Come acquisire i dati

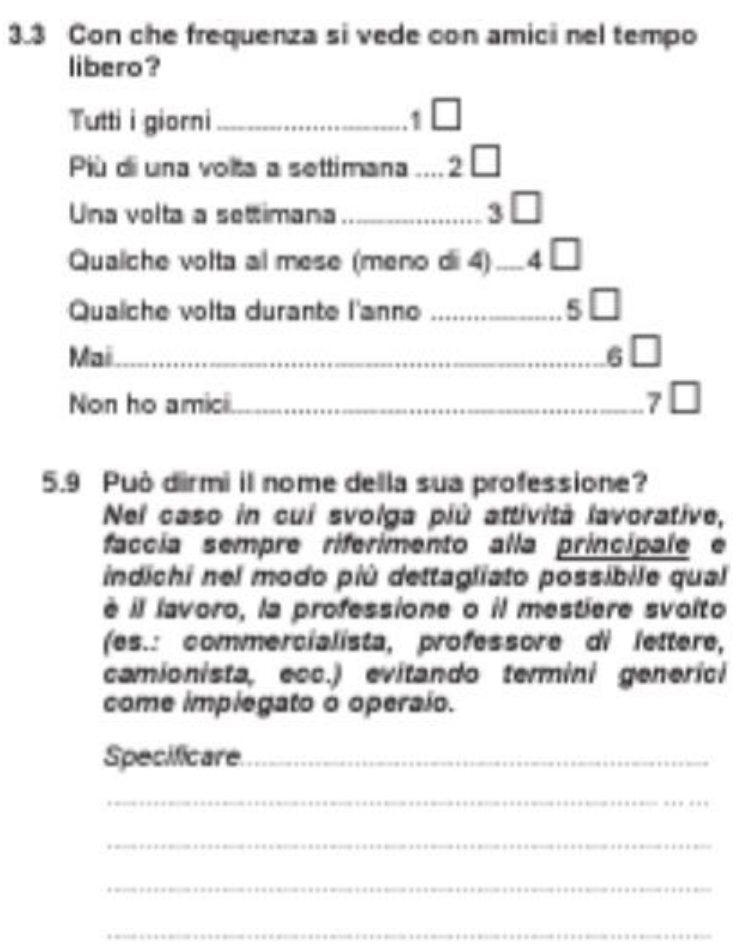
Il **questionario strutturato** prevede l'uso esclusivo o prevalente di risposte fisse precodificate ed è particolarmente idoneo nel caso di indagini su campionature molto ampie.

Il **questionario semistrutturato** viene formulato in modo da lasciare una certa libertà alle risposte dell'intervistato, seppure all'interno di griglie predefinite dal ricercatore.

Il **questionario non strutturato** fa invece ricorso a domande aperte, permettendo all'intervistato di dare una risposta il più possibile personale e non condizionata dall'ottica del ricercatore.

Il questionario non strutturato viene utilizzato soprattutto se l'obiettivo è soprattutto esplorativo, mentre con il questionario strutturato l'obiettivo è soprattutto la misurazione.

Esempi di domande



3.3 Con che frequenza si vede con amici nel tempo libero?

- Tutti i giorni ..... 1 ☐
- Più di una volta a settimana .... 2 ☐
- Una volta a settimana ..... 3 ☐
- Qualche volta al mese (meno di 4) .... 4 ☐
- Qualche volta durante l'anno ..... 5 ☐
- Mai ..... 6 ☐
- Non ho amici ..... 7 ☐

5.9 Può dirmi il nome della sua professione?
Nel caso in cui svolga più attività lavorative, faccia sempre riferimento alla principale e indichi nel modo più dettagliato possibile qual è il lavoro, la professione o il mestiere svolto (es.: commercialista, professore di lettere, camionista, ecc.) evitando termini generici come impiegato o operaio.

Specificare .....

5.4 Qual è il motivo principale per cui non ha lavorato da lunedì a domenica della scorsa settimana ?

- Cassa Integrazione Guadagni (ordinaria o straordinaria) ..... 01 ☐
- Ridotta attività dell'impresa per motivi economici e/o tecnici (esclusa CIG) ..... 02 ☐
- Sciopero ..... 03 ☐
- Vertenza sindacale, controversia di lavoro ..... 04 ☐
- Maltempo ..... 05 ☐
- Malattia, problemi di salute personali, infortunio ..... 06 ☐
- Ferie ..... 07 ☐
- Festività nella settimana ..... 08 ☐
- Orario variabile o flessibile (ad es. riposo compensativo) ..... 09 ☐
- Part time verticale ..... 10 ☐
- Studio o formazione non organizzata nell'ambito del proprio lavoro ..... 11 ☐
- Assenza obbligatoria per maternità ..... 12 ☐
- Assenza facoltativa fino all'ottavo anno del bambino (congedo parentale) ..... 13 ☐
- Motivi familiari (esclusa maternità obbligatoria e congedo parentale) ..... 14 ☐
- Mancanza di occasioni di maggior lavoro ..... 15 ☐
- Fa un lavoro occasionale ..... 16 ☐
- Fa un lavoro stagionale alle dipendenze ..... 17 ☐
- Altro motivo ..... 18 ☐

(specificare)

Qui possiamo vedere:

- ✓Una domanda chiusa (3.3), con le modalità di risposta già codificate.
- ✓Una domanda aperta (5.9), dove l'intervistato può rispondere liberamente.
- ✓Una domanda chiusa, con una modalità residua (5.4), dove si possono raccogliere ulteriori modalità di risposta.

Le variabili statistiche

Ogni unità presenta delle **caratteristiche** che vogliamo rilevare (ad es. il genere, l'età, l'occupazione, l'orientamento politico, ecc.).

Ogni caratteristica può assumere valori differenti tra le varie unità. In questo caso la caratteristica viene chiamata **variabile** (o anche **carattere**).

I valori che può assumere una variabile vengono chiamati **modalità**.

Le variabili statistiche

Le variabili si distinguono a seconda di come vengono espresse le modalità.

Se le modalità sono espresse in numeri cardinali, si chiamano **variabili quantitative** (o cardinali).

Altrimenti si chiamano **variabili qualitative** (o mutabili).

Ad esempio l'età è una variabile quantitativa, il comune di residenza è una variabile qualitativa.

Le variabili qualitative

Le variabili qualitative, a loro volta, si distinguono tra:

Mutabili rettilinee, che sono ordinabili secondo un determinato criterio (ad es. titolo di studio: primario, secondario, terziario).

Mutabili sconnesse, per le quali un ordinamento non ha motivi per essere preferibile ad un altro (ad es. settore di attività: agricoltura, industria, servizi).

Questa distinzione determina le operazioni statistiche che potremo fare su queste variabili.

La definizione operativa per creare questo tipo di variabili si chiama **classificazione**, ovvero raggruppare le unità in classi in base a una determinata caratteristica. Il ricercatore deve prevedere tutte le possibili classi (classificazione esaustiva).

Mutabili sconnesse

Proviamo a fare qualche esempio di **mutabili sconnesse**:

Genere, Corso di studio, Comune di nascita, stato civile, religione, nazionalità, ecc.

Tra le modalità di queste variabili possiamo solo verificare uguaglianza o diversità.

Mutabili rettilinee

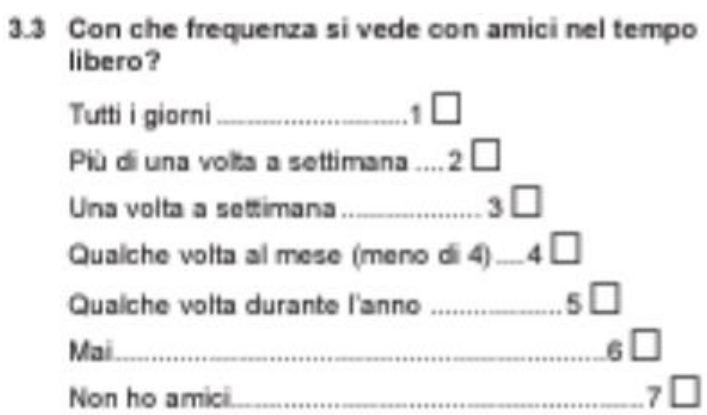
Proviamo a fare qualche esempio di **mutabili rettilinee**:

Titolo di studio, grado gerarchico militare, categoria sportiva, categorie contrattuali, ecc.

Tra le modalità di queste variabili, oltre a verificare uguaglianza o diversità, possiamo anche utilizzare un ordinamento e definire delle relazioni *maggiore di* o *minore di*.

La laurea è un titolo più elevato del diploma di scuola secondaria di secondo grado, il generale è un grado gerarchico più elevato del sergente, ecc.

Rivediamo gli esempi



3.3 Con che frequenza si vede con amici nel tempo libero?

- Tutti i giorni ..... 1 ☐
- Più di una volta a settimana .... 2 ☐
- Una volta a settimana ..... 3 ☐
- Qualche volta al mese (meno di 4) .... 4 ☐
- Qualche volta durante l'anno ..... 5 ☐
- Mai ..... 6 ☐
- Non ho amici ..... 7 ☐

5.4 Qual è il motivo principale per cui non ha lavorato da lunedì a domenica della scorsa settimana ?

- Cassa Integrazione Guadagni (ordinaria o straordinaria) ..... 01 ☐
- Ridotta attività dell'impresa per motivi economici e/o tecnici (esclusa CIG) ..... 02 ☐
- Sciopero ..... 03 ☐
- Vertenza sindacale, controversia di lavoro ..... 04 ☐
- Maltempo ..... 05 ☐
- Malattia, problemi di salute personali, infortunio ..... 06 ☐
- Ferie ..... 07 ☐
- Festività nella settimana ..... 08 ☐
- Orario variabile o flessibile (ad es. riposo compensativo) ..... 09 ☐
- Part time verticale ..... 10 ☐
- Studio o formazione non organizzata nell'ambito del proprio lavoro ..... 11 ☐
- Assenza obbligatoria per maternità ..... 12 ☐
- Assenza facoltativa fino all'ottavo anno del bambino (congedo parentale) ..... 13 ☐
- Motivi familiari (esclusa maternità obbligatoria e congedo parentale) ..... 14 ☐
- Mancanza di occasioni di maggior lavoro ..... 15 ☐
- Fa un lavoro occasionale ..... 16 ☐
- Fa un lavoro stagionale alle dipendenze ..... 17 ☐
- Altro motivo ..... 18 ☐

(specificare)

Qui possiamo vedere:

- ✓Una mutabile rettilinea (3.3), con le modalità di risposta che indicano una frequenza di incontri con gli amici calante.
- ✓Una mutabile sconnessa (5.4), dove non c'è un criterio che possa fornire un ordinamento.

Le variabili quantitative

Le **variabili quantitative** sono quelle le cui modalità vengono espresse con numeri cardinali.

Tra le modalità di queste variabili sono effettuabili le relazioni *maggiore di* o *minore di*, ma anche tutte le operazioni aritmetiche.

Sono ad esempio variabili quantitative: l'età, il numero di figli, la statura, il peso, il numero di esami sostenuti, ecc.

Le variabili quantitative

Le variabili quantitative si distinguono ulteriormente tra **variabili discrete** e **variabili continue**.

Nel caso di variabili discrete, le modalità sono espresse da numeri interi e tra una modalità e un'altra non ci sono stati intermedi. Ad esempio il numero di figli può essere 0, 1, 2,… ma non 1,3!

Si tratta quindi di variabili quantitative a proprietà discrete, esito di un **conteggio**.

Le variabili quantitative

Nel caso di variabili continue, le modalità sono espresse da numeri reali e le modalità possono essere in numero pressoché infinito, di norma aggregato in classi.

Si tratta quindi di variabili quantitative a proprietà continue, esito di una **misurazione**.

Le variabili quantitative

Tra le variabili quantitative distinguiamo anche tra **scale a intervalli** e **scale di rapporti**.

La scala a intervalli non ha uno zero assoluto che corrisponde ad una mancanza totale di proprietà.

La scala di rapporti ha invece uno zero non arbitrario che corrisponde alla mancanza di proprietà.

Le variabili quantitative

Un esempio tipico di scala a intervalli è la temperatura, dove si usano diverse scale (ad es. Celsius o Farenheit).

Lo zero in gradi Celsius corrisponde alla temperatura alla quale gela l'acqua.

Non è quindi un'assenza di temperatura!

Le differenze di grandezza possono quindi essere fatte solo in termini di distanze.

Ad esempio se in due città misuro una temperatura rispettivamente di 10° e 20° Celsius, non è corretto dire che in una città la temperatura è doppia dell'altra.

Se usassi la scala Farenheit, le temperature sarebbero di 50 e 68 gradi.

Altri esempi di scale a intervalli sono: l'anno di nascita, i voti (di un esame, di diploma di laurea, ecc.), latitudine e longitudine,…

Le variabili quantitative

Nelle scale di rapporti, siccome il punto zero è naturale si possono confrontare due modalità mettendole a rapporto.

Se un individuo ha due figli e un altro ne ha uno, posso affermare che il primo ha il doppio dei figli dell'altro.

Se un operaio ha uno stipendio netto mensile di 1500 euro e un dirigente ne ha uno di 4.500 euro, posso dire che il quest'ultimo percepisce uno stipendio pari a tre volte quello dell'operaio.

Esempi di scale di rapporti sono: età, reddito, statura, peso, spese per consumi, ammontare dei depositi sul C/C,…

Sentirete parlare anche di…

Si distinguono le variabili qualitative in base al numero di modalità che possono assumere:

Variabili dicotomiche (con due modalità di risposta come: si o no, presente o assente, ecc.).

Variabili politomiche (che possono assumere più di due modalità di risposta come corso di laurea, titolo di studio, ecc.).

Possiamo distinguere anche le variabili in base al ruolo assunto nella spiegazione dei fenomeni sociali. Possiamo ad esempi assumere una dipendenza tra due variabili, individuandone una come causa e l'altra come effetto.

Variabili indipendenti: la variabile individuata come causa (ad es. titolo di studio).

Variabili dipendenti. la variabile individuata come effetto (ad es. reddito mensile netto).

Variabili e analisi statistica

Per le finalità di questo corso ci concentreremo, parlando di analisi dei dati sulla seguente classificazione delle variabili:

- ✓Variabili qualitative sconnesse (con categorie non ordinate)
- ✓Variabili qualitative rettilinee (con categorie ordinate)
- ✓Variabili quantitative discrete
- ✓Variabili quantitative continue

Rappresentazione dei dati

La matrice dei dati, soprattutto se composta da tante variabili raccolte su tante unità statistiche, non è certo lo strumento più efficace per sintetizzare i risultati.

La forma più semplice per rappresentare una variabile è la distribuzione semplice.

Adesso facciamo un esempio pratico.

Rappresentazione dei dati

Supponiamo di aver raccolto i risultati degli studenti che hanno sostenuto l'esame di statistica sociale nell'anno accademico 2021-22.

Questi sono i risultati:

30, 27, 22, 24, 21, 19, 26, 18, 28, 21, 24, 22, 30, 28, 18, 19, 23, 26, 29, 27, 20, 30, 27, 26, 30, 30, 26, 24, 28, 27.

Rappresentazione dei dati

La prima cosa che possiamo fare è la **distribuzione delle frequenze assolute** della variabile **voto**.

In questo modo rappresento il numero di volte che si è presentata una modalità.

Voto	Frequenza assoluta
18	2
19	2
20	1
21	2
22	2
23	1
24	3
26	4
27	4
28	3
29	1
30	5
Totale	30

Rappresentazione dei dati

Generalizzando, se in un insieme di N casi una variabile X assume k modalità (x_1, x_2, …, x_j, …, x_k) di cui n_1, n_2, …, n_j, …, n_k sono le rispettive frequenze assolute, la tabella seguente rappresenta la distribuzione delle frequenze.

Modalità della variabile X	Frequenza assoluta
x_1	n_1
x_2	n_2
…	…
x_j	n_j
…	…
x_k	n_k
Totale	N

Rappresentazione dei dati

Spesso capita di dover mettere a confronto distribuzioni di frequenza della stessa variabile, risultato di due diverse rilevazioni, ad esempio la variabile voto nell'esame di statistica sociale nell'a.a. 2020-21 e 2021-22.

La tabella così com'è non è di facilissima interpretazione.

In questo caso è più utile la **distribuzione delle frequenze relative e percentuali** della variabile **voto**.

Voto	2021/22	2020/21
18	2	2
19	2	4
20	1	2
21	2	3
22	2	3
23	1	5
24	3	4
25	0	3
26	4	6
27	4	4
28	3	5
29	1	3
30	5	6
Totale	30	50

Rappresentazione dei dati

La distribuzione delle **frequenze relative** si ottiene dividendo la frequenza assoluta delle modalità per il numero totale dei casi.

La **frequenza percentuale** si ottiene moltiplicando per 100 la frequenza relativa.

Modalità della variabile X	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
x_1	n_1	n_1/N	n_1/N X 100
x_2	n_2	n_2/N	n_2/N X 100
…	…	…	…
x_j	n_j	n_j/N	n_j/N X 100
…	…	…	…
x_k	n_k	n_k/N	n_k/N X 100
Totale	N	1	100

Voto 2021/22	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
18	2	0,06667	6,666667
19	2	0,06667	6,666667
20	1	0,03333	3,333333
21	2	0,06667	6,666667
22	2	0,06667	6,666667
23	1	0,03333	3,333333
24	3	0,1	10
25	0	0	0
26	4	0,13333	13,33333
27	4	0,13333	13,33333
28	3	0,1	10
29	1	0,03333	3,333333
30	5	0,16667	16,66667
Totale	30	1	100

Voto 2020/21	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
18	2	0,04	4
19	4	0,08	8
20	2	0,04	4
21	3	0,06	6
22	3	0,06	6
23	5	0,1	10
24	4	0,08	8
25	3	0,06	6
26	6	0,12	12
27	4	0,08	8
28	5	0,1	10
29	3	0,06	6
30	6	0,12	12
Totale	50	1	100

La frequenza cumulata

Nel caso di variabili qualitative rettilinee o variabili quantitative per il ricercatore può essere utile conoscere con che frequenza si presentano le modalità di ordine inferiore o superiore a una specifica modalità.

Ad esempio, quanti studenti hanno preso più di un determinato voto.

Si utilizza la frequenza cumulata, ovvero il numero di casi che hanno un valore uguale o inferiore a quello della categoria di riferimento.

Ripetiamo che questo non si può utilizzare con le variabili qualitative sconnesse.

Distribuzioni cumulate

Guardando il nostro esempio, il 20%, ovvero 5 studenti su 30 hanno preso un voto inferiore o uguale a 20.

Per calcolare le frequenze cumulate bisogna iniziare dalla frequenza associata alla modalità più bassa e man mano sommando le frequenze delle modalità successive.

Generalizzando:

Frequenze assolute $N_j = n_1 + n_2 + \dots n_{j-1} + n_j$

Frequenze relative $F_j = f_1 + f_2 + \dots f_{j-1} + f_j$

Frequenze percentuali $P_j = p_1 + p_2 + \dots p_{j-1} + p_j$

Voto 2021/22	Frequenza cumulata	Frequenza cumulata relativa	Frequenza cumulata percentuale
18	2	0,07	6,7
19	4	0,13	13,3
20	5	0,17	16,7
21	7	0,23	23,3
22	9	0,30	30,0
23	10	0,33	33,3
24	13	0,43	43,3
25	13	0,43	43,3
26	17	0,57	56,7
27	21	0,70	70,0
28	24	0,80	80,0
29	25	0,83	83,3
30	30	1	100
Totale	30	1	100

Presentazione efficace di una tabella di frequenze relative e percentuali

- Compattare le informazioni. Fornire una rappresentazione sintetica di una variabile, senza rinunciare alle principali informazioni.
- Arrotondamento. Scegliere quante cifre decimali si vogliono presentare. I decimali devono avere un'importanza sostanziale.
- Quadratura. Per effetto dell'arrotondamento può accadere che il totale delle frequenze relative o percentuali non sia rispettivamente 1 o 100. Solitamente si aggiusta il numero più elevato in modo da sistemare il risultato.

Regole per la ricodifica delle variabili

- Approccio «semantico» legato agli scopi dell'analisi. Vale soprattutto per le variabili qualitative sconnesse.
- Approccio «numerico», legato alle frequenze di ogni nuova categoria.
- Per le variabili quantitative l'aggregazione in classi in sostanza significa decidere dove tagliare il «continuum» dei valori.

In generale è buona norma:

- Evitare di costruire classi con frequenze molto basse;
- Modulare l'ampiezza delle classi in funzione dei dati.

Regole per la ricodifica delle variabili

- Il numero di classi dev’essere abbastanza piccolo da fornire un’adeguate sintesi, ma abbastanza grande da garantire un buon livello di informazione.
- Ogni valore deve poter essere assegnato a una e una sola classe (classi disgiunte).
- Tutti i valori devono essere compresi nelle classi..

Distribuzioni aggregate

Quando si crea una distribuzione di frequenza si può decidere di ridurre il numero di modalità, rispetto a quelle della variabile analizzata.

Per le variabili qualitative si parla di **ricodifica**. Ad esempio se ho il Comune di nascita disporrei di circa 8.000 modalità di risposta, che posso aggregare per provincia, o numero di abitanti.

Per le variabili quantitative si parla di **aggregazione in classi**. Ad esempio posso aggregare l'età in classi d'età quinquennali. La scelta dell'ampiezza delle classi dipende dagli scopi della ricerca, ma deve mantenere in equilibrio da un lato la perdita di informazioni e dall'altra un numero eccessivo di categorie.

Per le variabili quantitative continue bisogna fare molta attenzione a non sovrapporre le classi.

Distribuzioni aggregate

Per la variabile aggregata ogni unità deve trovare collocazione in una e una sola modalità.

Bisogna evitare la sovrapposizione delle classi.

Ad esempio: in base al peso registrato, aggrego delle unità nelle seguenti classi:

Fino a 45 kg

Da 45 a 60 kg

Da 60 a 75 kg

Da 75 a 90 kg

Più di 90 kg

Per alcuni casi ci potrebbero essere dei dubbi.

Distribuzioni aggregate

Bisogna chiarire inequivocabilmente quali sono gli estremi delle classi. Si usano i seguenti simboli:

|— indica una classe chiusa a sinistra, che include il valore estremo di sinistra, ma non quello di destra.

—| indica una classe chiusa a destra, che include il valore estremo di destra, ma non quello di sinistra.

|—| classe chiusa, che include sia il valore estremo a destra che a sinistra.

Riscriviamo:

0 |— 45

45 |— 60

60 |— 75

75 |—|90

Più di 90 kg

Una proposta per leggere meglio i dati

Voto	Frequenza assoluta	Frequenza relativa	Frequenza percentuale	Frequenza cumulata
18 - 20	5	0,17	16,7	16,7
21 - 23	5	0,17	16,7	33,3
24 - 26	7	0,23	23,3	56,7
27 - 30	13	0,43	43,3	100,0

A. A. 2021/2022

Voto	Frequenza assoluta	Frequenza relativa	Frequenza percentuale	Frequenza cumulata
18 - 20	8	0,16	16,0	16,0
21 - 23	11	0,22	22,0	38,0
24 - 26	13	0,26	26,0	64,0
27 - 30	18	0,36	36,0	100,0

A. A. 2020/2021

Esercitiamoci assieme…

Un piccolo imprenditore svolge un'indagine sui propri collaboratori per comprendere se si trovano bene, se le retribuzioni sono considerate adeguate e se in questo periodo di particolare complessità qualcuno di essi si trova in difficoltà.

Chiede al responsabile del personale di raccogliere i dati e alla fine riceve una matrice dei dati, dalla quale abbiamo estratto qualche cella.

E' tutto chiaro?

Che tipi di variabili troviamo?

Sulla base dell'analisi univariata possiamo fare e a quali conclusioni possiamo arrivare?

Età	Titolo di studio	Ruolo	Reddito netto mensile medio	Comp. del nucleo fam.	Soddisfazione per la sit. econ.
46	Diploma	Operaio	3400	4	Poco
21	Diploma	Operaio	1600	2	Poco
48	Laurea	Dirigente	6200	3	Molto
37	Laurea	Dirigente	4300	3	Abbastanza
29	Diploma	Operaio	2400	2	Per niente
55	Licenza media	Operaio	2700	4	Per niente
41	Diploma	Impiegato	2800	1	Abbastanza
46	Diploma	Operaio	2100	3	Poco
63	Diploma	Impiegato	2600	2	Abbastanza
35	Diploma	Impiegato	2400	1	Abbastanza
51	Licenza media	Operaio	3500	3	Abbastanza
43	Licenza media	Operaio	2800	5	Per niente
26	Laurea	Operaio	1400	1	Poco