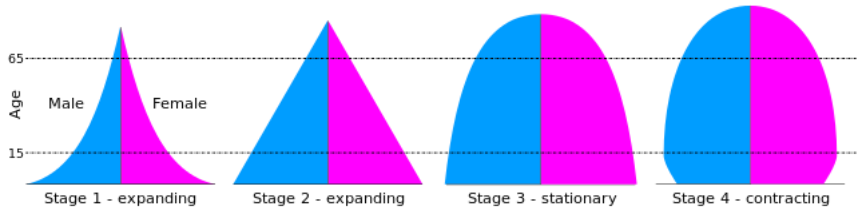
STATISTICA

La statistica è quella scienza che si preoccupa e studia i metodi per l'elaborazione dei dati con diverse modalità. La statistica si preoccupa di raccogliere e sintetizzare le informazioni per estrarre una conoscenza. Esempio di come la statistica è presente nella vita quotidiana: uno dei temi che coinvolge l'Italia e in particolare la Liguria è il fenomeno dell'invecchiamento della popolazione. L'ISTAT (ente preposto alla costruzione e al rilascio delle informazioni statistiche) ci ha detto a gennaio che la popolazione italiana è diminuita e che sono diminuite le nascite. Il diagramma sottostante si chiama **piramide dell'età** (o **della popolazione**) ed è il tipico diagramma utilizzato per descrivere una popolazione: ha più di 100 anni di storia ed è diviso in due parti, la parte a sinistra si riferisce ai maschi e la parte a destra si riferisce alle femmine.



Piramide dell'età nei seguenti anni:

- 1861: primo censimento del Regno d'Italia (stage 1).
- 1911: la popolazione cresce (stage 2)
- 1961: la piramide ha dei cambiamenti poiché c'è una transazione demografica ossia c'è un passaggio da un'economia rurale ad un'economia industriale che provocano una riduzione delle nascite (stage 3).
- 2010: dagli anni 50 la piramide si restringe, quindi si riducono le nascite ma si vive di più (stage 4).

I fenomeni che si studiano in statistica sono per lo più **fenomeni variabili** (es., l'andamento della popolazione, l'aumento della temperatura, la diffusione del coronavirus) cioè fenomeni che si possono manifestare in tanti modi diversi. i fenomeni che non presentano variabilità sono molto rari: anche quando abbiamo leggi deterministiche (es., leggi della fisica), ossia leggi per le quali uno stesso fenomeno produce sempre uno stesso risultato, se esaminiamo l'esperimento possiamo osservare che i risultati sono variabili perché mentre facciamo l'esperimento entrano in gioco molti elementi che influenzano il risultato. In ambito socio-economico ci occupiamo degli esseri umani che sono estremamente variabili dal punto di vista morfologico, caratteriale, etc. Questo si può spiegare bene con un'indagine che Banca Italia fa con scadenza triennale e che prende il nome di "**indagine sui bilanci delle famiglie**": ogni due anni Banca Italia seleziona un campione di famiglie. Fino al 2011 con cadenza decennale in Italia si faceva il censimento (operazione con la quale viene rilevata la popolazione di un territorio) della popolazione comprendeva circa 60 milioni di soggetti su tutto il territorio nazionale sui quali a una data specifica venivano rilevate determinate caratteristiche. Considerando che anche con il nuovo censimento per elaborare e rendere pubblici i dati ci vogliono 3 anni, il vecchio metodo di censimento era una procedura troppo complessa. Inoltre non essendo possibile per vari motivi intervistare tutta la popolazione si utilizza un **campione** (sottoinsieme di unità ridotto). Il campione viene usato, per es., nei sondaggi elettorali che vengono fatti intervistando circa un migliaio di persone e c'è un sito apposito sulla pagina della presidenza del consiglio dei ministri dove gli istituti di sondaggio sono obbligati a pubblicare la metodologia utilizzata per il sondaggio.

	A	B	C	D	E	F	G	H	I	J	K	L
1	nquest	sex	staciv	region	ncomp	eta	eta_class	condprof	nperc	titstud	y	yquart
2	173	1	3	18	1	64	4	3	1	3	49737,36	3
3	375	2	4	16	1	87	5	3	1	1	11320,94	0
4	629	1	4	5	2	73	5	3	2	2	40563,69	3
5	632	1	1	5	3	62	4	3	2	2	50735,08	3
6	633	2	4	5	1	78	5	3	1	1	19933,96	1
7	923	2	4	13	1	80	5	3	1	1	9110	0
8	1238	1	4	15	1	75	5	3	1	2	16217,2	0
9	1367	1	1	18	2	85	5	3	2	1	17351,66	1
10	1763	1	1	17	2	65	5	3	1	1	12840,7	0
11	1927	2	4	5	1	79	5	3	1	2	10505,7	0
12	1946	1	2	5	1	59	4	3	1	3	12187,52	0
13	2274	2	4	15	1	85	5	3	1	1	13360	0
14	2447	1	1	9	3	55	4	1	2	2	42202,37	3
15	2485	2	4	9	1	77	5	3	1	1	11407,54	0
16	2886	1	1	13	3	63	4	3	3	3	47125,77	3
17	3717	2	2	16	1	58	4	2	1	3	6000	0
18	4732	2	3	19	1	62	4	3	1	1	4700	0
19	5416	2	3	8	5	54	3	1	3	1	30065,39	2
20	7165	1	4	3	2	56	4	3	2	2	34900,72	2
21	7886	1	1	9	2	62	4	3	2	3	53774,82	3
22	11972	2	1	17	2	64	4	2	2	1	72544,84	3
23	20174	1	1	8	2	71	5	3	2	3	41170,82	3
24	20223	2	1	3	3	74	5	3	3	3	39179,3	2

Fig.1: le colonne (sex: sesso, staciv: stato civile) rappresentano le diverse caratteristiche, che sono dette variabili. Le righe si riferiscono ciascuna ad un'unità statistica diversa (**unità statistica**: unità elementare su cui vengono osservati i caratteri oggetto di studio. Un insieme di unità statistiche omogenee rispetto a una o più caratteristiche costituisce un collettivo statistico o una popolazione). Ogni variabile come genere, stato civile etc., sono riferite a quelle del capo famiglia.
Le famiglie totali prese in considerazione in questa tabella sono 8151.

Leggenda della tabella presente sopra:

- Nperc: numero percettori di reddito all'interno della famiglia;
- Ncomp: numero componenti del nucleo famigliare;
- Y: reddito annuo della famiglia;
- Nquest: numero questionario della famiglia. Il numero, per es., 173 è quello messo sul questionario compilato dalla famiglia. Questa variabile è irrilevante dal punto di vista statistico ma serve per distinguere una famiglia dall'altra.
- Sex: indica il genere e può assumere due valori diversi: femmina (codice num.: 2), maschio (codice num.: 1).
- Staciv: è lo stato civile ed ha quattro modalità: celibe – nubile, coniugato/a, divorziato/a, vedovo/a.
- Region: la regione è una variabile con 20 modalità perché 20 sono le regioni italiane.

Le variabili possono essere distinte:

1. Variabili qualitative: vengono misurate su:
    - Scala di misura nominale o sconnessa: se date due modalità della variabile è possibile affermare solo se queste sono uguali o diverse. Sono, per es., sesso, luogo di nascita, stato civile, religione, colore degli occhi etc. Tra le modalità di ciascuno di questi caratteri non è possibile stabilire un ordinamento e quindi le modalità possono essere elencate in modo del tutto arbitrario.
    - Scala di misura ordinale: se date due modalità della variabile è possibile solo dare un ordine, specificando che una precede l'altra. Caratteri ordinati sono quelli che esprimono un grado di soddisfazione (es., poco, molto), la posizione in una graduatoria, il titolo di studio (senza titolo, licenza elementare, licenza media, diploma, laurea, dottorato).
2. Variabili quantitative: vengono misurate su:
    - Scala di intervallo: sono molto rare e sono quelle nelle quali lo zero è fissato su base convenzionale. Un esempio è la temperatura misurata in gradi centigradi: lo zero che utilizziamo noi è uno zero convenzionale e non assoluto infatti è diverso se si misura in gradi Celsius o Fahrenheit.
    - Scala di rapporto: a differenza della scala di intervallo in questo caso possiamo fare un rapporto e quindi, per es., possiamo dire che una famiglia ha un reddito che è due volte più elevato di un'altra famiglia.

    Le variabili quantitative possono essere anche distinte come segue:
    - Variabili quantitative discrete: una variabile è discreta quando assume un numero intero di valori (i numeri interi sono quelli negativi e positivi, il reddito per esempio può essere anche negativo). Esempi: numero di figli, voto a un esame, numero di pezzi prodotti.
    - Variabili quantitative continue: variabili che possono assumere qualunque valore all'interno dell'intervallo. Esempi: peso e altezza.

N.B: La statistica si può distinguere in:

a. Statistica descrittiva (l'andamento nel tempo della temperatura, diffusione di una malattia nello spazio, piramide dell'età): io osservo un fenomeno e ne sintetizzo e descrivo le caratteristiche.
b. Inferenza: passiamo dal particolare al generale. Facendo questa operazione di inferenza siamo soggetti ad errore dovuto dal fatto che, per es., le 1000 persone che ho preso come campione potrebbero non essere una buona immagine della popolazione. L'inferenza dunque è l'insieme dei metodi e delle tecniche che ci consentono di gestire questo particolare errore.

Lavorare con il foglio elettronico della fig.1 è molto difficile dobbiamo quindi sintetizzare ciò che stiamo osservando. Gli strumenti di sintesi che si usano sono: costruire delle tabelle (quindi radunare tutte le 8151 famiglie in una tabella di piccole dimensioni), fare dei grafici, calcolare opportune misure (es., calcolare la media). Per es. come facciamo a sintetizzare la colonna del genere? Conto quante famiglie hanno un capofamiglia maschio e conto quante famiglie hanno un capofamiglia femmina. Se faccio questa operazione calcolo due quantità che vengono chiamate frequenze assolute (valore che corrisponde al numero di volte in cui è stato osservato un certo valore di una variabile). La frequenza assoluta è associata alla modalità: quindi io prendo la modalità e conto quante volte si è presentata. Tramite le frequenze, possiamo ottenere una distribuzione una rappresentazione molto più sintetica denominata distribuzione di frequenze.

Si consideri x, y, z (variabili) e $X_1, X_2, \dots, X_i, \dots, X_n$ (successione di tutti i valori osservati: è una colonna della tabella):

- Frequenze assolute $\boldsymbol{n_i}$: $\sum_{i=1}^{c} n_i = n$ $(i = 1,2, \dots, c \rightarrow c$ è il numero di modalità di una variabile);
- Frequenze relative $\boldsymbol{f_i}$: $f_i = n_i/n$
- Frequenze percentuali $\boldsymbol{p_i}$: $p_i = (n_i/n) \times 100$

Esempio di distribuzione di frequenza del "titolo di studio" (**fig.2**) creata partendo dalla modalità più bassa:

Titolo di studio	n_i	f_i	p_i	N_i	F_i	P_i
Lic. Elementare	2200	0,269906	26,99055	2200	0,269906	26,99055
Lic. Media	2827	0,346829	34,68286	5027	0,616734	61,67341
Maturità	2157	0,26463	26,46301	7184	0,881364	88,13642
Laurea	967	0,118636	11,86358	8151	1	100
Totale	8151	1	100			

N_i, F_i, P_i sono dette **frequenze cumulate** che non ci sono nella distribuzione di frequenza del genere perché viene calcolata su scala sconnessa. La distribuzione di frequenza della fig.2 viene detta distribuzione delle frequenze cumulate poiché la frequenza per una data classe è ottenuta come somma della corrispondente frequenza e di tutte quelle relative alle classi precedenti. Le frequenze cumulate si utilizzano solo in presenza di frequenze ordinabili. Le frequenze cumulate ci dicono il numero o la % di unità statistiche che presentano un valore pari o inferiore alla modalità corrispondente.
Consideriamo la i-esima frequenza cumulata:

- $N_i = n_1 + n_2 + \cdots + n_i = \sum_{g=1}^{i} n_g \rightarrow$ formula che ci permette di passare dalle frequenze alle frequenze cumulate.
- $n_i = N_i - N_{i-1} \rightarrow$ formula che ci permette di passare dalle frequenze cumulate alle frequenze non cumulate (per es. nella colonna delle n_i il valore 967 è dato da 8151-7184, 2200 invece è dato da 2200-0).

Se abbiamo un carattere misurato su scala continua la tabella di distribuzione di frequenza esplode, per es., se faccio la tabella con tutte le età misurate in anni ho una tabella molto grossa. Dunque la sintesi procede sempre in due passi:

1. **Il primo consiste nel prendere le modalità della variabile e dividerle in classi** (non considero più l'età in anni ma metto insieme tutti i soggetti che sono nati lo stesso anno, oppure invece che considerare classi di età annuali posso considerare classi di età quinquennali, o ancora posso considerare classi di età di ampiezza diversa). La **fig.3** è la rappresentazione della distribuzione di frequenza delle 8151 famiglie (fig.1) per classe di reddito ciascuna di ampiezza diversa. Non si va avanti di 10.000 euro in 10.000 euro perché la tabella verrebbe enorme (suddividere però le classi di 100.000 in 100.000 sarebbe comunque un problema perché, per es., metterei famiglie in povertà assoluta con famiglie ricche). Le classi di ampiezza diversa, dunque, esistono perché, per es., nei casi di classi di reddito basse devo essere molto dettagliato (es. c'è un enorme differenza tra una famiglia che ha un reddito di 10.000 euro e una famiglia che ha un reddito di 20.000 euro), al contrario, i comportamenti di due famiglie che hanno un reddito pari, una a 100.000 euro e l'altra a 150.000 euro, non sono molto diversi.

Reddito	n_i	f_i	p_i	N_i	F_i	P_i
Fino a 10.000	618	0,0758	7,58	618	0,0758	7,58
10.000-20.000	2171	0,2663	26,63	2789	0,3422	34,21
20.000-30.000	1955	0,2398	23,98	4744	0,5820	58,19
30.000-50.000	2212	0,2714	27,14	6956	0,8534	85,33
50.000-75.000	845	0,1037	10,37	7801	0,9571	95,70
75.000-100.000	218	0,0267	2,67	8019	0,9838	98,37
100.000-250.000	124	0,0152	1,52	8143	0,9990	99,89
>250.000	8	0,0010	0,10	8151	1,0000	99,99
Totale	8151	1	100	/	/	/

Es: F_i =0,8534 perché ho fatto 0,2714 (f_i)+0,5820 (valore di F_i precedente a 0,8534).

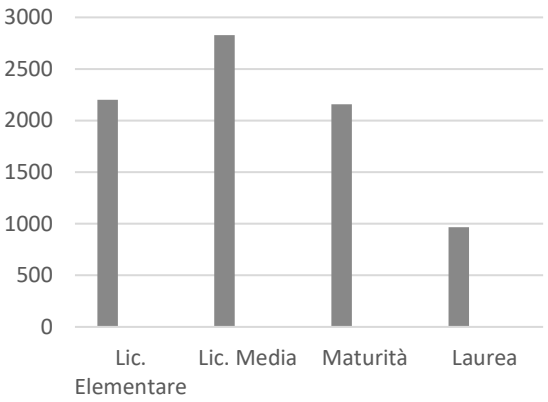
Se io dalla tavola (fig.1) passo alla distribuzione di frequenza (fig.2-3) faccio una sintesi, ma perdo le informazioni del singolo (anche se sono irrilevanti per la statistica che studia i fenomeni variabili).

2. **Il secondo procedimento consiste nella rappresentazione grafica delle distribuzioni semplici**. La trasformazione della distribuzione semplice da forma tabellare a immagine grafica ha senso se tale operazione riesce a rendere più evidenti e di facile lettura le caratteristiche della distribuzione della variabile sul collettivo preso in esame. I grafici sono bidimensionali (anche se in alcuni vengono aggiunti la tridimensionalità e la prospettiva).

Principali tipi di rappresentazione grafica:

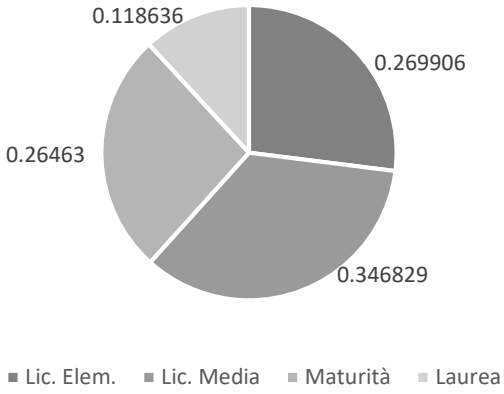
a. **Grafici a barre o a nastri**: in questi grafici ogni frequenza o intensità della distribuzione viene rappresentata o da una barra o da un nastro così da ottenere una successione di rettangoli aventi la stessa base (o altezza) e le altezze (o basi) proporzionali alle frequenze o quantità. I grafici a barre sono adatti a rappresentare le distribuzioni di

frequenza di caratteri quantitativi discreti (es. titolo di studio, voto esame, numero componenti di una famiglia). Grafico a barre della frequenza assoluta (n_i) della distribuzione di frequenza per titolo di studio:



b. **Cartogrammi**: per rappresentare le serie territoriali si utilizza un grafico chiamato cartogramma. Questo grafico ha come base una mappa sulla quale sono visibili i contorni delle aree geografiche o territoriali rispetto alle quali vengono analizzate le frequenze o le intensità di un carattere (per es., la popolazione residente, i nati, l'età media, il reddito medio, etc.). I cartogrammi a ripartizioni colorate sono dei cartogrammi in cui ogni area della carta è colorata in base alla distribuzione di frequenza.

c. **Diagramma a torta**: utili quando si vuole rappresentare la composizione di un aggregato. Con questo tipo di grafico è buona norma rappresentare distribuzioni con un numero di modalità non troppo elevato, poiché aumentando i settori circolari la loro dimensione diminuisce ed è più difficile poterli confrontare. Mentre il diagramma a barre può essere utilizzato per qualunque distribuzione di frequenza, il diagramma a torta è specifico per la distribuzione relativa. Un diagramma a torta è composto da un cerchio diviso in settori ciascuno dei quali è associato ad una modalità, e l'ampiezza del settore è proporzionale alla frequenza di questa modalità. È ideale per le frequenze relative perché sommano a 1 (la torta è il 100%). Diagramma a torta della f_i del titolo di studio:



d. **Istogramma**: è un grafico costituito da barre non distanziate, con basi uguali o diverse, dove ogni barra possiede un'area proporzionale alla corrispondente frequenza o quantità. In un istogramma con classi di ampiezza diversa, l'altezza *h* del rettangolo corrispondente a una classe viene chiamata densità e si ottiene come rapporto tra la frequenza e l'ampiezza della classe. L'ampiezza di classe $\Delta_i = x_{i+1} - x_i$. La densità di frequenza $d_i = n_i/\Delta_i$ (questo è per le frequenze assolute, ma se avessi le frequenze relative o percentuali basta sostituire, per es., n_i con f_i o p_i).