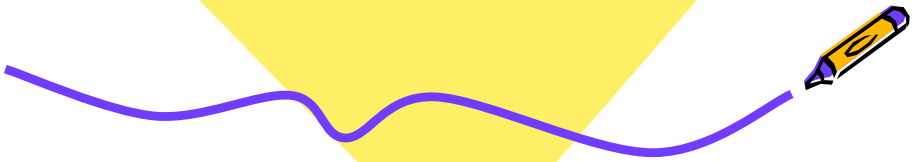
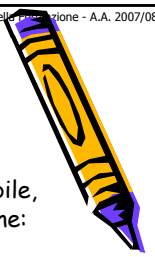
ANALISI DI UNA DISTRIBUZIONE

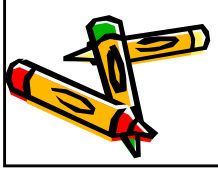
Tabelle e Grafici





ANALISI DI UNA DISTRIBUZIONE

- Punto di partenza: terminata una indagine, abbiamo una serie di dati osservati su un insieme di soggetti, sui quali abbiamo rilevato una variabile, ad es. abbiamo chiesto l'età a tutti gli iscritti a Scienze della Formazione:
  - supponiamo di avere osservato 2000 soggetti
    -> abbiamo una serie di 2000 numeri *interi*
    (se la variabile è misurata in *anni compiuti*)
  - supponiamo di fare lo stesso, questa volta sugli iscritti a Filosofia
    -> abbiamo un'altra serie di 500 numeri interi
- Domanda: Qual'è la facoltà più giovane ?
  - i dati non sono poi così tanti (potremmo averne molti di più, es. censimento), ma sempre troppi per capire qualcosa solamente "guardandoli"
  - per poter dare una risposta dobbiamo riuscire a *sintetizzarli* in qualche modo
- Come *descrivere* la distribuzione dell'età nelle due popolazioni ?



=> Organizziamo i dati in una tabella di frequenze

ANALISI DI UNA DISTRIBUZIONE

- **Tabella di Frequenze**

  Costruiamo una tabella dove riportiamo, per ogni modalità osservata della variabile età, il numero di soggetti la presentano

- **Frequenze assolute:**
  numero di unità statistiche che soddisfano una certa condizione, ovvero che presentano una certa modalità di risposta

$$n_i = frequenza\ osservata$$

- Osservazione: i valori osservati della variabile età risultano compresi nell'intervallo 19-35, ma se dovessimo rappresentare la distribuzione dell'intera popolazione italiana ?
- Il confronto risulta ancora poco evidente, per due ragioni:
  - le due popolazioni sono di dimensioni diverse
  - troppi livelli della variabile età

Frequenze Assolute		
Età	Scienze Formazione	Filosofia
19	350	80
20	300	70
21	250	60
22	200	55
23	150	70
24	180	60
25	200	30
26	80	20
27	130	10
28	60	0
29	25	5
30	50	0
31	10	0
32	5	0
33	0	15
34	5	25
35	5	0
Totale	2000	500

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DI UNA DISTRIBUZIONE

- **Frequenze relative**
  si calcolano dividendo le frequenze assolute per il numero totale di unità statistiche

$$frequenza\ relativa = \frac{frequenza\ assoluta}{numero\ di\ osservazioni}$$

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

i indica la riga ovvero la modalità i-esima
n(i) frequenza assoluta della riga i
N numero totale di osservazioni

- Vantaggio: le frequenze relative permettono di eliminare l'effetto della numerosità e quindi di confrontare le distribuzioni di popolazioni con numerosità diverse
- La somma delle frequenze relative è sempre uguale a 1

Frequenze Assolute e Relative				
Età	Scienze Formazione		Filosofia	
19	350	0,175	80	0,160
20	300	0,150	70	0,140
21	250	0,125	60	0,120
22	200	0,100	55	0,110
23	150	0,075	70	0,140
24	180	0,090	60	0,120
25	200	0,100	30	0,060
26	80	0,040	20	0,040
27	130	0,065	10	0,020
28	60	0,030	0	0,000
29	25	0,013	5	0,010
30	50	0,025	0	0,000
31	10	0,005	0	0,000
32	5	0,003	0	0,000
33	0	0,000	15	0,030
34	5	0,003	25	0,050
35	5	0,003	0	0,000
Totale	2000	1,000	500	1,000

ANALISI DI UNA DISTRIBUZIONE

- **Distribuzione di Frequenze in Classi**
  - per rendere più comprensibile la situazione, dobbiamo ridurre il numero di *livelli* della variabile considerati nell'analisi
  - ridefiniamo i livelli considerati per la variabile età, creando *classi di valori*
- Procediamo in questo modo:
  - suddividiamo l'intervallo contenente tutti i valori osservati, il **campo di variazione**, in un certo numero di sotto-intervalli, tutti di uguale ampiezza (es. 3 anni)
  - conteggiamo le frequenze assolute per le nuove classi di valori
  - ricalcoliamo le frequenze relative, verificando sempre che il totale sia pari a 1

Frequenze Assolute		
Età	Scienze Formazione	Filosofia
19-21	900	210
22-24	530	185
25-27	410	60
28-30	135	5
31-33	15	15
34-36	10	25
Totale	2000	500

Frequenze Relative		
Età	Scienze Formazione	Filosofia
19-21	0,4500	0,4200
22-24	0,2650	0,3700
25-27	0,2050	0,1200
28-30	0,0675	0,0100
31-33	0,0075	0,0300
34-36	0,0050	0,0500
Totale	1,0000	1,0000

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DI UNA DISTRIBUZIONE

- Come si scelgono le classi di valori, ovvero gli estremi degli intervalli:
  - la scelta è sempre arbitraria ma deve tenere conto di alcune indicazioni
  - la riduzione in classi opera una sintesi ma comporta una perdita di informazione
  - il numero di intervalli non deve quindi essere né troppo grande né troppo piccolo
  - l'ampiezza degli intervalli dovrebbe essere preferibilmente uguale per tutti
  - si determina il campo di variazione e lo si divide per il numero di intervalli desiderati

Frequenze Assolute		
Età	Scienze Formazione	Filosofia
19-21	900	210
22-24	530	185
25-27	410	60
28-30	135	5
31-33	15	15
34-36	10	25
Totale	2000	500

  - si stabilisce convenzionalmente se l'intervallo è chiuso a sinistra cioè comprende l'estremo inferiore, o invece chiuso a destra cioè comprende l'estremo superiore
  - si indica con A|--- B oppure con [A,B) un intervallo con estremo inferiore A incluso, ed estremo superiore B escluso
  - se la variabile è discreta, l'intervallo può comprendere entrambi gli estremi [A,B] avendo cura in questo caso di evitare sovrapposizioni tra le classi
  - in ogni caso la tabella deve indicare chiaramente quale metodo viene usato, per poter essere letta correttamente

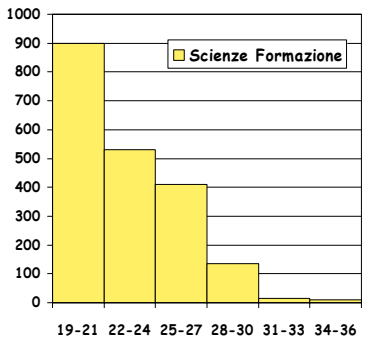
ANALISI DI UNA DISTRIBUZIONE

- Indicazioni per la costruzione di una tabella:
  - titolo/legenda che descriva in modo adeguato il contenuto
  - all'interno di un testo deve essere sempre numerata (es. Tabella 4)
  - se sono necessari ulteriori chiarimenti sulla sua costruzione occorre inserirle in nota sotto la tabella
  - tutte le righe e le colonne devono avere una etichetta appropriata e comprensibile
  - specificare le unità di misura adottate (es. %)
  - usare linee all'interno della tabella per agevolarne la lettura (es. intestazione, totali) ma senza esagerare
  - usare lo stesso numero di cifre decimali per tutti i dati dello stesso tipo, come quelli in una stessa colonna, e allineare sempre verticalmente le cifre decimali (a destra)
  - non esagerare con i decimali: usare il (minor) numero di cifre decimali necessario per non perdere informazioni rilevanti
    - se la variabile ha molti livelli o è continua, è opportuno raggruppare i dati in classi di valori, tenendo sempre conto che così facendo si perdono informazioni
    - i totali devono "tornare" ! (quando usiamo poche cifre decimali, gli errori di arrotondamento possono far si che il problema si ponga)

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DI UNA DISTRIBUZIONE

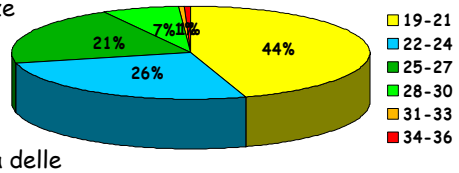
- **Rappresentazioni Grafiche**
  Per rendere immediatamente comprensibile l'andamento di una distribuzione, possiamo rappresentare graficamente i dati della tabella di frequenze
- **Istogramma di Frequenze**
- E' la rappresentazione grafica più utilizzata per la sua semplicità di realizzazione e di comprensione
- Per quanto facile da realizzare, deve rispettare dei principi ben precisi
- Le dimensioni delle barre verticali devono essere disegnate in modo da dare una rappresentazione corretta, e non falsata, della distribuzione

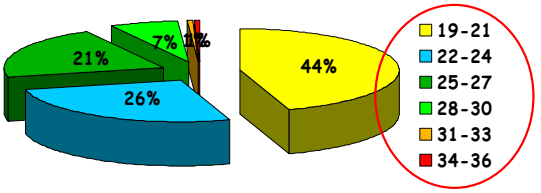


  - *base* e *altezza* dei rettangoli devono essere determinate in modo da rappresentare correttamente la tabella di frequenze :
  - le *aree* dei rettangoli devono essere proporzionali alle frequenze delle classi
  - la *base* deve essere proporzionale all'ampiezza della classe
  - l'*altezza* di ciascun rettangolo deve essere calcolata come rapporto tra frequenza della classe e ampiezza dell'intervallo, detto **densità** dell'intervallo

ANALISI DI UNA DISTRIBUZIONE

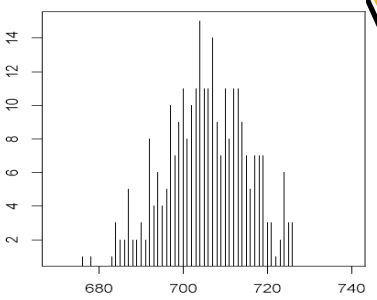
- **Torta**
- Fornisce una rappresentazione di una distribuzione alternativa all'istogramma
- La torta rappresenta sempre frequenze relative (%)
- Le "fette" della torta (le aree) sono proporzionali alle frequenze relative delle classi
- La torta non rende conto dell'ampiezza delle classi, quindi è una rappresentazione meno informativa dell'istogramma
- Esistono molte varianti nel disegnare le torte: una di queste è l'*esplosione* delle fette
- Per riportare nel grafico le modalità della variabile, in genere si utilizza una *legenda*





ANALISI DI UNA DISTRIBUZIONE

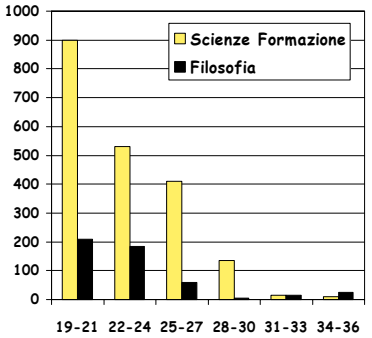
- **Istogramma di Frequenze**
- **Variabili Qualitative:**
  - una barra per ciascuna modalità di risposta
  - le barre si disegnano *separate* tra loro, come bastoncini di uguale spessore
  - l'altezza delle barre deve risultare proporzionale alle frequenze osservate



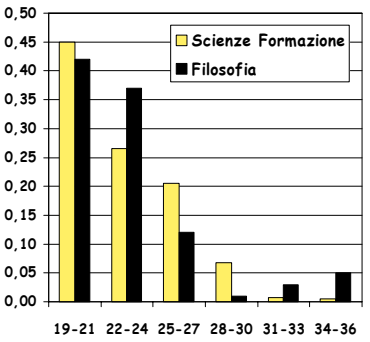
- **Variabili Quantitative Discrete:**
  - come per le qualitative, oppure ...
  - per raggiungere una maggior sintesi, si possono costruire classi di valori, come per le variabili continue
- **Variabili Quantitative Continue:**
  - si devono costruire le classi di valori, preferibilmente di uguale ampiezza
  - le barre si disegnano affiancate, senza spazi di separazione tra gli estremi degli intervalli
  - se le classi non sono tutte di uguale ampiezza, occorre fare attenzione, infatti le barre devono avere:
    - la base proporzionale all'ampiezza dell'intervallo della classe
    - l'AREA (e NON l'altezza) proporzionale alla frequenza della classe

ANALISI DI UNA DISTRIBUZIONE

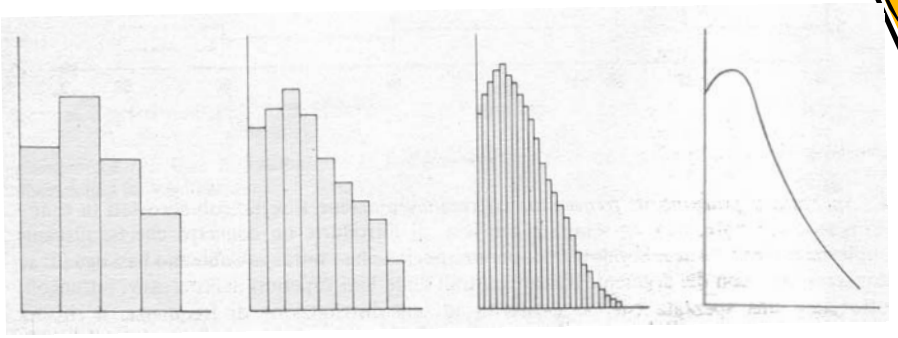
- **Istogrammi Affiancati**
- Si utilizzano per confrontare due o più distribuzioni
  - **di Frequenze Assolute:** riusciamo ad apprezzare soprattutto le diverse dimensioni delle due popolazioni



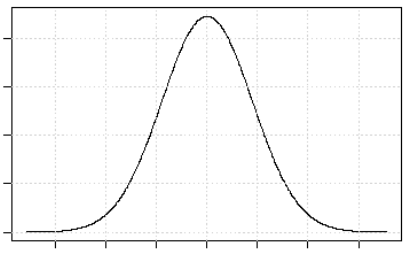
  - **di Frequenze Relative:** finalmente abbiamo un confronto efficace tra le due distribuzioni dell'età



ANALISI DI UNA DISTRIBUZIONE



- Al crescere del numero di osservazioni, e riducendo l'ampiezza degli intervalli, l'istogramma di frequenza tende a diventare una *curva*, che ben rappresenta la distribuzione della variabile



FACCIAMO UN PASSO AVANTI...

0011 0010 1010 1101 0001 0100 1011

formalizziamo le cose fin qui viste

ANALISI DI UNA DISTRIBUZIONE

- Formalizziamo il calcolo delle frequenze relative per una generica distribuzione:

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

$$N = n_1 + n_2 + \ldots + n_i + \ldots + n_k$$

- Abbiamo indicato con:
  - i : indice della modalità considerata
  - k : numero totale di modalità/classi
  - n: frequenza assoluta
  - f: frequenza relativa
  - N: numero totale di osservazioni

- f(i) indica la frequenza relativa della modalità (o classe) i-esima:
  in pratica della riga i-esima della tabella

Età	Frequenze Assolute	Frequenze Relative
x_1	n_1	$f_1 = \frac{n_1}{N}$
…	…	…
x_i	n_i	$f_i = \frac{n_i}{N}$
…	…	…
x_k	n_k	$f_k = \frac{n_k}{N}$
Totale	N	1

ANALISI DI UNA DISTRIBUZIONE

- **Sommatoria**: con la lettera Σ (sigma) si indica l'operazione di somma degli elementi specificati, allora possiamo scrivere:

$$N = \sum_{i=1}^{k} n_i = \underbrace{n_1 + n_2 + \ldots + n_i + \ldots + n_k}_{k}$$

si legge:
sommatoria, per i che va da 1 a k, di n(i)

- La "formula" descrive esattamente quello che si deve fare per calcolare la quantità: è solo un modo più veloce di scrivere una operazione rispetto alle parole
- Cominciamo allora a familiarizzare con questo modo di scrivere, ad esempio possiamo scrivere:

$$\sum_{i=1}^{k} f_i = 1$$

Età	Frequenze Assolute	Frequenze Relative
x_1	n_1	$f_1 = \frac{n_1}{N}$
…	…	…
x_i	n_i	$f_i = \frac{n_i}{N}$
…	…	…
x_k	n_k	$f_k = \frac{n_k}{N}$
Totale	N	1

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DI UNA DISTRIBUZIONE

- Per fare un po' di pratica con l'operatore sommatoria, dimostriamo che la somma delle frequenze relative è uguale a 1 :

$$\sum_{i=1}^{k} f_i = \underbrace{f_1 + f_2 + \ldots + f_i + \ldots + f_k}_{k}$$

$$= \frac{n_1}{N} + \frac{n_2}{N} + \ldots + \frac{n_i}{N} + \ldots + \frac{n_k}{N} =$$

$$= \frac{n_1 + n_2 + \ldots + n_i + \ldots + n_k}{N} =$$

$$= \frac{\sum_{i=1}^{k} n_i}{N} = \frac{N}{N} = 1$$

Età	Frequenze Assolute	Frequenze Relative
x_1	n_1	$f_1 = \frac{n_1}{N}$
…	…	…
x_i	n_i	$f_i = \frac{n_i}{N}$
…	…	…
x_k	n_k	$f_k = \frac{n_k}{N}$
Totale	N	1

ANALISI DI UNA DISTRIBUZIONE

- **Frequenze Cumulate Assolute**
  Numero di osservazioni minori o uguali ad un valore specificato della variabile

$$N_x = \sum_{i=1}^{x} n_i$$

- Es. Quanti soggetti hanno meno di 30 anni ?
- La tabella delle frequenze cumulate si costruisce sommando (cumulando), per ogni livello della variabile, le frequenze dei livelli inferiori fino a quello considerato
- E' un altro modo di rappresentare una distribuzione, che ha importanti applicazioni

- **Frequenze Cumulate Relative :**

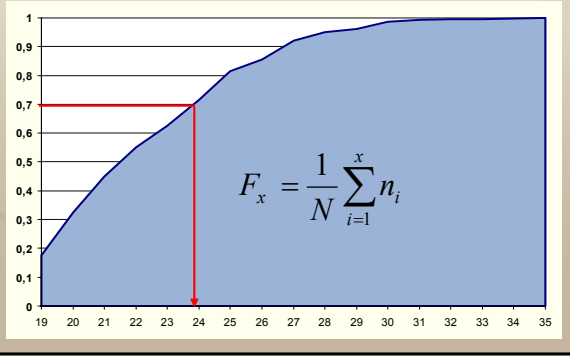
$$F_x = \sum_{i=1}^{x} f_i = \frac{1}{N} \sum_{i=1}^{x} n_i$$

Scienze Formazione		
Età	Frequenze Assolute	Freq. Ass. Cumulate
19	350	350
20	300	650
21	250	900
22	200	1100
23	150	1250
24	180	1430
25	200	1630
26	80	1710
27	130	1840
28	60	1900
29	25	1925
30	50	1975
31	10	1985
32	5	1990
33	0	1990
34	5	1995
35	5	2000
Totale	2000	

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

ANALISI DI UNA DISTRIBUZIONE

- **Funzione di ripartizione** (empirica)
- Le frequenze relative **cumulate** possono essere rappresentate graficamente mediante un istogramma di frequenze cumulate oppure con una linea ("spezzata").
- La curva risultante descrive la distribuzione in modo graficamente diverso rispetto all'istogramma, ma del tutto equivalente
- cioè fornisce la stessa informazione sulla distribuzione, e risponde in modo più immediato ad alcune domande (percentili)



Scienze Formazione		
Età	Frequenze Relative	Freq. Rel. Cumulate
19	0,1750	0,1750
20	0,1500	0,3250
21	0,1250	0,4500
22	0,1000	0,5500
23	0,0750	0,6250
24	0,0900	0,7150
25	0,1000	0,8150
26	0,0400	0,8550
27	0,0650	0,9200
28	0,0300	0,9500
29	0,0125	0,9625
30	0,0250	0,9875
31	0,0050	0,9925
32	0,0025	0,9950
33	0,0000	0,9950
34	0,0025	0,9975
35	0,0025	1,0000
Totale	1,0000	

L'OPERATORE SOMMATORIA Σ

0011 0010 1010 1101 0001 0100 1011

L'OPERATORE SOMMATORIA Σ

- Sia x(i) con i = 1, … , n una successione di numeri reali : $x_1 \ , \ x_2 \ , \ \dots \ , \ x_n$

  Si definisce sommatoria, per i che va da 1 a n, di x(i) l'operatore :

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \dots + x_n$$

- Proprietà:

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j = x_1 + x_2 + \dots + x_n$$

$$\sum_{i=1}^{n} c = c + c + \dots + c = n \cdot c$$

$$\sum_{i=1}^{n} c\,x_i = c\,x_1 + c\,x_2 + \dots + c\,x_n = c\,(x_1 + x_2 + \dots + x_n) = c \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i + y_i) = x_1 + y_1 + x_2 + y_2 + \dots + x_n + y_n = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

L'OPERATORE SOMMATORIA Σ

- Altre proprietà:

$$\sum_{i=1}^{n}(a+b\,x_i)=\sum_{i=1}^{n}a+\sum_{i=1}^{n}b\,x_i=n\,a+b\sum_{i=1}^{n}x_i$$

$$\sum_{i=1}^{n}(x_i+y_i)^2=\sum_{i=1}^{n}(x_i^2+y_i^2+2\,x_i\,y_i)=\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2+2\sum_{i=1}^{n}x_i\,y_i$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

L'OPERATORE SOMMATORIA Σ

- Errori da evitare (proprietà non valide) :

$$\sum_{i=1}^{n}(x_i\cdot y_i)\neq\sum_{i=1}^{n}x_i\cdot\sum_{i=1}^{n}y_i \qquad infatti,\ ad\ es.\ per\ \ n=2:$$

$$\sum_{i=1}^{2}(x_i\cdot y_i)=x_1\,y_1+x_2\,y_2\neq\sum_{i=1}^{2}x_i\cdot\sum_{i=1}^{2}y_i=(x_1+x_2)(y_1+y_2)$$

$$\sum_{i=1}^{n}\frac{x_i}{y_i}\neq\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}y_i}$$

$$(\sum_{i=1}^{n}x_i)^k\neq\sum_{i=1}^{n}(x_i)^k \qquad infatti,\ ad\ es.\ per\ \ n=2\ \ e\ \ k=2:$$

$$(\sum_{i=1}^{2}x_i)^2=(x_1+x_2)^2=x_1^2+x_2^2+2x_1\,x_2\neq\sum_{i=1}^{2}(x_i)^2=x_1^2+x_2^2$$