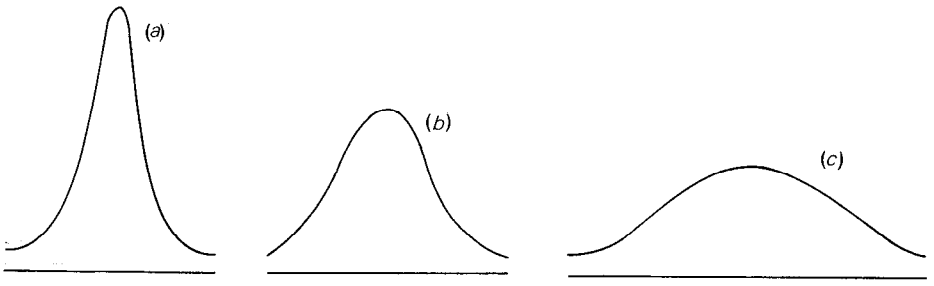
ANALISI DI UNA DISTRIBUZIONE

Variabilità

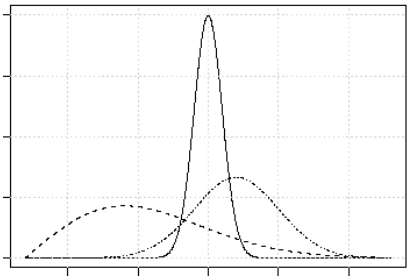
ANALISI DI UNA DISTRIBUZIONE

- Qualunque fenomeno empirico presenta una certa variabilità.
- In una popolazione, con riferimento ad una qualsiasi caratteristica, generalmente si osserva una variabilità tra soggetti: non c'è variabilità quando tutti i soggetti sono uguali (es. hanno tutti la stessa altezza, o lo stesso colore)
- Anche in un qualsiasi processo di misura riscontriamo sempre una variabilità: lo strumento non produce sempre lo stesso valore, anche misurando lo stesso oggetto, a causa del cosiddetto *errore di misura*.
- La variabilità osservata è in qualche senso la somma della variabilità *naturale* del fenomeno e della variabilità introdotta dallo strumento di osservazione.
- Dunque, quando i dati osservati non sono tutti uguali, la distribuzione presenta una certa variabilità, più o meno grande: cioè i dati sono più o meno dispersi intorno alla modalità che individua la tendenza centrale del fenomeno (media, mediana, …)



ANALISI DI UNA DISTRIBUZIONE

- **Indici di variabilità (o dispersione)**
- La rappresentazione grafica della distribuzione di frequenza ci permette di apprezzare immediatamente la diversa variabilità delle tre distribuzioni a confronto
- Come si può descrivere la variabilità di una distribuzione in modo sintetico, come abbiamo fatto per la tendenza centrale, cioè con un solo numero, immediatamente comprensibile, senza dover riportare tutta la distribuzione ?
- A questo scopo si introducono gli **indici di variabilità**, che forniscono una misura sintetica della dispersione dei dati
- La costruzione degli indici di variabilità è profondamente diversa in base alla natura della variabile:
  - variabili quantitative: indici di variabilità basati sui *valori* osservati
  - variabili qualitative: indici di mutabilità basati sulle *frequenze* osservate



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

ANALISI DI UNA DISTRIBUZIONE

- **Campo di variazione (range):** è la differenza tra il valore massimo e il valore minimo assunti dalla variabile

$$x_{\max} - x_{\min}$$

  - la variabile deve essere quantitativa, perché tale differenza abbia significato
  - facile da calcolare e da comprendere
  - è una misura molto grossolana della dispersione, infatti tiene conto solo di due valori, tutto il resto della distribuzione viene ignorata
  - due distribuzioni, a parità di campo di variazione, possono presentare forme e variabilità molto diverse
  - troppo sensibile a possibili valori anomali
- Si determina molto semplicemente: ordinando i singoli dati, oppure dalla tabella di frequenze

Es. range(SF) = 35 – 19 = 16

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

ANALISI DI UNA DISTRIBUZIONE

- **Scarto Interquartile:** è lo scarto tra il terzo e il primo quartile:

$$Q_3 - Q_1$$

  - L'idea è quella di individuare il range in cui cade il 50% dei casi *centrali* (più vicini alla mediana) della distribuzione
  - I quartili sono determinabili per variabili su scala *ordinale*; lo scarto interquartile, tuttavia, essendo una *differenza* tra valori, è calcolabile solo per variabili *quantitative*

- **Semiscarto Interquartile:** è la metà dello scarto interquartile

$$\frac{Q_3 - Q_1}{2}$$

  - è usato in alternativa al precedente, ma è ovviamente del tutto equivalente, cioè fornisce la stessa informazione

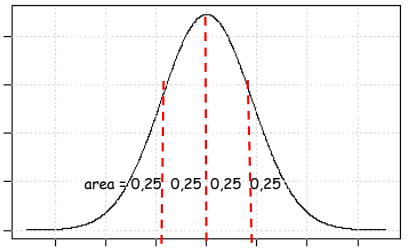
- Entrambi questi indicatori di variabilità si usano di solito in associazione alla mediana (quando come indice di centralità viene utilizzata la mediana).

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

ANALISI DI UNA DISTRIBUZIONE

- **I Quartili**
- Q1 (primo quartile) è un valore che lascia 25% dei casi a sinistra e 75% a destra: cioè un valore maggiore del 25% dei casi e minore di tutti i restanti casi (75%); in pratica è la mediana della "prima" metà dei dati
- Q2 (secondo quartile) è un valore maggiore del 50% dei casi e minore dei restanti casi (50%): in pratica è la mediana
- Q3 (terzo quartile) è un valore che lascia 75% dei casi a sinistra e 25% a destra, ovvero è un valore maggiore del 75% dei casi e minore di tutti i restanti casi (25%): in pratica è la mediana della "seconda" metà dei dati

- **I Percentili**
- I percentili Pn si definiscono operando, in modo del tutto analogo, la suddivisione della distribuzione in 100 parti:
  - Q1 = P25
  - Me = Q2 = P50
  - Q3 = P75



- Nella determinazione dei quartili ci possono essere dei margini di ambiguità, come vediamo nel prossimo esempio

Ambiguità nel calcolo dei quartili

Dati (già ordinati): 6,4 6,7 6,8 7,0 7,3 7,5 7,5 7,6 7,9 8,1

La mediana deve cadere tra 7,3 e 7,5. Tradizionalmente, si sceglie il punto centrale dell'intervallo, ovvero si pone mediana = 7,4.

La determinazione del primo (e del terzo) quartile è più ambigua. Il primo quartile dovrebbe lasciare sulla sinistra il 25% delle osservazioni, ovvero in questo caso 2,5 osservazioni. Questo è ovviamente impossibile da raggiungere esattamente. Esistono vari ragionamenti che possono essere utilizzati per *sciogliere* l'ambiguità. Ad esempio,

1. potremmo *decidere* di interpretare "lasciare a sinistra 2,5 osservazioni" come "posizionarsi sul punto intermedio tra la seconda e la terza osservazione (dal basso)" ovvero di *assegnare* al primo quartile il valore di 6,75. Allora, in maniera analoga potremmo *assegnare* al terzo quartile il valore di 7,75

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

ANALISI DI UNA DISTRIBUZIONE

2. oppure, potremmo *decidere* che il primo quartile deve dividere le osservazioni alla sinistra della mediana in due parti uguali. Quindi, poiché abbiamo alla sinistra della mediana 5 osservazioni, decidere di *porre* il primo quartile uguale al terzo dato dal basso (ovvero a 6,8). Argomentando in maniera analogo assegneremo al terzo quartile il valore 7,6 (= terza osservazione dal basso nel gruppo a destra della mediana).

Nessuna delle due scelte è migliore dell'altra. Si tenga comunque presente che, a meno di casi particolari, più il numero di osservazioni diventa grande, più le varie possibilità tendono ad avvicinarsi.

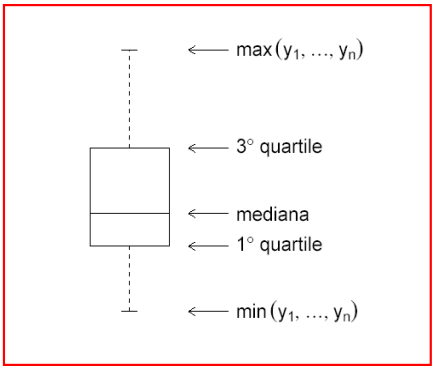
ANALISI DI UNA DISTRIBUZIONE

Quantili

- Generalizzano la mediana.
- L'idea alla base di un **quantile-p** dove $p \in [0,1]$ è di cercare un numero che sia più grande del $100 \times p\%$ dei dati osservati e più piccolo del restante $100 \times (1-p)\%$. Ad esempio, un quantile–0,1 deve essere un valore che lascia a sinistra il 10% delle osservazioni ed a destra il restante 90%. Si osservi che, per costruzione, $\hat{F}(y_p) \approx p$ dove con $\hat{F}(\cdot)$ abbiamo indicato la funzione di ripartizione empirica.
- I quantili con p uguale a 0,25, 0,50 e 0,75 vengono chiamati rispettivamente il primo, il secondo e il terzo **quartile**. Dividono la popolazione in quattro parti uguali. Si osservi che il 2° quartile coincide con la mediana. I quantili con $p = 0{,}01, \ldots, 0{,}99$ si chiamano **percentili**.

Diagrammi a scatola con baffi

- Il nome deriva dall'inglese (*box and whiskers plot* spesso, anche in italiano, abbreviato in *boxplot*).
- Forniscono una idea schematica di un insieme di dati basata sui quantili. Sono costituiti, come dice il nome, da una *scatola* e da due *baffi* costruiti in accordo al disegno sottostante.



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

ANALISI DI UNA DISTRIBUZIONE

- **Scarto Medio Assoluto (dalla Media):** è la media degli scarti, presi in valore assoluto, dalla media

$$\frac{\sum_i |x_i - M|}{n}$$

  - l'idea è quella di valutare se i dati si distribuiscono vicino o lontano dalla media calcolando tutti gli scarti e poi facendone la somma
  - gli scarti vengono presi in valore assoluto per evitare che scarti positivi e negativi si annullino
  - ricordiamo infatti che una delle proprietà della media aritmetica è proprio quella di annullare la somma degli scarti:

$$\sum_i (x_i - M) = 0$$

- Come si determina:
  è opportuno predisporre uno schema di calcolo come questo ->

i	x(i)	x(i)-M	|x(i)-M|
1	19	-5,25	5,25
2	20	-4,25	4,25
3	21	-3,25	3,25
4	22	-2,25	2,25
5	23	-1,25	1,25
6	24	-0,25	0,25
7	25	0,75	0,75
8	40	15,75	15,75
Totale	194	0,00	33,00
Media	24,25	0,00	4,13

ANALISI DI UNA DISTRIBUZIONE

- **Scarto Medio Assoluto dalla Mediana:** è la media degli scarti, presi in valore assoluto, dalla mediana

$$\frac{\sum_i |x_i - Me|}{n}$$

  - l'idea è quella di valutare se i dati si distribuiscono vicino o lontano dalla mediana, calcolando tutti gli scarti e poi sommandoli
  - gli scarti vengono presi in valore assoluto, per evitare che scarti positivi e negativi si compensino (anche se gli scarti dalla mediana non si annullano)

- Come si determina:
  - si determina la mediana: Me = 22,5
  - poi si prepara uno schema di calcolo simile al precedente ->

i	x(i)	x(i)-Me	|x(i)-Me|
1	19	-3,50	3,50
2	20	-2,50	2,50
3	21	-1,50	1,50
4	22	-0,50	0,50
5	23	0,50	0,50
6	24	1,50	1,50
7	25	2,50	2,50
8	40	17,50	17,50
Totale	194	14,00	30,00
Mediana	22,50	1,75	3,75

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

ANALISI DI UNA DISTRIBUZIONE

- **Scarto Medio Assoluto (dalla Media)** con dati in tabella di frequenze :

$$\frac{\sum_i |x_i - M| n_i}{n}$$

  - quando i dati sono organizzati in una tabella di frequenze, gli scarti in valore assoluto devono essere ponderati con le rispettive frequenze
  - ricordiamo che la proprietà della media aritmetica, nel caso ponderato diventa : $\sum_i (x_i - M) n_i = 0$

- Schema di calcolo:

x(i)	n(i)	x(i) n(i)	x(i)-M	[x(i)-M] n(i)	|x(i)-M| n(i)
19	300	5700	-2,76	-829,41	829,41
20	340	6800	-1,76	-600,00	600,00
21	250	5250	-0,76	-191,18	191,18
22	200	4400	0,24	47,06	47,06
23	150	3450	1,24	185,29	185,29
24	180	4320	2,24	402,35	402,35
25	200	5000	3,24	647,06	647,06
26	80	2080	4,24	338,82	338,82
Totale	1700	37000		0,00	3241,18
Media		21,76			1,91

ANALISI DI UNA DISTRIBUZIONE

- **Scarto Medio Assoluto dalla Mediana** con dati in tabella di frequenze

$$\frac{\sum_i |x_i - Me| n_i}{n}$$

  - quando i dati sono organizzati in una tabella di frequenze, gli scarti in valore assoluto devono essere ponderati con rispettive frequenze

- Schema di calcolo: per prima cosa occorre determinare la mediana, poi procediamo in modo del tutto analogo al precedente

x(i)	n(i)	f(i)	F(i)	x(i)-Me	[x(i)-Me] n(i)	|x(i)-Me| n(i)
19	300	0,18	0,18	-2,00	-600,00	600,00
20	340	0,20	0,38	-1,00	-340,00	340,00
21	250	0,15	0,52	0,00	0,00	0,00
22	200	0,12	0,64	1,00	200,00	200,00
23	150	0,09	0,73	2,00	300,00	300,00
24	180	0,11	0,84	3,00	540,00	540,00
25	200	0,12	0,95	4,00	800,00	800,00
26	80	0,05	1,00	5,00	400,00	400,00
Totale	1700	1,00			1300,00	3180,00
Mediana	21					1,87

ANALISI DI UNA DISTRIBUZIONE

- **Varianza:** è la media del quadrato degli scarti dalla media aritmetica

$$\sigma^2 = V(x) = \frac{\sum_i (x_i - M)^2}{n}$$

  - l'idea è sempre quella di valutare quanto i dati si distribuiscono vicino o lontano dalla Media attraverso gli scarti
  - nel caso della Varianza, per evitare che scarti positivi e negativi si compensino, vengono elevati al quadrato
  - la quantità a numeratore prende il nome di Devianza: la Varianza dunque è data dalla Devianza divisa per n

$$Var(x) = \frac{\sum_i (x_i - M)^2}{n} = \frac{Dev(x)}{n}$$

- Schema di calcolo: ->

i	x(i)	x(i)-M	[x(i)-M]^2
1	19	-5,25	27,56
2	20	-4,25	18,06
3	21	-3,25	10,56
4	22	-2,25	5,06
5	23	-1,25	1,56
6	24	-0,25	0,06
7	25	0,75	0,56
8	40	15,75	248,06
Totale	194	0,00	311,50
Media	24,25	0,00	38,94

ANALISI DI UNA DISTRIBUZIONE

- **Scarto Quadratico Medio (o Deviazione Standard)**: è dato dalla radice quadrata della Varianza, ovvero è la media quadratica degli scarti :

$$\sigma = SQM(x) = \sqrt{V(x)} = \sqrt{\frac{\sum_i (x_i - M)^2}{n}}$$

  - un difetto della Varianza è che risulta espressa in una scala diversa a quella della variabile (e della media): se la variabile è misurata in anni, la Varianza è espressa in anni al quadrato
  - Lo Scarto Quadratico Medio riporta la valutazione della variabilità nella stessa scala della media.

i	x(i)	x(i)-M	[x(i)-M]^2
1	19	-5,25	27,56
2	20	-4,25	18,06
3	21	-3,25	10,56
4	22	-2,25	5,06
5	23	-1,25	1,56
6	24	-0,25	0,06
7	25	0,75	0,56
8	40	15,75	248,06
Totale	194	0,00	311,50
Media	24,25	0,00	38,94

- Schema di calcolo:
  ovviamente, è lo stesso della Varianza:

$$\sigma = \sqrt{V(x)} = \sqrt{38{,}94} = 6{,}24$$

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

ANALISI DI UNA DISTRIBUZIONE

- **Varianza e Scarto Quadratico Medio** con dati in tabella di frequenze

$$\sigma^2 = V(x) = \frac{\sum_i (x_i - \bar{x})^2 n_i}{n} \qquad e \qquad \sigma = \sqrt{V(x)} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 n_i}{n}}$$

  - quando i dati sono organizzati in una tabella di frequenze, gli scarti in valore assoluto devono essere ponderati con le rispettive frequenze

- Schema di calcolo:

$$V(x) = 4{,}8$$

$$\sigma = \sqrt{4{,}8} = 2{,}19$$

x(i)	n(i)	x(i)n(i)	x(i)-M	[x(i)-M]^2	[x(i)-M]^2 n(i)
19	300	5700	-2,76	7,64	2293,08
20	340	6800	-1,76	3,11	1058,82
21	250	5250	-0,76	0,58	146,19
22	200	4400	0,24	0,06	11,07
23	150	3450	1,24	1,53	228,89
24	180	4320	2,24	5,00	899,38
25	200	5000	3,24	10,47	2093,43
26	80	2080	4,24	17,94	1435,02
Totale	1700	37000			8165,88
Media		21,76			4,80

ANALISI DI UNA DISTRIBUZIONE

- **Metodo di calcolo indiretto della Varianza**
- La Varianza è calcolabile come differenza tra la media dei quadrati e il quadrato della media:

$$V(x) = M(x^2) - [M(x)]^2$$

infatti:

$$V(x) = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n} = \frac{\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2}{n} =$$

$$= \frac{\sum x_i^2 - 2\bar{x}\sum x_i + n\,\bar{x}^2}{n} = \frac{\sum x_i^2}{n} - 2\,\bar{x}\,\frac{\sum x_i}{n} + \frac{n\,\bar{x}^2}{n} =$$

$$= M(x^2) - 2\,\bar{x}\cdot\bar{x} + \bar{x}^2 = M(x^2) - 2\,\bar{x}^2 + \bar{x}^2 = M(x^2) - \bar{x}^2$$

- Sfruttando questa relazione, il calcolo della Varianza risulta alquanto semplificato, soprattutto quando si ha a che fare con numeri decimali
- Evita infatti la necessità di calcolare tutti gli scarti dalla Media: si devono calcolare solo i quadrati dei valori osservati della variabile
- Si ottiene anche un risultato più preciso, in particolare se si opera "a mano", cioè con carta e penna (e calcolatrice), per il minor numero di errori di arrotondamento

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

ANALISI DI UNA DISTRIBUZIONE

- Metodo di calcolo indiretto della Varianza con dati in tabella di frequenze

$$V(x) = M(x^2) - [M(x)]^2 = \frac{\sum x_i^2 n_i}{n} - \left(\frac{\sum x_i n_i}{n}\right)^2$$

- Schema di calcolo del metodo indiretto:
  è più semplice, infatti oltre alla media (ponderata), occorre solo predisporre la colonna per calcolare la media dei quadrati (anch'essa ovviamente ponderata)

V(x) = 478,51 – (21,76)^2 = 4,8

x(i)	n(i)	x(i)n(i)	x(i)^2	x(i)^2 n(i)
19	300	5700	361	108300
20	340	6800	400	136000
21	250	5250	441	110250
22	200	4400	484	96800
23	150	3450	529	79350
24	180	4320	576	103680
25	200	5000	625	125000
26	80	2080	676	54080
Totale	1700	37000		813460
Media		21,76		478,51

ANALISI DI UNA DISTRIBUZIONE

- **Proprietà della Varianza** (e dello Scarto Quadratico Medio)
  - è uguale a zero quando tutti i valori sono uguali tra loro
  - cresce al crescere della variabilità, ovvero all'allontanarsi dei valori dalla media: non è limitata superiormente (non ha un max)
  - utilizza tutti i valori, quindi tiene conto anche dei valori estremi
  - i valori estremi incidono maggiormente di quelli prossimi alla media, perché gli scarti vengono elevati al quadrato: quindi uno scarto doppio di un altro pesa il quadruplo (e non il doppio) nel calcolo della varianza

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

ANALISI DI UNA DISTRIBUZIONE

- **Media e Varianza di una Variabile Dicotomica**
- Nel caso di una variabile dicotomica (si/no, accordo/disaccordo, 0/1), l'indice che si calcola per descrivere la distribuzione del fenomeno è semplicemente la **percentuale** di una delle due modalità: es. f(1)=0,65 (ovvero 65%)
- La percentuale, per una variabile dicotomica, è effettivamente la media della variabile: ci convinciamo di questo non appena pensiamo che i *valori* assunti dalla variabile siano 0 e 1:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{0+0+1+0+1+\ldots+1}{n} = \frac{n(1)}{n} = f(1) = \boxed{p}$$

- Per determinare la Varianza, in questo caso conviene usare il metodo di calcolo indiretto:

$$V(x) = M(x^2) - [M(x)]^2 = p - p^2 = \boxed{p\,(1-p)}$$

$$\textit{infatti:} \quad M(x^2) = \frac{\sum x_i^2}{n} = \frac{0^2+0^2+1^2+0^2+1^2+\ldots+1^2}{n} = p$$

ANALISI DI UNA DISTRIBUZIONE

- **Coefficiente di Variazione**: è dato dal rapporto tra lo Scarto Quadratico Medio e la Media

$$CV = \frac{\sigma}{\bar{x}} = \frac{SQM}{M}$$

  - E' un numero *puro*, cioè senza unità di misura.
  - E' utile per confrontare la variabilità di due distribuzioni con medie diverse: infatti lo stesso SQM=12 può indicare una variabilità molto piccola per una distribuzione con M=1000, ma una variabilità enorme per una distribuzione con M=0,8
  - A maggior ragione non possono essere confrontati SQM di distribuzioni di dati aventi diverse unità di misura.
  - Il CV è dunque utile per confrontare la variabilità di due insiemi di dati con scala diversa, o con diverse unità di misura
  - Non è limitato superiormente
  - Non è calcolabile quando M=0 : non si può dividere un numero per 0

- Esempio di calcolo:

  CV = SQM / M = 2,19 / 21,76 = 0,10

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

ANALISI DI UNA DISTRIBUZIONE

Gli studiosi di ecologia considerano la diversificazione delle specie che popolano un certo territorio come una proprietà fondamentale. Infatti, più le specie sono diversificate più è grande il patrimonio genetico e quindi più il sistema sarà capace di adattarsi a cambiamenti di qualsiasi origine. Viceversa, un territorio popolato da una sola specie è intrinsicamente fragile.

L'idea è fondamentalmente quella che è alla base della mutabilità. Si pensi, ad esempio, ad un lago e ai pesci che lo popolano. Se i pesci appartengono tutti alla stessa specie, allora la distribuzione dei pesci tra le varie specie assume la forma prevista dalla tabella di minima mutabilità. Viceversa, se il lago è popolato da più specie di pesci senza una specie particolarmente predominante (ovvero se la popolazione dei pesci è ben diversificata) allora la tabella che ci mostra come i pesci si ripartiscono tra le varie specie si avvicinerà a quella di massima mutabilità.

ANALISI DI UNA DISTRIBUZIONE

- **La Mutabilità**
- La variabilità di una variabile qualitativa è detta anche **mutabilità**
- Gli indici di mutabilità si basano sulle frequenze della distribuzione della variabile
- La situazione con variabilità minima (nulla) è quella in cui tutte le frequenze (100%) sono concentrate su un'unica modalità (la moda)
- A parità di numero di modalità distinte k, la distribuzione con maggiore dispersione è quella *uniforme*, con uguale frequenza per tutte le modalità osservate: cioè con f(i) = 1/k per ogni i

  es. se k = 5 -> f(i) = 1/k = 1/5 = 0,2

Età	n(i)	f(i)
19	0	0,0000
20	0	0,0000
21	0	0,0000
22	2000	1,0000
23	0	0,0000
Totale	2000	1,0000

Età	n(i)	f(i)
19	400	0,2000
20	400	0,2000
21	400	0,2000
22	400	0,2000
23	400	0,2000
Totale	2000	1,0000

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

ANALISI DI UNA DISTRIBUZIONE

- **Indice di Mutabilità di Gini:**

$$G = \sum_{i=1}^{k} f_i \,(1 - f_i\,) = 1 - \sum_{i=1}^{k} f_i^2$$

infatti:

$$G = \sum_{i=1}^{k} f_i \,(1 - f_i\,) = \sum_{i=1}^{k} (f_i\, - f_i^2) \;=\; \sum_{i=1}^{k} f_i\, - \sum_{i=1}^{k} f_i^2 \;=\; 1 - \sum_{i=1}^{k} f_i^2$$

- In assenza di variabilità, tutte le frequenze sono concentrate su un'unica modalità (la moda), quindi: k=1, f(1)=1 e quindi l'indice G vale: G = 1 – f(1)² = 0
- Più la distribuzione è dispersa, ovvero al crescere del numero di modalità distinte k, maggiore è il valore dell'indice
- A parità di k, la distribuzione con maggiore dispersione è quella *uniforme*, che presenta uguale frequenza (1/k) per tutte le modalità osservate.
  In tale situazione l'indice assume il suo valore massimo, pari a:

$$\max(G) = 1 - \sum_{i=1}^{k} \frac{1}{k^2} = 1 - k\,\frac{1}{k^2} = 1 - \frac{1}{k} = \frac{k-1}{k}$$

ANALISI DI UNA DISTRIBUZIONE

- **Indice di Mutabilità di Gini normalizzato:**

$$\tilde{G} = \frac{G}{\max(G)} = G / \frac{k-1}{k} = \frac{k}{k-1} G$$

- E' compreso tra 0 e 1: infatti si tratta di una quantità divisa per il valore massimo che tale quantità può assumere

- Es. di calcolo: k = 16

$$G = \sum_{i=1}^{k} f_i \ (1 - f_i \ ) = 1 - \sum_{i=1}^{k} f_i^2$$

G = 1 - 0,11 = 0,89
~
G = 0,89 * 16/15 = 0,95

Età	n(i)	f(i)	f(i)^2
19	350	0,1750	0,0306
20	300	0,1500	0,0225
21	250	0,1250	0,0156
22	200	0,1000	0,0100
23	150	0,0750	0,0056
24	180	0,0900	0,0081
25	200	0,1000	0,0100
26	80	0,0400	0,0016
27	130	0,0650	0,0042
28	60	0,0300	0,0009
29	25	0,0125	0,0002
30	50	0,0250	0,0006
31	10	0,0050	0,0000
32	5	0,0025	0,0000
~~33~~	~~0~~	~~0,0000~~	~~0,0000~~
34	5	0,0025	0,0000
35	5	0,0025	0,0000
Totale	2000	1,0000	0,1100

ANALISI DI UNA DISTRIBUZIONE

- **Indice di Entropia:**

$$\tilde{H} = \frac{H}{\max(H)} = \frac{-\sum_{i=1}^{k} f_i \ \log(f_i \ )}{\log(k)}$$

- A numeratore troviamo H, l'*Entropia* della distribuzione di frequenze: in fisica è un indice del *disordine* in un sistema (caos) , nella teoria dell'informazione è una misura di informazione o incertezza
- A denominatore troviamo il max che H può assumere in caso di massima dispersione: risulta pari al logaritmo del numero di modalità distinte k
- E' una misura compresa tra 0 e 1 (0=nessuna variabilità, 1=max variabilità)
- Schema di calcolo:
  utilizzando il log in base 2 abbiamo:
  ~
  H = - (-3,3919)/log(16) = 3,3919/4 = 0,85

Età	n(i)	f(i)	log f(i)	f(i) log f(i)
19	350	0,1750	-2,5146	-0,4401
20	300	0,1500	-2,7370	-0,4105
21	250	0,1250	-3,0000	-0,3750
22	200	0,1000	-3,3219	-0,3322
23	150	0,0750	-3,7370	-0,2803
24	180	0,0900	-3,4739	-0,3127
25	200	0,1000	-3,3219	-0,3322
26	80	0,0400	-4,6439	-0,1858
27	130	0,0650	-3,9434	-0,2563
28	60	0,0300	-5,0589	-0,1518
29	25	0,0125	-6,3219	-0,0790
30	50	0,0250	-5,3219	-0,1330
31	10	0,0050	-7,6439	-0,0382
32	5	0,0025	-8,6439	-0,0216
~~33~~	~~0~~	~~0,0000~~	~~-~~	~~-~~
34	5	0,0025	-8,6439	-0,0216
35	5	0,0025	-8,6439	-0,0216
Totale	2000	1,0000		-3,3919

ANALISI DI UNA DISTRIBUZIONE

- **Proprietà dell'Entropia (H):**

$$H = -\sum_{i=1}^{k} f_i \ \log(f_i)$$

- L'Entropia (H) assume il suo valore massimo quando la dispersione (disordine) dei dati è massima, quindi nel caso di distribuzione *uniforme*, quando si registra uguale frequenza (1/k) per tutte le (k) modalità osservate:

$$\max(H) = -\sum_{i=1}^{k} \underbrace{\frac{1}{k}\log\frac{1}{k}}_{\text{costante rispetto ad i}} = -k\frac{1}{k}\log\frac{1}{k} = -\log\frac{1}{k} \underset{\substack{\uparrow \\ \text{proprietà del log}}}{=} -(-\log k) =$$

$$= \log k$$

LOGARITMO

- Il logaritmo in base b (b > 0) di un numero reale x, si definisce come l'esponente da dare a b per ottenere x:

$$y = \log_b x \quad \Leftrightarrow \quad b^y = x \quad \textit{ovvero}: \quad x = b^{\log_b x}$$

- Le basi più comunemente utilizzate sono 2, 10 ed **e** = 2,7183... (numero di Nepero): il logaritmo in base **e** è detto neperiano o anche *naturale*

- Proprietà (valgono qualunque sia la base del log):

$$\log_b b = 1$$

$$\log 1 = 0$$

$$\log(x\ y) = \log x + \log y$$

$$\log(x + y) \neq \log x + \log y$$

$$\log(x^k) = k \log x$$

$$\log(1/x) = \log(x)^{-1} = -\log x$$

$$\log(x/y) = \log x - \log y$$