Università di Ferrara
fondata nel 1391

Elementi di statistica medica

I dati e la loro rappresentazione

Dati ordinati in ranghi

- In alcune situazioni, una serie di osservazioni è disposta in ordine decrescente in relazione alla grandezza.
- A ciascuna osservazione viene assegnato un numero che corrisponde alla specifica posizione nella sequenza.
- Esempio: cause di morte in Italia.

Dati ordinati in ranghi

Tassi standardizzati per sesso e causa	Maschi	
CAUSE DI MORTE	Numero decessi	
	2000	2002
2) Tumori	91.853	94.139
Tumori maligni dello stomaco	*6.492*	6.312
Tumori maligni del colon, retto e ano	*8.807*	*9.216*
Tumori maligni della trachea, bronchi e polmoni	*25.503*	*26.370*
7) Diabete mellito	6.998	7.034
8) Malattie del sistema nervoso	5.922	6.100
1) Malattie del sistema circolatorio	107.411	106.615
Infarto del miocardio	*20.072*	*21.671*
Disturbi circolatori dell'encefalo	*27.380*	*26.385*
3) Malattie dell'apparato respiratorio	21.904	19.763
6) Malattie dell'apparato digerente	12.980	12.611
5) Cause accidentali e violente	15.774	15.406
4) Altre cause	16.547	16.460
TOTALE	279.389	278.129

Dati discreti

- Sono importanti sia l'ordine che la grandezza.
- I numeri sono realmente misurabili e possono assumere solo valori specifici che differiscono per quantità fisse e che escludono valori intermedi.
- Il risultato di un'operazione aritmetica su due valori discreti non è necessariamente un valore discreto.

Dati discreti

Incidenti e persone infortunate secondo la conseguenza, per anno e mese - Anno 2007

ANNI	Totale incidenti			Incidenti mortali		
	N	Morti	Feriti	N	Morti	Feriti
2004	243.490	6.122	343.179	5.548	6.122	4.710
2005	240.011	5.818	334.858	5.271	5.818	4.096
2006	238.124	5.669	332.955	5.178	5.669	4.189
2007	230.871	5.131	325.850	4.718	5.131	3.741

Dati continui

- I dati che rappresentano quantità misurabili, ma che non si limitano ad assumere determinati valori (come i numeri interi) sono noti come dati continui.
- In tutti i casi sono possibili valori frazionari.
- Avendo senso il misurare la distanza fra due osservazioni, è possibile applicare operazioni aritmetiche.
- Il solo fattore limitante per un'osservazione continua è il **grado di accuratezza**.

Dati continui

Tab. 1 Situazione attuale emissioni delle 2 centrali policombustibili del polo chimico, torce, autobotti (a) – Flussi di massa, espressi in ton/anno

	CTE1 (b)	CTE2 (c)	Torce e Autob. (d)	Totale
Macroinquinanti				
NOx	67,5	921,6	121,1	1110,2
CO	33,8	460,8	120,4	615,0
SOx	229,5	2027,5	0,1	2257,1
Particolato filtr.le (PM10) (e)	6,8	92,2	0,4	99,4
Particolato cond.le (f)	n.m.	n.m.	n.m.	n.m.
Microinquinanti (g)				
Metalli pesanti (As-Ba-Be-Cd-Cr-Co-Cu-Mn-Mo-Ni-Pb-Se-V-Zn)		590,7		
Sostanze organiche				
Formaldeide		119,6		
Toluene		22,5		
Xileni		0,4		
Acetaldeide		n.r.		
Etilbenzene		0,2		
Benzene		0,8		
Naftalene		4,1		
IPA		4,0		
Particolato totale (filtrabile + condensabile)		252,4		

Fonte: HERA, Ferrara

La matrice dei dati

• I dati codificati di una rilevazione statistica, effettuata su **n** unità statistiche con riferimento a **p** variabili, sono raccolti in una tabella definita "**matrice dei dati**"

N.	Sesso	Titolo di studio	Età (anni)	Peso (Kg)	N. ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
…	…	…	…	…	…
n	F	Diploma	60	55	6

La matrice dei dati

Ogni riga rappresenta una **unità statistica**.

N.	Sesso	Titolo di studio	Età	Peso Kg	N. ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
…	…	…	…	…	…
n	F	Diploma	60	55	6

La matrice dei dati

Ogni colonna rappresenta una **variabile** .

N.	Sesso	Titolo di studio	Età (anni)	Peso (Kg)	N. ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
…	…	…	…	…	…
n	F	Diploma	60	55	6

La matrice dei dati

La matrice dei dati contiene tutte le informazioni analitiche relative a ciascuna unità statistica.

• La visualizzazione della matrice dei dati (soprattutto quando i dati sono molti – molte unità statistiche / molte variabili) non consente di cogliere immediatamente gli aspetti salienti del fenomeno che si è interessati a studiare.

• Occorre, perciò, effettuare una **sintesi**.

Sintesi dei dati

Ruolo della statistica nella metodologia della ricerca

- Fasi nello studio di una POPOLAZIONE :
  - **Schematizzazione**
  - **Osservazione**
  - **Descrizione**

Schematizzazione

- La schematizzazione, consiste nella definizione del fenomeno, nell'individuazione della collettività in cui esso si realizza e nella scelta delle caratteristiche della collettività che interessano

Per esempio: se si intende studiare la propensione al consumo di alcolici (= il fenomeno) nella popolazione europea per la fascia di età compresa tra i 14 e i 30 anni (= collettività) si potrebbe decidere di rilevare la spesa media per alcolici in una settimana (= caratteristica di interesse)

Osservazione

- L'**osservazione** è la raccolta, l'ordinamento e la classificazione del materiale di osservazione.
- Al termine di questa fase, alla collettività di individui si sostituisce un insieme di dati, o **rappresentazione dei dati**, su cui si può operare con procedimenti matematici.

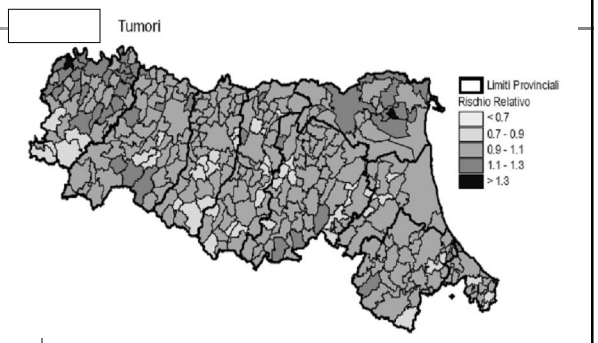
Descrizione

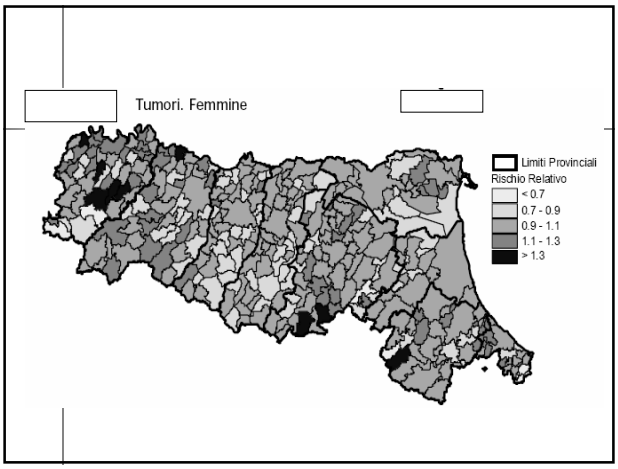
- Nella fase di descrizione vengono impiegati appositi indici per descrivere il fenomeno studiato.
- Esempio: frequenza di morte per tumore in Emilia-Romagna.

Numero assoluto dei decessi, tassi grezzi e tassi standardizzati di mortalità calcolati con metodo di standardizzazione diretto (x 100.000 abitanti) per AUSL di residenza relativi all'anno 2004 (popolazione di riferimento: RER 1998). Totale

Azienda di residenza	Totale morti	Tasso grezzo	Tasso standardizzato	Errore standard
Piacenza	1.122	409,93	368,14	11,04
Parma	1.569	379,74	359,76	9,13
Reggio Emilia	1.444	296,53	315,29	8,34
Modena	1.977	299,61	313,32	7,08
Bologna	3.006	366,49	337,43	6,18
Imola	400	322,42	311,00	15,60
Ferrara	1.450	414,55	365,94	9,68
Ravenna	1.235	338,02	308,98	8,83
Forlì	648	365,25	335,21	13,23
Cesena	537	277,01	292,29	12,69
Rimini	869	302,86	327,48	11,16
Regione	*14.257*	*343,43*	*332,46*	*2,8*

Mappe di mortalità per comune. Livello di rischio relativo di morte stimato rispetto alla media regionale. Mortalità generale. Periodo 1998-2003





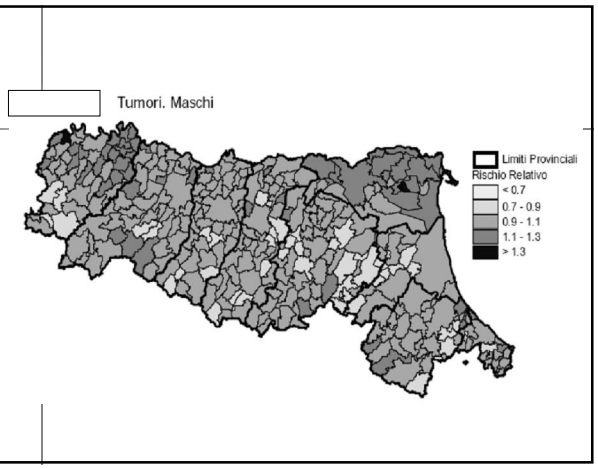


Tabelle di contingenza o frequenza

•Le tabelle di contingenza sono tabelle utilizzate in statistica per rappresentare e analizzare le **relazioni tra due o più variabili**.

•In esse si riportano le **frequenze** congiunte delle variabili.

•Riportano le classi o categorie di valori, insieme alle **frequenze assolute** (numero degli elementi) entro ciascuna categoria, **frequenze relative** (percentuali), **frequenze cumulate** assolute e relative (percentuali).

FREQUENZA ASSOLUTA

- La frequenza assoluta (della classe) è il numero di unità che appartengono ad una classe.

Classi di peso	Numero soggetti (FREQUENZA ASSOLUTA)
<60 kg	2
61 - 70 kg	7
71 - 80 kg	12
> 80 kg	4
TOTALE	25

FREQ. RELATIVA E FREQ. PERCENTUALE

Classi di peso	FREQUENZA ASSOLUTA	FREQUENZA RELATIVA	FREQUENZA PERCENTUALE
<60 kg	2	=2/25 =0,08	8%
61 - 70 kg	7	=7/25 =0,28	28%
71 - 80 kg	12	=12/25 =0,48	48%
> 80 kg	4	=4/25 =0,16	16%
TOTALE	25	=25/25 =1,00	100%

Tabelle di contingenza o frequenza

Il caso più semplice è quello delle **tabelle tetracoriche**, in cui ciascuna delle due variabili assume solo due possibili valori, ad esempio:

Colore degli occhi\Colore dei Capelli	Biondi	Non Biondi	Totale
Chiari	21	19	40
Non chiari	9	51	60
Totale	30	70	100

Rappresentazione dei dati

VARIABILE QUALITATIVA RILEVATA CON SCALA NOMINALE: **TIPOLOGIA FAMILIARE**

Principali tipologie familiari 1995 (in migliaia)	frequenze assolute	frequenze relative	frequenze percentuali
Persone sole	4275	0.205	20.5%
Genitore solo con figli	1689	0.081	8.1%
Coppie senza figli	4338	0.208	20.8%
Coppie con figli	9948	0.477	47.7%
Altre famiglie	605	0.029	2.9%
Totale	20855	1.000	100.0%

Rappresentazione dei dati

Principali tipologie familiari 1995 (in migliaia)	frequenze assolute	frequenze relative	frequenze percentuali
Persone sole	4275	0.205	20.5%
Genitore solo con figli	1689	0.081	8.1%
Coppie senza figli	4338	0.208	20.8%
Coppie con figli	9948	0.477	47.7%
Altre famiglie	605	0.029	2.9%
Totale	20855	1.000	100.0%

$$\text{frequenze relative} = \frac{\text{frequenze assolute}}{\text{totale}}$$

$$\text{frequenze percentuali} = \text{frequenze relative} \times 100$$

Rappresentazioni grafiche

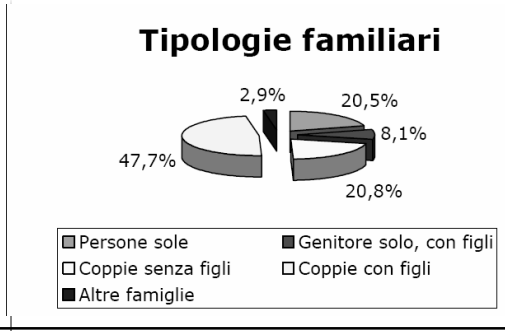
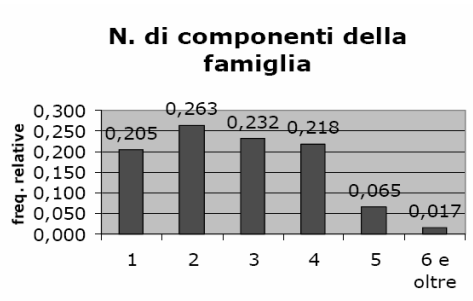


Tabelle di frequenza

Famiglie per numero di componenti 1995 (in migliaia)				
Componenti	f_i	f_i %	F_i	F_i %
1	4281	20,5%	4281	20,5%
2	5493	26,3%	9774	46,8%
3	4845	23,2%	14619	70,0%
4	4553	21,8%	19172	91,8%
5	1358	6,5%	20530	98,3%
6 e oltre	355	1,7%	20885	100,0%
Totale	**20885**	**100,0%**		

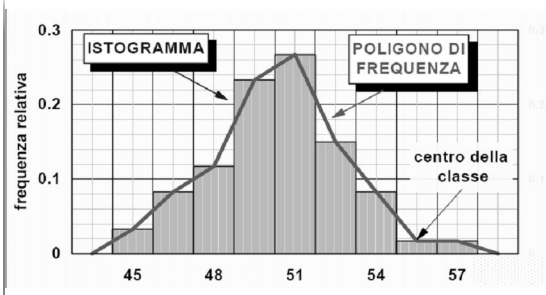
Istogramma



Istogrammi e poligoni di frequenza

- Negli istogrammi e nei poligoni di frequenza **le frequenze sono proporzionali all'area** (delimitata dalla linea spezzata che li costituisce ed inclusa tra due valori reali sull'asse orizzontale) **e non all'altezza della figura**.
- Ovviamente, quando le classi hanno tutte la stessa ampiezza, l'area è proporzionale anche all'altezza.
- I valori riportati sull'asse verticale indicano la densità di frequenza per una prefissata ampiezza di classe.

Istogramma



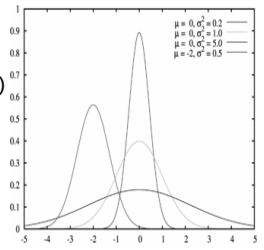
Istogrammi e poligoni di frequenza

- Riportando sulle ordinate la frequenza relativa ed usando un'unità di misura unitaria per l'intervallo di classe sulle ascisse, l'area sottesa dall'istogramma e dal poligono di frequenza è PARI A 1 (la frequenza relativa cumulata è pari a 1 all'ultima classe).
- La probabilità ha valori numerici compresi tra 0 e 1 (estremi compresi): corrispondenza tra il valore numerico dell'area sottesa dall'istogramma ed il valore numerico della probabilità.

Curva di Gauss

- La variabile casuale Normale (detta anche variabile casuale Gaussiana, curva di Gauss, Campana di Gauss, curva degli errori, curva a campana, ogiva) è una variabile casuale continua con due parametri, indicata tradizionalmente con:

$$N(\mu;\sigma^2)$$



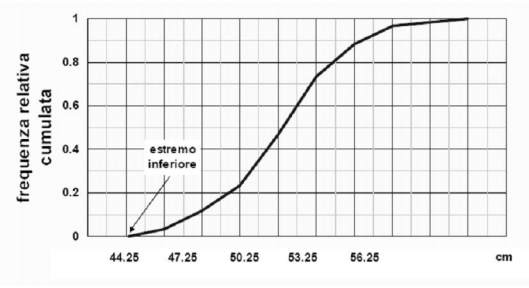
- Si tratta di una delle più importanti variabili casuali.

• Karl Friedrich Gauss descrisse la Normale studiando il moto dei corpi celesti. Altri la usavano per descrivere fenomeni anche molto diversi come i colpi di sfortuna nel gioco d'azzardo o la distribuzione dei tiri attorno ai bersagli. Da qui i nomi curva di Gauss e curva degli errori.

• Fu Francis Galton a intuire che la curva in questione, da lui detta anche ogiva, poteva essere applicata a fenomeni anche molto diversi, e non solo ad "errori". Questa di idea di curva per descrivere i "dati" in generale portò ad usare il termine Normale, in quanto rappresentava uno substrato normale ovvero la norma per qualsiasi distribuzione presente in natura.

• Nel tentativo di confrontare curve diverse, Galton - in mancanza di strumenti adeguati - si limitò ad usare due soli parametri: la **media** e la **varianza** , dando così inizio alla statistica parametrica

Ogiva di Galton / CURVA SIGMOIDE



Ruolo della statistica nella metodologia della ricerca

Fasi nello studio di **PARTE** di una popolazione:

- Schematizzazione
- Formulazione di ipotesi
- Osservazione
- Descrizione
- Inferenza

Inferenza

- L'inferenza (o induzione) è quell'insieme di procedure che fa risalire dalla descrizione del campione a quella dell'insieme più ampio (popolazione) e che permette una verifica delle ipotesi formulate.
- Inferire è quindi trarre una conclusione. Inferire X significa concludere che X è vero; un'inferenza è la conclusione tratta da un insieme di fatti o circostanze. Gran parte dello studio della logica esplora la validità o non validità di inferenze e implicazioni.

… E quando qualcuno vi propone di credere a una proposizione voi dovete prima esaminare se essa è accettabile, perché la nostra ragione è stata creata da Dio, e ciò che piace alla nostra ragione non può non piacere alla ragione divina, sulla quale peraltro sappiamo solo quello che, per analogia e spesso per negazione, ne inferiamo dai procedimenti della nostra ragione. …

(Guglielmo da Baskerville in Il nome della rosa, pag. 139, Umberto Eco)