STATISTICA MEDICA

Libri di testo consigliati:

Daniel W., “Biostatistica”, III edizione, EDISES, 2019.

Norman G., Streiner D., Capelli G., d’Abramo G., “Biostatistica”, Casa editrice ambrosiana, Milano, 2000.

Alcune definizioni della statistica

•Un ramo della matematica applicata che si occupa della raccolta e dell'interpretazione dei dati quantitativi e dell'uso della teoria delle probabilità per la stima di parametri di una popolazione.

•Lo studio scientifico dei dati numerici basato sui fenomeni naturali.

•La procedura matematica per descrivere le probabilità e la distribuzione casuale o non-casuale della materia o del verificarsi degli eventi.

•Una serie di teoremi matematici che aiuta ad analizzare i dati attribuendo significatività ai risultati.

•Una raccolta di metodi per raccogliere, organizzare, riassumere, analizzare e interpretare i dati, e per trarre conclusioni basate su di essi.

La scienza è l'arte di raccogliere, riassumere ed analizzare dati soggetti a variazione casuale (Biology Online)

METODOLOGIA CLINICA

Necessita di:

- **Quantificazione**
- **Formalizzazione matematica**

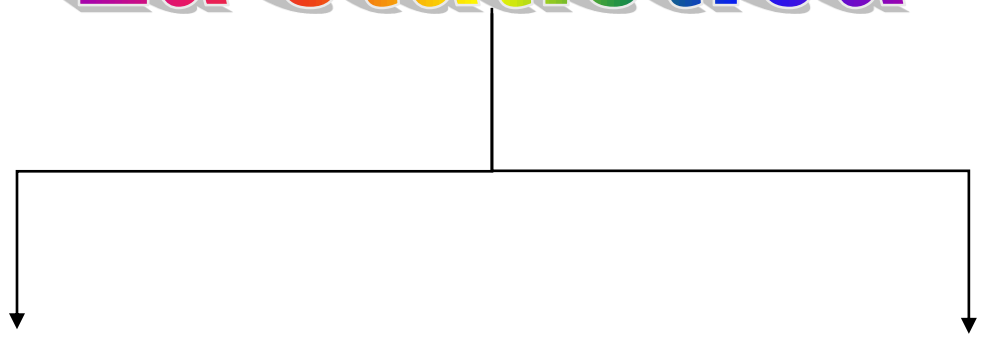
EPIDEMIOLOGIA

Ha come oggetto lo studio della distribuzione delle malattie in una popolazione e dei fattori che la influenzano e fornisce i dati che sono di guida al procedimento clinico

STATISTICA

E' il mezzo oggettivo per la pianificazione delle indagini e l'interpretazione dei risultati

La statistica



Descrittiva

Si occupa della presentazione e sintesi dei dati

Inferenziale

Permette di trasferire le informazioni ottenute su un campione all'intera popolazione

La **variabile** è ciò che viene osservato o misurato e può assumere uno tra una serie definita di possibili ***valori***

FONTI DI DATI EPIDEMIOLOGICI E STATISTICI

Le fonti di dati correnti si basano su un sistema misto con finalità generali, di carattere socio- demografiche, economiche e sanitarie

La raccolta dei dati avviene a livello locale, regionale e nazionale con modalità e frequenza di rilevazione specifiche in base alle quali si possono distinguere:

- ***fonti universali e continue*** **(mortalità e sue cause, infortuni sul lavoro, SDO…)**
- ***fonti universali e sporadiche*** **(censimenti)**
- ***fonti campionarie e continue*** **(registri di patologia, registri tumori….)**
- ***fonti campionarie e sporadiche*** **(indagini osservazionali, sperimentali)**

SPECULARE

UNIVERSO ← PARAMETRI

PROGRAMMARE

Campionamento

Disegno Sperimentale

Stima di intervallo

Test di ipotesi

DECIDERE

CAMPIONE → STATISTICHE

DESCRIVERE

LE FASI DELL'INDAGINE STATISTICA

PIANIFICAZIONE

Formulazione obiettivi - Analisi letteratura - Identificazione variabili - Scelta della metodologia - Campionamento

↓

RILEVAZIONE

Acquisizione delle informazioni

↓

ELABORAZIONE

Sintesi dei dati grezzi in dati derivati

↓

PRESENTAZIONE

Esposizione dei risultati in tabelle, grafici, indici,…

↓

INTERPRETAZIONE

Spiegazione dei risultati

La pianificazione di uno studio

Fase preliminare

- Formulazione obiettivi
- Analisi della letteratura
- Definire l'unità statistica
- Definire l'unità di rilevazione
- Identificazione variabili e sistemi di rilevazione
- Identificazione dei confondenti

Campionamento

- Definizione della popolazione obiettivo
- Scelta del campione
- Determinazione campionaria

Metodologia

- Scelta della tipologia di studio
- Scelta dei metodi di rilevazione
- Identificazione risorse disponibili e necessarie
- Autorizzazioni
- Valutazione tempi di esecuzione

Studio pilota

- Addestramento del personale
- Test del questionario
- Accettabilità dei partecipanti

Rilevazione dei dati

Preparazione dei questionari

- Schede di rilevazione
- Codifiche

Elaborazione dei dati

Sintesi dei dati grezzi

- Determinazioni di indici, funzioni, modelli

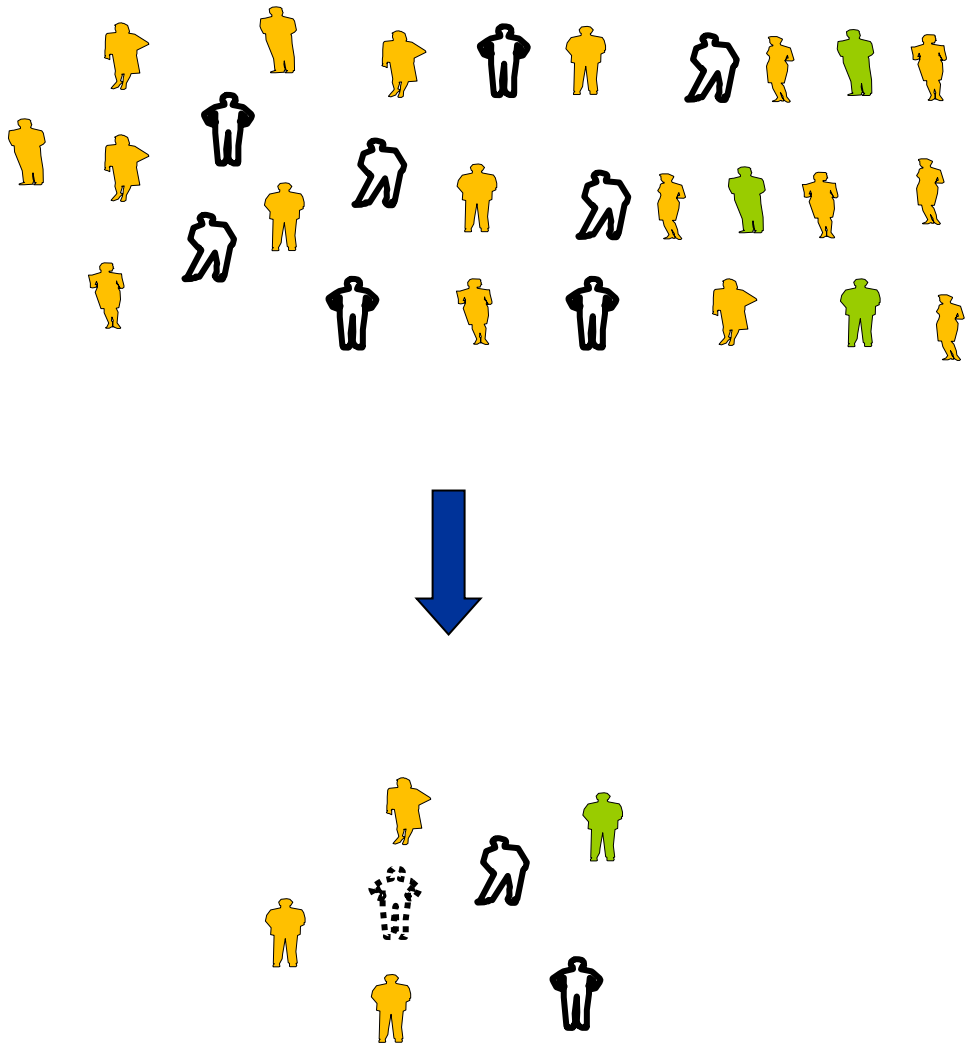
Presentazione dei risultati

Esposizione

- Presentazione dei risultati usando indici, grafici, tabelle

Interpretazione

- Spiegazione e discussione dei risultati ottenuti

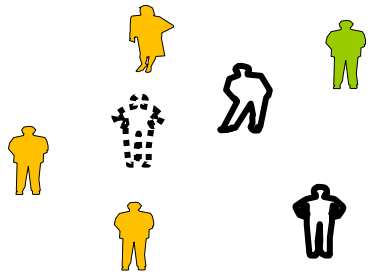


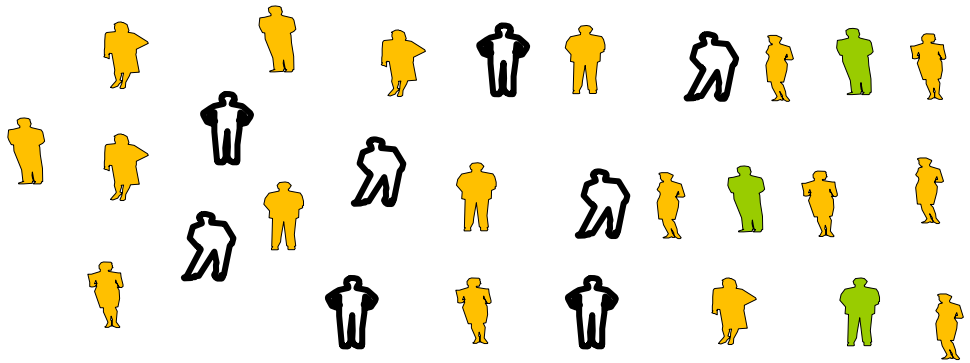
Quanto ciò che rileviamo su un campione rispecchia ciò che avviene nella popolazione?
Oppure:
Con che probabilità le misure rilevate sul campione

stime

sono i veri valori della popolazione

parametri





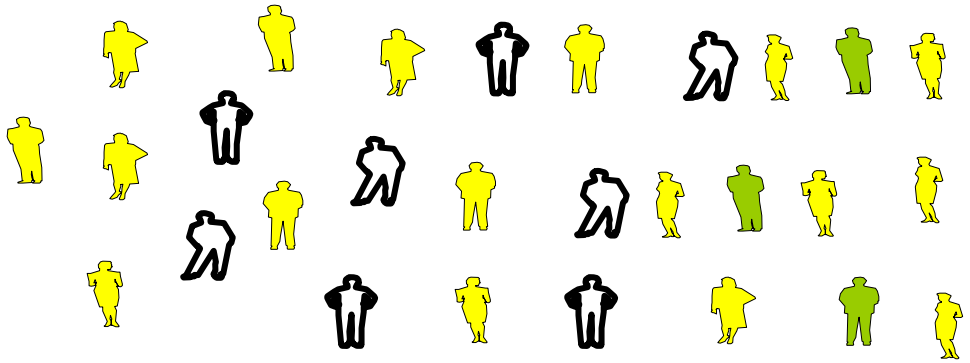
Media
Varianza
Conteggi
Percentuali
…

Stime

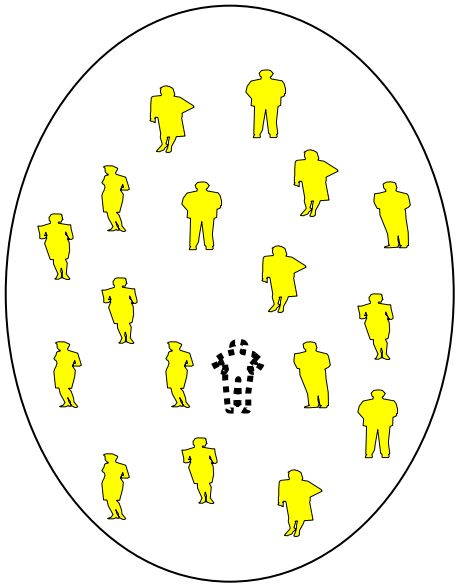
Leggi della probabilità
Statistica inferenziale

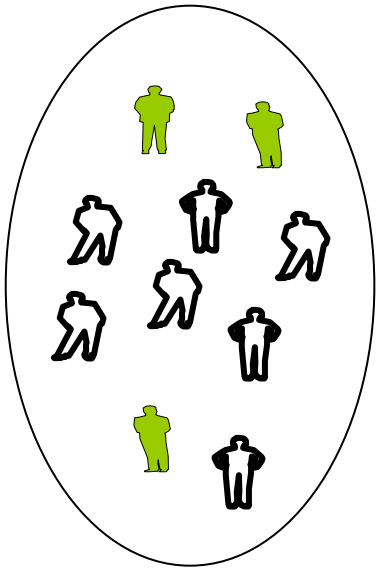
Parametri

Il campionamento



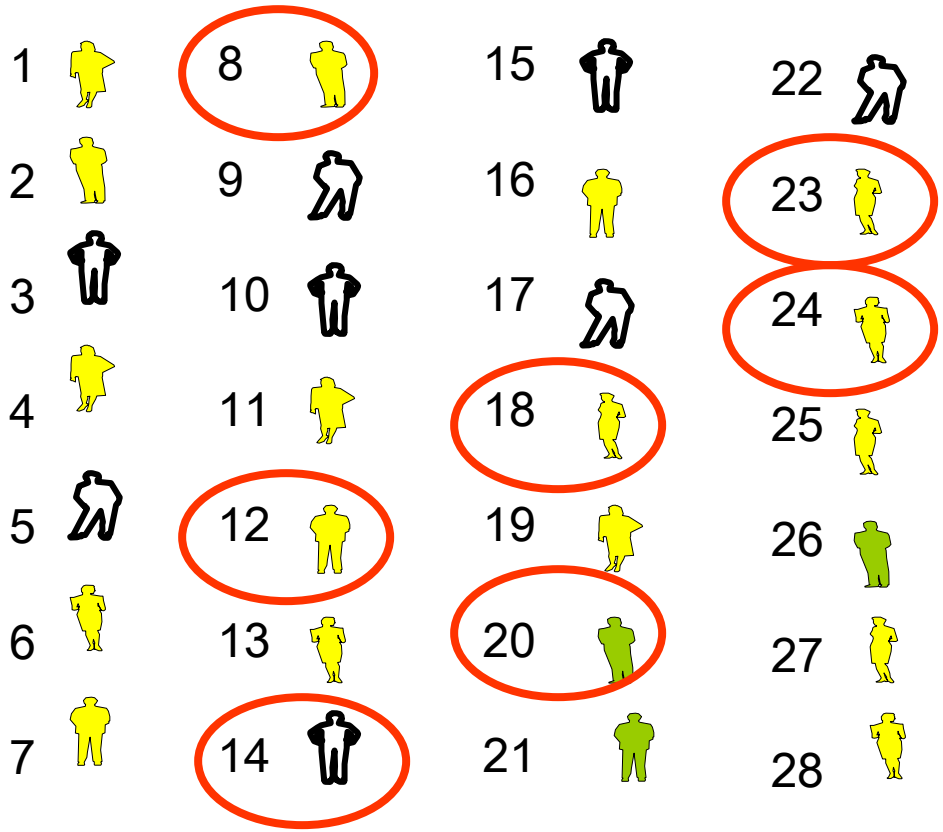
Consiste nel prelevare dalla popolazione un gruppo di elementi di **adeguata** numerosità e **completezza** rispetto alle caratteristiche della popolazione





Il campione casuale semplice

La caratteristica principale di questa tecnica è determinata dal fatto che **tutti gli elementi della popolazione hanno la stessa probabilità di entrare a far parte del campione**



Si deve disporre di un elenco ordinato e numerato

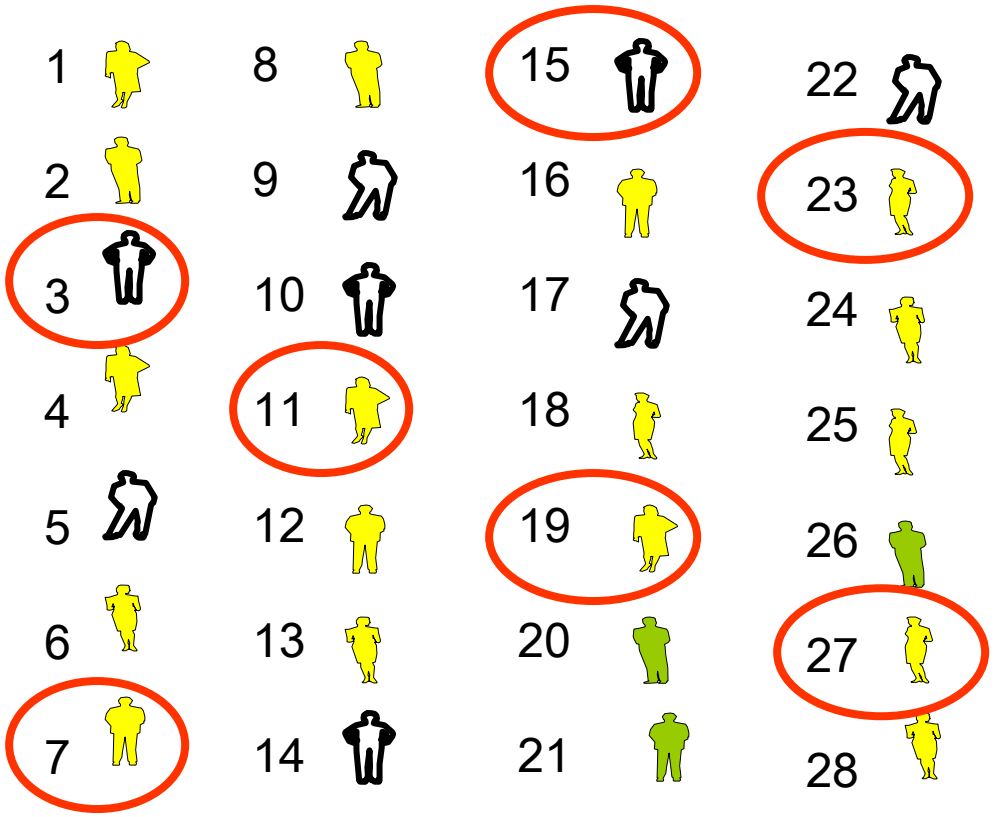
Si estraggono tanti numeri quanti sono gli elementi da campionare

Si selezionano gli individui identificati da quel numero

8 18 23 12 14 20 24

Il campione sistematico

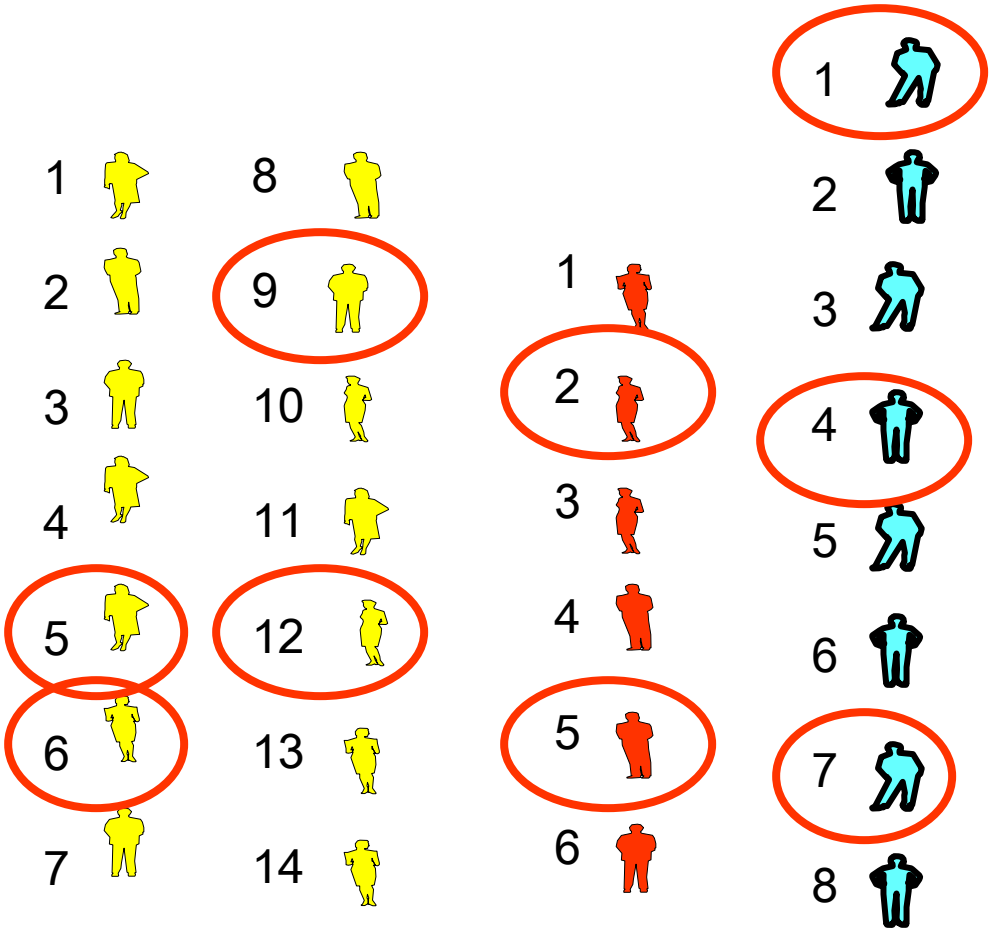
Il campione si costituisce procedendo con l'estrazione degli elementi secondo un intervallo regolare.



Bisogna determinare la **frazione di campionamento**:
Supponiamo che il campione deve essere di 7 elementi (n), allora 28/7=4 da cui FC=N/n.
Devo estrarre 1 soggetto ogni 4.
Estraggo in maniera casuale un numero tra 1 e 4, ad esempio 3: questo è il punto di partenza.
Il campione sarà costituito da: 3, 7, 11, 15, 19, 23, 27

Il campione stratificato

Consiste nel suddividere gli elementi di una popolazione in più sottogruppi omogenei ed estrarre un campione casuale semplice da ogni sottogruppo. Questo campionamento consente una maggiore precisione delle stime.



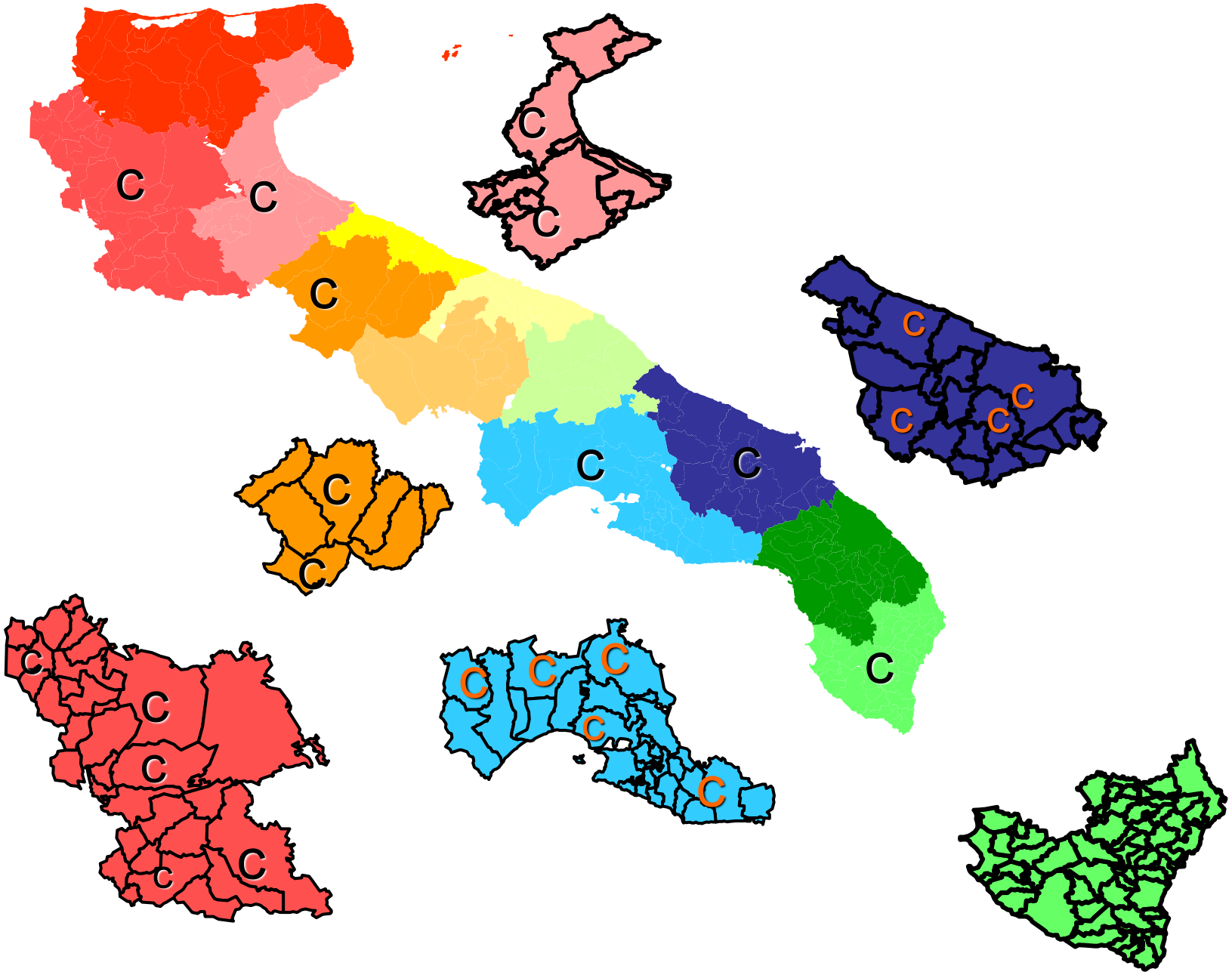
Il campione a stadi

Consiste nel giungere alla costituzione del campione con una procedura di estrazioni casuali che riducono progressivamente la popolazione iniziale.

Supponiamo di estrarre un campione di assisti del servizio sanitario regionale la popolazione di partenza è il totale dei residenti nella regione; si eseguono uno dopo l'altro i seguenti campioni casuali:

1. delle ASL
2. dei comuni che fanno parte delle ASL precedentemente estratte
3. dei quartieri che costituiscono il comune
4. delle vie
5. dei palazzi
6. delle famiglie
7. del soggetto

Questo metodo si usa spesso quando le popolazioni di partenza sono molto grandi e non si dispone di un elenco ordinato degli elementi della popolazione


C
C
C

Variabile: qualunque caratteristica che possa assumere valore diverso in tempi, spazi, persone differenti

Se misurabile ed espressa con un valore numerico
QUANTITATIVA

Continua: misurabile su una scala continua, il valore numerico dipende dalle caratteristiche dello strumento di misura adottato (peso, pressione arteriosa, statura, età)

Discreta: assume solo valori interi (frequenza cardiaca, numero di sigarette, numero di carie)

Se espressa con un aggettivo o sostantivo
QUALITATIVA

Nominale: il valore della variabile è espresso da un aggettivo o da un sostantivo (sesso, presenza/assenza di una cardiopatia, gruppo sanguigno)

Ordinale: il valore della variabile è espresso da aggettivi o sostantivi in cui è possibile riconoscere un criterio di ordinamento (classe di scompenso, scala del dolore)

Variabili quantitative continue

- Peso (kg)
- Pressione arteriosa (mmHg)
- Età (anni)
- Statura (cm)

Variabili quantitative discrete

- Frequenza cardiaca (batt./min)
- Numero di sigarette
- Numero di carie

Variabili qualitative

- Sesso
- Presenza / assenza di cardiopatia dilatativa
- Classe di scompenso NYHA
- Gruppo sanguigno (A, AB, B, 0)

Scale di misura

- **Nominale**: la caratteristica viene espressa da un attributo, senza una effettiva misurazione (gruppo sanguigno, sesso…)

- **Ordinale**: la caratteristica pur esprimendo una qualità consente un ordinamento dei risultati (indici dello stato di salute)

- **Rango**: gli elementi possono essere ordinati in base alla grandezza delle osservazioni, per cui è possibile costruire una graduatoria (scala dell'ansia etc…)

- **Quantitativa discreta**: la misura è effettuata solo per numeri interi

- **Quantitativa continua**: la misura è effettuata per tutti i possibili valori in un intervallo continuo

GUARDARE I DATI

Verificare la validità dei dati inseriti:
Presenza di valori troppo bassi (outlyer)
Presenza di valori troppo alti (outlyer)
Presenza di valori mancanti (missing data: celle vuote, codice 9 o 99 o 999, simbolo *, simbolo “-”, simbolo “.”)
Concentrazione di osservazioni su pochi valori
Etc…

Studio della distribuzione di frequenza della variabile sia quantitativa che qualitativa

Tabella di frequenza

N	
Test entrata	Totale
0	161
0,09	2
0,11	1
0,13	4
0,16	1
0,18	2
0,2	6
0,23	4
0,25	1
0,27	5
0,29	1
0,32	3
0,36	1
0,39	1
0,41	1
0,43	1
0,45	1
0,46	1
0,48	2
0,52	4
0,55	1
0,64	6
0,66	2
0,69	3
0,75	1
0,78	2
0,85	1
0,96	1
1,01	3
1,05	1
1,35	1
1,38	1
1,51	1
1,54	1
1,60	1
Totale complessivo	229

Si vuole valutare il livello di alcol ematico mediante il test all'etilometro. Di lato si osservano i singoli valori osservati.

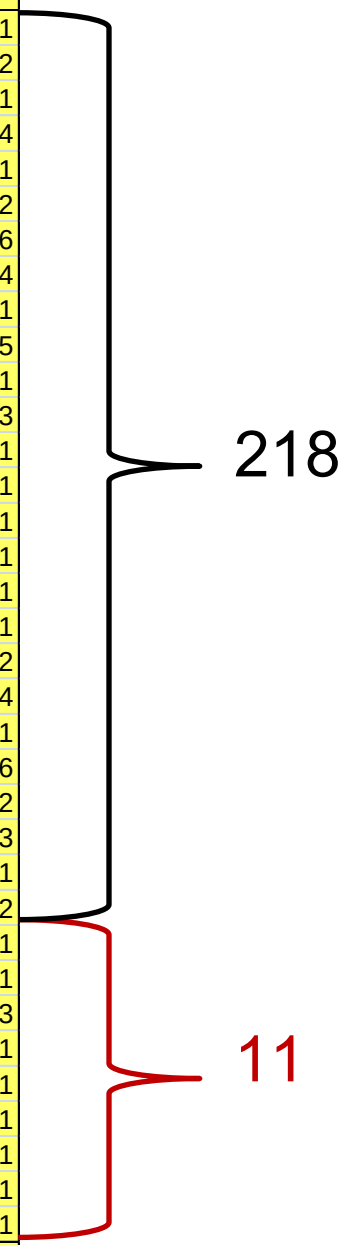
Per sintetizzare ed esporre i risultati con una tabella conviene renderla più piccola, determinando delle "classi di alcolemia".

Se si conoscono uno o più valori di "cut-off" si possono utilizzare questi per determinare le classi:

Es. Guida con tasso alcolemico superiore a 0,8g/l
→ sospensione patente per almeno 6 mesi

Si possono scegliere 2 classi:
-Alcolemia < 0,8 g/l
-Alcolemia > 0,8 g/l

N	
Test entrata	Totale
0	161
0,09	2
0,11	1
0,13	4
0,16	1
0,18	2
0,2	6
0,23	4
0,25	1
0,27	5
0,29	1
0,32	3
0,36	1
0,39	1
0,41	1
0,43	1
0,45	1
0,46	1
0,48	2
0,52	4
0,55	1
0,64	6
0,66	2
0,69	3
0,75	1
0,78	2
0,85	1
0,96	1
1,01	3
1,05	1
1,35	1
1,38	1
1,51	1
1,54	1
1,60	1
Totale complessivo	229



Livello di alcol	Frequenza assoluta
0-0,8	218
0,81-1,6	11
Totale	229

Puglia

Livello di alcol	Frequenza assoluta
0-0,8	218
0,81-1,6	11
Totale	229

Basilicata

Livello di alcol	Frequenza assoluta
0-0,8	385
0,81-1,6	15
Totale	400

Dove si beve di più ?

Puglia

Livello di alcol	Frequenza assoluta
0-0,8	218
0,81-1,6	11
Totale	229

Basilicata

Livello di alcol	Frequenza assoluta
0-0,8	385
0,81-1,6	15
Totale	400

Dove si beve di più ?

Livello di alcol	Frequenza assoluta	Frequenza relativa
0-0,8	218	0,952
0,81-1,6	11	0,048
Totale	229	1

Livello di alcol	Frequenza assoluta	Frequenza relativa
0-0,8	385	0,963
0,81-1,6	15	0,038
Totale	400	1

N	
Test entrata	Totale
0	161
0,09	2
0,11	1
0,13	4
0,16	1
0,18	2
0,2	6
0,23	4
0,25	1
0,27	5
0,29	1
0,32	3
0,36	1
0,39	1
0,41	1
0,43	1
0,45	1
0,46	1
0,48	2
0,52	4
0,55	1
0,64	6
0,66	2
0,69	3
0,75	1
0,78	2
0,85	1
0,96	1
1,01	3
1,05	1
1,35	1
1,38	1
1,51	1
1,54	1
1,60	1
Totale complessivo	229

Si vuole valutare il livello di alcol ematico mediante il test all'etilometro. Di lato si osservano i singoli valori osservati.

Per sintetizzare ed esporre i risultati con una tabella conviene renderla più piccola, determinando delle "classi di alcolemia".

Per determinare il numero di classi (K) può essere utile la formula di Sturges:
$K=1+3,322(\text{Log } n)$
Dove n è la numerosità campionaria.

L'ampiezza (w) della classe sarà data da:
$W=R/K$
Dove R è il range ossia la differenza tra il valore più grande e quello più piccolo presenti nei dati

Per determinare il numero di classi (K) si usa la formula di Sturges:

$$K=1+3{,}322(\text{Log } n)=1+3{,}322(\text{Log } 229)=8{,}8$$

dove n è la numerosità campionaria.

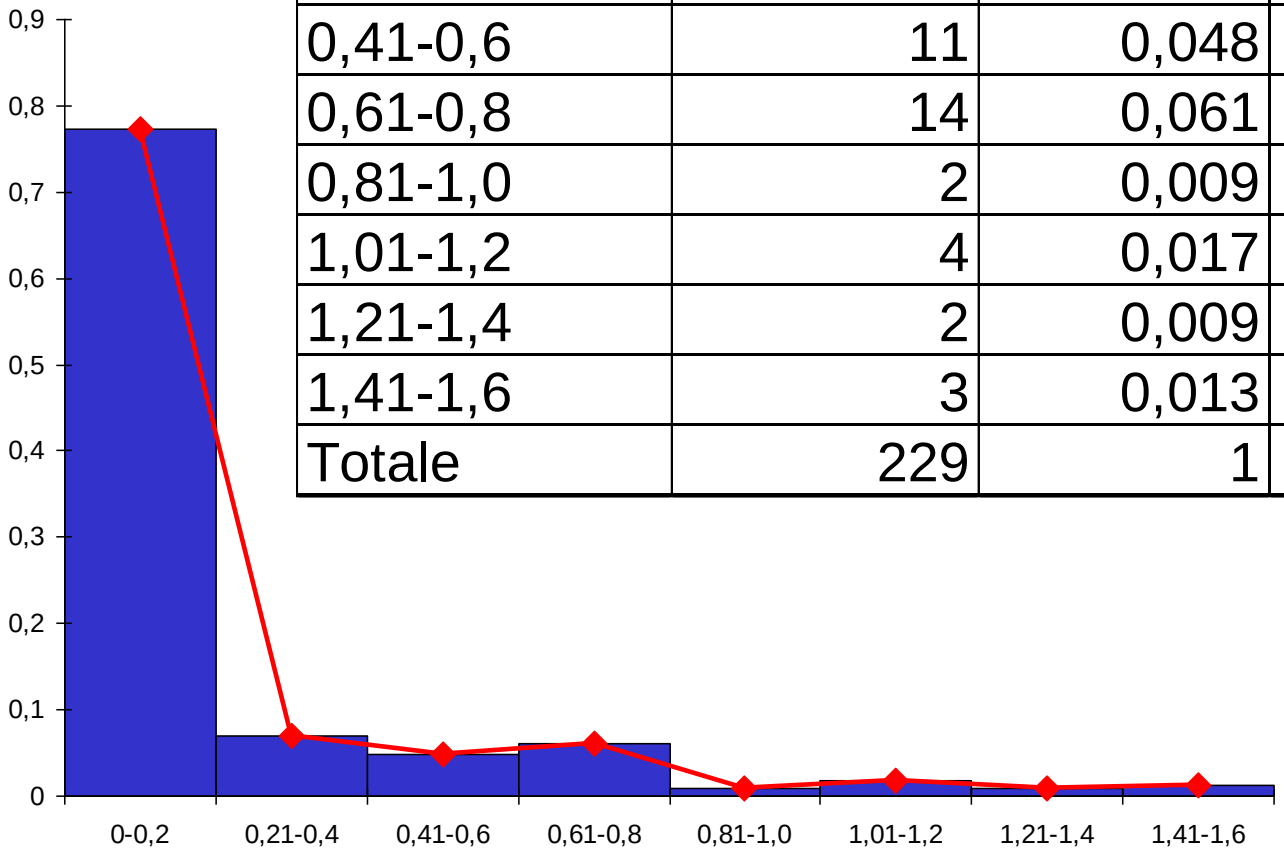
l'ampiezza (w) della classe sarà data da:

$$W= R / K=(1.60-0)/8=0.2$$

dove **R è il range** ossia la differenza tra il valore più grande e quello più piccolo

Il risultato della formula di Sturges relativo all'esempio in questione è mostrato in tabella e nel grafico sottostante

Livello di alcol	Frequenza assoluta	Frequenza relativa	Frequenza cumulativa assoluta	Frequenza cumulativa relativa
0-0,2	177	0,773	177	0,773
0,21-0,4	16	0,070	193	0,843
0,41-0,6	11	0,048	204	0,891
0,61-0,8	14	0,061	218	0,952
0,81-1,0	2	0,009	220	0,961
1,01-1,2	4	0,017	224	0,978
1,21-1,4	2	0,009	226	0,987
1,41-1,6	3	0,013	229	1,000
Totale	229	1		



DESCRIVERE I DATI

INDICI SINTETICI:

Tendenza centrale: media

mediana

moda

Variabilità: varianza - deviazione standard

coefficiente di variazione

range

differenza interquartile (differenza tra 25° e 75° percentile)

I valori degli indici possono essere determinati per strato (qualora vi fossero variabili per le quali abbia senso stratificare) e presentati per mezzo di tabelle

INDICI DI TENDENZA CENTRALE

Media

Aritmetica:
utilizzabile sempre, in particolare quando i dati seguono una distribuzione normale

$$\bar{x} = \frac{\sum x_i}{n}$$

Geometrica:
utilizzata per dati con distribuzione log-normale

$$\bar{x}_g = \frac{\sum \log x_i}{N}$$

$$m_g = \sqrt[n]{\prod x_i}$$

Armonica indicata quando i dati sono relativi alla misura tempo

$$\bar{x}_a = \frac{N}{\sum 1/x_i}$$

Moda:
valore più frequente

Mediana: valore che bipartisce un'insieme di dati ordinati

Determinazione della MEDIANA

Altezza	Frequenza Assoluta	Frequenza Relativa	Frequenza cumulativa Assoluta	Frequenza cumulativa Relativa
168	8	0,03	8	0,03
169	9	0,03	17	0,06
170	14	0,05	31	0,10
171	19	0,06	50	0,17
172	26	0,09	76	0,25
173	37	0,12	113	0,38
174	38	0,13	151	0,50
175	45	0,15	196	0,65
176	39	0,13	235	0,78
177	21	0,07	256	0,85
178	18	0,06	274	0,91
179	26	0,09	300	1,00
Totale	**300**	**1**		

MISURE DI DISPERSIONE

Range — Differenza tra valore massimo e minimo

Differenza interquartile — Differenza tra valore del 75° e del 25° percentile

Varianza — Misura media dei quadrati dello scostamento dei singoli valori dalla media aritmetica

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Deviazione standard

$$\sqrt{S^2}$$

Coefficiente di variazione = C.V. = (Deviazione standard/media)x100

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Gradi di libertà

Ovvero il numero di determinazioni che posso scegliere liberamente meno i vincoli

Statura	Media ($\bar{X}$)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
170	171	-1	1
172		1	1
176		5	25
166		-5	25
Totale		**0**	**52**

$$S^2 = \frac{52}{3} = 17{,}3$$

Vincolo: $\sum (x_i - \bar{x}) = 0$

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Gradi di libertà

Ovvero il numero di determinazioni che posso scegliere liberamente meno i vincoli

Statura	Media ($\bar{X}$)	($X_i - \bar{X}$)
170	171	-1
172		1
176		5
166		-5
Totale		**0**

Vincolo: $\sum(x_i - \bar{x}) = 0$

Statura	Media ($\bar{X}$)	($X_i - \bar{X}$)
171	171	0
178		7
160		-11
?		?
Totale		**0**

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Gradi di libertà

Ovvero il numero di determinazioni che posso scegliere liberamente meno i vincoli

Statura	Media ($\bar{X}$)	($X_i - \bar{X}$)
170	171	-1
172		1
176		5
166		-5
Totale		**0**

Vincolo: $\sum(x_i - \bar{x}) = 0$

Statura	Media ($\bar{X}$)	($X_i - \bar{X}$)
171	171	0
178		7
160		-11
Totale		**0**

+4 !!!!! → Statura = 175

$$\sum \left(x_i - \bar{x} \right) = 0$$

$$\sum \left(x_i - \bar{x} \right) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} =$$

$$= \sum x_i - n \frac{\sum x_i}{n} = 0$$

DESCRIVERE I DATI

RAPPRESENTAZIONI CON GRAFICI

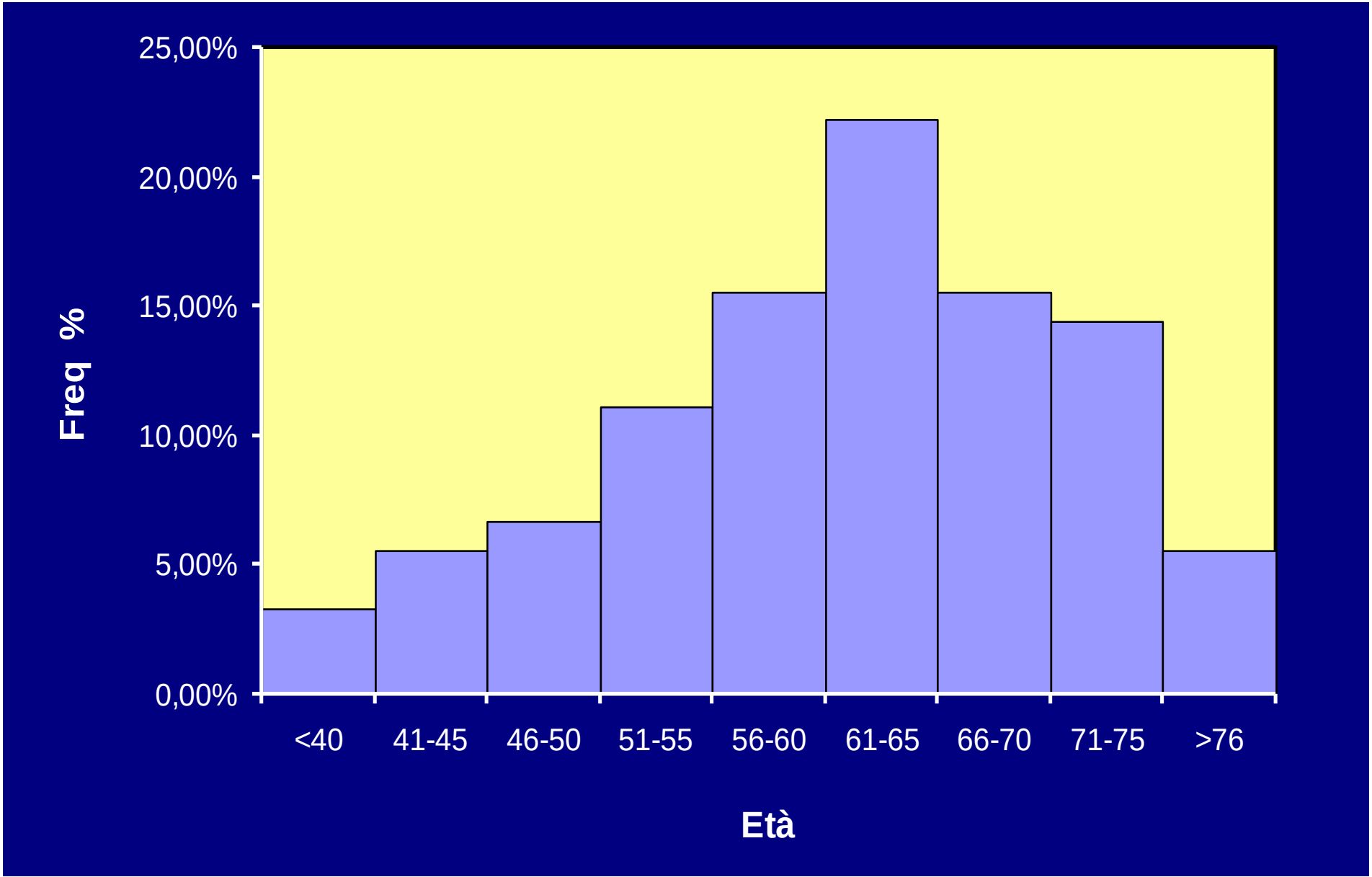
ISTOGRAMMI: sull'asse x c'è la variabile continua suddivisa in classi e sull'asse y la frequenza (più correttamente quella "relativa", cioè la percentuale) con cui quella classe si presenta

Età

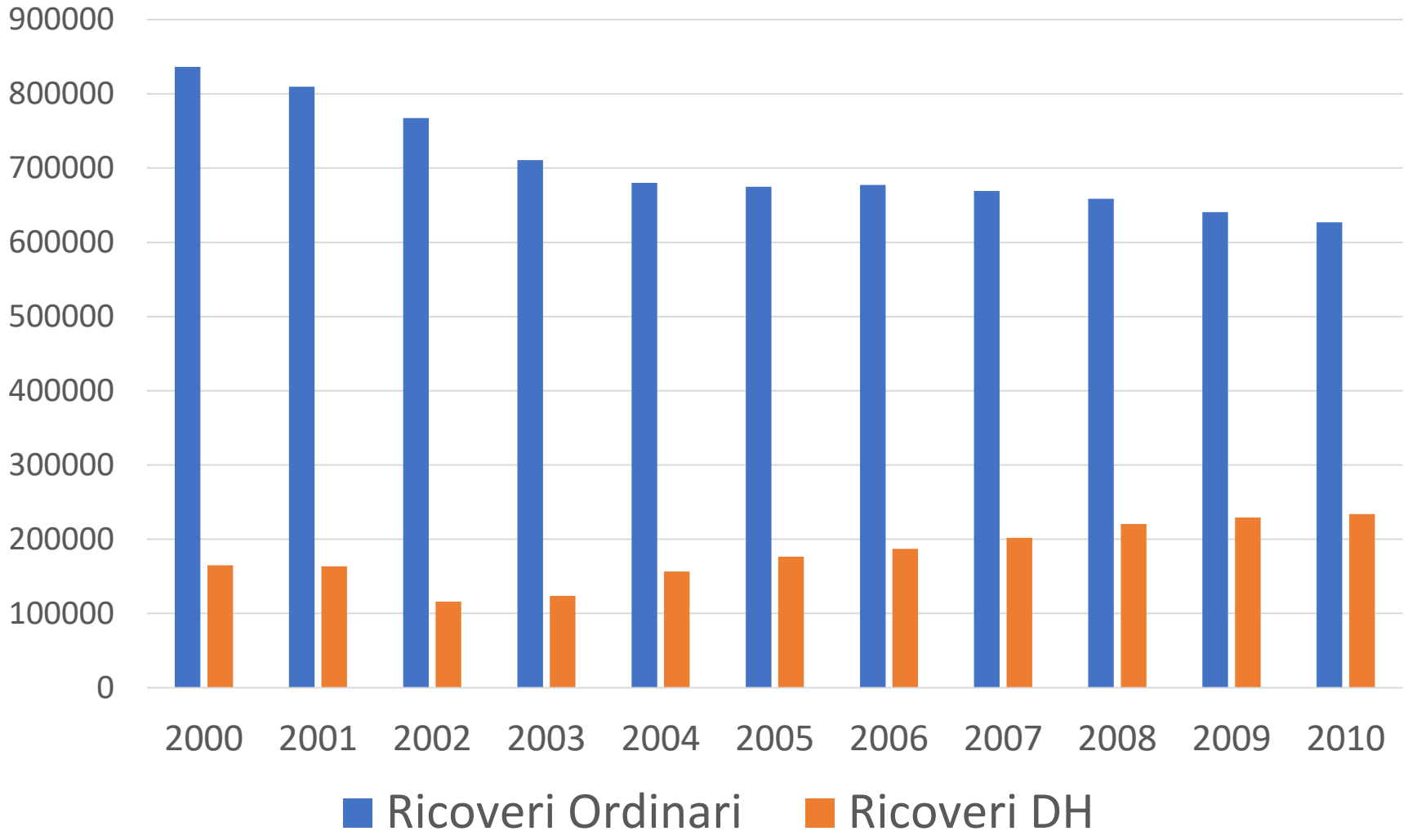
Peso

Frequenza cardiaca

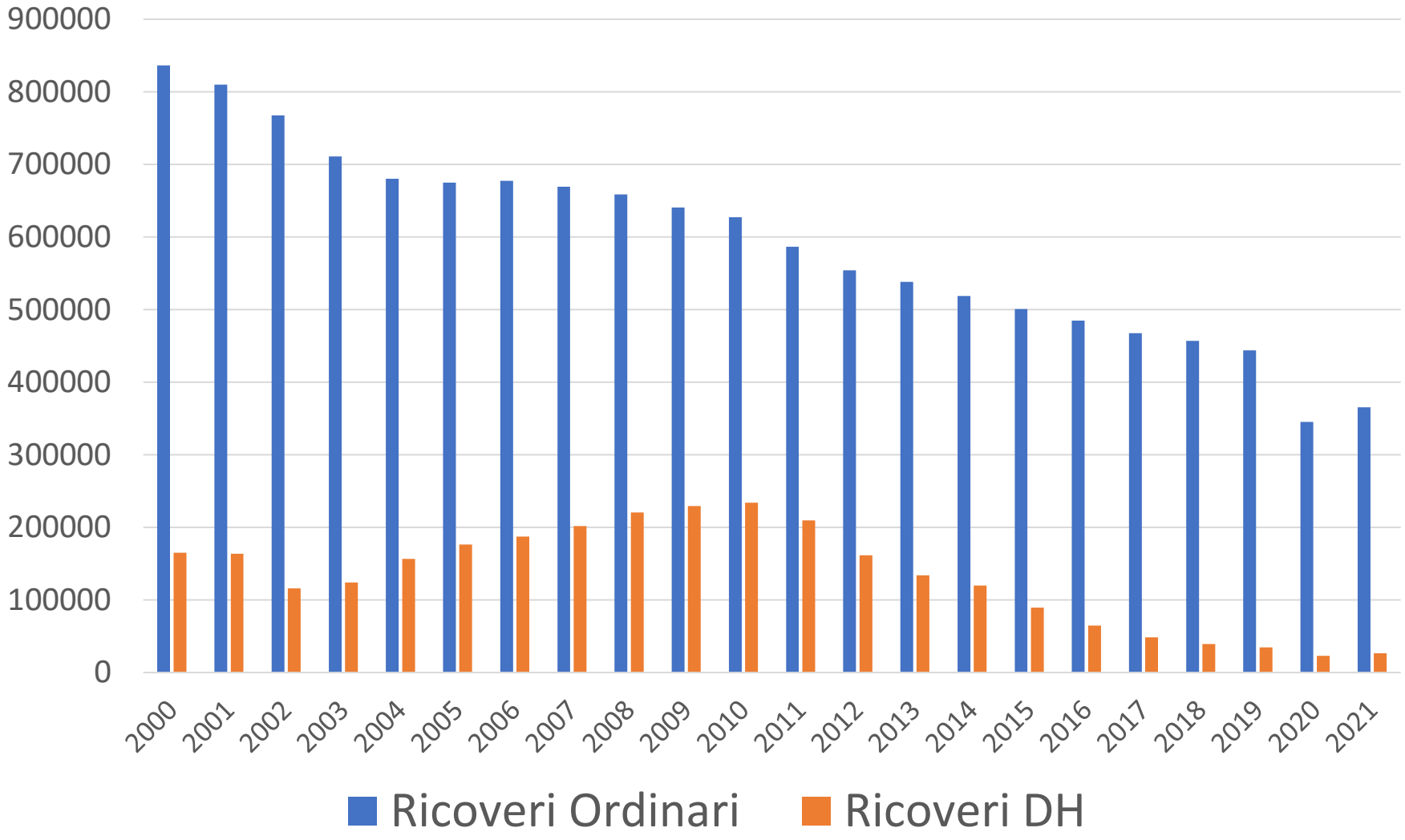
DISTRIBUZIONE DEI PAZIENTI PER CLASSI DI ETA'.



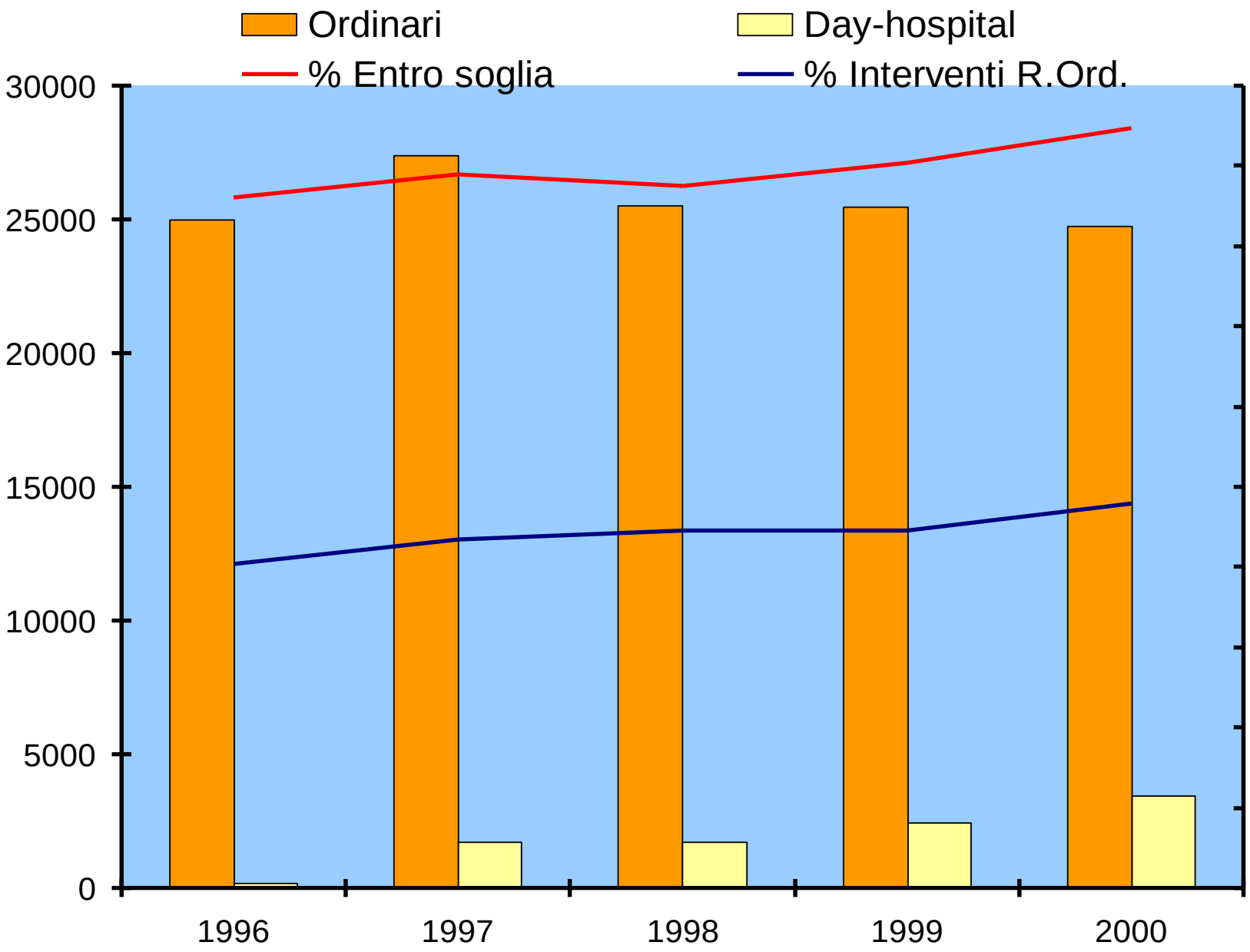
Distribuzione dei ricoveri per tipo di ricovero
Frequenze assolute 2000-2010



Distribuzione dei ricoveri per tipo di ricovero
Frequenze assolute 2000-2021



Distribuzione dei ricoveri nel quinquennio 96-00 e percentuale di ricoveri con DRG chirurgico



DESCRIVERE I DATI

RAPPRESENTAZIONI CON GRAFICI

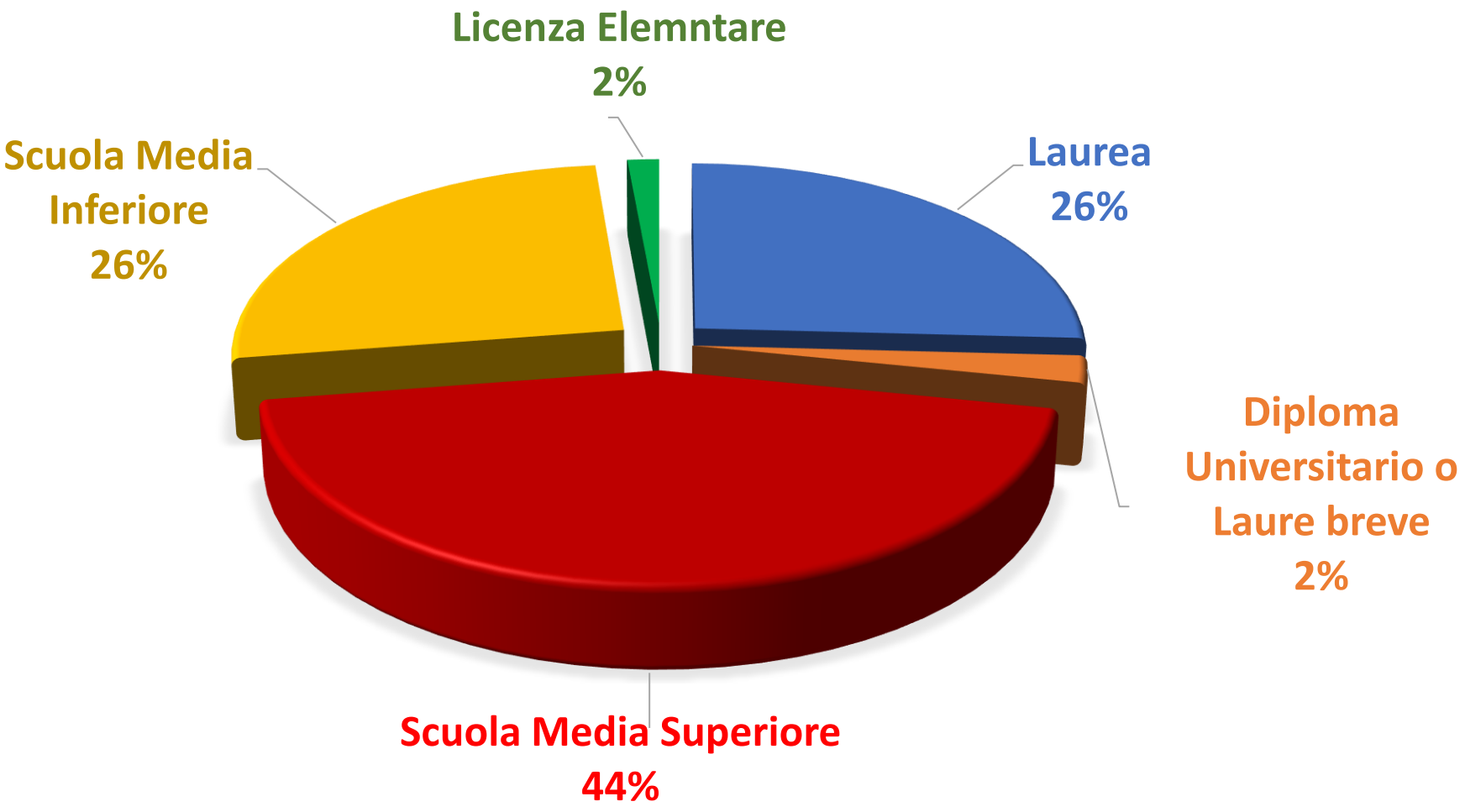
DIAGRAMMI A TORTA: l'intera circonferenza rappresenta il 100%, ciascuno spicchio indica la percentuale con cui si presenta un carattere. E' indicato per le variabili qualitative

Sesso

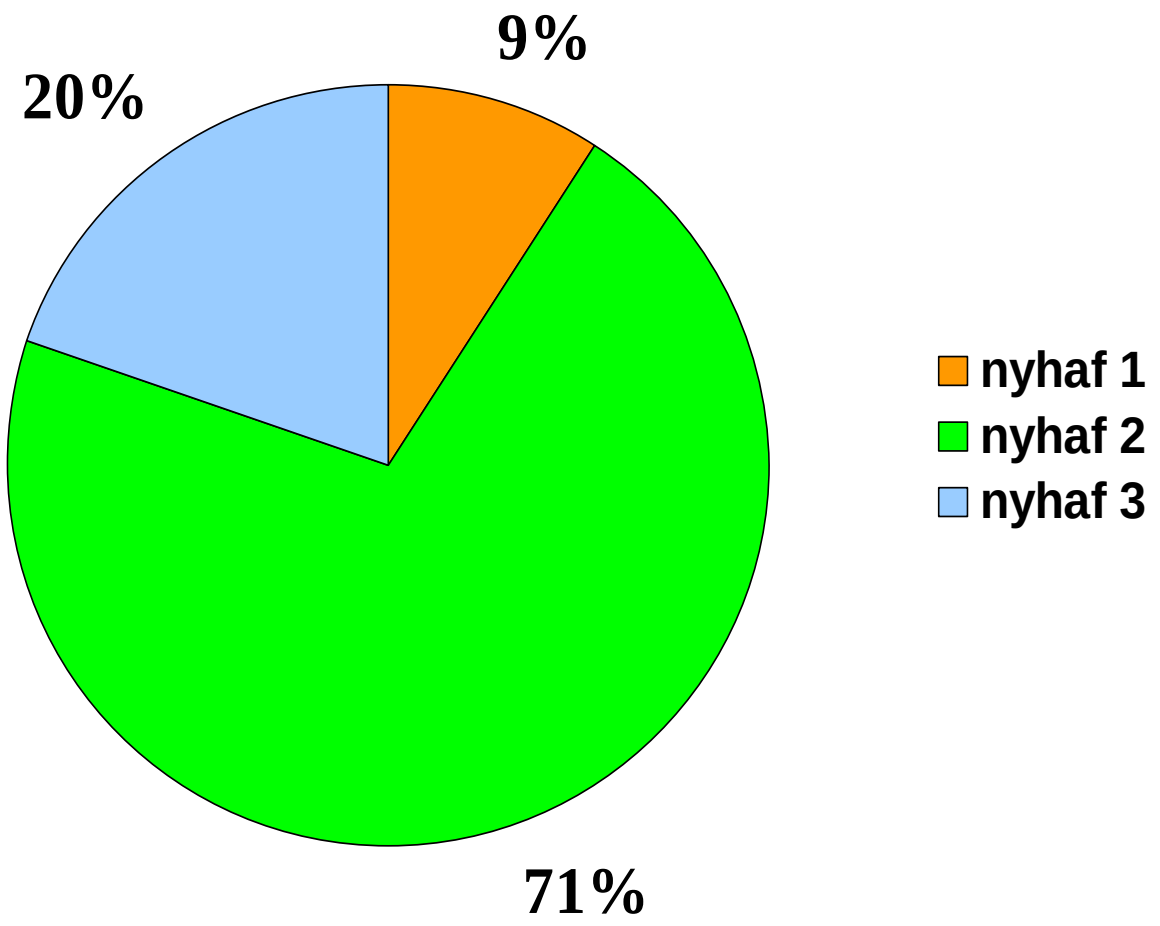
Trattamento

Classe NYHA…

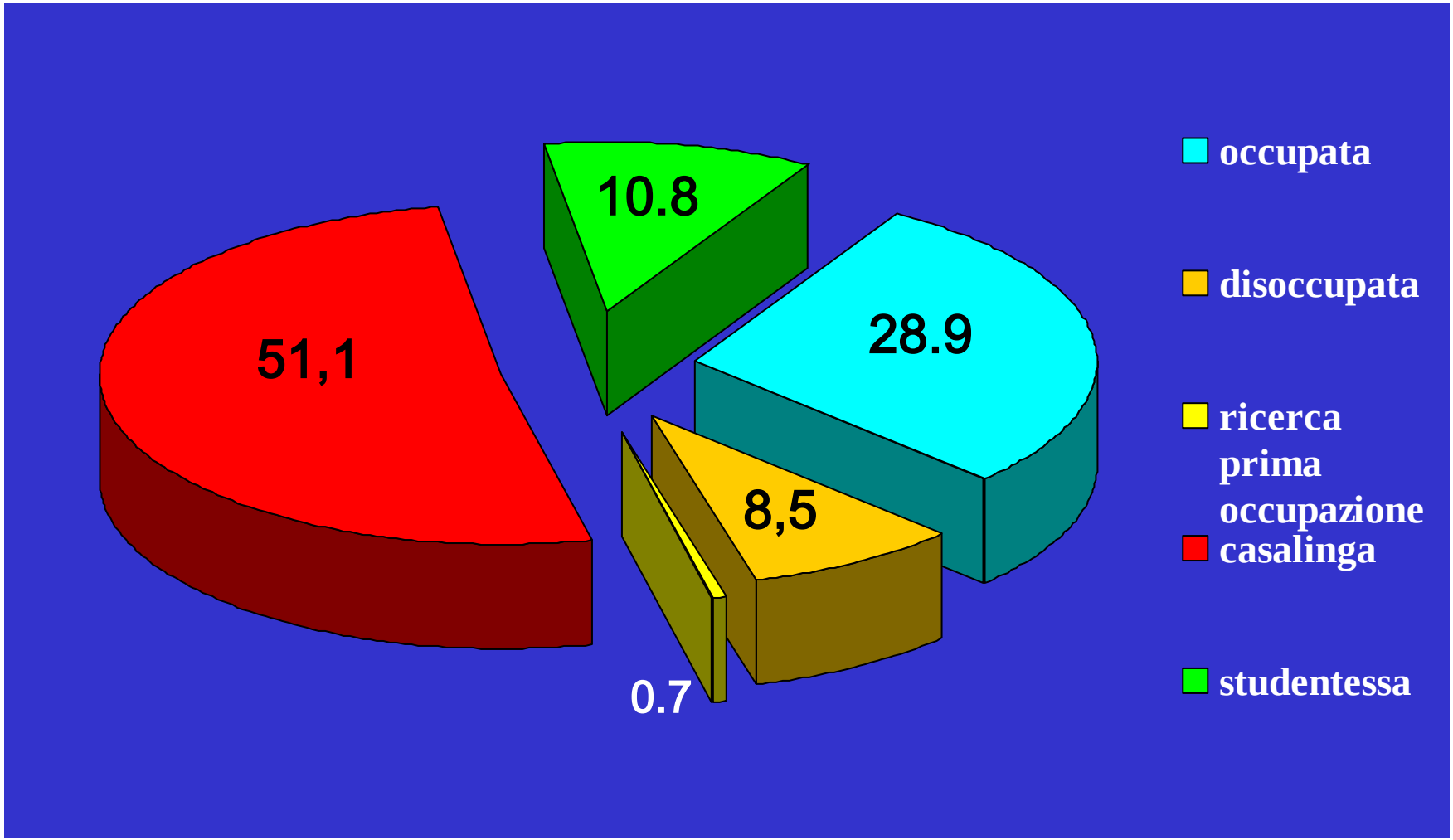
Distribuzione dei parti per titolo di studio della madre – Anno 2021



DISTRIBUZIONE DEI PAZIENTI PER CLASSE NYHA ALLA FINE DELLO STUDIO.



IVG e condizione professionale anno 2001



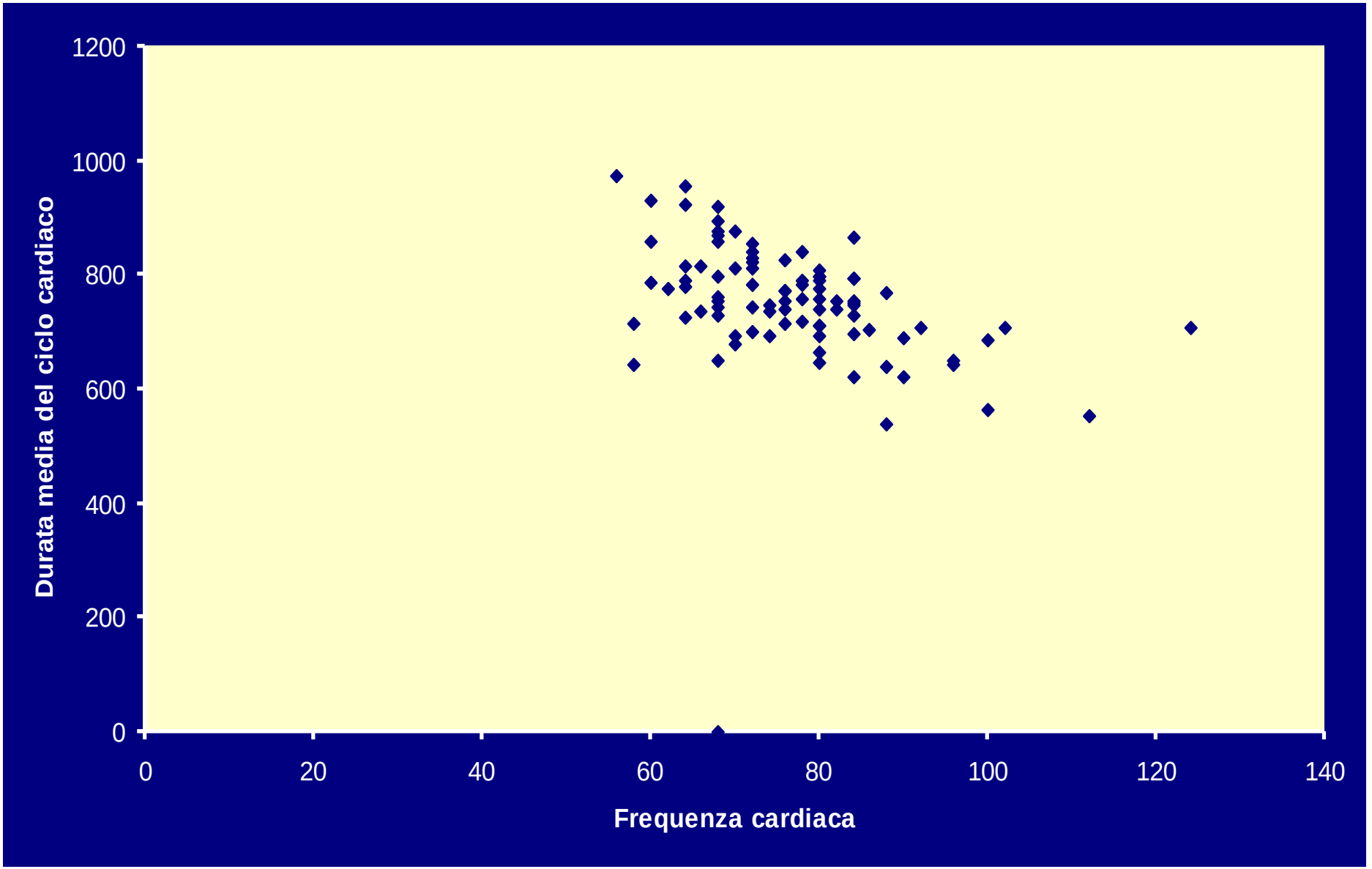
DESCRIVERE I DATI

RAPPRESENTAZIONI CON GRAFICI

DIAGRAMMI A DISPERSIONE: utili per valutare le relazioni tra variabili quantitative; sull'asse x e sull'asse y ci sono le due variabili di cui si vuole studiare la relazione.

Il grafico mostra come ciascuna osservazione si colloca nel piano cartesiano in relazione ai valori delle due variabili

DIAGRAMMA A DISPERSIONE PER L'ANALISI DELLA RELAZIONE TRA FREQUENZA CARDIACA E DURATA MEDIA DEL CICLO CARDIACO.

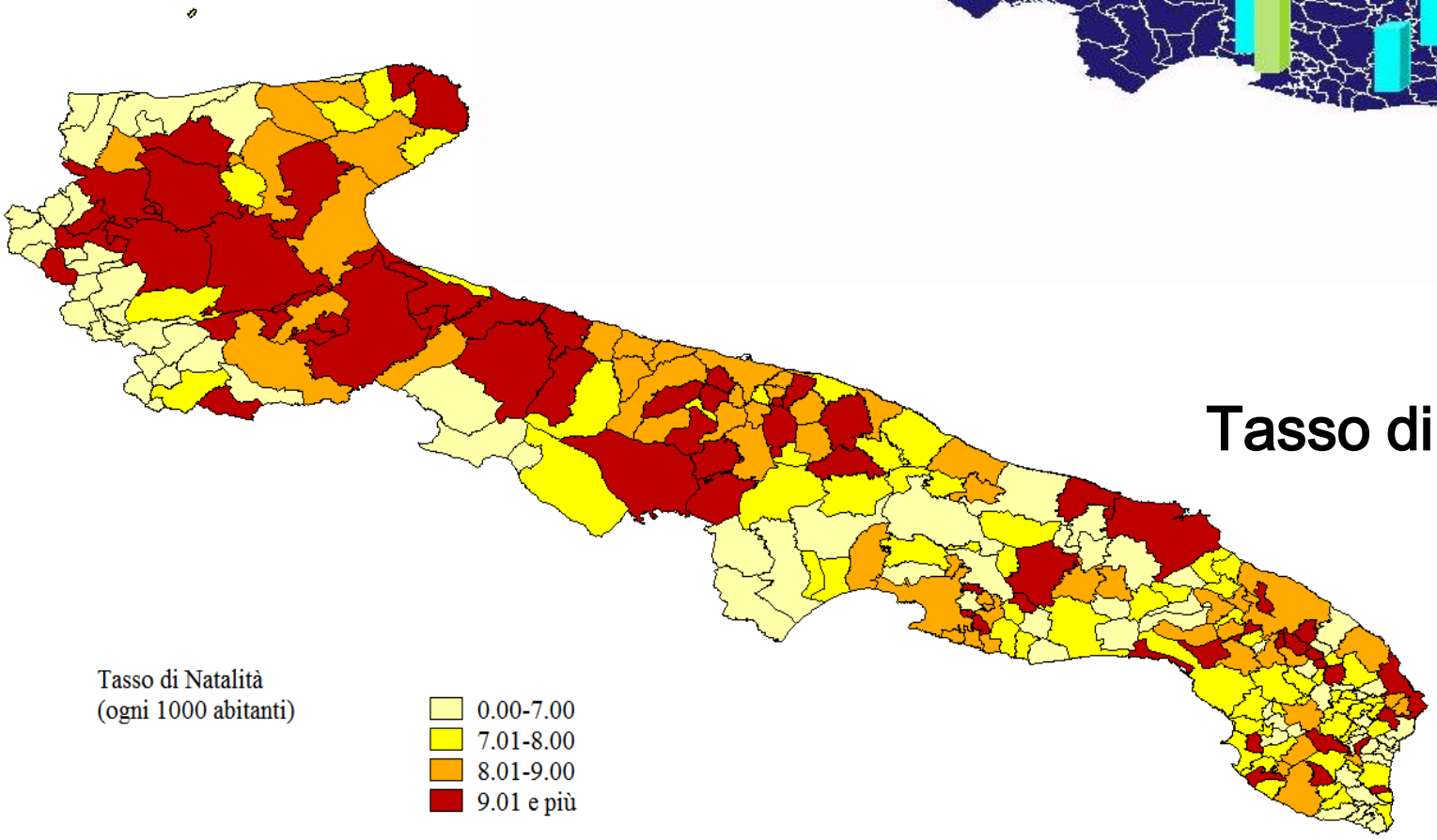


DESCRIVERE I DATI

RAPPRESENTAZIONI CON CARTOGRAMMI

Distribuzione geografica dei posti letto

Tasso di Natalità



Distribuzione geografica dei ricoveri importati ed esportati - anno 2005