www.sce.unimore.it

Scienze della Comunicazione
e dell'Economia

DATI E LORO RAPPRESENTAZIONE GRAFICA

Luca La Rocca
luca.larocca@unimore.it

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Una definizione di statistica capace di cogliere diversi aspetti salienti della disciplina è quella fornita dal Dizionario di Italiano on-line della Garzanti Linguistica:

> *analisi quantitativa dei fenomeni collettivi che hanno attitudine a variare, allo scopo di descriverli e di individuare le leggi o i modelli che permettono di spiegarli e di prevederli;*
> `http://garzantilinguistica.sapere.it.`

In pratica: sulla base di opportuni dati relativi a un fenomeno di interesse, un'analisi statistica cerca di rispondere alle domande che qualcuno (un astronomo, un antropologo, un biologo, uno psicologo, un economista, un sociologo, un medico, un fisico, un chimico, un cittadino...) si pone sul fenomeno in questione.

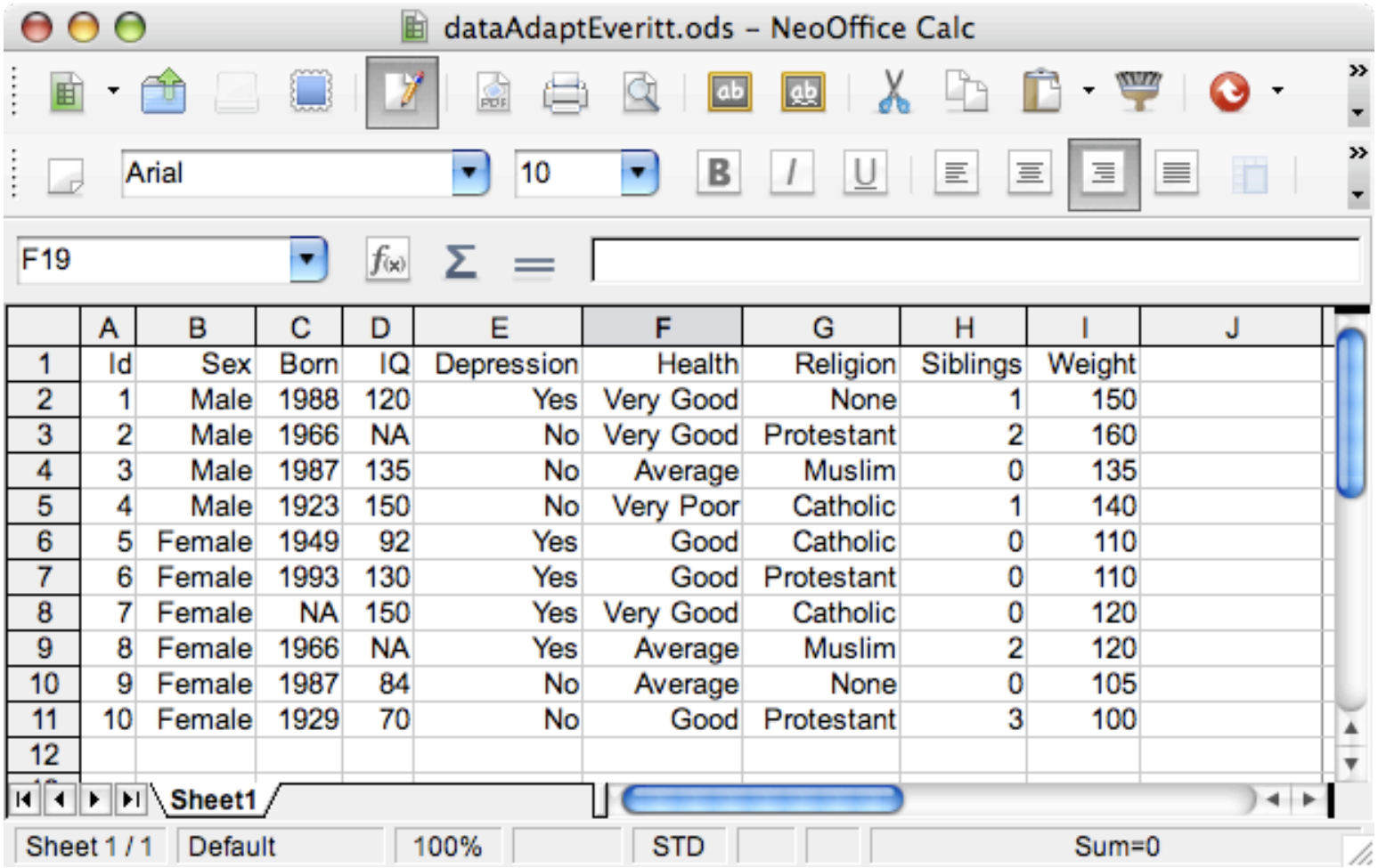
Un insieme di dati tipicamente consiste dei valori osservati di un certo numero di caratteri (es. altezza, peso, ...) su un certo numero di unità statistiche (es. individui, automobili, ...) appartenenti a un collettivo (es. cittadinanza, parco auto, ...) di interesse.

Gli insiemi di dati ammettono una rappresentazione canonica sotto forma di matrice dei dati: una riga per ogni unità statistica e una colonna per ogni carattere...

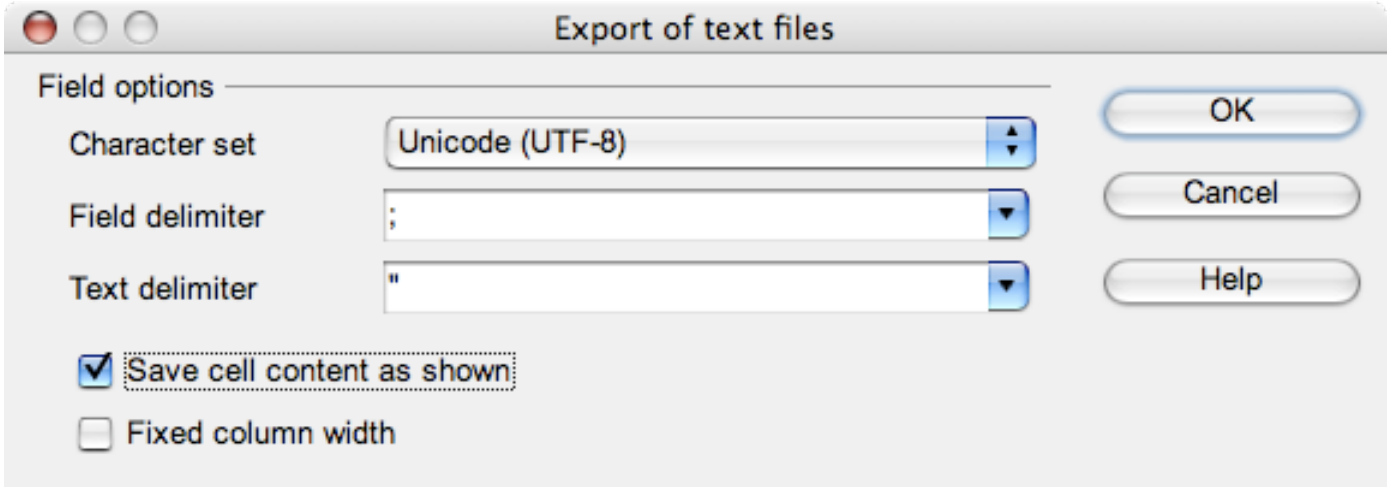
Un'ipotetica matrice dei dati (adattata da Everitt, 2005, p. 2)

Id	Sex	Born	IQ	Depression	Health	Religion	Siblings	Weight (lb)
1	Male	1988	120	Yes	Very Good	None	1	150
2	Male	1966	NA	No	Very Good	Protestant	2	160
3	Male	1987	135	No	Average	Muslim	0	135
4	Male	1923	150	No	Very Poor	Catholic	1	140
5	Female	1949	92	Yes	Good	Catholic	0	110
6	Female	1993	130	Yes	Good	Protestant	0	110
7	Female	NA	150	Yes	Very Good	Catholic	0	120
8	Female	1966	NA	Yes	Average	Muslim	2	120
9	Female	1987	84	No	Average	None	0	105
10	Female	1929	70	No	Good	Protestant	3	100

Tipicamente, in concreto, la matrice dei dati sarà salvata in un foglio di calcolo. . .



	A	B	C	D	E	F	G	H	I	J
1	Id	Sex	Born	IQ	Depression	Health	Religion	Siblings	Weight	
2	1	Male	1988	120	Yes	Very Good	None	1	150	
3	2	Male	1966	NA	No	Very Good	Protestant	2	160	
4	3	Male	1987	135	No	Average	Muslim	0	135	
5	4	Male	1923	150	No	Very Poor	Catholic	1	140	
6	5	Female	1949	92	Yes	Good	Catholic	0	110	
7	6	Female	1993	130	Yes	Good	Protestant	0	110	
8	7	Female	NA	150	Yes	Very Good	Catholic	0	120	
9	8	Female	1966	NA	Yes	Average	Muslim	2	120	
10	9	Female	1987	84	No	Average	None	0	105	
11	10	Female	1929	70	No	Good	Protestant	3	100	
12										




```
"Id";"Sex";"Born";"IQ";"Depression";"Health";"Religion";"Siblings";"Weight"
1;"Male";1988;120;"Yes";"Very Good";"None";1;150
2;"Male";1966;"NA";"No";"Very Good";"Protestant";2;160
3;"Male";1987;135;"No";"Average";"Muslim";0;135
4;"Male";1923;150;"No";"Very Poor";"Catholic";1;140
5;"Female";1949;92;"Yes";"Good";"Catholic";0;110
6;"Female";1993;130;"Yes";"Good";"Protestant";0;110
7;"Female";"NA";150;"Yes";"Very Good";"Catholic";0;120
8;"Female";1966;"NA";"Yes";"Average";"Muslim";2;120
9;"Female";1987;84;"No";"Average";"None";0;105
10;"Female";1929;70;"No";"Good";"Protestant";3;100
```

Se la matrice dei dati è salvata in formato CSV (Comma Separated Value) con i campi separati da `;` è immediato caricarla in R:

> *R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL* `http://www.R-project.org`.

Si tratta di un software open source, disponibile per diversi sistemi operativi, affidabile, flessibile ed espandibile (esiste un archivio in continua crescita di librerie opzionali che implementano una vasta gamma di analisi statistiche).

R funziona "a linea di comando"...

Il seguente estratto dalla R Console mostra come si possa leggere la matrice dei dati sulla depressione dal file `dataAdaptEveritt.csv` nell'oggetto `X`, usando Id per i nomi di riga, e poi visualizzare `X`:

```
> X <- read.csv2("dataAdaptEveritt.csv", row.names = "Id")
> X
       Sex Born  IQ Depression    Health   Religion Siblings Weight
1     Male 1988 120        Yes Very Good       None        1    150
2     Male 1966  NA         No Very Good Protestant        2    160
3     Male 1987 135         No   Average     Muslim        0    135
4     Male 1923 150         No Very Poor   Catholic        1    140
5   Female 1949  92        Yes      Good   Catholic        0    110
6   Female 1993 130        Yes      Good Protestant        0    110
7   Female   NA 150        Yes Very Good   Catholic        0    120
8   Female 1966  NA        Yes   Average     Muslim        2    120
9   Female 1987  84         No   Average       None        0    105
10  Female 1929  70         No      Good Protestant        3    100
```

R può leggere dati in diversi formati: si digiti `help(read.csv2)` al prompt della R console per maggiori informazioni.

Le unità elementari che costituiscono un collettivo statistico devono essere omogenee rispetto a una o più caratteristiche di interesse (altrimenti non avremmo interesse a considerarle tutte assieme).

Negli ipotetici dati di Everitt le unità statistiche sono esseri umani per i quali si vuole studiare, ad esempio, il fenomeno della depressione: quali fattori la determinano?

In altri casi le unità statistiche possono essere automobili (es. studio del parco auto circolanti in relazione al fenomeno dell'inquinamento da gas di scarico) oppure ospedali (es. analisi delle prestazioni del servizio sanitario nazionale) oppure...

La definizione dell'unità statistica deve essere non ambigua e questo a volte richiede che sia molto dettagliata.

Inoltre, al variare delle esigenze di studio, la definizione dell'unità statistica può variare nel tempo (nel qual caso il confronto fra studi successivi diviene problematico).

Consideriamo, per esempio, la definizione di famiglia usata dall'ISTAT (Istituto Nazionale di Statistica) per il Censimento (Generale della Popolazione e delle Abitazioni) nel 1981 e successivamente nel 2001 (recependo il nuovo regolamento anagrafico del 1989): per famiglia si è inteso. . .

. . . un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, affiliazione, tutela o da vincoli affettivi, coabitanti e aventi dimora abituale nello stesso comune, che normalmente provvedono al soddisfacimento dei loro bisogni mediante la messa in comune di tutto o parte del reddito di lavoro o patrimoniale da esse percepito. Una famiglia può essere costituita anche di una sola persona la quale provvede in tutto o in parte con i propri mezzi di sussistenza al soddisfacimento dei bisogni individuali. Sono considerate facenti parte della famiglia, come membri aggregati ad essa, anche le persone addette ai servizi domestici, nonché le altre persone che, a qualsiasi titolo, convivono abitualmente con la famiglia stessa (Borra & Di Ciaccio, 2008, p. 2).

. . . un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o affettivi, coabitanti ed aventi dimora abituale nello stesso comune (anche se non sono ancora iscritte nell'anagrafe della popolazione residente del comune medesimo). Una famiglia può essere costituita anche da una sola persona. L'assente temporaneo non cessa di appartenere alla propria famiglia sia che si trovi presso altro alloggio (o convivenza) dello stesso comune, sia che si trovi in un altro comune. La definizione di famiglia adottata per il censimento è quella contenuta nel regolamento anagrafico. Il personale di servizio della famiglia (domestici, collaboratori familiari, ecc.) che dimori abitualmente nella abitazione, costituisce famiglia a sé stante, sempreché tra i componenti la famiglia e il personale suddetto non vi siano legami di alcun tipo fra quelli compresi nella definizione già citata (Borra & Di Ciaccio, 2008, p. 2).

I collettivi statistici si classificano in

- collettivi di stato (es. le famiglie residenti in Italia al 21 ottobre 2001, gli esercizi commerciali di Reggio Emilia alle ore 24 del 20 maggio 2006, . . .)
- collettivi di movimento (es. i passeggeri di Scandinavian Airlines dal 1 agosto 2004 al 31 luglio 2005, le automobili vendute in Italia nel 2006, . . .)

a seconda che siano definiti con riferimento a un istante di tempo o a un periodo di tempo.

Si usa il termine popolazione per denotare il collettivo oggetto di studio nel suo complesso, mentre il termine campione denota il collettivo formato dalle sole unità osservate (quelle per le quali si hanno dati); si dice numerosità campionaria il numero di unità osservate.

La statistica descrittiva tratta l'analisi dei dati come se questi rappresentassero esaustivamente il fenomeno studiato, ovvero supponendo la coincidenza di popolazione e campione.

La statistica inferenziale tratta l'estensione dei risultati dell'analisi dei dati dal campione alla popolazione di cui questo è immagine e, più in generale, studia la loro generalizzazione a fini esplicativi o previsivi.

Intanto nella perfida Albione...

...Ben Goldacre unpicks bad science to explain good science at `http://www.badscience.net`

...David Spiegelhalter and collegues try to make sense of chance, risk, luck, uncertainty and probability at `http://understandinguncertainty.org`

...the Royal Statistical Society campaigns to make Britain better with numbers and statistics at `http://www.getstats.org.uk`

...Nigel Hawkes directs a pressure group whose aim is to detect and expose the distortion and misuse of statistical information, and identify those responsible, at `http://www.straightstatistics.org`

Le rilevazioni di dati si distinguono in

- rilevazioni sperimentali
- rilevazioni osservazionali

a seconda che sia possibile o meno controllare (modificare) le condizioni sotto le quali si svolge il processo di osservazione e le caratteristiche delle unità statistiche osservate.

Tipicamente si hanno rilevazioni sperimentali quando si opera in laboratorio (es. fisica, chimica e in certi casi medicina) mentre si hanno rilevazioni osservazionali quando si opera sul campo (es. scienze economiche e sociali, epidemiologia).

In un contesto sperimentale, le condizioni sotto le quali si svolge il processo di osservazione e le caratteristiche delle unità statistiche osservate si riassumono in un certo numero di fattori rilevanti per il fenomeno studiato.

In particolare, si dicono sperimentali quei fattori di cui l'esperimento si propone di verificare l'effetto (es. trattamento farmacologico in una prova clinica) e di stratificazione quei fattori che descrivono le unità sperimentali (es. peso ed età del paziente).

Si noti l'uso del termine fattore come sinonimo di carattere in un contesto sperimentale (Borra & Di Ciaccio, 2008, p. 7).

I fattori sperimentali e di stratificazione si possono controllare direttamente per mezzo di un opportuno disegno sperimentale.

Per esempio i pazienti che partecipano a una prova clinica possono essere suddivisi in quattro strati sulla base del loro peso e della loro età (“giovani leggeri”, “giovani pesanti”, “anziani leggeri”, “anziani pesanti”) dopo di che, all’interno di ogni strato, alcuni pazienti riceveranno il nuovo farmaco di cui si vuole verificare l’efficacia, altri un placebo.

La scelta dei pazienti da trattare sarà lasciata al caso, ovvero si ricorrerà a randomizzazione, per controllare indirettamente i fattori trascurati dal disegno sperimentale (es. abitudini alimentari del paziente).

In un contesto osservazionale, un'indagine statistica si limita a rilevare alcuni caratteri di interesse in un campione della popolazione studiata, senza interferire con i fenomeni che in tale popolazione si manifestano.

Si parla di indagine totale o censimento quando il campione coincide con l'intera popolazione, altrimenti si parla di indagine campionaria.

L'indagine campionaria riduce tempi e costi, permettendo di rilevare dati di buona qualità su molti caratteri, al prezzo di dovere ricorrere a tecniche di inferenza statistica per "assicurarsi" che i risultati ottenuti si estendano all'intera popolazione: questo è possibile, se il campione è scelto in modo opportuno (casualmente), mediante un'opportuna valutazione probabilistica del meccanismo di scelta del campione.

Esempi italiani di indagini totali sono

- il Censimento Generale dell'Agricoltura
- il Censimento Generale della Popolazione e delle Abitazioni
- il Censimento Generale dell'Industria e dei Servizi

svolti dall'Istituto Nazionale di Statistica, ISTAT, con cadenza decennale; le ultime edizioni dei primi due sono state svolte nel 2010 e 2011, rispettivamente, la prossima edizione del terzo sarà svolta nel 2012.

> *L'Istituto Nazionale di Statistica è un ente di ricerca pubblico. Presente nel Paese dal 1926, è il principale produttore di statistica ufficiale a supporto dei cittadini e dei decisori pubblici. Opera in piena autonomia e in continua interazione con il mondo accademico e scientifico (*`http://www.istat.it`*).*

Dal 1989 l'ISTAT coordina il SISTAN (Sistema Statistico Nazionale) formato da una rete di operatori statistici pubblici e privati cui sono affidate le rilevazioni statistiche di interesse pubblico stabilite dal Programma Statistico Nazionale (triennale con aggiornamento annuale):

`http://www.sistan.it`

A livello locale l'USCI (Unione Statistica Comuni Italiani) si pone come interlocutore tecnico privilegiato del SISTAN:

`http://www.usci.it`

A livello europeo la missione di fornire informazione statistica di qualità all'Unione Europea è affidata a EUROSTAT (Ufficio di Statistica della Commissione Europea):

`http://epp.eurostat.ec.europa.eu`

A U.S. example of sample survey is the GSS (General Social Survey), an ongoing study of the National Opinion Research Center (NORC) at the University of Chicago, Illinois, started in 1972 and containing...

> *...a standard 'core' of demographic and attitudinal questions, plus topics of special interest. The GSS is the largest project funded by the Sociology Program of the National Science Foundation and, except for the U.S. Census, the most frequently analyzed source of information in the social sciences (`http://gss.norc.org`).*

The NORC was established in 1941 by Harry H. Field, at the University of Denver, Colorado, as an independent survey research organization to provide the public and government with objective measures of public opinion.

A glossary of statistical terms in a number of languages is published on the Web at `http://isi.cbs.nl/glossary` by the International Statistical Institute (ISI), which is an old scientific association seeking...

> *...to develop and improve statistical methods and their application through the promotion of international activity and co-operation (`http://isi-web.org`).*

All'ISI è affiliata, in Italia, la Società Italiana di Statistica (SIS), il cui scopo fondamentale è quello di promuovere le scienze statistiche e le loro applicazioni (`http://www.sis-statistica.it`); per esempio la SIS cerca di migliorare la consapevolezza nell'uso delle statistiche tramite tramite la pubblicazione di un magazine online:

`http://sis-statistica.it/magazine`

Anche a fronte di una ricca produzione di dati da parte della statistica ufficiale (altre due fonti importanti da una prospettiva italiana sono `http://www.bancaditalia.it` e `http://www.oecd.org`) è chiaro che spesso e volentieri il fenomeno di interesse non potrà essere studiato sulla base di dati già raccolti (dati secondari) ma sarà necessario avvalersi di dati raccolti appositamente (dati primari) per rispondere alle domande di ricerca che interessa affrontare.

Per esempio, se si vuole studiare la fedeltà all'insegna dei clienti di un certo ipermercato, occorre (fare) intervistare alcuni di loro sulla base di un opportuno questionario le cui domande permettano di misurare, fra l'altro, le diverse componenti della fiducia; tipicamente si giungerà a tali domande sulla base di una ricerca qualitativa preliminare (es. focus group) e di un pretest del questionario.

Passaggi critici di un'indagine statistica sono:

- l'individuazione della popolazione di riferimento (idealmente con una lista completa, esatta e aggiornata delle sue unità);
- scelta del piano di campionamento (casuale semplice, casuale stratificato, . . .);
- scelta della tecnica di rilevazione (intervista diretta, indagine postale, intervista telefonica, . . .);
- registrazione dei dati (su supporto informatico) e loro verifica.

Occorre tenere presente che la qualità dei risultati di un'analisi statistica è limitata dalla qualità dei dati analizzati.

La definizione di un carattere statistico deve specificare i valori che questo può assumere, ovvero le sue possibili modalità.

Si richiede che le modalità di un carattere statistico siano esaustive e non sovrapposte: comunque si prenda un'unità statistica nella popolazione di interesse, deve essere possibile stabilire in modo univoco la modalità assunta dal carattere su tale unità.

È possibile (in pratica frequente) che per alcune unità statistiche del campione analizzato non si conosca la modalità assunta da uno o più caratteri; si parla in questo caso di dati mancanti e ci si avvale della modalità ausiliaria NA (Not Available).

I dati mancanti riducono la numerosità campionaria effettiva.

Con riferimento agli ipotetici dati per lo studio della depressione:

- il carattere Depression ammette due sole modalità (Yes, No);
- il carattere Health ammette cinque modalità (Very Poor, Poor, Average, Good, Very Good) di cui quattro osservate;
- il carattere Siblings ammette, in linea di principio, infinite modalità (i numeri interi positivi);
- il carattere Weight ammette, in linea di principio, infinite modalità (i numeri reali positivi).

Si noti il ruolo speciale giocato dal carattere Id che assume una modalità differente su ogni unità statistica e in questo modo la identifica (nella realtà potrebbe essere il codice fiscale).

Quali modalità per il carattere Religion?

La scelta delle modalità di un carattere dipende dalla popolazione studiata (gli ipotetici dati per lo studio della depressione fanno riferimento alla prima delle due scelte seguenti):

- un carattere Religion con modalità Cattolica, Ebraica, Musulmana, Nessuna, Ortodossa, Protestante e Altra potrebbe andare bene per la popolazione tedesca;
- per la popolazione giapponese potrebbe essere meglio un carattere Religion con modalità Cristiana, Buddista, Nessuna, Scintoista e Altra.

In ogni caso, per non perdere dettagli importanti, si deve fare in modo che la modalità Altra risulti minoritaria.

Si è soliti classificare i caratteri sulla base delle modalità che possono assumere:

> *un carattere si dice quantitativo, se le sue modalità sono espresse numericamente, altrimenti si dice qualitativo;*
> *un carattere quantitativo è anche detto variabile, un carattere qualitativo è anche detto mutabile.*

Si noti che R usa il termine factor per indicare una mutabile.

Nell'esempio sulla depressione:

- Depression, Religion e Health sono caratteri qualitativi (mutabili);
- Siblings, Born e Weight sono caratteri quantitativi (variabili).

Va da sé che è sempre possibile (e spesso utile) codificare in forma numerica le modalità di un carattere qualitativo.

Es. si può porre Yes = 1 e No = 0 per Depression e Very Poor = 1, Poor = 2, Average = 3, Good = 4 e Very Good = 5 per Health.

Tuttavia, evidentemente, la codifica è arbitraria.

Es. si può porre Yes = 1 e No = 2 per Depression e Very Poor = 1, Poor = 2, Average = 4, Good = 6 e Very Good = 7 per Health.

In effetti, in R, un `factor` non è altro che una variabile corredata di un'opportuna descrizione della codifica.

Un carattere qualitativo si dice ordinato, o misurato su scala ordinale, quando è dato un ordine sull'insieme delle sue modalità: o due modalità sono uguali, o una precede l'altra.

Un carattere qualitativo si dice sconnesso, o misurato su scala nominale, quando non è dato alcun ordine sull'insieme delle sue modalità: o due modalità sono uguali, o sono diverse.

Nell'esempio sulla depressione il carattere Religion è misurato su scala nominale, mentre Health è misurato su scala ordinale.

Si noti che la codifica numerica di un carattere misurato su scala ordinale deve rispettarne l'ordinamento delle modalità (se non vuole essere ingannevole).

Per un carattere dicotomico (che può assumere due sole modalità) come Sex e Depression la distinzione tra scala nominale e scala ordinale non è interessante.

Pur in assenza di un ordinamento delle modalità di un carattere, vi può essere una nozione di prossimità fra le stesse: es. le modalità Cattolica, Protestante e Ortodossa di Religion sono vicine fra loro (e potrebbero essere aggregate nell'unica modalità Cristiana, se lo si ritenesse utile ai fini dell'analisi).

I caratteri ordinati si distinguono in

- caratteri rettilinei (es. Titolo di Studio con modalità Nessun Titolo, Diploma di Maturità, Laurea Triennale, Laurea Specialistica, Dottorato)
- caratteri ciclici (es. Direzione del Vento con modalità Nord, Nord-Est, Est, Sud-Est, Sud, Sud-Ovest, Ovest, Nord-Ovest)

a seconda che l'ordinamento delle modalità sia quello dei punti di una retta o quello dei punti di una circonferenza.

Il carattere Health di Everitt è un altro esempio di carattere rettilineo, il Mese di Nascita un altro esempio di carattere ciclico.

Un carattere quantitativo si dice misurato su scala a intervalli quando non esiste uno zero assoluto (es. Born nell'esempio sulla depressione).

Il tipico esempio di carattere misurato su scala a intervalli è la Temperatura in gradi Celsius, il cui zero (punto di congelamento dell'acqua a pressione atmosferica) è convenzionale. Infatti, se si considera la Temperatura in gradi Fahrenheit

$$F = \frac{9}{5}C + 32,$$

0 gradi Celsius corrispondono a 32 gradi Fahrenheit.

Le modalità di un carattere misurato su scala a intervalli sono definite a meno di una trasformazione affine positiva: $Y = aX + b$, $a > 0$, $b \in \mathbb{R}$.

Un carattere quantitativo si dice misurato su scala di rapporti quando esiste uno zero assoluto.

Nell'esempio sulla depressione, tali sono Weight e

$$Age = 2009 - Born$$

perché l'assenza di massa corporea e la nascita sono zeri assoluti.

Le modalità di un carattere misurato su scala di rapporti sono definite a meno di una trasformazione lineare positiva: $Y = aX$, $a > 0$.

Per esempio 150 lb sono lo stesso che $0.4536 \times 150 = 68.04$ Kg e 21 anni sono lo stesso che $12 \times 21 = 252$ mesi.

Per un carattere misurato su scala a intervalli ha senso considerare le differenze, ma non i rapporti:

- se un radiatore alla temperatura di 40 gradi Celsius si trova in una stanza dove ci sono 10 gradi Celsius, la differenza di 30 gradi Celsius origina il flusso di calore che riscalda la stanza;
- d'altra parte non è vero che 40 gradi Celsius sono il quadruplo di 10 gradi Celsius (se usassimo i gradi Fahrenheit avremmo il radiatore a 104 gradi e la stanza a 50 gradi...);
- analogamente si può ragionare per la data di un evento (che evidentemente dipende dal calendario adottato);
- se poi si vuole misurare una temperatura su scala di rapporti, allora c'è il Kelvin.

Per un carattere misurato su scala di rapporti ha senso considerare le differenze e i rapporti:

- se ho 3 grammi d'oro su un piatto e 1 grammo d'oro sull'altro, la differenza $3 - 1 = 2$ grammi è la massa da aggiungere al secondo piatto per riequilibrare la bilancia;
- inoltre 3 grammi d'oro sono il triplo di 1 grammo d'oro, perché occorre triplicare la quantità di oro nel secondo piatto per riequilibrare la bilancia;
- analogamente si può ragionare per il tempo trascorso (sostituendo la bilancia con un orologio);
- il raddoppio della temperatura in Kelvin di un gas perfetto a volume costante corrisponde a un incremento "unitario" della sua entropia.

Le scale di misura (nominale, ordinale, a intervalli, per rapporti) introdotte da Stevens (1947) sono da alcuni indicate come discriminanti ai fini della scelta del metodo statistico.

Come discusso da Velleman & Wilkinson (1993) questo approccio all'analisi dei dati può essere pericoloso, perché rischia di essere troppo rigido rispetto alla varietà del mondo reale.

Occorre tuttavia tenere sempre presente la natura dei dati che si stanno analizzando, assieme agli obiettivi dell'analisi.

Per esempio, se l'allenatore di una squadra scolastica afferma di avere assegnato casualmente i numeri di maglia, ma le matricole si lamentano di avere ricevuto numeri alti (che non gradiscono) e si vuole stabilire se ciò sia frutto del caso o di una discriminazione, come andranno considerati i numeri di maglia?

Se per l'allenatore sono solo etichette, dunque su scala nominale, per le matricole sono quantomeno su scala ordinale. . . in pratica, se non si adotta il punto di vista delle matricole, non si può rispondere alla domanda. In questo caso è il problema a determinare la scala di misura (in un altro contesto il numero di maglia avrebbe probabilmente l'unico ruolo di identificatore dell'unità statistica).

Nell'ambito delle ricerche di marketing è prassi comune domandare a un campione di clienti di esprimere quanto siano d'accordo con un certo numero di affermazioni, avvalendosi di una scala ordinale del tipo

Per Niente, Molto Poco, Poco, Relativamente, Abbastanza, Molto, Del Tutto

Successivamente vi è interesse a ridurre le numerose affermazioni con cui i clienti hanno espresso il loro grado di accordo a un limitato numero di fattori sottostanti (es. attenzione alla funzionalità del prodotto, attenzione all'aspetto del prodotto, . . .)

Questa riduzione può ottenersi mediante un'analisi fattoriale, se si "promuovono" le risposte su una scala a intervalli: 1, 2, 3, 4, 5, 6, 7.

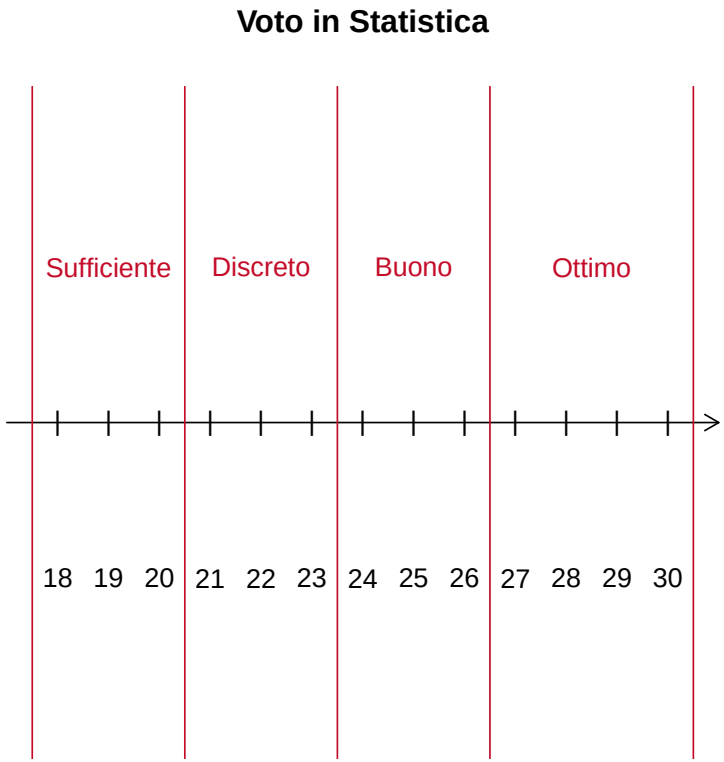
La promozione del grado di accordo da ordinale a intervalli è utile, se non si dimentica che:

- la codifica adottata (1, 2, 3, 4, 5, 6, 7) presuppone che tra le modalità Molto Poco e Poco vi sia la stessa distanza che c'è tra Relativamente e Abbastanza;
- la promozione non implica che Relativamente sia il doppio di Molto Poco (ovvero non è una promozione a scala di rapporti); infatti la codifica 0, 1, 2, 3, 4, 5, 6 sarebbe altrettanto legittima e ci direbbe che Relativamente è il triplo di Molto Poco. . . ne segue che è privo di senso affermare che un cliente è il 25% più d'accordo (es. soddisfatto) di un altro, a meno di non formulare un'ulteriore ipotesi (es. che Per Niente sia uno zero assoluto... opinabile).

D'altra parte, in alcune circostanze, può essere utile "retrocedere" un carattere quantitativo su scala ordinale, al fine di ottenere una sintesi che faciliti la comprensione del fenomeno studiato.

In particolare può essere utile rimpiazzare "molte" modalità numeriche con "pochi" intervalli tra loro disgiunti, operando quella che si dice una suddivisione in classi del carattere; si noti che la scelta degli intervalli comporta un certo grado di arbitrarietà.

Si consideri per esempio il Voto in Statistica: tenendo conto che le insufficienze non vengono registrate e trascurando l'eventuale lode, si possono raggruppare gli esiti in 4 classi, Sufficiente (18, 19, 20), Discreto (21, 22, 23), Buono (24, 25, 26) e Ottimo (27, 28, 29, 30), individuate dagli intervalli di estremi 17.5, 20.5, 23.5, 26.5, 30.5...



Suddivisione in 4 classi

Si dice ampiezza di una classe la lunghezza dell'intervallo che la definisce:

- l'ampiezza della classe Sufficiente è pari a $20.5 - 17.5 = 3$;
- l'ampiezza della classe Ottima è pari a $30.5 - 26.5 = 4$.

Come si vede, non è necessario che tutte le classi abbiano la stessa ampiezza; si richiede invece che le classi siano disgiunte ed esaustive, in modo da definire delle nuove modalità.

Si noti come l'arrotondamento dei voti all'unità intera consenta di non preoccuparsi dell'eventualità che un valore osservato coincida con un estremo comune a due intervalli.

In generale occorrerà precisare se gli intervalli siano aperti a destra o aperti a sinistra, vale a dire se i loro estremi appartengano all'intervallo seguente o a quello precedente.

Per esempio, fissate due soglie a 120 lb e 150 lb, le modalità del carattere Weight di Everitt si possono suddividere in:

- $90 \vdash 120$, $120 \vdash 150$ e $150 \vdash 180$, con la modalità 120 lb nella seconda classe e la modalità 150 lb nella terza classe;
- $90 \dashv 120$, $120 \dashv 150$ e $150 \dashv 180$, con la modalità 120 lb nella prima classe e la modalità 150 lb nella seconda classe.

Si noti che in questo esempio tutte le classi hanno la stessa ampiezza (30 lb).

Confronto tra intervalli aperti a destra (dx) e sinistra (sx)

Id	Weight (lb)	Weight (dx)	Weight (sx)
1	150	$150 \vdash 180$	$120 \dashv 150$
2	160	$150 \vdash 180$	$150 \dashv 180$
3	135	$120 \vdash 150$	$120 \dashv 150$
4	140	$120 \vdash 150$	$120 \dashv 150$
5	110	$90 \vdash 120$	$90 \dashv 120$
6	110	$90 \vdash 120$	$90 \dashv 120$
7	120	$120 \vdash 150$	$90 \dashv 120$
8	120	$120 \vdash 150$	$90 \dashv 120$
9	105	$90 \vdash 120$	$90 \dashv 120$
10	100	$90 \vdash 120$	$90 \dashv 120$

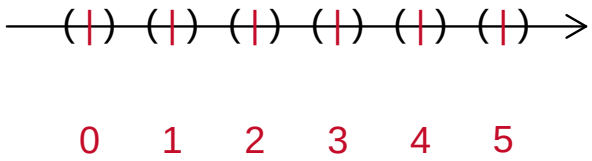
La distinzione tra caratteri rettilinei e ciclici si applica anche ai caratteri quantitativi (es. la direzione del vento misurata come angolo rispetto al nord è un carattere quantitativo ciclico).

I caratteri quantitativi si distinguono anche in

- caratteri discreti (es. Siblings, Numero di Figli, . . .)
- caratteri continui (es. Weight, Altezza, . . .)

a seconda che le loro modalità siano o meno isolate (lo sono se per ognuna di esse posso trovare un segmento della retta reale che la contiene senza contenere altre modalità).

Numero di Figli



Altezza (cm)

Le modalità di un carattere discreto sono in numero finito o, al più, costituiscono un'infinità numerabile (possono essere messe in corrispondenza biunivoca con un sottoinsieme dei numeri naturali, ovvero possono essere "contate"): es. i Genitori Laureati possono essere 0, 1 o 2, mentre per il Numero di Figli e i Siblings non c'è un massimo teorico. . . anche se magari se ne può trovare uno pratico.

In effetti, da un punto di vista applicativo, i caratteri discreti sono quelli che provengono da un'attività di conteggio; si noti che per tali caratteri non interessa la distinzione tra scala a intervalli e scala di rapporti.

Il punto di partenza di un'analisi statistica è generalmente la matrice dei dati di interesse, le cui colonne elencano le modalità assunte dai caratteri rilevati sulle unità del campione.

A tale elencazione si dà il nome di distribuzione unitaria

- semplice, se ci si riferisce a un singolo carattere (colonna della matrice dei dati)
- multipla, se ci riferisce a un insieme di più di caratteri (al limite l'intera matrice dei dati).

Si veda per esempio l'ipotetica matrice dei dati per lo studio della depressione...

Id	Sex	Born	IQ	Depression	Health	Religion	Siblings	Weight (lb)	Age (2009)
1	Male	1988	120	Yes	Very Good	None	1	150	21
2	Male	1966	NA	No	Very Good	Protestant	2	160	43
3	Male	1987	135	No	Average	Muslim	0	135	22
4	Male	1923	150	No	Very Poor	Catholic	1	140	86
5	Female	1949	92	Yes	Good	Catholic	0	110	60
6	Female	1993	130	Yes	Good	Protestant	0	110	16
7	Female	NA	150	Yes	Very Good	Catholic	0	120	NA
8	Female	1966	NA	Yes	Average	Muslim	2	120	43
9	Female	1987	84	No	Average	None	0	105	22
10	Female	1929	70	No	Good	Protestant	3	100	80

. . . qui sopra riportata, per comodità, completa delle modalità assunte dal carattere $Age = 2009 - Born$.

La distribuzione unitaria riflette fedelmente il manifestarsi del fenomeno studiato nel campione osservato, ma non consente (a meno che il campione non abbia numerosità davvero esigua) di cogliere immediatamente le caratteristiche salienti del fenomeno stesso.

Per esempio ci si può chiedere:

- vi sono più maschi o più femmine?
- si tratta di persone giovani o anziane?
- sono individui prevalentemente sani o malati?

Solo la risposta alla prima domanda (più femmine) è immediata, nonostante la numerosità campionaria sia davvero esigua.

Si dice frequenza assoluta della modalità x del carattere X il numero di unità sulle quali il carattere X assume la modalità x.

Per esempio la frequenza assoluta di Male (Sex) nei dati sulla depressione è pari a 4. Negli stessi dati (Weight) la frequenza assoluta di 110 lb è pari a 2, quella di 135 lb è pari a 1.

Un carattere con "poche" modalità è utilmente riassunto dalla sua distribuzione di frequenza assoluta (semplice), ovvero dall'elenco delle sue modalità e delle rispettive frequenze assolute...

Per esempio la distribuzione di frequenza assoluta di Sex è rappresentata dalla seguente tabella:

Sex	Freq.
Female	6
Male	4
Total	10

R fornisce il comando `table` per costruire la distribuzione di frequenza assoluta di un carattere:

```
> sort(table(X$Sex), decreasing = TRUE)

Female   Male
     6      4
```

Qualora le modalità siano ordinate (es. derivino dalla suddivisione in classi di un carattere quantitativo) si richiede che esse siano elencate in ordine (crescente/descrescente); per esempio la distribuzione di frequenza assoluta di Health è rappresentata dalla tabella

Health	Freq.
Very Poor	1
Poor	0
Average	3
Good	3
Very Good	3
Total	10

che evidenzia come si tratti in prevalenza di individui sani.

Per un carattere con “molte” modalità come Age ci si può avvalere di una suddivisione in classi:

Age	Freq.
$0 \vdash 30$	4
$30 \vdash 60$	2
$60 \vdash \infty$	3
Total	9

In questo modo si evidenzia come vi siano sia giovani che anziani.

Si noti che il totale delle osservazioni è in questo caso 9 (e non 10) per via di un dato mancante (l’età della settima unità statistica); si noti anche che l’ultimo intervallo è illimitato (a destra).

Come tutte le sintesi, il passaggio da una distribuzione unitaria alla corrispondente distribuzione di frequenza comporta una perdita di informazione: non è più possibile stabilire quali unità statistiche assumono una certa modalità di un certo carattere...

...il che, per ragioni di privacy, può anche essere un vantaggio.

Se i caratteri di interesse sono due, si dirà distribuzione di frequenza assoluta (doppia) l'elenco incrociato delle loro modalità e delle rispettive frequenze assolute:

	Sex		
Depression	Female	Male	Total
No	2	3	5
Yes	4	1	5
Total	6	4	10

La distribuzione di frequenza doppia di X e Y consente di studiare l'associazione tra X e Y (oltre che X e Y individualmente); si vedano più avanti le distribuzioni condizionate (profili riga e colonna).

Il comando `table` di R può produrre anche distribuzioni doppie:

```
> addmargins(table(X$Depression,X$Sex))

      Female Male Sum
  No       2    3   5
  Yes      4    1   5
  Sum      6    4  10
```

In generale possiamo considerare la distribuzione di frequenza assoluta (multipla) di due o più caratteri, con crescenti difficoltà di rappresentazione al crescere del numero di caratteri.

Per esempio in R è disponibile un oggetto `HairEyeColor` contente la distribuzione tripla di Sex, Hair e Eye in un campione di 592 studenti di statistica della University of Delaware...

```
> HairEyeColor
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

Idealmente `HairEyeColor` è stato ottenuto dalle variabili Hair, Eye e Sex mediante il comando `table` (in realtà è un oggetto built-in).

Fate of passengers on the maiden voyage of the ocean liner “Titanic”, summarized according to four categorical variables (in a flat table):

```
> ftable(Survived ~ ., data = Titanic)
                 Survived  No Yes
Class Sex    Age
1st   Male   Child            0   5
             Adult          118  57
      Female Child            0   1
             Adult            4 140
2nd   Male   Child            0  11
             Adult          154  14
      Female Child            0  13
             Adult           13  80
3rd   Male   Child           35  13
             Adult          387  75
      Female Child           17  14
             Adult           89  76
Crew  Male   Child            0   0
             Adult          670 192
      Female Child            0   0
             Adult            3  20
```

Quando si voglia confrontare la distribuzione di un carattere in due campioni con diversa numerosità, conviene utilizzare le frequenze relative, o le frequenze percentuali, al posto di quelle assolute.

Per un carattere X con modalità $x_1, x_2, \ldots, x_{k-1}, x_k$ aventi frequenze assolute $n_1, n_2, \ldots, n_{k-1}, n_k$ la frequenza relativa di x_i è definita come

$$f_i = \frac{n_i}{n},$$

dove $n = n_1 + \cdots + n_k$ è la numerosità campionaria (effettiva).
La frequenza percentuale di x_i non è altro che

$$p_i = 100 \times f_i.$$

Completando, per esempio, la tabella del carattere Health con le frequenze relative e quelle percentuali

Health	Freq.	Rel.	%
Very Poor	1	0.1	10
Poor	0	0.0	0
Average	3	0.3	30
Good	3	0.3	30
Very Good	3	0.3	30
Total	10	1.0	100

si vede come queste facciano riferimento a una numerosità campionaria convenzionale (pari a 1 e 100 rispettivamente).

Frequenze relative e percentuali sono in realtà la stessa cosa, come si evidenzia osservando che % è solo un modo di scrivere “diviso cento”: es. $10\% = 10/100 = 0.1$.

Da un punto di vista matematico conviene lavorare con numeri tra zero e uno, mentre da un punto di vista concettuale si può preferire fare riferimento a un collettivo di cento unità...

...o magari di mille unità, scrivendo $‰ = 1/1000$, se non di un milione di unità (a seconda della situazione); per esempio converrà esprimere la prevalenza di una malattia rara in “casi per milione”.

Quale che sia la numerosità campionaria convenzionale scelta (1, 100, ...) questa permette di confrontare campioni con diversa numerosità; per esempio (Pace & Salvan, 1996, p. 65) la tabella seguente confronta la distribuzione per sesso della popolazione residente in Italia in due diversi censimenti (dati in migliaia, confini attuali):

	Anno 1861			Anno 1981		
Sesso	Freq.	Rel.	%	Freq.	Rel.	%
M	13399	0.5089	50.89	27506	0.4863	48.63
F	12929	0.4911	49.11	29051	0.5137	51.37
Totale	26328	1.0000	100.00	56557	1.0000	100.00

L'inversione di maggioranza, già evidente in termini assoluti, è meglio apprezzata (e quantificata) in termini relativi (da -1.78% a $+2.74\%$).

Spesso interessa confrontare due (sotto)campioni individuati condizionando al valore di un altro carattere:

Depression|Sex						
	Sex					
Depression	Female		Male		Total	
No	0.33	(2)	0.75	(3)	0.50	(5)
Yes	0.67	(4)	0.25	(1)	0.50	(5)
Total	1.00	(6)	1.00	(4)	1.00	(10)

di modo che il 67% delle donne (ipoteticamente osservate) è depresso...

Sex|Depression						
	Sex					
Depression	Female		Male		Total	
No	0.40	(2)	0.60	(3)	1.00	(5)
Yes	0.80	(4)	0.20	(1)	1.00	(5)
Total	0.60	(6)	0.40	(4)	1.00	(10)

. . . laddove l'80% delle persone depresse (ipoteticamente osservate) è donna: attenzione a non confondere queste due percentuali!

Prosecutor's fallacy: poiché i colpevoli, visti fuggire, hanno tutti i capelli neri, quest'uomo con i capelli neri è colpevole. . . =:-O

Un po' di terminologia per le distribuzioni di frequenza relativa rappresentate da tabelle a doppia entrata:

- si dice distribuzione congiunta la distribuzione doppia;
- si dicono distribuzioni marginali le distribuzioni semplici;
- si dicono profili riga e profili colonna le distribuzioni condizionate, a seconda di come sono rappresentate.

La tabella a pag. 71 rappresenta dunque come profili colonna le distribuzioni condizionate di Depression dato Sex, assieme alla distribuzione marginale di Depression, mentre la tabella a pag. 72 rappresenta come profili riga le distribuzioni condizionate di Sex dato Depression, assieme alla distribuzione marginale di Sex.

La tabella seguente rappresenta invece...

. . . la distribuzione congiunta di Depression e Sex, assieme alle loro distribuzioni marginali:

	Sex		
Depression	Female	Male	Total
No	0.2	0.3	0.5
Yes	0.4	0.1	0.5
Total	0.6	0.4	1.0

Si noti che non si tratta di altro che della tabella a pag. 63 dove tutti i valori sono stati divisi per 10 (numerosità campionaria).

In virtù di inevitabili errori di arrotondamento, può capitare che le frequenze relative non sommino a uno (le frequenze percentuali non sommino a cento). Per esempio:

Age	Freq.	Rel.	%
$0 \vdash 30$	4	0.44	44
$30 \vdash 60$	2	0.22	22
$60 \vdash \infty$	3	0.33	33
Total	9	0.99	99

Questa situazione non va confusa con il caso in cui gli intervistati possano dare più risposte a una stessa domanda...

. . . per esempio Il Sole 24 Ore di lunedi 12 febbraio 2007 riporta le richieste di un campione nazionale di presidenti di Ordini e Collegi professionali in merito al futuro ruolo delle loro associazioni:

Richiesta	%
Formazione continua	80.8
Orientamento deontologico	36.9
Certificazione delle competenze	35.5
Innovazione organizzativa	22.3
Internazionalizzazione	19.5

Questa non è una distribuzione di frequenza (semmai sono cinque distribuzioni di frequenza per cinque distinte variabili dicotomiche).

Se X è un carattere ordinale (es. ottenuto per suddivisione in classi di un carattere quantitativo) ha senso e può essere utile considerare frequenze cumulate (dirette):

- la frequenza cumulata assoluta della modalità x è il numero di unità statistiche che assumono modalità minori o uguali a x;
- le frequenze cumulate relative (percentuali) si ottengono scalando la numerosità campionaria a 1 (100).

Se $x_1 < x_2 < \cdots < x_{k-1} < x_k$ sono le modalità di X, denoteremo la frequenza cumulata di x_i con N_i, se assoluta, con F_i, se relativa, con P_i, se percentuale ($i = 1, \ldots, k$).

Ai fini del calcolo converrà prima trovare

$$N_i \quad = \quad n_1 + \cdots + n_i$$

per $i = 1, \ldots, k$ (si noti che $N_1 = n_1$ e $N_k = n$) quindi ottenere

$$F_i = \frac{N_i}{N_k} \qquad \& \qquad P_i = 100 \times F_i$$

per la modalità x_i; per esempio troveremo

i	Age	n_i	N_i	f_i	F_i	p_i	P_i
1	$0 \vdash 30$	4	4	0.44	0.44	44	44
2	$30 \vdash 60$	2	6	0.22	0.67	22	67
3	$60 \vdash \infty$	3	9	0.33	1.00	33	100
	Total	9		0.99		99	

Naturalmente si può considerare la distribuzione di frequenza cumulata anche per un carattere qualitativo, purché ordinato, come per esempio il carattere Health:

i	Health	n_i	N_i	N_i^{rev}	f_i	F_i	F_i^{rev}	p_i	P_i	P_i^{rev}
1	Very Poor	1	1	10	0.1	0.1	1.0	10	10	100
2	Poor	0	1	9	0.0	0.1	0.9	0	10	90
3	Average	3	4	9	0.3	0.4	0.9	30	40	90
4	Good	3	7	6	0.3	0.7	0.6	30	70	60
5	Very Good	3	10	3	0.3	1.0	0.3	30	100	30
	Total	10			1.0			100		

Sulla base delle frequenze cumulate inverse (ottenute invertendo l'ordine delle modalità e denotate con "rev" ad apice) si può affermare che il 90% dei soggetti non sta male e il 60% dei soggetti sta bene.

Per un carattere qualitativo non ordinato può comunque avere senso considerare le frequenze cumulate relative ordinando le modalità in base alle frequenze (assolute) osservate...

```
> reliTable <- sort(table(X$Religion), decreasing = TRUE)
> reliTable

  Catholic Protestant     Muslim       None
         3          3          2          2
> cumsum(reliTable)/sum(reliTable)
  Catholic Protestant     Muslim       None
       0.3        0.6        0.8        1.0
```

...evidenziando così il fatto che le due religioni più diffuse "coprono" il sessanta percento dei pazienti (nel campione).

Nell'ambito dei caratteri quantitativi, si parla di carattere trasferibile quando ha senso immaginare che un'unità statistica ceda parte del suo carattere (al limite tutto) a un'altra unità statistica.

Un esempio tipico di carattere trasferibile è il reddito. Un altro esempio è il numero di automobili possedute. Esempi di caratteri non trasferibili sono l'età e il peso (se non si parla di bagaglio...).

Dei caratteri trasferibili interessa studiare la concentrazione, ovvero la non uniformità della sua distribuzione nel collettivo studiato (es. poche persone spesso detengono la maggior parte del reddito).

I primi 12 paesi per valore delle esportazioni emiliano-romagnole
(Regione Emilia Romagna, 2006, dati in milioni di euro)

Paese	Export	Paese	Export
Germania	4390	Fed. Russa	1021
Francia	4322	Belgio	948
Stati Uniti	4066	Paesi Bassi	934
Spagna	2561	Austria	840
Regno Unito	2396	Grecia	807
Svizzera	1068	Giappone	710

Per un carattere trasferibile ha senso e può essere utile considerare la distribuzione di quantità rispetto a una classificazione delle unità; per esempio i primi dodici paesi per valore delle esportazioni emiliano-romagnole si possono classificare in

- Unione Europea (Germania, Spagna, Francia, Regno Unito, Belgio, Paesi Bassi, Austria, Grecia) ed Extra UE (Stati Uniti, Svizzera, Fed. Russa, Giappone).
- Export $0 \vdash 1000$ (Giappone, Grecia, Austria, Paesi Bassi, Belgio) Export $1000 \vdash 3000$ (Fed. Russa, Svizzera, Regno Unito, Spagna) ed Export $3000 \vdash \infty$ (Stati Uniti, Francia, Germania).

La tabella seguente riporta la distribuzione di quantità dell'export rispetto alla classificazione UE / Extra UE (criterio esterno):

	Export	%
UE	17206	71.48
Extra UE	6865	28.52
Totale	24071	100.00

In pratica è la distribuzione unitaria dell'export (eventualmente espresso in percentuale del totale) nelle unità aggregate date dalla classificazione.

La tabella seguente riporta la distribuzione di quantità dell'export per la classificazione in base all'export stesso (criterio interno):

Export	Quantità	Q. Cum. Inv.	%	% Cum. Inv.	# Paesi
$0 \vdash 1000$	4243	24071	17.63	100.00	5
$1000 \vdash 3000$	7048	19828	29.28	82.37	4
$3000 \vdash \infty$	12780	12780	53.09	53.09	3
Totale	24071		100.00		12

Le quantità cumulate inverse mostrano che il 25% dei paesi (Stati Uniti, Francia e Germania) riceve più del 50% delle esportazioni, mentre il 58% dei paesi (quelli con $Export \geq 1000$) ne riceve più dell'80%; le esportazioni sono dunque, in qualche misura, concentrate.

Si dice serie storica una successione di osservazioni nel tempo; la seguente tabella ne mostra due (ovvero una doppia).

Per capita GDP (Gross Domestic Product) based on PPP (Purchasing-Power-Parity) in current international dollars.

Country	2007	2008	2009	2010	2011	2012	2013	2014
Italy	30478.85	30580.83	29273.91	29079.57	29254.64	29889.11	30766.92	31784.96
UK	35601.15	36522.90	35286.03	34881.40	35401.91	36540.60	37838.03	38807.83

Estimates Start After 2008. Source: International Monetary Fund, World Economic Outlook Database, April 2009.

Mediante opportune serie storiche (un'opportuna serie storica multipla) si può studiare l'evoluzione nel tempo di un fenomeno di interesse.

Qui ci focalizziamo su serie storiche quantitative, ma possono essere interessanti anche serie storiche qualitative (es. partito di maggioranza al 1 gennaio dal 1990 al 2010).

Data una serie storica (semplice) Y e due istanti di tempo s e t, il numero indice (semplice)

$$I_{t/s} = \frac{Y_t}{Y_s}$$

esprime la variazione relativa di Y da s a t; spesso si usa $100 \times I_{t/s}$.

Per esempio, con riferimento alla tabella precedente[1], si trova

$$I^{ITA}_{08/07} = \frac{30580.83}{30478.85} = 1.003346$$

e similmente $I^{UK}_{08/07} = 1.025891$, vale a dire una variazione percentuale dello 0.33% per l'Italia e del 2.6% per il Regno Unito: $\frac{Y_t - Y_s}{Y_s} = \frac{Y_t}{Y_s} - 1$.

[1] In generale non è opportuno valutare variazioni nel tempo sulla base di valori PPP, ma in questo caso particolare si vuole solo illustrare la nozione di numero indice.

Non affrontiamo il problema della costruzione di numeri indici complessi per sintetizzare le variazioni contemporanee di più serie storiche.

L'informazione contenuta in una serie storica può essere espressa mediante una serie di numeri indici in base fissa, scegliendo un opportuno istante di tempo come riferimento:

Country	2007	2008	2009	2010	2011	2012	2013	2014
Italy	99.67	100.00	95.73	95.09	95.66	97.74	100.61	103.94
UK	97.48	100.00	96.61	95.51	96.93	100.05	103.60	106.26

confrontiamo le economie di Italia e Regno Unito "facendo 100" il PIL pro capite a parità di potere d'acquisto nell'ultimo anno di osservazione (calcolando cioè $100 \times I_{t/2008}$, per t da 2007 a 2014, separatamente per le due serie storiche).

Per cambiare base in una serie di numeri indici a base fissa è sufficiente procedere come se i numeri indici fossero i valori della serie originale, mentre per ricostruire quest'ultima è necessario conoscere uno dei suoi valori (diciamo quello di base).

L'informazione contenuta in una serie storica può essere espressa mediante una serie di numeri indici in base mobile, scegliendo come riferimento, per ogni istante di tempo, l'istante precedente:

Country	2007	2008	2009	2010	2011	2012	2013	2014
Italy	NA	100.33	95.73	99.34	100.60	102.17	102.94	103.31
UK	NA	102.59	96.61	98.85	101.49	103.22	103.55	102.56

confrontiamo le economie di Italia e Regno Unito in termini di variazioni percentuali rispetto all'anno precedente (calcolando $100 \times I_{t/(t-1)}$, per t da 2008 a 2014, separatamente per le due serie storiche).

Va da sé che i numeri indice in base mobile dipendono dalla frequenza delle osservazioni (es. annuale/trimestrale per il PIL nei Conti Economici dell'Istat $\Rightarrow$ variazioni tendenziali/congiunturali).

Passare da base fissa a base mobile è immediato, mentre passare da base mobile a base fissa richiede una successione di moltiplicazioni (Borra & Di Ciaccio, 2008, p. 109).

Per un approfondimento sul confronto tra Italia e UK in termini di PIL si rinvia a un articolo di Sandro Brusco sul blog noiseFromAmerika

`http://www.noisefromamerika.org/index.php/articoli/1612`

dove è anche presente un collegamento ai dati qui analizzati.

Assai spesso è utile trasformare la rappresentazione tabellare di una distribuzione in un'immagine grafica che permetta all'occhio di leggere con facilità le caratteristiche salienti della distribuzione.

Un buon grafico (Borra & Di Ciaccio, 2008, p. 36) dovrebbe essere:

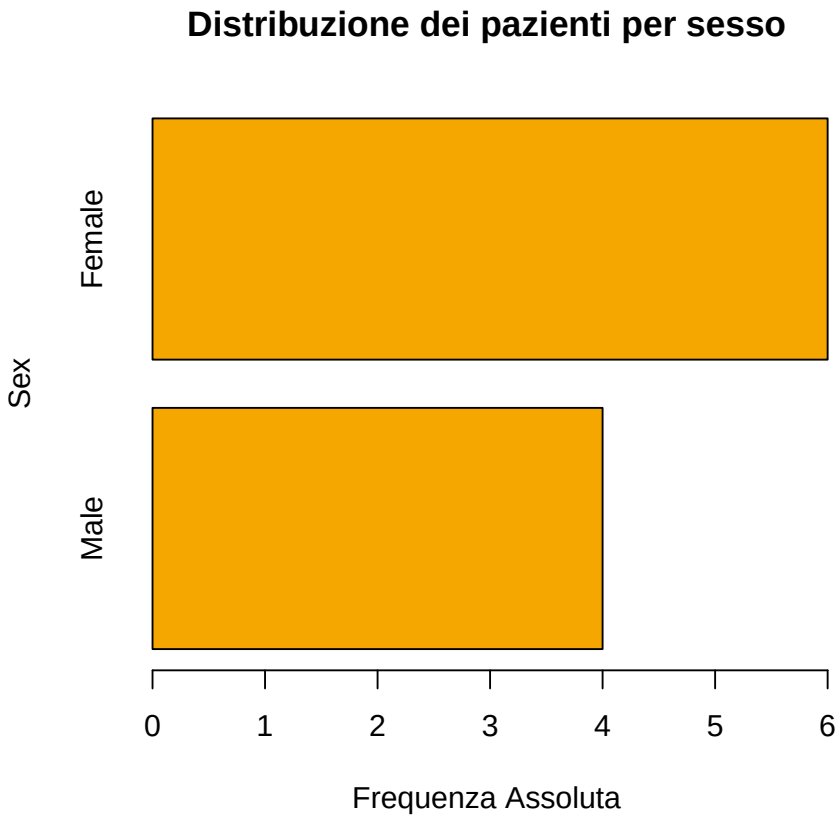
- accurato (con tutti i dettagli importanti della distribuzione);
- semplice (senza elementi superflui, quali possono essere tridimensionalità e prospettiva);
- chiaro (di facile lettura, anche prescindendo dal testo che lo discute);
- di bell'aspetto (armonioso nelle proporzioni e nei colori);
- strutturato (con una gerarchia di elementi).

Per rappresentare la distribuzione di frequenza di un carattere qualitativo su scala nominale (o dicotomico) si può usare un grafico a nastri: ogni modalità sarà rappresentata da un nastro (barra orizzontale) avente lunghezza proporzionale alla frequenza della modalità.

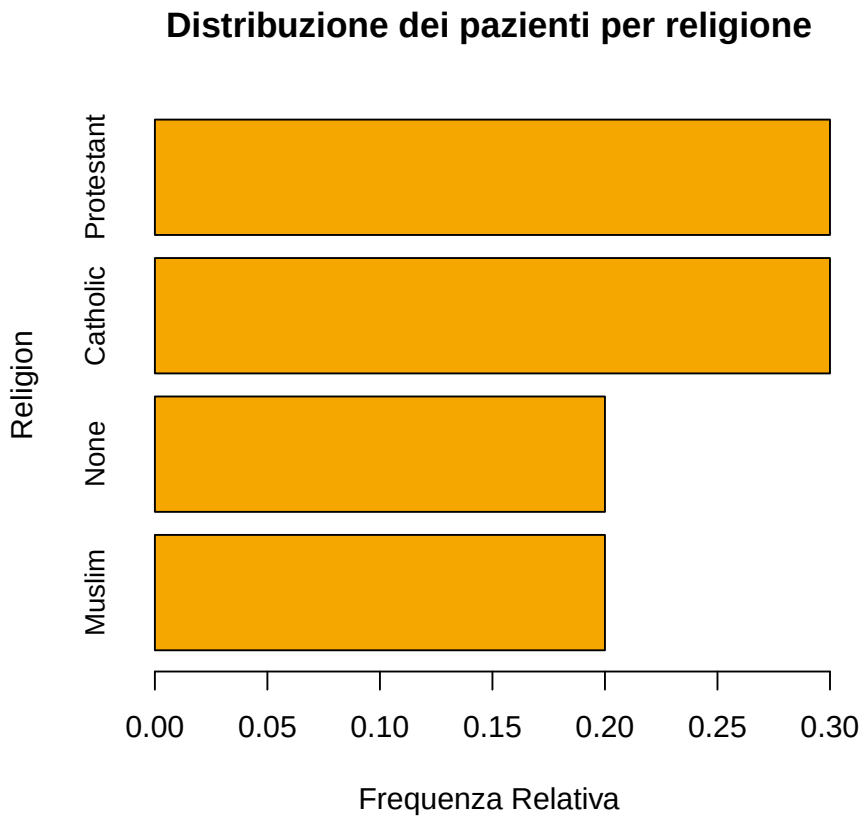
In un grafico a nastri:

- i nastri hanno tutti la stessa larghezza;
- i nastri sono distanziati tra loro;
- i nastri sono generalmente ordinati dal più lungo al più corto.

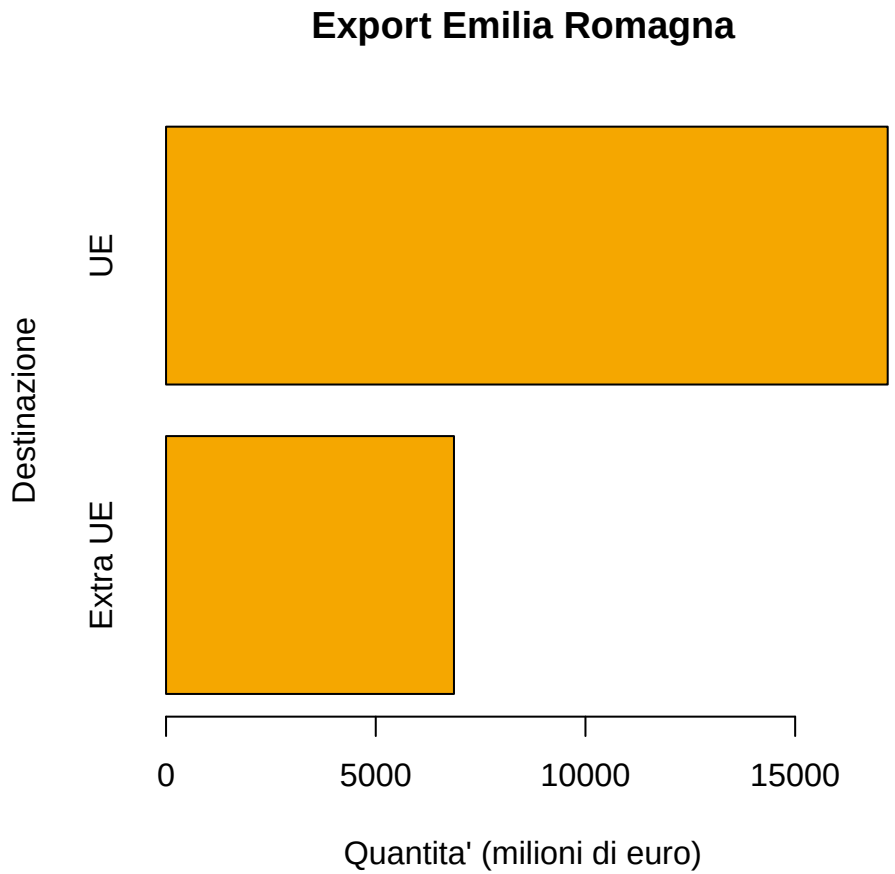
Si può usare un grafico a nastri anche per rappresentare una distribuzione di quantità rispetto a una classificazione non ordinata (o dicotomica).




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
barplot(sort(table(X$Sex)),
        horiz = TRUE,
        col = colE,
        main = "Distribuzione dei pazienti
                per sesso",
        xlab = "Frequenza Assoluta",
        ylab = "Sex")
```




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
relifreq <- table(X$Religion)
barplot(sort(relifreq)/sum(relifreq),
        horiz = TRUE,
        col = colE,
        main = "Distribuzione dei pazienti
                per religione",
        xlab = "Frequenza Relativa",
        ylab = "Religion",
        cex.names = 0.9)
```


Export Emilia Romagna
Destinazione
UE
Extra UE
0
5000
10000
15000
Quantita' (milioni di euro)

Qualora si voglia rappresentare la distribuzione di un carattere qualitativo su scala nominale in due o più collettivi a confronto, si può ricorrere a un grafico a nastri multipli: per ogni modalità saranno riportati nel grafico tanti nastri quanti sono i collettivi.

In alternativa ci si può avvalere di un grafico a nastri suddivisi: ogni collettivo sarà rappresentato da un nastro suddiviso in tante parti quante sono le modalità del carattere, ognuna avente lunghezza proporzionale alla modalità rappresentata.

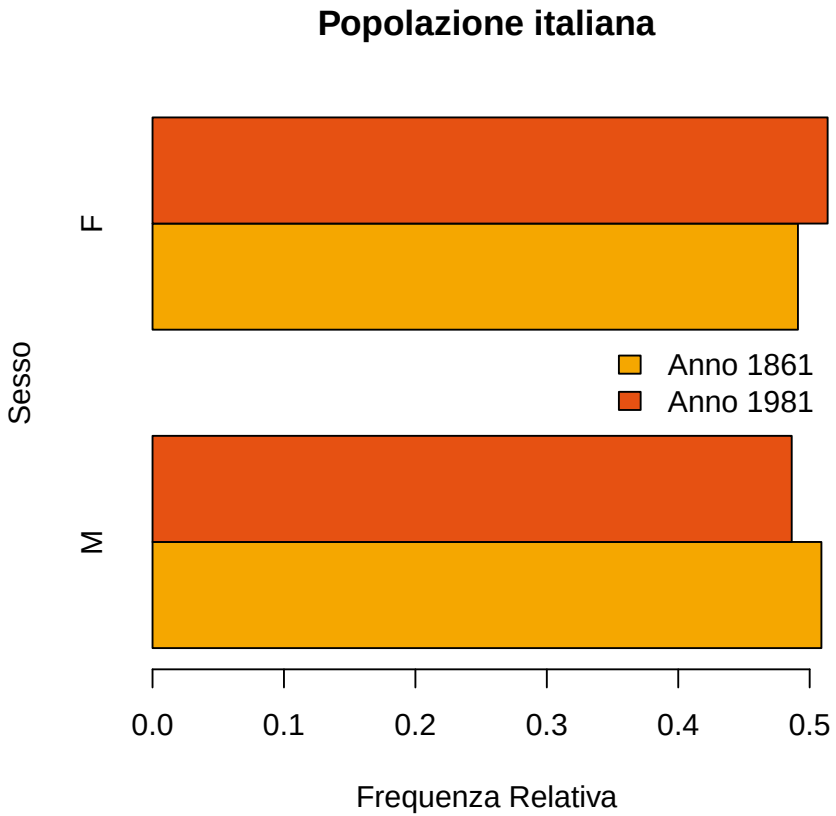
Se il carattere ha tre o più modalità è preferibile usare un grafico a nastri multipli, perché in un grafico a nastri suddivisi le modalità rappresentate in posizione interna sono difficilmente confrontabili (es. nel grafico più avanti i Blue Eyes sono più frequenti tra i Red Hair o tra i Black Hair?)

(Preliminarmente) costruiamo in R una matrice contenente i dati sulla popolazione residente in Italia riportati da Pace & Salvan...

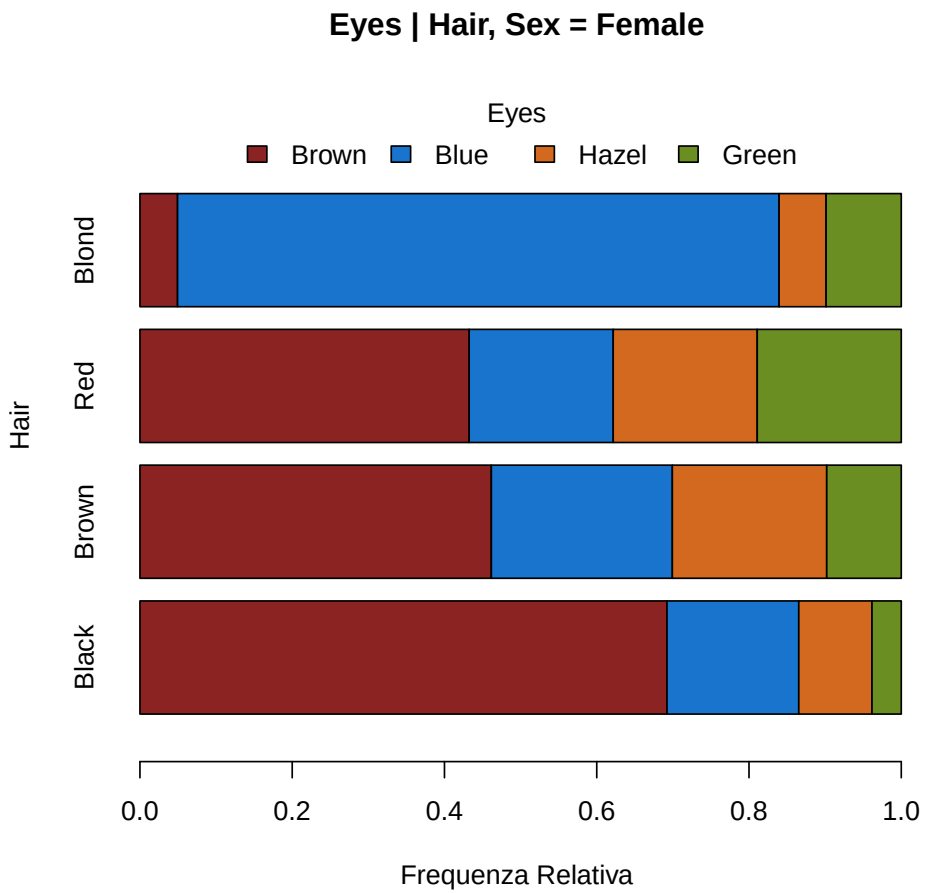
```
> popit <- matrix(c(13399, 27506, 12929, 29051), 2, 2)
> rownames(popit) <- c("Anno 1861", "Anno 1981")
> colnames(popit) <- c("M", "F")
> popit
              M     F
Anno 1861 13399 12929
Anno 1981 27506 29051
```

...e calcoliamo le frequenze relative di uomini e donne nel vecchio e nel nuovo censimento:

```
> relpopit <- prop.table(popit, 1)
> relpopit
                   M         F
Anno 1861 0.5089259 0.4910741
Anno 1981 0.4863412 0.5136588
```



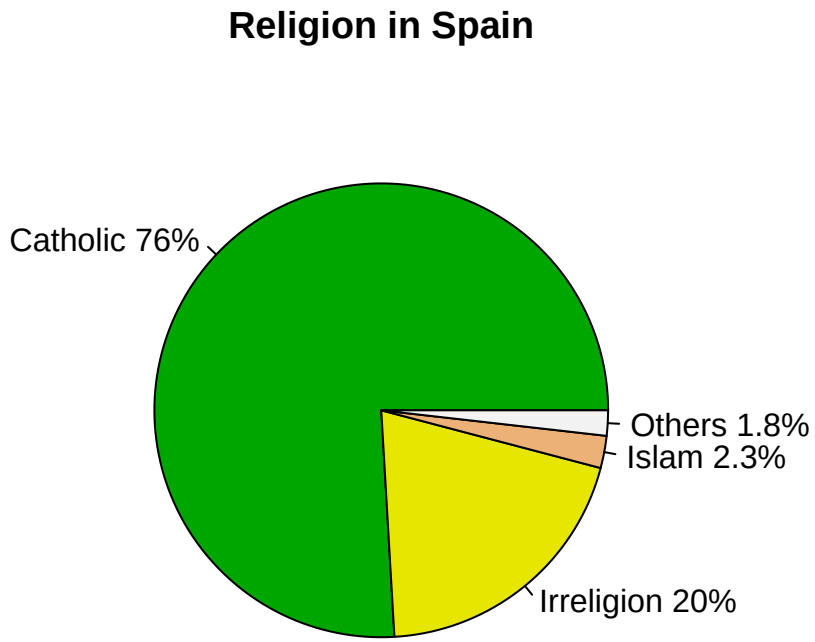

```
colC <- rgb(0.9023438,0.3164062,0.0703125)
colE <- rgb(0.9609375,0.6562500,0.0000000)
barplot(relpopit,
        beside = TRUE,
        horiz = TRUE,
        col = c(colE, colC),
        main = "Popolazione italiana",
        xlab = "Frequenza Relativa",
        ylab = "Sesso")
legend("right",
       legend = rownames(relpopit),
       fill = c(colE, colC),
       bty = "n")
```


Eyes | Hair, Sex = Female
Eyes
Brown
Blue
Hazel
Green
Hair
Blond
Red
Brown
Black
0.0
0.2
0.4
0.6
0.8
1.0
Frequenza Relativa

Una rappresentazione grafica alternativa[2] per un carattere qualitativo osservato in un singolo collettivo è il grafico a torta: ogni modalità sarà rappresentata da un settore circolare (spicchio o fetta) con angolo al centro proporzionale alla corrispondente frequenza.

La figura seguente mostra (sulla base di dati la cui attendibilità non può essere data per scontata) come la Spagna possa dirsi un paese a maggioranza cattolica con una cospicua minoranza di non credenti...

[2] Si tratta di una rappresentazione grafica accattivamente, per la sua concretezza nell'esprimere un tutto come somma di parti, che però presenta dei limiti (discussi più avanti) in termini di facilità di lettura; va da sé che per sfruttare la concretezza di questa rappresentazione le parti devono esser disgiunte (cosa che per esempio non accade nella copertina del Venerdì di Repubblica del 29 febbraio 2008) oltre che "poche".



Source: Wikipedia (Religion in Spain) 23 Jan 2010

Ci sono buone ragione per evitare i grafici a torta:

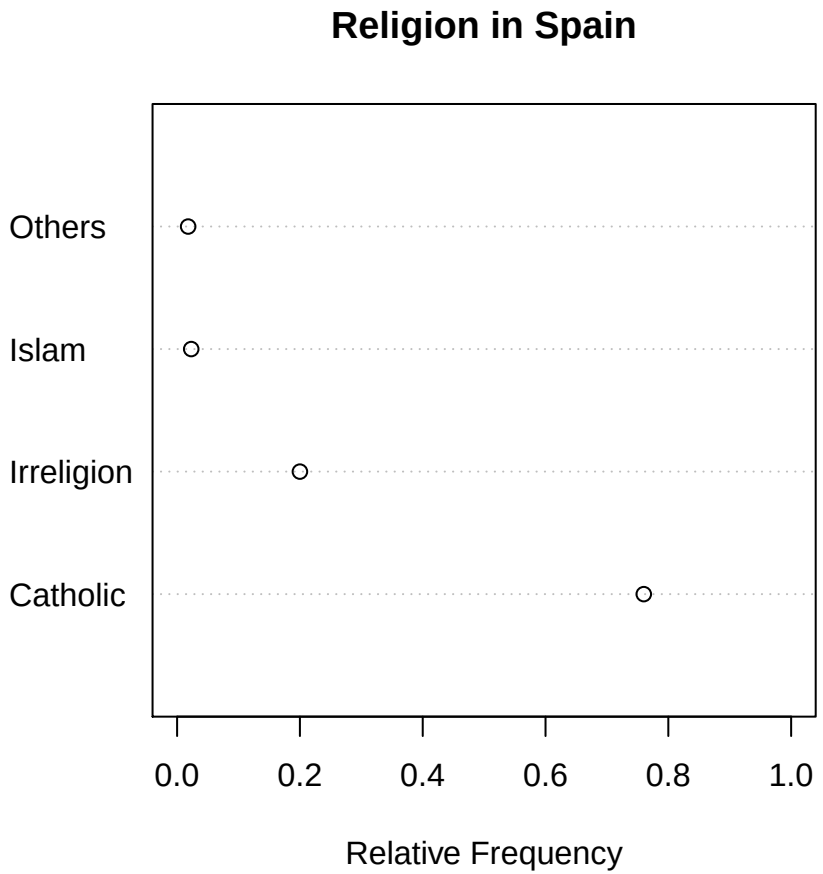
> *Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements (Cleveland, 1985, p. 264)*[3]

Come riportato nella pagina di aiuto del comando `pie` di R

> *This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists*

e il ragionamento resta valido nel confronto tra grafici a torta e grafici a nastri (i dot chart, di cui segue per completezza un esempio, sono una variante "minimalista" dei grafici a nastri proposta da Cleveland).

[3] Cleveland, W. S. (1985). The elements of graphing data. Wadsworth, Monterey.


Religion in Spain
Others
Islam
Irreligion
Catholic
0.0
0.2
0.4
0.6
0.8
1.0
Relative Frequency

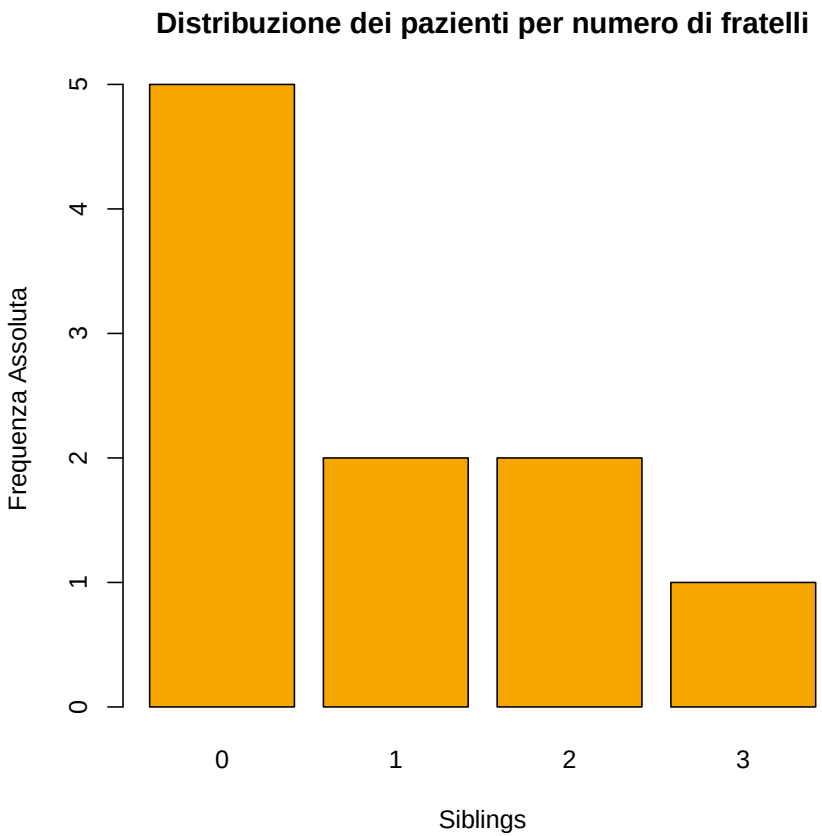
Per rappresentare la distribuzione di frequenza di un carattere qualitativo su scala ordinale o quantitativo discreto si può usare un grafico a barre: ogni modalità sarà rappresentata da una barra (verticale) avente lunghezza proporzionale alla frequenza della modalità.

In un grafico a barre:

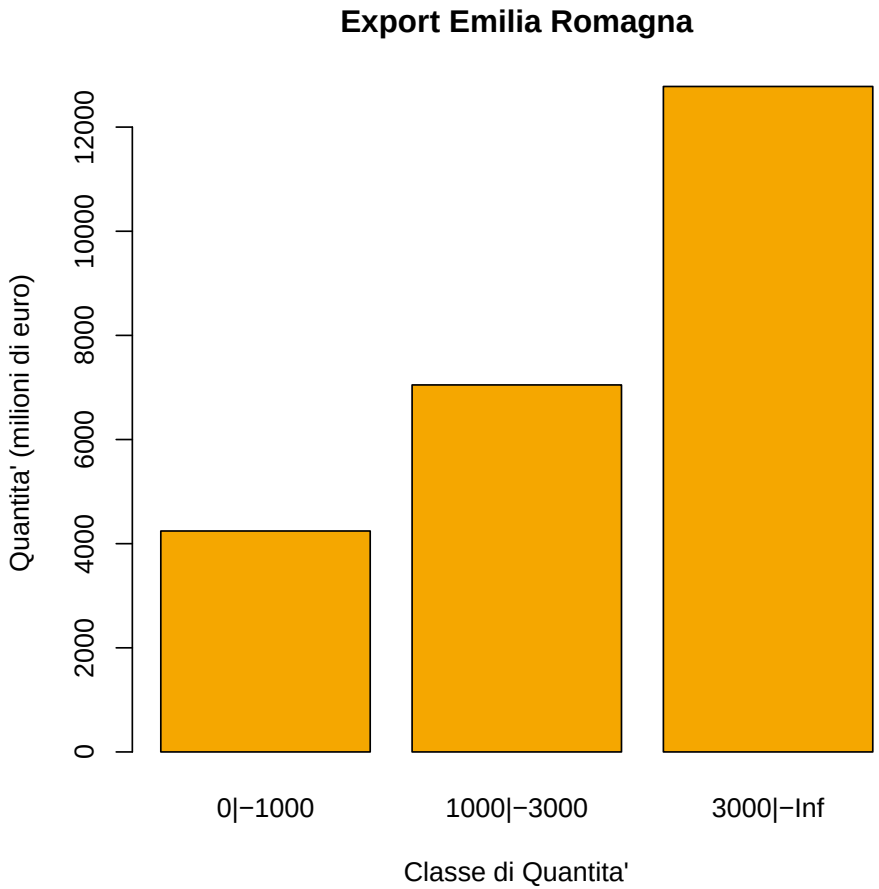
- le barre hanno tutte la stessa larghezza;
- le barre sono distanziate tra loro;
- le barre sono ordinate secondo l'ordine delle modalità.

Si può usare un grafico a barre anche per rappresentare una distribuzione di quantità rispetto a una classificazione ordinata.


Distribuzione dei pazienti per stato di salute
Frequenza Relativa
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Very Poor
Poor
Average
Good
Very Good
Health



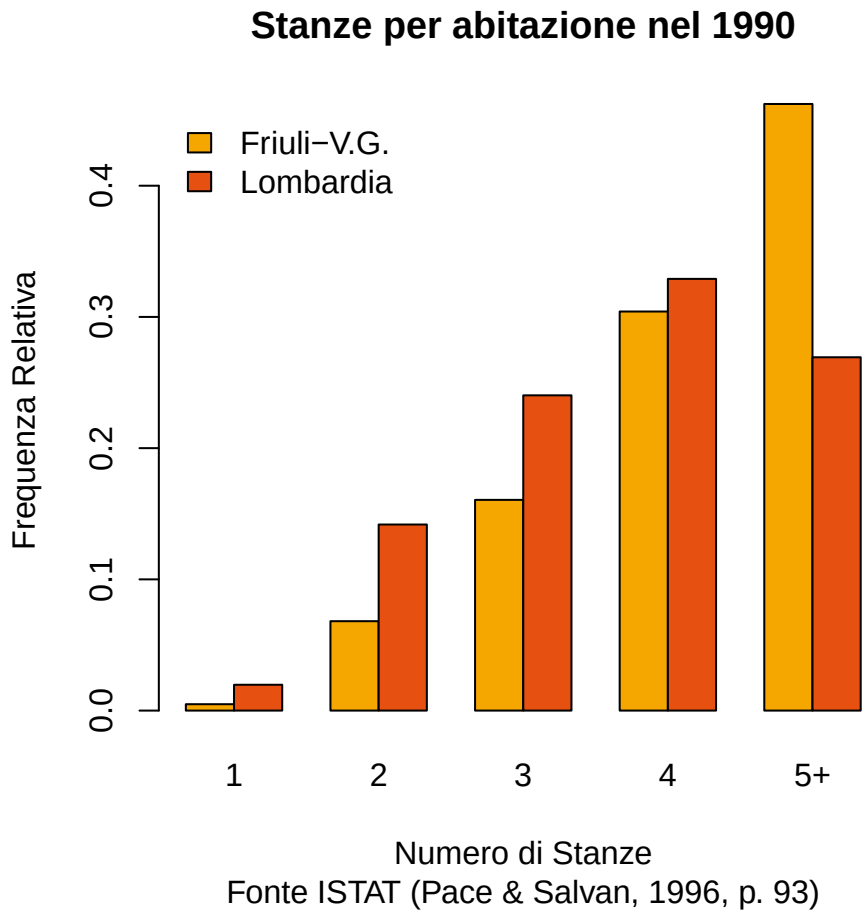

```
colE <- rgb(0.9609375,0.6562500,0.0000000)
barplot(table(X$Siblings),
        horiz = FALSE,
        col = colE,
        main = "Distribuzione dei pazienti
                per numero di fratelli",
        xlab = "Siblings",
        ylab = "Frequenza Assoluta")
```


Export Emilia Romagna
Quantita' (milioni di euro)
0
2000
4000
6000
8000
10000
12000
0|−1000
1000|−3000
3000|−Inf
Classe di Quantita'

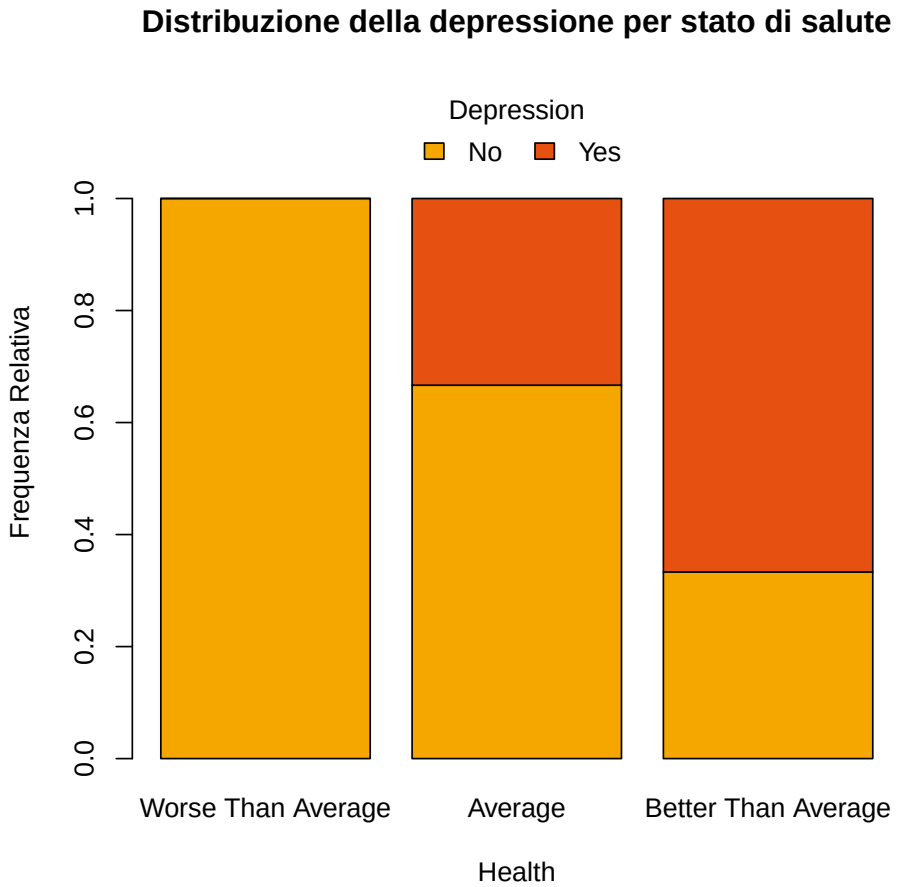
Va da sé che l'uso di nastri (barre orizzontali) per caratteri su scala nominale e di barre (verticali) per caratteri su scala ordinale è più che altro un'utile convenzione per distinguere i due casi.

Ci si può avvalere di un grafico a barre multiple per rappresentare la distribuzione di un carattere qualitativo su scala ordinale o quantitativo discreto in due o più collettivi a confronto.

Un grafico a barre suddivise permette di confrontare la distribuzione di un carattere (preferibilmente dicotomico) in due o più collettivi definiti da un carattere ordinato.



Fonte ISTAT (Pace & Salvan, 1996, p. 93)

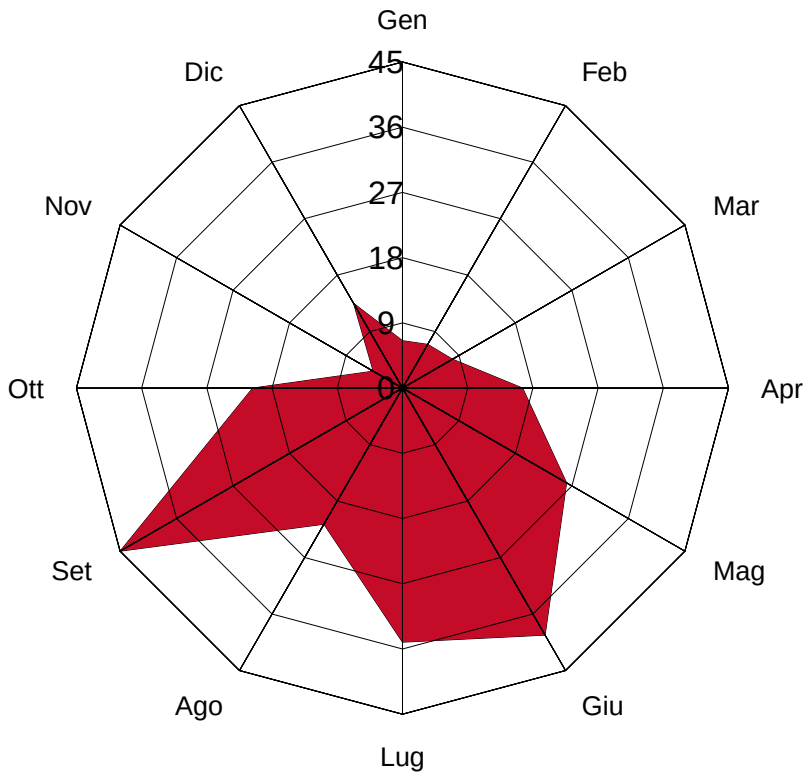

Distribuzione della depressione per stato di salute
Depression
No
Yes
Frequenza Relativa
0.0
0.2
0.4
0.6
0.8
1.0
Worse Than Average
Average
Better Than Average
Health

Per rappresentare la distribuzione di frequenza di un carattere qualitativo ciclico (o una distribuzione di quantità rispetto a una classificazione ciclica[4]) si può usare un grafico radar: si suddividerà l'angolo giro in tante parti uguali quante sono le modalità del carattere (classi della classificazione) e, sui raggi così ottenuti, si tracceranno dei segmenti di lunghezza proporzionale alle rispettive frequenze (quantità); infine, per maggiore chiarezza, si campirà l'area del poligono individuato dagli estremi dei segmenti.

La figura seguente mostra la distribuzione dei matrimoni per mese, in Italia, nel 2006 (fonte: ISTAT, `http://demo.istat.it`)...

[4] es. fatturato per mese

Matrimoni in Italia nel 2006 (in migliaia)



Per rappresentare la distribuzione di frequenza di un carattere quantitativo continuo suddiviso in classi si può usare un istogramma: ogni classe sarà rappresentata da una barra (verticale) avente per base l'intevallo che la definisce e area proporzionale alla sua frequenza.

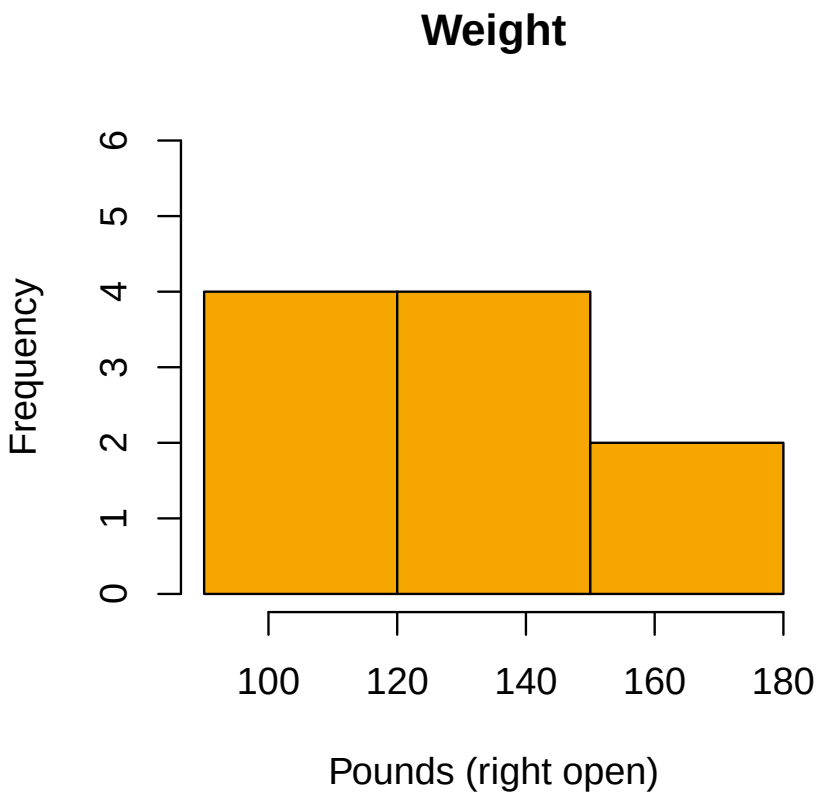
In un istogramma:

- le barre possono avere o meno tutte la stessa larghezza, a seconda che le classi abbiano o meno tutte la stessa ampiezza;
- in generale è l'area (e non l'altezza) delle barre a essere proporzionale alla frequenza;
- le barre sono adiacenti tra loro.

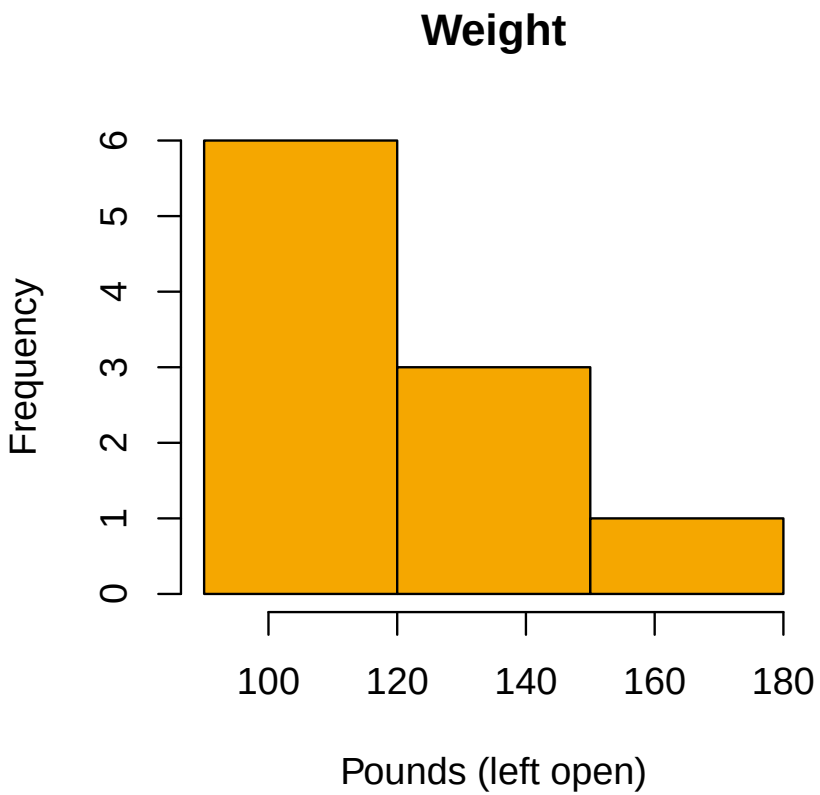
Se un istogramma si basa su classi tutte della stessa ampiezza, le sue barre avranno tutte la stessa larghezza e le loro altezze (oltre che le loro aree) saranno proporzionali alle rispettive frequenze: si parlerà di istogramma a basi regolari.

In questo caso l'unica differenza con un grafico a barre è che le barre sono adiacenti invece che distanziate (in modo da ricordare che si tratta di un carattere continuo).

Un istogramma dipende chiaramente dalla suddivisione in classi su cui si basa e in particolare dal fatto che si usino intervalli aperti a destra o a sinistra per definirla.




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
hist(X$Weight,
     breaks = c(90, 120, 150, 180),
     right = FALSE, # not right closed
     col = colE,
     main = "Weight",
     xlab = "Pounds (right open)",
     ylim = c(0, 6))
```




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
hist(X$Weight,
     breaks = c(90, 120, 150, 180),
     right = TRUE, # right closed
     col = colE,
     main = "Weight",
     xlab = "Pounds (left open)"),
     ylim = c(0, 6))
```

Se un istogramma si basa su classi di ampiezza diversa, le sue barre avranno larghezza diversa e le loro altezze non saranno più proporzionali alle rispettive frequenze.

In generale le altezze delle barre di un istogramma saranno proporzionali alle cosiddette densità di classe:

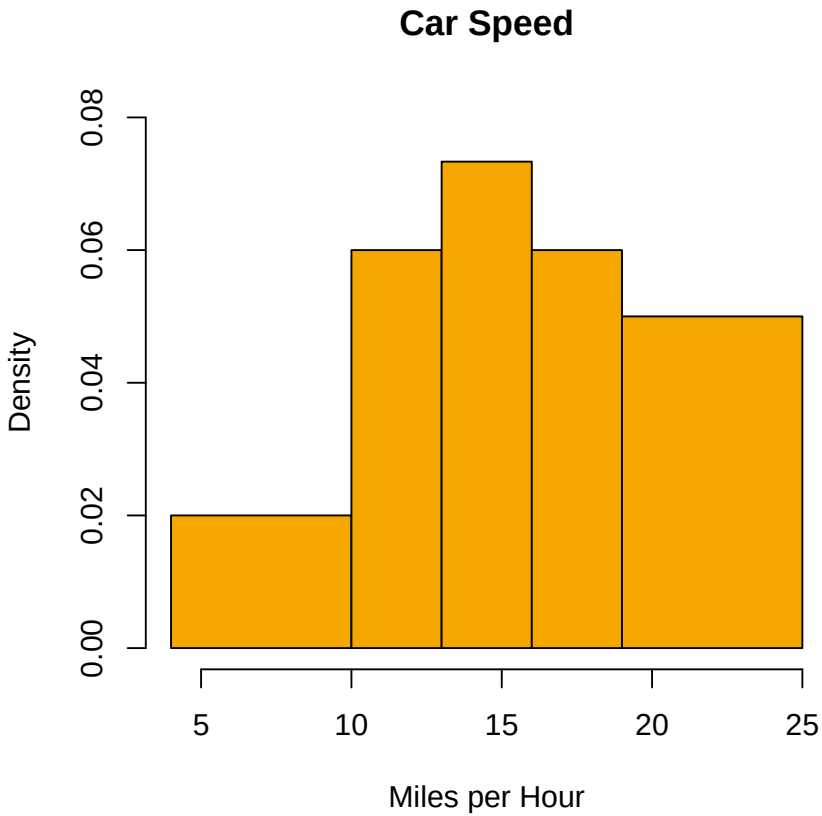
$$h_i = \frac{f_i}{a_i},$$

dove f_i e a_i sono rispettivamente la frequenza relativa e l'ampiezza dell'i-esima classe ($i = 1, \ldots, k$ per un istogramma con k classi).

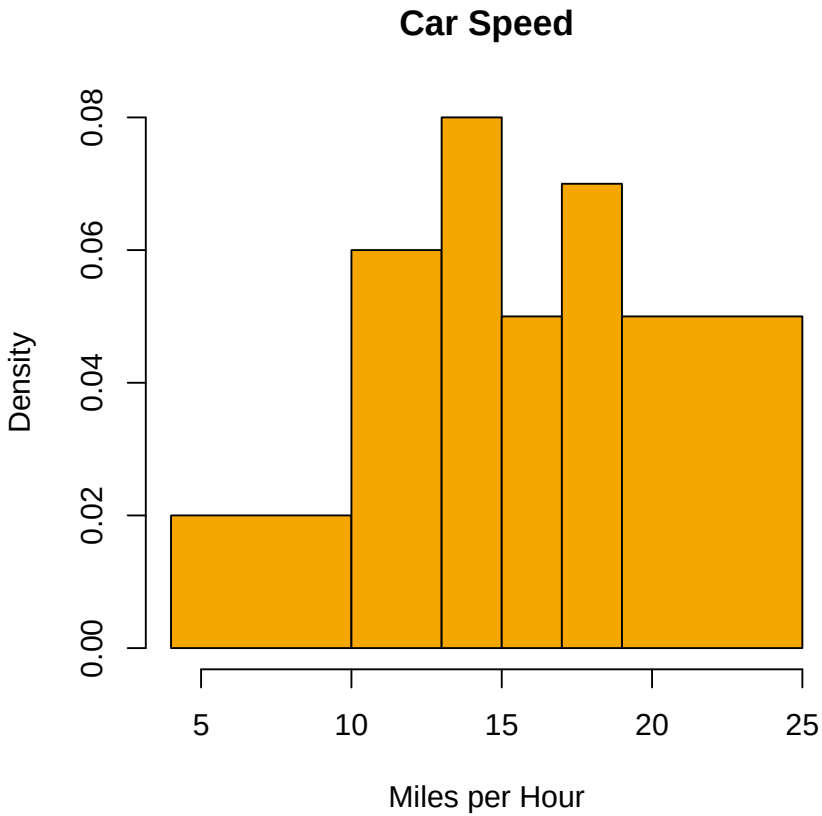
Per quanto riguarda l'unità di misura, se per esempio il carattere rappresentato è in Kg, le densità saranno in % / Kg ("affollamento").

Il dataset Cars di R contiene la velocità (Speed, in miglia per ora) e la distanza di arresto (Dist, in piedi) per un campione di 50 auto degli Anni Venti; la tabella seguente riporta una distribuzione di frequenza del carattere Speed (4 e 25 sono la minima e la massima velocità rilevate) completata con le ampiezze e le densità di classe.

Speed	Freq.	%	Bin Width (mph)	Density (% / mph)
4 ⊢ 10	6	12	6	2.00
10 ⊢ 13	9	18	3	6.00
13 ⊢ 16	11	22	3	7.33
16 ⊢ 19	9	18	3	6.00
19 ⊢ 25	15	30	6	5.00
Total	50	100	21	




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
hist(cars$speed,
     breaks = c(4, 10, 13, 16, 19, 25),
     right = FALSE, # not right closed
     col = colE,
     main = "Car Speed",
     xlab = "Miles per Hour",
     ylim = c(0, 0.08))
```




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
hist(cars$speed,
     breaks = c(4, 10, 13, 15, 17, 19, 25),
     right = FALSE, # not right closed
     col = colE,
     main = "Car Speed",
     xlab = "Miles per Hour",
     ylim = c(0, 0.08))
```

Speed	Freq.	%	Bin Width (mph)	Density (% / mph)
4 ⊢ 10	6	12	6	2.00
10 ⊢ 13	9	18	3	6.00
13 ⊢ 15	8	16	2	8.00
15 ⊢ 18	8	16	3	5.33
18 ⊢ 20	7	14	2	7.00
20 ⊣ 25	12	24	5	4.80
Total	50	100	21	

Come si vede l'istogramma dipende da quali e quanti intervalli si scelgono per suddividere il carattere in classi; in ogni caso, per costruzione, l'area sottesa a un istogramma è pari a 1 (100%).

Posto di avere a disposizione la distribuzione unitaria del carattere, come sceglieremo quali e quanti intervalli usare?

Conviene senz'altro sperimentare diverse soluzioni, esplorando la distribuzione del carattere studiato e cercando un istogramma di bell'aspetto che ne sia rappresentativo.

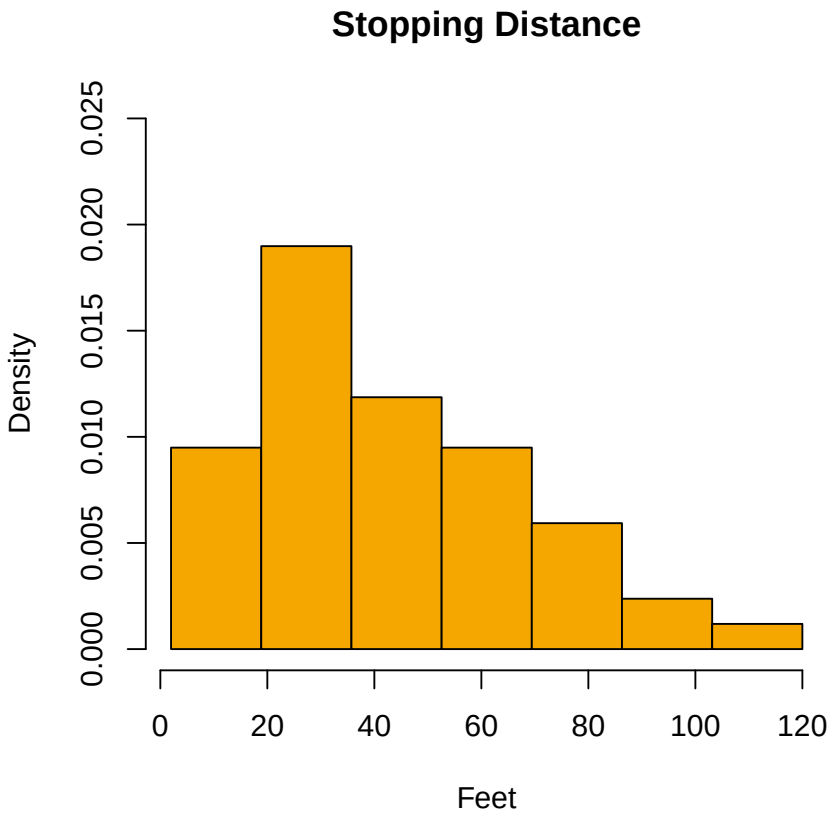
La regola di Sturges, risalente al 1926 e basata su un criterio di armonia delle proporzioni, suggerisce di usare

$$k \simeq 1 + \log_2 n$$

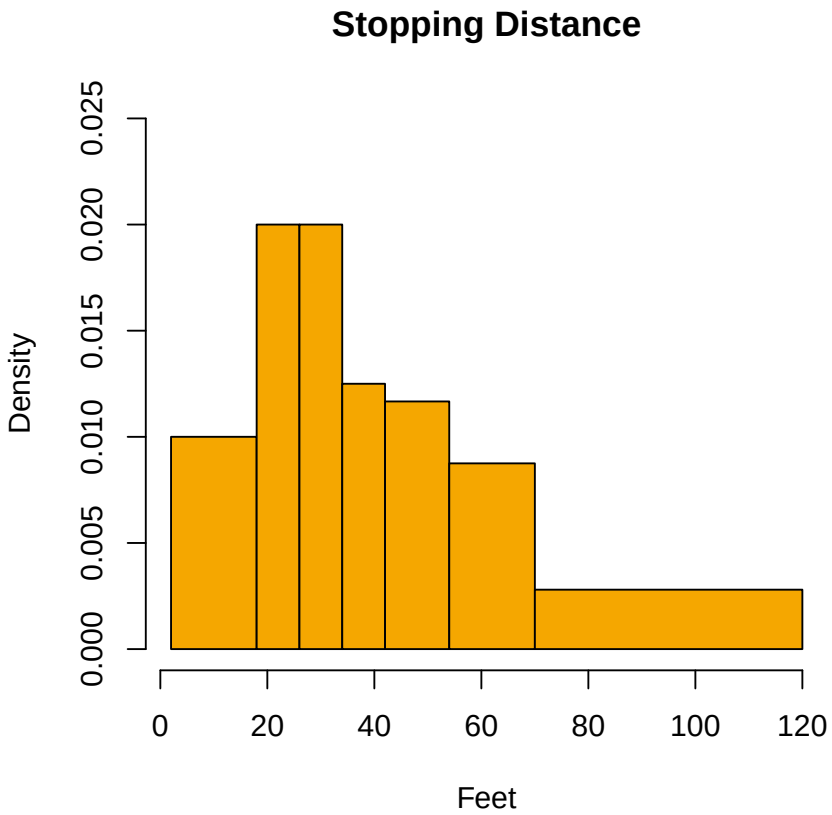
classi per n osservazioni. Il suo impiego è piuttosto diffuso, in pratica, sebbene in teoria vi siano regole migliori (secondo criteri di stima della densità). Per $n = 50$ si ottiene $k \simeq 6.6$ (es. $k = 7$).

Una volta che si sia fissato il numero delle classi, si può procedere in (almeno) due modi:

- si può optare per classi tutte della stessa ampiezza (nel qual caso non è indispensabile calcolarne le densità, ma è comunque utile avere un grafico con area sottesa pari a 100%);
- si può cercare di fare in modo che via sia in ogni classe all'incirca lo stesso numero di osservazioni (la densità di una classe con poche osservazioni è molto sensibile al variare dei sui estremi, se la variazione comporta l'ingresso/uscita di una o più unità).



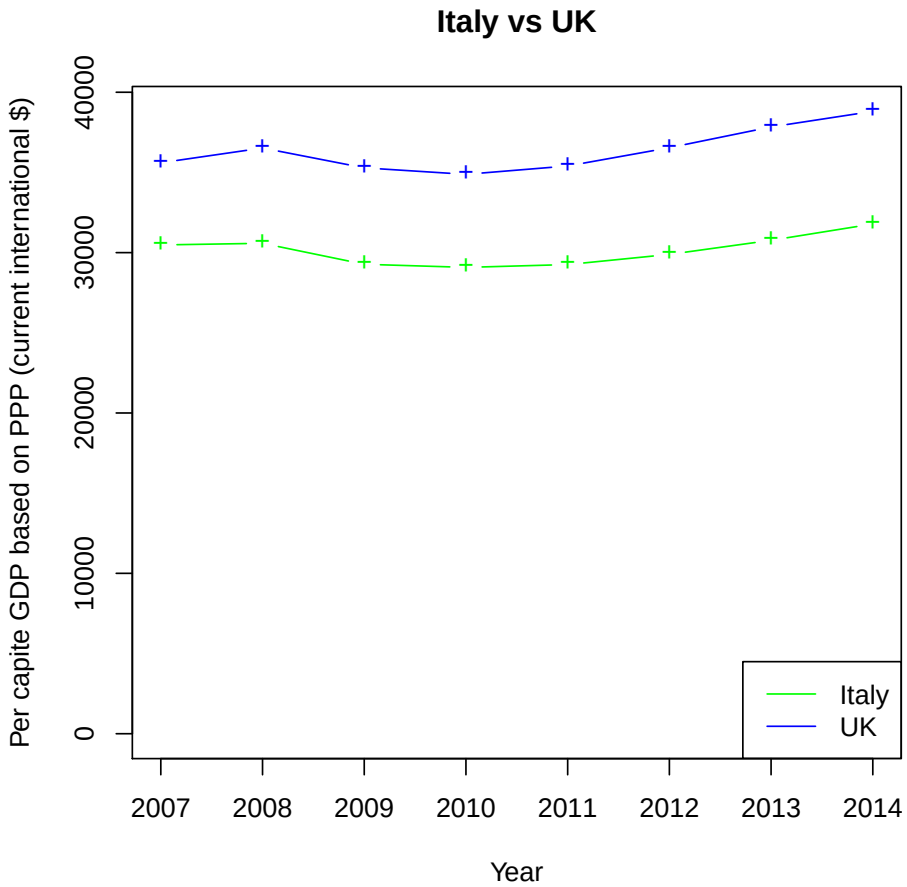

```
colE <- rgb(0.9609375,0.6562500,0.0000000)
hist(cars$dist,
     breaks = seq(2, 120, len = 8),
     freq = FALSE,
     col = colE,
     main = "Stopping Distance",
     xlab = "Feet",
     ylim = c(0, 0.025))
```




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
hist(cars$dist,
     breaks = quantile(cars$dist,
                       seq(0, 1, len = 8)),
     col = colE,
     main = "Stopping Distance",
     xlab = "Feet",
     ylim = c(0, 0.025))
```

Per rappresentare una serie storica (quantitativa) si può usare un diagramma cartesiano: ogni osservazione sarà rappresentata da un simbolo (es. una croce) centrato nel punto che ha come coordinate, rispetto a un'opportuna coppia di assi cartesiani ortogonali, l'istante di tempo in cui l'osservazione si colloca e il valore osservato; le coppie di simboli relativi a osservazioni successive saranno collegate con dei segmenti (di linea) in modo da realizzare una (linea) spezzata che suggerisca all'occhio lo scorrere del tempo.

La figura seguente confronta l'andamento del PIL di Italia e Regno Unito nelle previsioni del Fondo Monetario Internazionale...


Italy vs UK
Per capite GDP based on PPP (current international $)
0
10000
20000
30000
40000
2007
2008
2009
2010
2011
2012
2013
2014
Year
Italy
UK

Se si vuole rappresentare graficamente l'evoluzione nel tempo di una distribuzione di frequenza, ci si può avvalere di un grafico ad aree: si rappresentano le serie storiche delle frequenze cumulate (rispetto a un ordine più o meno arbitrario delle modalità) mediante delle spezzate (come in un diagramma cartesiano) e si campiscono con colori diversi le aree comprese tra spezzate successive.

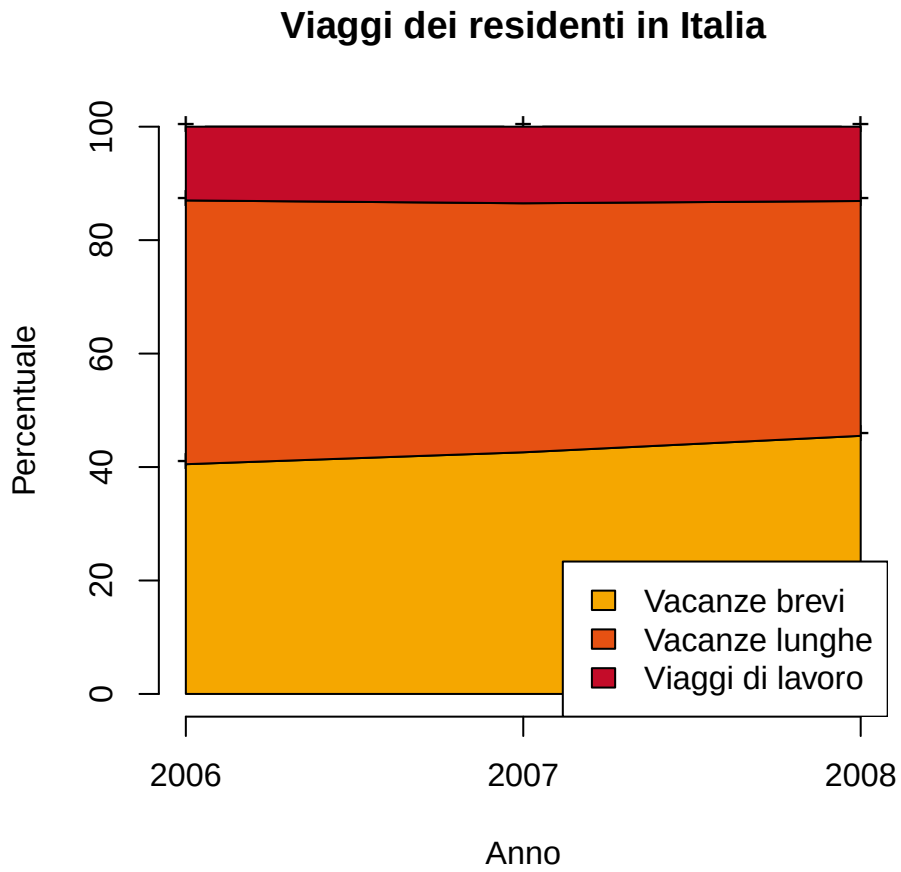
La tabella seguente (comunicato stampa ISTAT del 18 febbraio 2009) classifica i viaggi (con pernottamento) dei residenti in Italia in vacanze brevi (da 1 a 3 notti) vacanze lunghe (almeno 4 notti) e viaggi di lavoro, riportando i dati per gli anni 2006, 2007 e 2008. . .

Viaggi e notti per tipologia del viaggio: valori in migliaia e composizioni percentuali.

	Vacanza Corta		Vacanza Lunga		Lavoro		Totale	
Anno	F. Ass.	F. Perc.	F. Ass.	F. Perc.	F. Ass.	F. Perc.	F. Ass.	F. Perc.
2006	43662	40.5	50228	46.5	14006	13.0	107895	100.0
2007	47911	42.6	49262	43.9	15196	13.5	112369	100.0
2008	55919	45.5	50891	41.4	16128	13.1	122938	100.0

I dati relativi al 2008 sono preliminari (provvisori).

Il successivo grafico ad aree mostra come nel tempo aumenti la percentuale di vacanze brevi a scapito di quelle lunghe. . .


Viaggi dei residenti in Italia
Percentuale
0
20
40
60
80
100
2006
2007
2008
Anno
Vacanze brevi
Vacanze lunghe
Viaggi di lavoro

Si parla di dati spaziali quando si ha una collezione di osservazioni relative a località o aree geografiche; la seguente tabella costituisce un esempio del secondo tipo (tratto dagli indicatori socio-sanitari regionali dell'ISTAT).

Per rappresentare graficamente un insieme di dati spaziali (relativi ad aree geografiche) con osservazioni quantitative (o comunque ordinali) si può usare un cartogramma (a ripartizioni colorate):
si campiscono le aree di un'opportuna mappa con colori di intensità tanto maggiore (o minore) quanto più grandi sono i valori osservati, avendo fissato tonalità e saturazione (`http://it.wikipedia.org/wiki/Hue_Saturation_Brightness`);
si veda per esempio la Figura 2.10.1 di Borra & Di Ciaccio (2008).

Spesa sanitaria pubblica per funzione economica e regione nel 2007 (euro correnti pro capite)

Regione	Spesa a Gestione Diretta	Altre Spese[5]	Spese in Convenzione	Totale
Piemonte	982	114	613	1709
Valle d'Aosta	1276	136	502	1914
Lombardia	801	99	733	1633
Trentino-Alto Adige	1236	132	536	1904
Veneto	901	108	628	1638
Friuli-Venezia Giulia	1115	117	481	1714
Liguria	1118	128	635	1881
Emilia-Romagna	1035	124	539	1697
Toscana	1092	129	466	1687
Umbria	1055	122	480	1657
Marche	994	110	497	1601
Lazio	938	111	876	1925
Abruzzo	929	119	682	1730
Molise	1033	103	811	1947
Campania	868	106	688	1663
Puglia	853	93	695	1641
Basilicata	993	107	553	1653
Calabria	976	125	707	1808
Sicilia	873	113	680	1666
ITALIA	936	111	657	1703

Fonte: ISTAT, Contabilità Nazionale

[5] Servizi amministrativi, interessi passivi, imposte e tasse, premi di assicurazione, contribuzioni diverse.


BORRA, S. & DI CIACCIO, A. (2008).
Statistica: Metodologie per le Scienze Economiche e Sociali (Seconda Edizione).
McGraw-Hill, Milano.


EVERITT, B. (2005).
An R and S-PLUS® Companion to Multivariate Analysis.
Springer-Verlag, London.


PACE, L. & SALVAN, A. (1996).
Introduzione alla Statistica, I: Statistica Descrittiva.
CEDAM, Padova.


REGIONE EMILIA ROMAGNA (2006).
I Numeri dell'Emilia Romagna.
CLEUB, Bologna.


STEVENS, S. S. (1946).
On the theory of scales of measurement.
Science **103**, 677–680.


VELLEMAN, P.F. & WILKINSON, L. (1993).
Nominal, ordinal, interval and ratio typologies are misleading.
The American Statistician **47**, 65–72.