Data Analysis

Riassunti numerici e prime rappresentazioni grafiche

Nicola Torelli

18/3/2021

Contents

I dati **1**
Popolazione e unità statistiche 1
Analisi esplorativa dei dati 3

Tipi di dati **3**
Variabili statistiche 3
La matrice dei dati 5

Tabelle di frequenza **8**
Tabelle di frequenza semplici 8
Tabelle a doppia entrata 10

I dati

Nel seguito faremo riferimento al termine **dati** intendendo, in generale, un insieme di informazioni, di solito piuttosto ampio, raccolte secondo una varietà di schemi e per scopi di diversa natura, relativamente a un insieme di **unità**.

L'aspetto di cui ci occuperemo nel seguito è quello di mettere in luce aspetti interessanti relativamente al collettivo di unità ed evidenziare tendenze e pattern caratteristici mediante specifiche tecniche di analisi, incluso il ricorso a rappresentazioni grafiche.

Questo fase viene denominata "analisi esplorativa dei dati", e talvolta si omette l'aggettivo esplorativo, oppure "analisi descrittiva". L'uso di appropriati riassunti numerici e rappresentazioni grafiche che caratterizza l'EDA (esploratory data analysis secondo il termine coniato da W. Tukey nel secolo scorso) ha una storia piuttosto lunga. L'aspetto essenziale è che in tale fase si vorrebbe che siano i dati stessi a parlare, senza ricorrere a assunzioni specifiche, e a rivelare caratteristiche salienti e interessanti.

Popolazione e unità statistiche

Non ci dilungeheremo qui sulla importante questione che riguarda i modi con cui sono raccolti i dati e lo scopo per il quale sono raccolti.

I dati possono essere raccolti per diversi motivi:

- potrebbero essere raccolti a supporto della ricerca scientifica in diversi ambiti,

- potrebbere derivare dalla gestione di un processo. Ad esempio i dati raccolti dalla pubblica amministrazione nel gestire un servizio o quelli raccolti automaticamente a seguito dell'utilizzo di un software o nella gestione di un sito web.

La distinzione posta può in molti casi rivelarsi più sfumata di quanto qui proposto e spesso è possibile utilizzare un dato amministrativo anche a scopo di ricerca scientifica. Ovviemente occorre essere consapevoli che in quest'ultimo caso potrebbero esserci notevoli limiti e risultati cui si perviene non sono sufficientemente solidi e generali se usati a fini scientifici.

In generale, va detto che l'obiettivo ultimo è quello di conoscere le caratteristiche di una **popolazione**.

Una popolazione è una collettività e gli elementi di tale collettività sono detti **unità statistiche**, sono esempi

- la popolazione degli italiani di sesso maschile con oltre 18 anni al 01/01/2012;
- le famiglie italiane al 01/01/2012;
- i 218 comuni del FVG;
- i clienti di un negozio
- coloro che accedono a un sito web.

La popolazione può essere finita (ad es. la popolazione italiana) o infinita (ad es. tutte le persone affette da una patologia, oggi o in futuro).

Dati e ricerca scientifica

Se i dati vengono raccolti a scopo di ricerca scientifica allora è necessario utilizzare accorgimenti enlla raccolta dei dati che li rendano idonei a supportare le congetture che si avanzano o le ipotesi scientifiche proposte. Congetture o ipotesi che potrebbero riguardare anche un collettivo, ovvero una popolazione, anche più ampia di quella da cui ho tratto i dati.

Occorre che i dati vengano raccolti utilizzando protocolli che permettano di generalizzare quello che emergerà dalla loro analisi. La statistica, e in particolare quella inferenziale, stabilisce criteri e regole perchè si possa attribuire ai dati raccolti un valore scientifico. A tal fine si distingue fra:

1. dati ottenuti secondo disegni sperimentali controllati;
2. dati osservazionali.

I primi sono quelli che consentono in modo più coerente di investigare sulla esistenza di relazioni causali fra fattori (stabilire se un farmaco o una cura sono efficace per curare una malattia, stabilire se un processo di produzione cosnete di ottenere elementi di maggiore qualità, etc.). Tuttavia non sempre è possibile raccogliere i dati secondo disegni sperimentali controllati.

I secondi sono disponibili in un numero di casi forse più ampio. Essi vengono spesso raccolti in indagini o rilevazioni statistiche che sono di tipo: - totale (o censuario) se osservo tutte le unità della popolazione (è appunto il caso del censimento), - o parziale (cosa inevitabile se la popolazione è infinita) quindi osservando solo alcuni elementi della popolazione.

Inoltre possono esservi diversi schemi di osservazione (sezionale o trasversale, prospettivo o longitudinale, retrospettivo, etc.).

Nel caso della rilevazione parziale è cruciale per poter poi fare affermazioni generali, che valgano quindi per l'intera popolazione e non solo per i dati osservati, che la raccolta dei dati avvenga secondo schemi che li rendano rappresentativi. La migliore garanzia è offerta da una selezione (campionamento) degli elementi da osservare che segua criteri di selezione casuale. Il tema del campionamento, in particolare della selezione casuale, come presupposto per potre trarre conclusioni scientificamente valide dalla osservazione parziale di una popolazione è parte integrante della teoria statistica e non verrà qui discusso oltre. In numerosi casi sarà inoltre necessario assumere anche che la naturale variabilità che si osserverà nei dati sia rappresentabile con opportuni modelli teorici (ad esempio assumendo che una data caratteristica assuma valori che sono rappresentabili secondo un modello sticastico gaussiano).

Nel caso i dati rilevati anche con schemi di campionamento validi si è in un contesto osservazionale e poter trarre conclusioni sulla esistenza di una relazione fra le variabili, da interpretare in senso causale come nell'esempio dei dati sperimentali, è di solito molto più complesso e occorre ricorrere ad assunzioni non verificabile empiricamente.

Dati amministrativi

Molti dei dati che si raccolgono derivano dalla gestione di un processo (ad esempio i dati raccolti dalla pubblica amministrazione nel gestire un servizio, o quelli contenuti in una base di dati per gestire un processo aziendale). Sono quindi raccolti non per scopi scientifici. Anche essi si riferiscono a un collettivo (ad es. tutte le transazioni effettuate in un dato giorno da una banca) per cui può esservi un forte interesse a esaminare le caratteristiche di tale collettivo senza alcuna pretesa di generalizzare ad altre popolazioni o a trarre evidenze scientifiche.

Non è infrequente perlatro il caso in cui dati di diversa natura vengano combinati per avere quadri sempre più completi di un fenomeno.

Analisi esplorativa dei dati

L'analisi esplorativa dei dati (AED) o data analysis o analisi descrittiva non si pone come obiettivo quello di ricavare conclusioni su un aggregato diverso da quello osservato (cosa indispensabile in contesti scientifici). Per cui pur essendo rilevanti lo scopo con cui i dati sono raccolti, la modalità e lo schema di rilevazione dei dati, al fine di fornire un contesto alle analisi, l'attenzione è sulle tecniche per rappresentare sinteticamente i dati (anche con opportune tecniche grafiche di visualizzazione) così da mettere in evidenza alcune caratteristiche essenziali ai fini di monitorare un fenomeno, effettuare confronti, elaborare congetture da sottoporre poi ad analisi più accurate.

Le conclusioni che si traggono non voglione avere carattere di generalità e non si vuole estendere quanto si osserva sull'insieme di dati disponibile a popolazioni più ampie utilizzando apparati formali (come quello della statistica inferenziale dove si riesce a misurare anche l'attendibilità delle conclusioni che si traggono).

Tuttavia i pattern osservati nei dati sono evidenze utili seppure riferibili asclusivamente all'insieme di dati osservato. Si noti che se i dati si riferiscono a un'intera popolazione (come per il censimento) ottenere una sintesi degli stessi efficace è informazione valida per l'intera popolazione.

Non si fa quindi riferimento a priori a modelli stocastici che potrebbero aver generato i dati come nel caso dell'inferenza statistica o all'esistenza di relazioni speciali fra le quantità osservate. Si prescinde inoltre dall'idea che i dati siano "perfetti" e si ammette che essi possano essere sporchi, inaccurati, osservati in modo incompleto e con errori. Per cui l'analisi dei dati che introdurremo potrà includere un fase non banale di "pulizia" dei dati preliminare alla fase di analisi esplorativa.

Tipi di dati

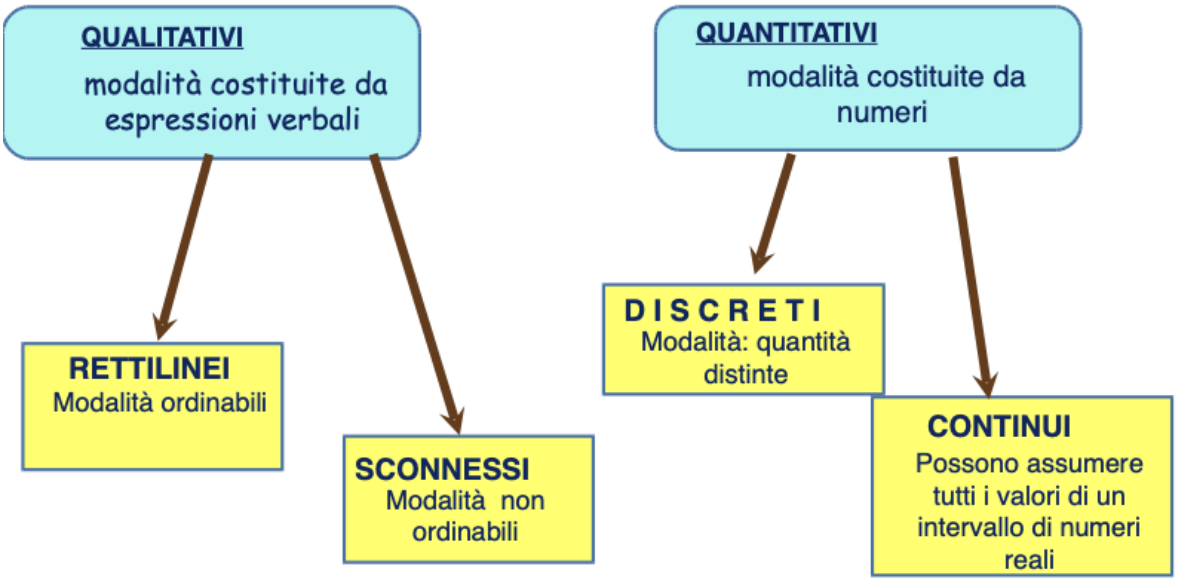
Variabili statistiche

Un dato statistico è il risultato della rilevazione (misurazione/osservazione) di un qualche **variabile** o **carattere** su un'**unità statistica** appartenente a una elemento di una popolazione.

variabile o carattere: ogni aspetto elementare oggetto di rilevazione nelle unità statistiche del collettivo.

Nel seguito, i termini carattere e variabile verranno usati in modo interscambiabile.

Tipi di variabili (o caratteri)



Variabili qualitative

- Una variabile è **qualitativa** se i valori che può assumere, detti *modalità*, si presentano espressi in forma verbale;
  - una variabile qualitativa è sconnessa se le sue modalità non implicano una graduazione;
  - una variabile qualitativa è ordinale se le sue modalità implicano una graduazione;
- le modalità possono essere predefinite a priori;
- a volte, in rilevazioni con questionari, le modalità vengono desunte a posteriori a partire dalla descrizione dettagliata dello stato della singola unità relativamente al carattere in questione.

Le variabili qualitative ai fine delle analisi dei dati che verranno condotte con R sarà opportuno definirle come **fattori** (qualitativi)

Variabili quantitative

- Una variabile è **quantitativa** se assume valori espressi in forma numerica che corrispondono a una misurazione o a un conteggio;
- rispetto ai valori che possono assumere
  - una variabile quantitativa è **discreta** se l'insieme dei valori numerici che assume è finito oppure numerabile;
  - una variabile quantitativa è **continua** se l'insieme dei valori numerici che assume è, almeno concettualmente, associabile con i valori di un intervallo reale, limitato o illimitato.

NB. Per la limitata precisione utilizzabile nel rilevare le misure, la distinzione tra variabile discreta e continua è di fatto convenzionale.

La matrice dei dati

La più semplice forma con cui rappresentare i dati relativi ad alcune variabili, diciamo p, su un collettivo di n unità è la **matrice dei dati**. Ovvero una matrice che ha n righe e p colonne. Così che una riga rappresenta i dati raccolti per una generica unità e una colonna contiene il vettore di valori osservati su ciascuna variabile per l'insieme delle unità.

Di solito n è molto maggiore di p e l'obiettivo dell'analisi dei dati è quello di analizzare le colonne della matrice:

- se si prende in esame una variabile (colonna) per volta si parla di analisi di una singola variabile o **analisi univariata**
- se si prendono in esame più variabili (più colonne) congiuntamente si parla di analisi **bivariata** nel casi di due variabili o **multivariata** se considero più di due colonne congiuntamente.

Si noti che le righe della matrice di dati (ovvero le unità statistiche) non hanno di solito un ordinamento particolare. E ai fini dell'analisi potrei scambiare una riga con un'altra senza che questo abbia alcun rilievo nella successiva analisi.

Vi sono però casi in cui i dati sono ordinati. Può accadere ad esempio che siano ordinabili rispetto a una variabile che rappresenti il tempo in cui avviene la raccolta del dato. Si pensi al caso in cui ogni riga contiene le informazioni raccolte in un dato giorno su alcune variabili. L'ordine temporale ha ovviamente importanza per cui i dati conviene che vengano ordinati secondo tale variabile e che se ne tenga conto nell'analisi. In tal caso i dati costituiscono delle **serie temporali**. Analogamente potrei tenere conto del fatto che i dati sono riferiti a un determinato contesto spaziale e avere delle coordinate che si riferiscono a luogo in cui dati sono osservati. Anche in tal caso si potrebbe tenere conto del fatto che i dati osservati possano esser ordinati secondo un criterio che tenga conto del riferimento spaziale (**serie spaziali o territoriali**).

Il data frame in R

Abbiamo già visto che in R la struttura del "data frame" ha una rappresentazione coerente con la matrice dei dati.

Leggiamo i dati dal data frame contenuto in `neonati.csv`:

```
neo<- read.csv("neonati.csv", sep=";", header=TRUE)
head(neo)
str(neo)
```

```
##   Peso settimane  fuma
## 1 2940        38     S
## 2 2420        36     S
## 3 2760        39     S
## 4 2440        35     S
## 5 3301        42     S
## 6 2715        36     S
## 'data.frame':    32 obs. of  3 variables:
##  $ Peso     : int  2940 2420 2760 2440 3301 2715 3130 2928 3446 2957 ...
##  $ settimane: int  38 36 39 35 42 36 39 39 42 39 ...
##  $ fuma     : chr  "    S" "    S" "    S" "    S" ...
```

Con riguardo al tipo di variabili le tre variabili del data frame sono:

- Peso : quantitativa continua

- settimane: quantitativa discreta
- fuma: variabile qualitativa (fattore)

```
is.factor(neo$fuma)
neo$fuma<-factor(neo$fuma)
str(neo)
```

```
## [1] FALSE
## 'data.frame':    32 obs. of  3 variables:
##  $ Peso     : int  2940 2420 2760 2440 3301 2715 3130 2928 3446 2957 ...
##  $ settimane: int  38 36 39 35 42 36 39 39 42 39 ...
##  $ fuma     : Factor w/ 2 levels "    N","    S": 2 2 2 2 2 2 2 2 2 2 ...
```

E' buona regola ai fini delle analisi successive trasformare la variabile qualitativa in un fattore.

Vediamo altri due esempi di data frame e del tipo di variabili che esse contengono:

```
library(MASS)
data("Cars93")
str(Cars93)
```

```
## 'data.frame':    93 obs. of  27 variables:
##  $ Manufacturer      : Factor w/ 32 levels "Acura","Audi",..: 1 1 2 2 3 4 4 4 4 5 ...
##  $ Model             : Factor w/ 93 levels "100","190E","240",..: 49 56 9 1 6 24 54 74 73 35 ...
##  $ Type              : Factor w/ 6 levels "Compact","Large",..: 4 3 1 3 3 3 2 2 3 2 ...
##  $ Min.Price         : num  12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 ...
##  $ Price             : num  15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...
##  $ Max.Price         : num  18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3 ...
##  $ MPG.city          : int  25 18 20 19 22 22 19 16 19 16 ...
##  $ MPG.highway       : int  31 25 26 26 30 31 28 25 27 25 ...
##  $ AirBags           : Factor w/ 3 levels "Driver & Passenger",..: 3 1 2 1 2 2 2 2 2 2 ...
##  $ DriveTrain        : Factor w/ 3 levels "4WD","Front",..: 2 2 2 2 3 2 2 3 2 2 ...
##  $ Cylinders         : Factor w/ 6 levels "3","4","5","6",..: 2 4 4 4 2 2 4 4 4 5 ...
##  $ EngineSize        : num  1.8 3.2 2.8 2.8 3.5 2.2 3.8 5.7 3.8 4.9 ...
##  $ Horsepower        : int  140 200 172 172 208 110 170 180 170 200 ...
##  $ RPM               : int  6300 5500 5500 5500 5700 5200 4800 4000 4800 4100 ...
##  $ Rev.per.mile      : int  2890 2335 2280 2535 2545 2565 1570 1320 1690 1510 ...
##  $ Man.trans.avail   : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 1 1 1 ...
##  $ Fuel.tank.capacity: num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
##  $ Passengers        : int  5 5 5 6 4 6 6 6 5 6 ...
##  $ Length            : int  177 195 180 193 186 189 200 216 198 206 ...
##  $ Wheelbase         : int  102 115 102 106 109 105 111 116 108 114 ...
##  $ Width             : int  68 71 67 70 69 69 74 78 73 73 ...
##  $ Turn.circle       : int  37 38 37 37 39 41 42 45 41 43 ...
##  $ Rear.seat.room    : num  26.5 30 28 31 27 28 30.5 30.5 26.5 35 ...
##  $ Luggage.room      : int  11 15 14 17 13 16 17 21 14 18 ...
##  $ Weight            : int  2705 3560 3375 3405 3640 2880 3470 4105 3495 3620 ...
##  $ Origin            : Factor w/ 2 levels "USA","non-USA": 2 2 2 2 2 1 1 1 1 1 ...
##  $ Make              : Factor w/ 93 levels "Acura Integra",..: 1 2 4 3 5 6 7 9 8 10 ...
```

Il data frame è disponibile nel package `MASS` (che occorre eventualmente installare). Se si vogliono maggiori dettagli sulle variabili si cerchi l'help su Cars93

Sapreste classificare le variabili presenti nel dataset?

Infine vediamo ancora le variabili nel dataframe `AutoBi` nel package `insuranceData`

```
library(insuranceData)
data("AutoBi")
str(AutoBi)
```

```
## 'data.frame':    1340 obs. of  8 variables:
##  $ CASENUM : int  5 13 66 71 96 97 120 136 152 155 ...
##  $ ATTORNEY: int  1 2 2 1 2 1 1 1 2 2 ...
##  $ CLMSEX  : int  1 2 1 1 1 2 1 2 2 1 ...
##  $ MARITAL : int  NA 2 2 1 4 1 2 2 2 2 ...
##  $ CLMINSUR: int  2 1 2 2 2 2 2 2 2 2 ...
##  $ SEATBELT: int  1 1 1 2 1 1 1 1 1 1 ...
##  $ CLMAGE  : int  50 28 5 32 30 35 19 34 61 NA ...
##  $ LOSS    : num  34.94 10.892 0.33 11.037 0.138 ...
```

Trasformare le variabili in fattori e ricodificare

Nell'ultimo caso le variabili hanno codici numerici, anche quando sono concettualmente variabili qualitative, e sono infatti di tipo numerico (numeric o integer). E' opportuno per le successive analisi decidere quali sono in realtà fattori qualitativi e trasformarle oltre che magari assegnare e denominare i livelli del fattore del fattore. Ad esempio, la variabile `AutoBi$MARITAL` è in realta una variabile qualitativa (un fattore) con 4 modalità (=1 if married, =2 if single, =3 if widowed, and =4 if divorced/separated):

```
AutoBi$MARITAL<- factor(AutoBi$MARITAL)
levels(AutoBi$MARITAL)<-c("married", "single", "widowed", "divorced")
str(AutoBi$MARITAL)
```

```
##  Factor w/ 4 levels "married","single",..: NA 2 2 1 4 1 2 2 2 2 ...
```

Si noti che con levels si possono anche ricodificare i livelli riducendoli a 3 ad esempio:

```
levels(AutoBi$MARITAL)<-c("married", "single", "previouslymarried", "previouslymarried")
levels(AutoBi$MARITAL)
```

```
## [1] "married"           "single"            "previouslymarried"
```

Trasformare le variabili quantitative in fattori: classi di valori

Un'operazione che viene spesso effettuata è quella di trasformare una variabile quantitativa in una variabile in classi. A questo fine si può utilizzare la funzione `cut()` che di fatto converte una variabile numerica in un fattore.

Vediamo un esempio. Consideriamo la variabile numerica (quantitativa `CLMAGE`).

```
range(AutoBi$CLMAGE, na.rm=TRUE)
CLMAGEclass<-cut(AutoBi$CLMAGE, breaks=6)
str(CLMAGEclass)
levels(CLMAGEclass)
```

```
## [1]  0 95
##  Factor w/ 6 levels "(-0.095,15.8]",..: 4 2 1 3 2 3 2 3 4 NA ...
## [1] "(-0.095,15.8]" "(15.8,31.7]"   "(31.7,47.5]"   "(47.5,63.3]"
## [5] "(63.3,79.2]"   "(79.2,95.1]"
```

Con la funzione `cut()` abbiamo quindi definito un fattore (ordinato) che sostituisce al valore numerico la classe di valori entro cui questo valore cade. Col parametro braek abbiamo specificato che vogliamo 6 classi e

in questo caso R crea 6 classi di uguale ampeazza fra il minimo e il massimo (attenzione a trattare gli NA nelle successive analisi: essi resteranno tali anche dopo la trasformazione in classi della variabile).

Sarebbe possibile anche definire le classi di ampiezza diversa (cosa che vedremo più avanti è opportuna in molti casi). In tal caso si passa al parametro breaks un vettore di valori che costituiscono le delimitazioni delle classi. Ad esempio:

```
CLMAGEclass<-cut(AutoBi$CLMAGE, breaks=c(-1,15,24,36,50,95))
str(CLMAGEclass)
levels(CLMAGEclass)
```

```
##  Factor w/ 5 levels "(-1,15]","(15,24]",..: 4 3 1 3 3 3 2 3 5 NA ...
## [1] "(-1,15]" "(15,24]" "(24,36]" "(36,50]" "(50,95]"
```

Potremmo aggiungere la nuova variabile al data frame `AutoBi`

```
AutoBi<-cbind(AutoBi,CLMAGEclass)
str(AutoBi)
```

```
## 'data.frame':    1340 obs. of  9 variables:
##  $ CASENUM    : int  5 13 66 71 96 97 120 136 152 155 ...
##  $ ATTORNEY   : int  1 2 2 1 2 1 1 1 2 2 ...
##  $ CLMSEX     : int  1 2 1 1 1 2 1 2 2 1 ...
##  $ MARITAL    : Factor w/ 3 levels "married","single",..: NA 2 2 1 3 1 2 2 2 2 ...
##  $ CLMINSUR   : int  2 1 2 2 2 2 2 2 2 2 ...
##  $ SEATBELT   : int  1 1 1 2 1 1 1 1 1 1 ...
##  $ CLMAGE     : int  50 28 5 32 30 35 19 34 61 NA ...
##  $ LOSS       : num  34.94 10.892 0.33 11.037 0.138 ...
##  $ CLMAGEclass: Factor w/ 5 levels "(-1,15]","(15,24]",..: 4 3 1 3 3 3 2 3 5 NA ...
```

La notazione `(liminf-limsup]` indica che la classe è chiusa a destra è aperta a sinistra. Quindi una unità che mostra esattamente il valore `liminf` non sarà incluso nella classe mentre se assume il valore esatto `limsup` sarà nella classe.

Si noti che se le classi non coprono tutti i valori numerici del vettore questi saranno definiti come NA.

Si può anche usare come estremo superiore (inferiore) di una classe il valore speciale `Inf` o `-Inf`.

Tabelle di frequenza

Tabelle di frequenza semplici

La più semplice elaborazione per una variabile qualitativa (e per una variabile quantitativa trasformata in fattore utilizzando le classi di valori) è costituita dalla **tabella di frequenza**.

Si noti che per i fattori qualitativi sono noti i possibili valori che possono assumere, anche se non conosciamo tutti i valori presenti in un dato set di dati.

La funzione principale per ottenere la tabella di frequenza per un fattore è `table()`.

Vedremo che lo stesso comando consente di gestire anche la costruzione di tabelle di frequenza per più fattori qualitativi congiuntamente ottenendo così le **tabelle a più entrate** o **tabelle di contingenza**.

Il suo utilizzo nella forma più semplice permette di ottenere una tabella ove si conta il numero di casi corrispondenti a ciascuna modalità di un fattore.

Si consideri l'esempio seguente, che utilizza ancora il set di dati AutoBi.

Come si può leggere dalla documentazione in R, la variabile ATTORNEY è un vettore numerico che indica se l'attore è rappresentato da un procuratore (=1) o meno (=2). Conviene prima costruire un fattore e quindi utilizzare 'table' per ottenere il numero di colloro che sono rappresentati da un avvocato.

```
  AutoBi$ATTORNEY<- factor(AutoBi$ATTORNEY)
  levels(AutoBi$ATTORNEY) = c("yes", "no")
  tabella<-table(AutoBi$ATTORNEY)
  tabella
```

```
##
## yes  no
## 685 655
```

La stessa cosa possiamo fare per la variabile AutoBi$MARITAL che avevamo già trasformato in fattore

```
  tabella1<-table(AutoBi$MARITAL)
  tabella1
```

```
##
##           married            single previouslymarried
##               624               650                50
```

La funzione table() restituisce un oggetto della classe table. Di fatto è un array la cui dimensione dipende dalla dimensione della tabella. in questo caso è quindi possibile utilizzarlo come un vettore con tanti elementi quante sono le modalità della variabile.

I valori indicati in corrispondenza delle modalità della variabili sono detti **frequenze assolute** e la tabella rappresenta quella che è detta **distribuzione di frequenza**.

Spesso tuttavia, soprattutto se si vogliono fare confronti con tabelle ottenute per collettivi di diversa dimensione, conviene passare alle **frequenze relative** ottenute come rapporto fra le frequenze assolute e il totale dei casi del collettivo (esso potrebbe essere ottenuto con la funzione length() applicata alla variabile ma occorre prestare attenzione al fatto che potrebbero esserci NA).

```
#tabella di frequenze relative
freqrel<-table(AutoBi$MARITAL)/length(AutoBi$MARITAL)
freqrel
sum(freqrel)
```

```
##
##           married            single previouslymarried
##        0.46567164        0.48507463        0.03731343
## [1] 0.9880597
```

Si noti che la somma non dà 1 come dovrebbe.

La funzione margin.table (la cui utilità risulterà meglio più avanti) applicata all'oggetto generato da table() fornisce il numero totale di casi (meno il numero di valori mancanti NA):

```
totale<-margin.table(table(AutoBi$MARITAL))
totale
# usiamo questo come divisore
freqrel<-table(AutoBi$MARITAL)/totale
freqrel
sum(freqrel)
# ora la somma da 1 come dovrebbe
```

```
## [1] 1324
##
##           married            single previouslymarried
```

```
##        0.47129909        0.49093656        0.03776435
## [1] 1
```

Si noti che il totale dei casi validi è diverso da 1340 perchè sono presenti alcuni NA. Quindi la tabella di frequenze relative ottenuta dividendo per la lunghezza del vettore in questo caso fornisce un valore non corretto.

Infine esiete anche al funzione `prop.table()` chensi applica ad oggetti di tipo `table` e che consente di ottenere direttamente la versione della tabella con le frequenze relative

La funzione `table()` in realtà può essere applicata su qualsiasi variabile numerica o su una variabile carattere. Si noti tuttavia:

1. che non ha senso utilizzarla su una variabile numerica quantitativa continua che presumibilmente ha tanti valori distinti quanti sono i casi. Conviene passare alla rappresentazione in classi che abbiamo già introdotto. Ad esempio, consideriamo il caso dell'età

```
table(CLMAGEclass)
```

```
## CLMAGEclass
## (-1,15] (15,24] (24,36] (36,50] (50,95]
##     144     280     269     299     159
```

2. che se utilizzata su una variabile numerica discreta ha senso soprattutto se la variabile assume pochi valori distinti. Se vi sono molti valori discreti distinti allora convine trattarla come nel caso della variabile continua suddividendo in classi.
3. l'uso di `table` su un fattore anziché su un vettore di caratteri rende evidente quando per alcuni livelli del fattore non vi sono osservazioni. Si consideri ad esempio il codice seguente

```
AutoBi$CLMSEX <- factor(AutoBi$CLMSEX)
str(AutoBi$CLMSEX)
a <- AutoBi[AutoBi$CLMSEX==1, ]$CLMSEX
table(a)
```

```
##  Factor w/ 2 levels "1","2": 1 2 1 1 1 2 1 2 2 1 ...
## a
##   1   2
## 586   0
```

L'output mostra che non ci sono femmine in `a`, come previsto avendo selezionato solo le righe per cui il sesso era maschile.

Tabelle a doppia entrata

Il comando `table()` può essere utilizzato anche per costruire **tabelle a più entrate** (dette anche di contingenza) che rappresentano la distribuzione congiunta di due o più variabili categoriali (fattori).

Si considerino le variabili `ATTORNEY` e `CLMSEX`. Codifichiamo come fattore la seconda e rinominaiamo i livelli della variabile di genere in `"F"` e `"M"`:

```
AutoBi$CLMSEX<-factor(AutoBi$CLMSEX)
levels(AutoBi$CLMSEX) <- c("M", "F")
str(AutoBi$ATTORNEY)
 str(AutoBi$CLMSEX)
```

```
##  Factor w/ 2 levels "yes","no": 1 2 2 1 2 1 1 1 2 2 ...
##  Factor w/ 2 levels "M","F": 1 2 1 1 1 2 1 2 2 1 ...
```

Possiamo usare la funzione `table()` per calcolare la tabella a doppia entrata per la coppia di variabili:

```
tab1 <- table(AutoBi$CLMSEX,AutoBi$ATTORNEY)
tab1
```

```
##
##     yes  no
##   M 325 261
##   F 352 390
```

Le tabelle a doppia entrata sono costituite dalle frequenze assolute congiunte (ovvero quante unità mostrano congiuntamente una determinata coppia di modalità).

Si noti che è possibile ottenere, facendo le somme per riga e per colonna, rispettivamente le tabelle di frequenza delle variabili coinvolte prese singolarmente. Quando ricaviamo le distribuzioni semplici nel contesto di un'analisi congiunta (doppia in questo caso) ci si riferisce ad esse come **distribuzioni marginali**.

Si possono estrarre le distribuzioni marginali di riga pernedendo l'oggetto creato da `table` e utilizzando la funzione `margin.table()` in cui il secondo parametro indica se si vuole fare il totale di riga o di colonna. Ad esempio,

```
margin.table(tab1,1)
```

```
##
##   M   F
## 586 742
```

e

```
margin.table(tab1,2)
```

```
##
## yes  no
## 677 651
```

Soprattutto nel caso delle tabelle a doppia entrata è utile ricorrere alle frequenze relativa. Ve ne sono di tre tipi e possono essere ottenute con la funzione già introdotta `prop.table()`:

1. Le **frequenze relative congiunte**

```
prop.table(tab1)
```

```
##
##           yes        no
##   M 0.2447289 0.1965361
##   F 0.2650602 0.2936747
```

2. Le **frequenze relative di riga**

```
prop.table(tab1, 1)
```

```
##
##           yes        no
##   M 0.5546075 0.4453925
##   F 0.4743935 0.5256065
```

Si noti che ogni riga somma a 1

3. Le **frequenze relative di colonna**

```
prop.table(tab1, 2)
```

```
##
##           yes        no
```

```
##   M 0.4800591 0.4009217
##   F 0.5199409 0.5990783
```

Si noti che ogni colonna somma a 1

Le frequenze relative di riga o di colonna sono anche dette **frequenze relative condizionate**.

Quali frequenze relative considerare in una tabella a doppia entrata?

Per meglio comprendere la lettura di una tabella a doppia entrata facendo riferimento alle distribuzioni condizionate e per decidere se procedere alla lettura per riga o per colonna, si consideri ancora un esempio.

L'uso di una tabella congiunta aggiunge molto all'analisi delle variabili prese singolarmente in quanto consente di mettere in luce se le due variabili siano **associate**. In effetti la tabella a doppia entrata è esposta per evidenziare che la conoscenza di una variabile aumenta le nostre informazioni sulla seconda.

Nella maggior parte dei casi si ha in mente quale sia il ruolo delle due variabili:

- una delle due variabili è quella oggetto di interesse: la chiameremo **variabile dipendente** o **variabile risposta**
- l'altra variabile è quella la cui conoscenza consentirebbe di aumentare le nostre informazioni sulla variabile risposta, la chiameremo **variabile indipendente** o **fattore esplicativo**.

Se si ha in mente tale lettura allora la tabella da esaminare è quella delle **frequenze relative condizionate** in relazione alla variabile indipendente.

Chiariremo con un esempio. Consideriamo le due variabili `ATTORNEY` e `LOSS`. Se le analizziamo congiuntamente ci attendiamo che vi sia una relazione far le due variabili. La prima lettura potrebbe essere qualla in cui siamo interessati a capire perchè ci si rivolga a un avvocato.

La variabile risposta è quindi il fattore qualitativo `ATTORNEY`. Potremmo voler esplorare la congettura che ci si rivolga all'avvocato più frequentemente se il danno subito `LOSS` è maggiore e tale variabile è assunta come fattore eplicativo. A tal fine analizziamo le due variabili congiuntamente ma prima trasformiamo la variabile numerica `LOSS` in un fattore utilizzando le classi di valori. Infine mettiamo affianco alla tabella la distribuzione marginale della variabile risposta. Si noti che utilizziamo classi di valori di ampiezza diversa (vedremo più avanti perchè è opportuno)

```
AutoBi$LOSSclass<-cut(AutoBi$LOSS,breaks=c(0,0.5,2,4,8,1100))
#
#
table(AutoBi$ATTORNEY)  # frequenze assolute della variabile marginale ATTORNEY
tot<-prop.table(table(AutoBi$ATTORNEY))
#
#
tot                     # frequenze relative della variabile marginale ATTORNEY
tabella1<-table(AutoBi$ATTORNEY, AutoBi$LOSSclass)
#
#
tabella1                # tabella di frequenze congiunte
mio<-prop.table(tabella1, 2)
cbind(mio, tot)         # tabella di frequenze condizionate di colonna (cioè condizionatamente a  LOSS)
                        # con aggiunta la distribuzione marginale
```

```
##
## yes  no
## 685 655
##
##      yes       no
```

```
## 0.511194 0.488806
##
##       (0,0.5] (0.5,2] (2,4] (4,8] (8,1.1e+03]
##   yes      49     104   263   147         122
##   no      239     220   133    48          15
##       (0,0.5]   (0.5,2]     (2,4]     (4,8] (8,1.1e+03]      tot
## yes 0.1701389 0.3209877 0.6641414 0.7538462   0.8905109 0.511194
## no  0.8298611 0.6790123 0.3358586 0.2461538   0.1094891 0.488806
```

Si vede chiaramente che la distribuzione percentuale del ricorso all'avvocato cambia se ci condizioniamo alle classi di `LOSS`.

A tal fine si confrontino le frequneze relative di `yes` nelle diverse colonne e si osservi che per danni bassi (sotto i 500 dollari) solo il 17% ricorre all'avvocato mentre nelle classi di danno superiore ai 4800 dollari si arriva al 75% o addirittura all'89% per l'ultima classe. Le colonne sono le distribuzioni della variabile `ATTORNEY` **condizionate** alla variabile `LOSS` (opportunamente categorizzata). I dati quaindi mostrano che la il ricorso all'avvocato "dipende" dal danno.

Per confontare le distribuzioni condizionate (che saranno tante quante sono le colonne ovvero le modalità della variabile di condizionamento) è opportuno utilizzare la distribuzione di frequenze relative. Se si guarda alla distribuzione di frequenze assolute o alla distribuione di frequenze relative congiunte è più difficile effettuare i confronti visto che in ogni categoria della variabile `LOSS`troverò un numero diverso di soggetti.

Si ricorre molto più frequentemente all'avvocato se il danno è maggiore. La tabella rivela quindi che le due variabili sono **associate**. In particolare poi avendo svolto l'analisi della variabile `ATTORNEY` condizionatamente a `LOSS`, diremo che il ricorso all'avvocato dipende dall'ammontare del danno.

Si noti che se invece le distribuzioni condizionate di "ÀTTORNEY"' fossero tutte simili (al limite uguali) tra loro, e simili di conseguenza alla distribuzione marginale dedurremmo che le due variabili non sono associate (sono cioè **indipendenti**). Si noti che il termine "simili" lascia aperta la questione di misurare le differenze fra le distribuzioni condizionate secondo criteri opportuni e di decidere quando le differenze fra le condizionate possono ritenersi trascurabili. Tale discorso apre il campo a considerazioni che richiedono elementi di statistica inferenziale.

Si noti che se si invertisse il ruolo delle variabili si avrebbe una lettura alternativa. La scelta di quale sia appropriata dipende dal contesto dell'analisi. Tuttavia se vi fosse indipendenza nel senso esposto soprà questa esisterebbe qualunque sia la lettura

Tabelle di frequenze a più entrate

Nulla vieta di costruire tabelle di frequenze multiple. Tuttavia la lettura di una tabella di frequenza multipla è spesso complessa per cui è raro che si vada oltre 3 variabili. Ad esempio immaginiamo di aggiungere nella'analisi la variabile `CLMSEX`.

```
tabella3<-table(AutoBi$ATTORNEY, AutoBi$LOSSclass, AutoBi$CLMSEX)
tabella3
```

```
## , ,  = M
##
##
##       (0,0.5] (0.5,2] (2,4] (4,8] (8,1.1e+03]
##   yes      25      52   127    65          56
##   no      103      83    55    14           6
##
## , ,  = F
##
##
```

```
##       (0,0.5] (0.5,2] (2,4] (4,8] (8,1.1e+03]
##   yes      24      52   131    81          64
##   no      133     136    78    34           9
```

Come si vede vengono rappresentate tabelle a due entrate per ciascun livello della terza. Si noti anche che la struttura dell'oggetto generato da `table()` è un array tridimensionale. Infatti se si scrive `tabella3[2,3,1]` si ottiene 55 che è la frequenza assoluta di coloro che non hanno avvocato, sono nella classe di danno fra 2 e 4 e sono femmine.

Si noti che in tal caso il concetto di distribuzione marginale iniza a essere più complessoperchè potrei considerare distribuzioni marginali di sottoinsiemi delle variabili considerate. Cioè avrei, nell'esempio,DA tre distribuzioni marginali bivariate e tre marginali univariate.