LEZIONE_2

2022-10-06

INDICE

INTRODUZIONE

- DISTINZIONE tra ANALISI
  - UNIVARIATE: 1 CARATTERISTICA
  - MULTIVARIATE: 2+ CARATTERISTICHE
- PRIMA di un'analisi INFERENZIALE è opportuno svolgerne una DESCRITTIVA iniziale, per comprendere il CONTESTO TEORICO

ANALISI ESPLORATIVA

- **TIPOLOGIA DI DATI**
  - **OSSERVAZIONALI / SPERIMENTALI**
  - **CAMPIONARI / CENSITI**
  - **VARIABILI STATISTICHE**
- **INDIVIDUARE UNITÀ STATISTICHE**
  - **DATI MANCANTI**
  - **DATI SPORCHI**
- **PULIZIA DATI**
  - **CODIFICA**
  - **ORGANIZZAZIONE**
- **METODI GRAFICI**

VARIABILI STATISTICHE

IDENTIFICANO LE PROPRIETÀ DELLE UNITÀ STATISTICHE

MODALITÀ

SONO I VALORI CHE UNA VARIABILE PUÒ ASSUMERE

NOTAZIONI

- **Y = variabile generica**
- **y = modalità generica**
- **Y' = dominio dei valori ammessi da Y**
  - **Sy = {y1, ... , yj} con j<=N**
    - per qualsiasi j != i –> yj != yi
  - **N = numero di unità statistiche considerate**

TIPOLOGIE

- **VARIABILI QUALITATIVE (CATEGORIALI) –> STRINGHE TESTUALI:**
  - **SCONNESSE (NOMINALI): non è possibile individuare un'ordine "naturale"**
    - religione
    - colore degli occhi
    - genere
  - **ORDINALI: è possibile identificare un ordine**
    - livello di istruzione
    - gerarchie
  - **DICOTOMICHE: se | Y' | = 2**
    - i due valori ammessi possono essere codificati come 0 e 1, risparmiando memoria e preservando la quantità di informazione
- **VARIABILI QUANTITATIVE (NUMERICHE):**
  - **DISCRETE: se Y' è un insieme FINITO**
  - **CONTINUE: se Y' possiede un range continuo di valori, e ogni valore è valido**
  - **INTERVALLI: non esiste uno 0 arbitrario**
  - **RAPPORTI: esiste uno 0 arbitrario**

GERARCHIA

1. **QUANTITATIVE CONTINUE**
2. **QUANTITATIVE DISCRETE**
3. **QUALITATIVE ORDINALI**
4. **QUALITATIVE NOMINALI**

SALENDO DI LIVELLO GERARCHICO SI AUMENTA LA QUANTITÀ DI INFORMAZIONE

DISTRIBUZIONI DI FREQUENZA

FREQUENZE ASSOLUTE

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
Y = mtcars$cyl  # SCEGLIAMO LA VARIABILE Y = mtcars$cyl
(Sy = unique(Y))  # SUPPORTO DELLA VARIABILE Y
```

```
## [1] 6 4 8
```

```
(J = length(Sy))   # LUNGHEZZA DEL SUPPORTO
```

```
## [1] 3
```

```
(N = length(Y))     # NUMERO DI UNITÀ STATISTICHE
```

```
## [1] 32
```

```
(cylFreqTable = group_by(mtcars,cyl) %>% summarise(frequency = n())  ) # ANALOGO A --> (cylFreqTable =
```

```
## # A tibble: 3 x 2
##     cyl frequency
##   <dbl>     <int>
## 1     4        11
## 2     6         7
## 3     8        14
```

```
# LA SOMMA DI TUTTE LE FREQUENZE ASSOLUTE = NUMERO DI UNITÀ STATISTICHE
sum(cylFreqTable$frequency) == N
```

```
## [1] TRUE
```

VARIABILE QUANTITATIVA CONTINUA –> J QUASI UGUALE A N

È OPPORTUNO CREARE DELLE CLASSI DI MODALITÀ E CONTARE LE OCCORRENZE IN ESSE

CLASSI DI MODALITÀ

- RAPPRESENTANO DEI SOTTOINSIEMI DEL RANGE DI UNA VARIABILE
- DEVONO ESSERE NE TROPPE NE TROPPO POCHE
  - NUMERO OTTIMALE CIRCA N^(1/2)
- DEVONO RAPPRESENTARE DEGLI INTERVALLI DISGIUNTI
  - [yo,y1) –> y1 escluso
  - [y1,y2) –> y2 escluso

```
Sy = levels(factor(mtcars$mpg))
length(Sy)
```

```
## [1] 25
```

```
# ci sono tanti possibili valori ammessi da mpg (Miles Per Gallon)
(mpgRange = max(mtcars$mpg)-min(mtcars$mpg) )# ==> range(mtcars$mpg)
```

```
## [1] 23.5
```

```
(classList = split(mtcars$mpg, cut(mtcars$mpg, length(Sy)^0.5)) )
```

```
## $`(10.4,15.1]`
## [1] 14.3 10.4 10.4 14.7 13.3 15.0
##
## $`(15.1,19.8]`
##  [1] 18.7 18.1 19.2 17.8 16.4 17.3 15.2 15.5 15.2 19.2 15.8 19.7
##
## $`(19.8,24.5]`
## [1] 21.0 21.0 22.8 21.4 24.4 22.8 21.5 21.4
##
## $`(24.5,29.2]`
## [1] 27.3 26.0
##
## $`(29.2,33.9]`
## [1] 32.4 30.4 33.9 30.4
```

```
(classFreqTable = data.frame(
  "freq"= unlist(lapply(classList, length))
))
```

```
##             freq
## (10.4,15.1]    6
## (15.1,19.8]   12
## (19.8,24.5]    8
## (24.5,29.2]    2
## (29.2,33.9]    4
```

```
sum(classFreqTable$freq) == length(mtcars$mpg)
```

```
## [1] TRUE
```

FREQUENZE RELATIVE

INDICA IL RAPPORTO TRA LA FREQUENZA ASSOLUTA E IL NUMERO TOTALE DI UNITÀ STATISTICHE

```
# p1 = f1 / sum([f1,f2,...,fj]) j<=N --> N = n° unità
#    = f1 / n
```

```
(cylRelFreqTable = table(mtcars$cyl)/length(mtcars$cyl)*100)  # *100 finale serve a mostrare i valori p
```

```
##
##      4      6      8
## 34.375 21.875 43.750
```

```
cylRelFreqTable[[1]]/100 * length(mtcars$cyl) == cylFreqTable[cylFreqTable$cyl==4,]$frequency
```

```
## [1] TRUE
```

```
# calcola la frequenza assoluta a partire da quella relativa e la confronta con quella assoluta effetti
```

FREQUENZE CUMULATE

- **VARIABILI ORDINABILI, QUALITATIVE E QUANTITATIVE**
- Fi = Frequenza assoluta con cui si presentano modalità con ordini <= i-esimo ordine
  - Dato F = {f1,. . . ,fj} j = | Sy | <= N n° unità
  - F1 = f1 –> Fj = N
- Pi = Frequenza relativa cumulata, analoga a Fi
  - Dato P = {p1, . . . , pj} –> | P | = | F |
  - IPOTESI: P1 = p1 –> Pj = N

```
(cylCumFreq = cumsum(table(mtcars$cyl)))    # FREQUENZE CUMULATE
```

```
##  4  6  8
## 11 18 32
```

```
ll = nrow(cylFreqTable) # = nrow(cylFreqTable)
if (   (cylCumFreq[[1]] == cylFreqTable[1,"frequency"]) &&
       (cylCumFreq[[ll]] == N)    # N definito in precedenza come n° di unità
) {
  print("IPOTESI CONFERMATA")
}
```

```
## [1] "IPOTESI CONFERMATA"
```

```
ok=TRUE
for(i in 1:nrow(cylFreqTable)){
   if(cylCumFreq[[names(cylCumFreq)[i]]] != sum(cylFreqTable[1:i,"frequency"])){
     ok = FALSE;
     break;
   }
}
if (ok) {
  print("cumsum WORKS!!!")
} else {
  print("SOMETHING IS WRONG!!!")
}
```

```
## [1] "cumsum WORKS!!!"
```