

1 Statistica descrittiva e inferenziale

2 Frequenze

3 Indici di tendenza centrale

4 Credits

1 Statistica descrittiva e inferenziale

- La statistica descrittiva
- Tecniche

2 Frequenze

3 Indici di tendenza centrale

4 Credits

Statistica descrittiva e inferenziale

- **Statistica descrittiva**
 - Riassume, descrive, esplora i dati osservati
 - Prima fase nella valutazione delle proprietà psicometriche di uno strumento
- **Statistica inferenziale**
 - Usa i dati di un **campione** per fare inferenze sulla **popolazione**
 - Confronti tra gruppi, valutazione di interventi, validazione di test

- È utile per controllare, descrivere ed esplorare i dati
- Serve a riassumere i dati attraverso indici statistici, tabelle e grafici
- **Non esiste una buona analisi statistica, senza una buona analisi descrittiva**

Tecniche

- **Distribuzione di frequenza e grafici:** procedure per descrivere tutti i dati in modo conveniente
- **Indici di tendenza centrale/indici di posizione:** singoli numeri che descrivono “dove” si colloca la distribuzione dei punteggi, dove è situato il “centro di gravità” dei punteggi stessi. Nel caso dei quantili (quartili, decili, percentili) si tratta di valori che dividono un insieme di dati ordinati in intervalli uguali, e rappresentano le soglie sotto le quali cade una certa percentuale dei dati.

1 Statistica descrittiva e inferenziale

2 **Frequenze**

3 Indici di tendenza centrale

4 Credits

1 Statistica descrittiva e inferenziale

2 Frequenze

- Dati
- Notazioni fondamentali
- Le frequenze assolute semplici
- Le frequenze assolute cumulate
- Le frequenze relative semplici
- Le frequenze relative cumulate
- Visualizzazione

3 Indici di tendenza centrale

4 Crediti

Dati

Indichi il suo grado di accordo con l'obbligo di installare sul computer aziendale un software di controllo che registri le ore di attività. Risponda utilizzando la seguente scala: per niente, poco, abbastanza, completamente.

Il dataset raccoglie i dati rilevati su 20 intervistati sulla base della seguente codifica: 1= per niente, 2=poco, 3=abbastanza, 4=completamente

```
# installo  
#> install.packages("dplyr")  
# carico pacchetti  
library(dplyr)  
# carico il dataset  
load("data/data_hr.rda")
```

```
tibble(data_hr) #visualizzo il dataset
```

```
# A tibble: 20 x 3
```

	id	reparto	accordo
	<int>	<chr>	<dbl>
1	1	HR	1
2	2	HR	3
3	3	Commerciale	3
4	4	IT	1
5	5	Commerciale	4
6	6	Commerciale	3
7	7	Commerciale	3
8	8	HR	1
9	9	Commerciale	2
10	10	Commerciale	4
11	11	Produzione	3
12	12	IT	1
13	13	HR	3
14	14	Produzione	3
15	15	Produzione	4

Verifico il tipo di dato codificato:

```
str(data_hr)
```

```
'data.frame':  20 obs. of  3 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ reparto: chr  "HR" "HR" "Commerciale" "IT" ...
 $ accordo: num  1 3 3 1 4 3 3 1 2 4 ...
```

Trasformo **id** e **reparto** a fattore, attraverso la funzione `as.factor()`

```
data_hr$id = as.factor(data_hr$id)
data_hr$reparto = as.factor(data_hr$reparto)
```

Trasformo **accordo** a fattore, attribuendoci delle labels, quindi attraverso la funzione `factor()`:

```
data_hr$accordo = factor(data_hr$accordo,  
                           levels=c(1,2,3,4),  
                           labels=c("per_niente", "poco", "abbastanza", "molto"))
```

```
data_hr$accordo
```

```
[1] per_niente abbastanza abbastanza per_niente molto      abbastanza  
[7] abbastanza per_niente poco      molto      abbastanza per_niente  
[13] abbastanza abbastanza molto      poco      molto      poco  
[19] abbastanza molto  
Levels: per_niente poco abbastanza molto
```

```
# Controllo il tipo di dato  
str(data_hr$accordo)
```

```
Factor w/ 4 levels "per_niente","poco",...: 1 3 3 1 4 3 3 1 2 4 ...
```

```
head(data_hr$accordo,n=4) # le prime 4 unità statistiche
```

```
[1] per_niente abbastanza abbastanza per_niente  
Levels: per_niente poco abbastanza molto
```

```
tail(data_hr$accordo,n=2) # le ultime 2 unità statistiche
```

```
[1] abbastanza molto  
Levels: per_niente poco abbastanza molto
```

```
n = nrow(data_hr) # numerosità campionaria
```


Notazioni fondamentali

- Sia X la variabile grado di accordo
- Sia X_j la modalità j -esima di X , dove $j = 1 \dots 4$
- Sia n il totale delle unità statistiche ($n = 20$)

Le frequenze assolute semplici

- La frequenza assoluta semplice di una modalità è il numero naturale di unità statistiche che presentano tale modalità
- La generica frequenza assoluta semplice associata alla modalità j si indica con il simbolo f_j :

f_j = numero di unità statistiche con modalità j

Ad esempio nel nostro caso, come calcolo la frequenza assoluta semplice della modalità 2 (ie. poco)?

```
# vettore di TRUE e FALSE, TRUE quando accordo è poco
data_hr$accordo=="poco"
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
[13] FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE
```

```
# Sommo tutti i TRUE, perchè TRUE = 1
sum(data_hr$accordo=="poco")
```

```
[1] 3
```

$f_2 = 3$, indica che 3 rispondenti hanno espresso la modalità poco d'accordo.

Le frequenze assolute semplici

Se vogliamo calcolare le frequenze assolute semplici per ogni modalità, possiamo utilizzare la funzione `table()`, oppure la funzione `count()`

```
table(data_hr$accordo)
```

per_niente	poco	abbastanza	molto
4	3	8	5

```
data_hr |> count(accordo, name = "fj")
```

	accordo	fj
1	per_niente	4
2	poco	3
3	abbastanza	8
4	molto	5

Le frequenze assolute cumulate

- La frequenza assoluta cumulata di una modalità è la somma delle frequenze assolute semplici delle modalità precedenti alla modalità data più la frequenza assoluta semplice della modalità data
- La generica frequenza assoluta cumulata associata alla modalità j si indica con il simbolo F_j

$$F_j = \sum_{i \leq j} f_i$$

Quanti hanno espresso un grado di accordo inferiore o uguale a poco d'accordo?

```
data_hr |> count(accordo, name = "fj")
```

```
      accordo fj
1 per_niente  4
2      poco   3
3 abbastanza  8
4      molto  5
```

Le frequenze assolute cumulate

```
# sommo quanti hanno risposto per niente a quanti poco  
sum(data_hr$accordo=="per_niente") + sum(data_hr$accordo=="poco")
```

```
[1] 7
```

Ad esempio, $F_2 = 7$, indica che 7 rispondenti hanno espresso un grado di accordo inferiore o uguale a poco d'accordo

Le frequenze assolute cumulate

Attraverso la funzione `cumsum()` possiamo calcolare la somma cumulata, in questo caso le frequenze assolute cumulate

```
table(data_hr$accordo)
```

per_niente	poco	abbastanza	molto
4	3	8	5

```
cumsum(table(data_hr$accordo))
```

per_niente	poco	abbastanza	molto
4	7	15	20

Le frequenze assolute cumulate

Attraverso la funzione `cumsum()` possiamo calcolare la somma cumulata, in questo caso le frequenze assolute cumulate

```
data_hr |> count(accordo, name = "fj")
```

	accordo	fj
1	per_niente	4
2	poco	3
3	abbastanza	8
4	molto	5

```
data_hr |> count(accordo, name = "fj") |>  
  mutate(Fj = cumsum(fj)) #aggiungo una variabile
```

	accordo	fj	Fj
1	per_niente	4	4
2	poco	3	7
3	abbastanza	8	15
4	molto	5	20

Le frequenze relative semplici

- La frequenza relativa semplice è data dal rapporto tra la **frequenza assoluta semplice di tale modalità** e **il numero totale** di unità statistiche osservate
- La generica frequenza relativa semplice associata alla modalità j si indica con il simbolo p_j

$$p_j = \frac{f_j}{n}$$

Come calcolo la frequenza assoluta semplice?

```
table(data_hr$accordo)
```

per_niente	poco	abbastanza	molto
4	3	8	5

Le frequenze relative semplici

Qual'è il numero totale di unità statistiche?

```
n
```

```
[1] 20
```

```
nrow(data_hr)
```

```
[1] 20
```

```
length(data_hr$accordo)
```

```
[1] 20
```

Qual'è frequenza relativa semplice associata alla modalità 3?

```
sum(data_hr$accordo=="abbastanza")/n
```

[1] 0.4

Ad esempio, $p_3 = 0.40$ indica che il 40% dei rispondenti ha manifestato un grado d'accordo pari a abbastanza.

Qual'è frequenza relativa semplice associata alla tutte le modalità?

```
table(data_hr$accordo)/n
```

per_niente	poco	abbastanza	molto
0.20	0.15	0.40	0.25

Una frequenza relativa semplice varia sempre tra 0 e 1.

Le frequenze relative cumulate

- La frequenza relativa cumulata di una modalità è la somma delle frequenze relative semplici delle modalità precedenti alla modalità data più la frequenza relativa semplice della modalità data
- La generica frequenza relativa cumulata associata alla modalità j si indica con il simbolo P_j

$$P_j = \sum_{i \leq j} p_i$$

Come prima, attraverso la funzione `cumsum()`, possiamo calcolare le frequenze relative **cumulate**

```
# frequenze relative semplici  
table(data_hr$accordo)/n
```

per_niente	poco	abbastanza	molto
0.20	0.15	0.40	0.25

```
# frequenze relative cumulate  
cumsum(table(data_hr$accordo))/n
```

per_niente	poco	abbastanza	molto
0.20	0.35	0.75	1.00

Ad esempio, $P_2 = 0.35$, indica che il 35% dei rispondenti ha espresso un grado di accordo inferiore o uguale a poco d'accordo.

Una frequenza relativa cumulata varia sempre tra 0 e 1.

Le frequenze relative cumulate

Indice j	Modalità	Xj	fj	Fj	pj	Pj
1	per_niente	1	4	4	0.20	0.20
2	poco	2	3	7	0.15	0.35
3	abbastanza	3	8	15	0.40	0.75
4	molto	4	5	20	0.25	1.00

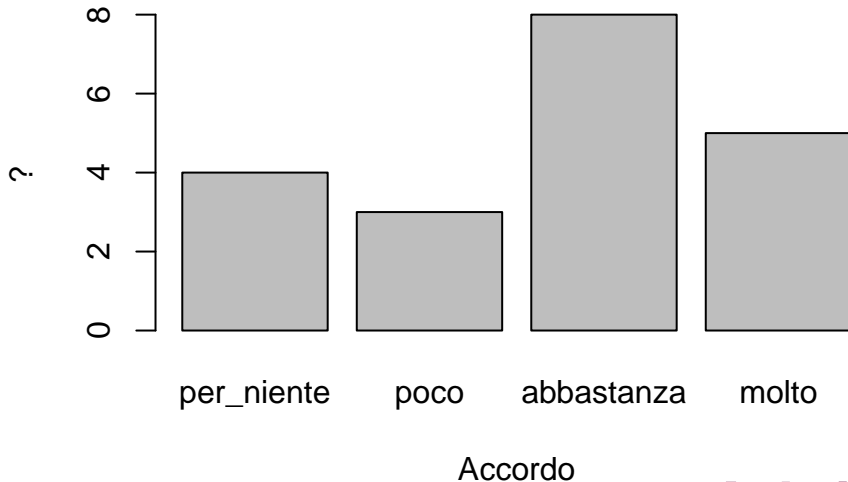
- ① La frequenza assoluta cumulata riferita all'ultima modalità ($j \in 1 \dots k$) è pari al numero totale delle unità statistiche:

$$F_k = n$$

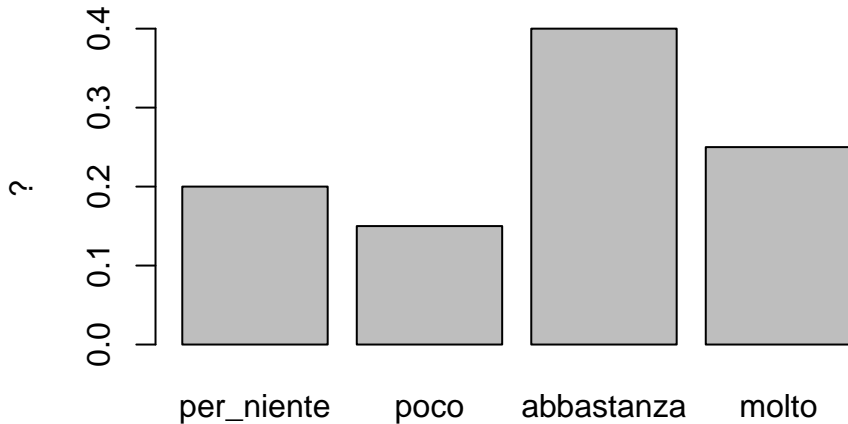
- ② La frequenza relativa cumulata riferita all'ultima modalità ($j \in 1 \dots k$) è pari 1:

$$P_k = 1$$

```
barplot(table(data_hr$accordo), ylab="?",  
        xlab="Accordo") # diagramma a barre
```

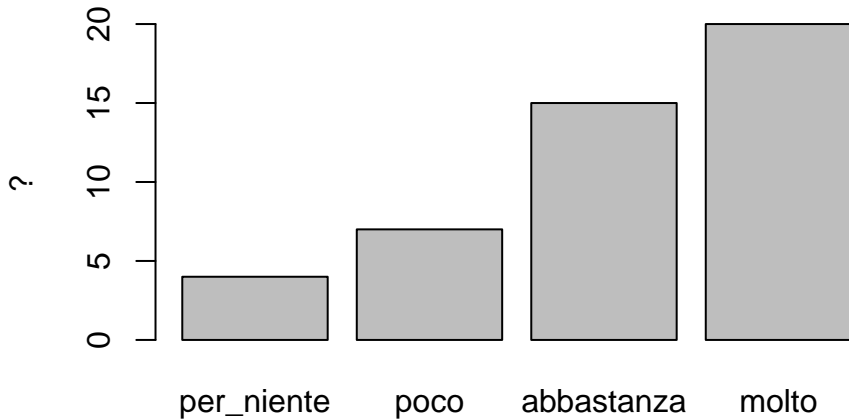


```
barplot(table(data_hr$accordo)/n, ylab="?",  
        xlab="Accordo") # diagramma a barre
```



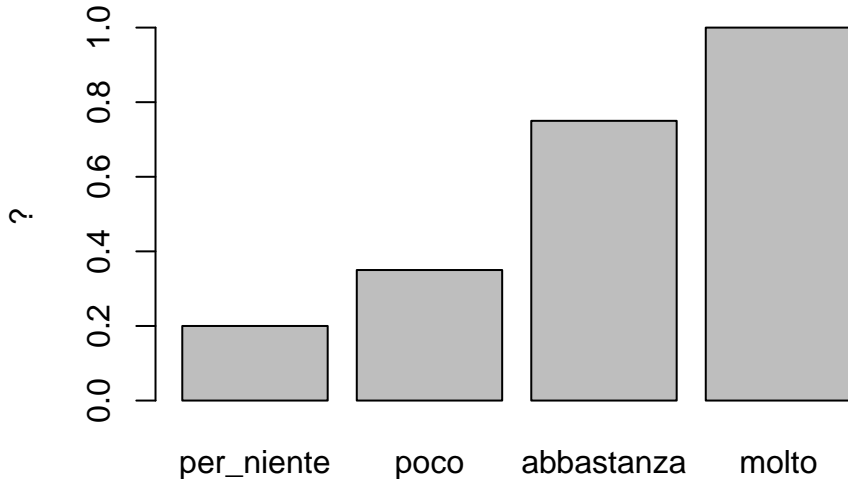
Accordo

```
barplot(cumsum(table(data_hr$accordo)), ylab="?",  
        xlab="Accordo") # diagramma a barre
```



Accordo

```
barplot(cumsum(table(data_hr$accordo))/n, ylab="?",  
        xlab="Accordo") # diagramma a barre
```



1 Statistica descrittiva e inferenziale

2 Frequenze

3 **Indici di tendenza centrale**

4 Credits

1 Statistica descrittiva e inferenziale

2 Frequenze

3 Indici di tendenza centrale

- Dati
- La moda
- La media aritmetica
- La mediana
- Indici di tendenza centrale e scale di misura

4 Credits

Indici di tendenza centrale

Un indice di tendenza centrale è un valore che descrive e riassume il centro di una distribuzione di dati

Dati

I dati che seguono rappresentano un esempio simulato di valutazioni di prestazione in un test di **problem solving** raccolte su $n = 20$ dipendenti appartenenti a quattro reparti (Produzione, Commerciale, HR, IT). Il punteggio è dato dal numero di risposte corrette su 10 ed è ricondotto a quattro fasce (A,B,C,D).

Prestazione	Risposte corrette	Numero
Intervento immediato (D)	0-4	0
Richiesta di attenzione (C)	5-6	2
Buono (B)	7-9	17
Ottimo (A)	10	1

La moda

La moda di una distribuzione di dati rilevati sulla variabile X , è la modalità che si presenta con la massima frequenza.

Ad esempio, rispetto ai dati relativi “alle capacità di problem solving”, la moda è

Prestazione	Risposte corrette	Numero
Intervento immediato (D)	0-4	0
Richiesta di attenzione (C)	5-6	2
Buono (B)	7-9	17
Ottimo (A)	10	1

Può accadere che data una distribuzione di dati, esista più di una modalità a cui è associata la massima frequenza. In questi casi non esiste un'unica Moda, e si parlerà a seconda del caso di distribuzione bi-modale (con due mode), tri-modale (con tre mode)

Proviamo a calcolare la moda attraverso i comandi imparati per il calcolo delle frequenze:

```
# carico il data set
load("data/data_ps.rda")
```

Secondo voi qual'è la variabile d'interesse?

Che funzione posso usare per ottenere la moda?

```
#visualizzo la struttura/le variabili
```

```
str(data_ps)
```

```
'data.frame': 20 obs. of 15 variables:
```

```
$ id      : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
$ reparto      : chr  "HR" "HR" "Commerciale" "IT" ...
```

```
$ accordo          : num  1 3 3 1 4 3 3 1 2 4 ...
```

```
$ p1          : int  1 1 1 0 1 0 1 1 0 1 ...
```

```
$ p2          : int  1 0 1 1 1 1 1 1 1 1 ...
```

```
$ p3          : int  1 1 1 1 1 1 0 1 0 1 ...
```

```
$ p4 : int 0 1 1 0 0 0 1 0 1 1 ...
```

```
$ p5          : int  1 1 1 1 1 1 1 1 1 1 1 ...
```

```
$ p6          : int  1 1 1 1 0 1 0 1 1 1 ...
```

```
$ p7          : int  1 1 1 1 1 1 1 1 1 1 1 ...
```

```
$ p8          : int  1 1 1 0 1 1 1 1 1 1 ...
```

```
$ p9          : int  1 0 0 1 1 1 1 1 0 1 ...
```

```
$ p10          : int  1 1 1 0 1 1 0 1 1 1 ...
```

```
$ problem_solving_0_10: int 9 8 9 6 8 8 7 9 7 10 ...
```

```
$ fascia_ps      : chr  "Sufficiente (B)" "Sufficiente (B)" "Sufficiente
```

```
table(data_ps$fascia_ps)
```

Criterio raggiunto (A)	Richiesta di attenzione (C)
1	2
Sufficiente (B)	
17	

La media aritmetica

La media aritmetica di una distribuzione di dati rilevati sulla variabile X , è data dalla somma dei dati divisa per il numero di unità statistiche:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

```
[1] 9 8 9 6 8 8 7 9 7 10 8 8 8 9 9 7 9 7 5 9
```

Calcolo della media

$$\bar{X} = \frac{\sum_{i=1}^{20} X_i}{20}$$

```
n = length(data_ps$problem_solving_0_10)
n
```

[1] 20

```
sum(data_ps$problem_solving_0_10)/n
```

[1] 8

```
mean(data_ps$problem_solving_0_10)
```

[1] 8

Proprietà della media

- La media varia tra il minimo e il massimo dei dati osservati
- La media è l'unico valore che, se sostituito a tutti i dati osservati, non cambia la loro somma (e quindi la loro media)

Proprietà della media

- Moltiplicando ciascun dato per una costante k si otterrà una media pari alla moltiplicazione tra la media dei dati originali e la costante k :

$$\overline{X} = \frac{\sum(kX_i)}{n} = k\overline{X}_{\text{dati originali}}$$

```
k = 10
x1 = data_ps$problem_solving_0_10
media1 = mean(x1)

x2 = data_ps$problem_solving_0_10 * k
media2 = mean(x2)

media2/media1
```

```
[1] 10
```

Proprietà della media

- La somma degli scarti tra i dati rilevati e la media è pari a 0:

$$\overline{X} = \sum (X_i - \overline{X}) = 0$$

```
media = mean(data_ps$problem_solving_0_10)
round(sum(data_ps$problem_solving_0_10-media),10)
```

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Calcolo della mediana: caso n dispari

- Se n (la numerosità dei dati raccolti) è dispari il valore centrale della serie ordinata dei dati è la mediana.
- La posizione i del dato corrispondente alla mediana è dato dalla seguente formula:

$$i = \frac{n + 1}{2}$$

Come trovo la mediana?

```
x_order[i] # estraggo l'elemento in posizione i
```

[1] 8

```
median(data_dispari$problem_solving_0_10)
```

[1] 8

Calcolo della mediana: caso n pari

- Se n è pari non esiste un valore della serie di dati che possa essere definito centrale.
- Si potrà unicamente dire che la mediana è compresa tra i valori aventi le seguenti posizioni “centrali”.

$$i_{inf} = \frac{n}{2} \quad i_{sup} = \frac{n}{2} + 1$$

9 9 9 9 10

Come trovo la mediana?

```
# estraggo gli elementi in posizione i_inf e i_sup
x_order[i_inf:i_sup]
```

[1] 8 8

In questo esempio il valore in posizione i_{inf} equivale al valore in posizione i_{sup} , quindi possiamo dire che la mediana = 8.

Nel caso in cui ottenessimo differenti valori, concluderemmo che la mediana è compresa tra quei valori.

$$X_{Mdn} = \frac{X_{inf} + X_{sup}}{2}$$

La mediana

```
voto = c(24, 29, 30, 22, 22, 26) #creo vettore voti
n = length(voto) # quante osservazioni ho?
voto_ordinato = sort(voto) #ordino gli elementi
voto_ordinato
```

[1] 22 22 24 26 29 30

```
voto_ordinato[n/2] #  $i_{inf} = n/2$ 
```

[1] 24

```
voto_ordinato[(n/2) + 1] #  $i_{sup} = (n/2) + 1$ 
```

[1] 26

$$(24+26)/2$$

[1] 25

Potremmo concludere che la mediana stia tra 24 e 26, e quella stimata sia 25.

Proprietà della mediana

- La mediana è poco influenzata (al contrario della media) da valori estremamente grandi o piccoli presenti nella distribuzione dei dati. Per questo viene detta stimatore “robusto”

Indici di tendenza centrale e scale di misura

Indice	Variabile nominale	Variabile ordinale	Variabile quantitativa
Moda	SI	SI	SI
Mediana	NO	SI	SI
Media	NO	NO	SI

1 Statistica descrittiva e inferenziale

2 Frequenze

3 Indici di tendenza centrale

4 Credits

1 Statistica descrittiva e inferenziale

2 Frequenze

3 Indici di tendenza centrale

4 Credits

