

Statistica descrittiva con fogli di calcolo

Stoianov, Ceccato

Distribuzioni di probabilità empirica

1) OSSERVAZIONI

campione X :

N osservazioni $\{x_1, x_2, \dots, x_N\}$

scala di misura Y

K livelli $[y_1, \dots, y_K]$

Esempio: il sesso $\{M, F\}$ di 30 persone: [M, F, F, F, F, M, F, M, M, F, M, M ...]

2) NUMERO di OSSERVAZIONI

frequenza f_i del livello $y_i \in Y$ nel campione X :

il numero di osservazioni di y_i nel X

probabilità empirica p_i di un livello $y_i \in Y$ nel campione X : $p_i = f_i / N$

3) DISTRIBUZIONE del NUMERO di OSSERVAZIONI

distribuzione di frequenze $F(y)$ nel campione X :

l'insieme di frequenze f_i di ciascun livello $y_i \in Y$ nel campione X

distribuzione di probabilità empirica $P(y)$ nel campione X :

l'insieme di probabilità empirica p_i di ciascun livello $y_i \in Y$ nel campione X

Frequenze

Codice_studente	Voto
1	27
2	30
3	26
4	29
5	26
6	27
7	25
8	25
9	25
10	29



Voto	f
25	3
26	2
27	2
28	0
29	2
30	1

Categorie abbinare

- Se abbiamo **dati numerici con tanti livelli**, rischiamo di avere pochi osservazioni per ciascun livello ... Che cosa fare ?
- Soluzione: definire una **scala S derivata** dalla scala originale Y, con un limitato numero di livelli (bin) $\{s_1, s_2, \dots, s_M\}$.
- Nella nuova scala S, ciascuno livello s_i raggruppa livelli $\{y_{i1}, y_{i2}, \dots, y_{ik}\}$
- In una scala ad intervalli (ad esempio tempo, spazio, ecc), il numero di livelli raggruppati in ciascun nuovo livello s_i deve essere uguale, per rispettare la omogeneità della scala originale.
- ESEMPIO:
Raggruppare i 1000 livelli 0 – 1000 mm in una scala con 10 livelli.

Distribuzioni di frequenze di dati numerici nei Fogli Elettronici

Se abbiamo dati X di tipo numerico (**scala ad intervalli**), possiamo utilizzare la funzione **frequency()** per calcolare la distribuzione dei dati X .

1. Avendo le **osservazioni** X in una colonna (es.: B2:B21)
2. Inserire i **livelli** S in un'altra colonna (es: D2:D11)
3. Applicare la funzione **frequency**(vettore_osserv.; vettore_categorie)
 - Selezionare una colonna per il **risultato** (es., E2:E11)
 - Scrivere '=frequency('
 - Selezionare / riferire le **osservazioni**; scrivere ';'.
 - Selezionare / riferire le **categorie**; scrivere ')'
 - Premere '**Ctrl-Maiusc-Invio**' (solo '**Invio**' calcola una frequenza solo!!)

Si nota:

- I livelli indicati definiscono una nuova scala S per i dati.
- La frequenza f_i di ciascun livello s_i riporta il numero di osservazioni con valori $(s_{i-1} \dots s_i]$.

Distribuzioni: calcolo e rappresentazione grafico

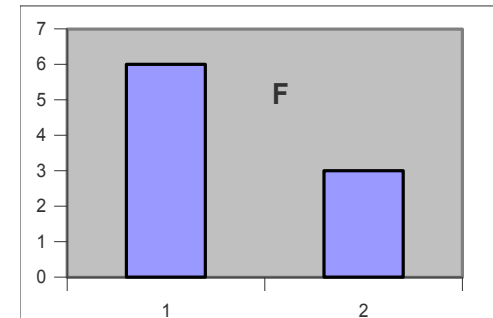
Grafici:

1. Pochi livelli: **Istogramma**
(diagramma a barre)

Nome	sezzo
Ivan	1
Sara	2
Tomi	1
Piero	1
Linda	2
Maria	2
Stefano	1
Leo	1
Boris	1

Sesso	F	p()
1	6	0.6667
2	3	0.375

=FREQUENZA(B2:B10;D2:D3)
=E2/CONTA.NUMERI(B2:B10)

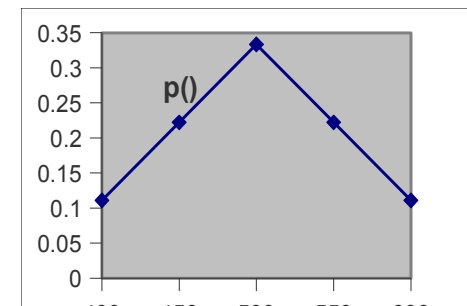


2. Tanti livelli: **Grafico a linea**

Nome	RT
Ivan	380
Sara	430
Tomi	515
Piero	475
Linda	460
Maria	560
Stefano	420
Leo	510
Boris	497

RT	F	p()
400	1	0.1111
450	2	0.2222
500	3	0.3333
550	2	0.2222
600	1	0.1111

=FREQUENZA(B2:B10;D2:D6)



Esercizio: probabilità empirica

- File S3_RT.xls
- Calcolare le **frequenze** degli RT (bin a distanza di 50)
- Calcolare la **distribuzione di probabilità empirica**
- **Fare grafica** della distribuzione di probabilità.

Tendenze centrali

1. Se abbiamo dati in una colonna (es.: B2:B21)
1. **La media** = somma divisa per il numero dei punti.
 - selezionare una cella per il **risultato** (di solito sotto i dati) (es: B23)
 - scrivere '=sum(';
 - selezionare le **osservazioni**; ')'; Premere '**Invio**'
 - in un'altra cella, dividere la somma per il numero dei dati (**count()**)
3. **Mediana o moda**: ordinare i dati e trovare:
 - la categoria centrale (mediana)
 - più-frequente (moda)(oppure utilizzando distribuzioni di frequenze: la categoria più numerosa)

Variabilità

- **Range / Intervallo** (distanza tra i valori estremi): $\max(X) - \min(X)$
 - **Somma dei quadrati delle distanze dalla media (scarti)** $SS_x = \sum (x_i - \bar{X})^2$
 - **Varianza della popolazione** (la media degli scarti²): $\sigma_x^2 = \frac{\sum (x_i - \bar{X})^2}{n}$
- $$\sigma_x^2 = \frac{\sum (x_i - \bar{X})^2}{n} = \frac{\sum (x_i^2 - 2x_i \bar{X} + \bar{X}^2)}{n} = \frac{\sum x_i^2}{n} - \bar{X}^2 = \overline{X^2} - \bar{X}^2$$
- **Deviazione standard (Scarto quadratico medio)**: $\sigma_x = \sqrt{\sigma_x^2}$

Calcolo della varianza / dev.st

I. La **varianza** (della popolazione):

Algoritmo **A**: la media del quadrato delle scarti

1. Calcolare la media M
2. Calcolare gli scarti dalla media $D_i=(X_i-M)$
3. Calcolare il quadrato degli scarti $D_i^2=D_i*D_i$.
4. Calcolare la media del quadrato delle distanze **var** = media(D_i^2).

$$Var_x = \frac{\sum (x_i - \bar{X})^2}{n}$$

Algoritmo **B**: la diff. tra la media dei quadrati e il quadrato della media:

1. Calcolare la media $M1$
2. Calcolare i quadrati: X_i^2
3. Calcolare la media dei quadrati $M2$
4. Calcolare il quadrato della media $M1*M1$
5. Calcolare la differenza **var**= $M2-M1*M1$.

$$Var_x = \overline{X^2} - \bar{X}^2$$

II. La **dev. standard**:

dst=sqrt(var)

$$\sigma_x = \sqrt{Var_x}$$

Covarianza

- Due variabili aleatorie (stocastiche) X e Y possono co-variare. (per.es.: studenti con voto alto in Matematica hanno alto voto in Informatica.)
- La **covarianza** – la media del prodotto di X e Y normalizzati con loro medie – esprime il grado di dipendenza lineare tra X e Y:

$$\begin{aligned}Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(Y)E(X)\end{aligned}$$

- se **positivo**: al **crescere** di x in media **cresce** anche y,
- se **negativo** al **crescere** di x in media **decrescere** y.
- Si nota: **la covarianza di una variabile con se stessa = la varianza !**

Algoritmo **A**:

1. Calcolare le medie M_x e M_y
2. Calcolare il prodotto degli scarti $D_i = (X_i - M_x)(Y_i - M_y)$
3. Calcolare la media $Cov = \text{media}(D_i)$.

Algoritmo **B**:

1. Calcolare i prodotti $X_i Y_i$
2. Calcolare le medie M_x , M_y , M_{xy}
3. Calcolare la differenza **Cov** = $M_{xy} - M_x M_y$.

Funzioni per la statistica descrittiva

Funzione

Sintassi

(ad esempio, per le osservazioni in A2:A150)

- MEDIA `AVERAGE(A2:A150)`
- MEDIANA `MEDIAN(A2:A150)`
- *VARIANZA della popolazione* `VAR(A2:A150)`
- SCARTO quadr.med. (*pop.*) `STDEV(A2:A150)`
- COVARIANZA tra 2 campioni `COVAR(A2:A150; B2:B150)`

Esercizio: medie e variabilità

- File dati: **S3_RT**
calcolare **con e senza** funzioni statistiche:
la media, la varianza, e la dev.st. dei RT
- File dati: **S2_Math**
calcolare covarianza tra *add_time*, *sub_time*

Tabelle di contingenza (tabelle pivot)

- **Serve:** a sintetizzare una caratteristica dei dati rispetto ad altre loro caratteristiche
- **I dati:** una serie di casi con varie caratteristiche (tabella), di cui alcuni sono causali (indipendenti) ed altri sono dipendenti.
- **Tabella pivot** ad una-, due-, o più- entrate (i fattori causali A, B, \dots) in cui:
 - gli elementi della riga (colonna) codificano i livelli delle categorie causali A (B),
 - le celle contengono una misura di sintesi della variabile dipendente X per ciascuna combinazione $A_i B_j$ dei livelli dei fattori A e B
 - l'ultima riga/colonna codifica la misura di sintesi della variabile dipendente per ciascun livello del fattore A (B).
- **Misure di sintesi:**
 - numero di casi
 - somma / media / dev.st. dei valori X_i per ciascuna cella.
 - altri
 - i valori possono essere riportati come % rispetto $A_i / B_j / Totale$

Esempio: i dati: {età, condizione, accuratezza};
la tabella: accuratezza (condizione, età)

Esempio pivot

oggett	s_type	age	IQ
C1	control	126	98
C2	control	132	124
C3	control	136	111
C4	control	131	91
C5	control	125	107
C6	control	129	105
C7	control	132	94
C8	control	129	107
C9	control	127	105
C10	control	138	115
C11	control	127	109
C12	control	126	96
C13	control	130	98
C14	control	134	118
C15	control	135	91
C16	control	141	127
NF1	number fact	130	98
NF2	number fact	132	101

Conteggio di s_type	
s_type	
control	16
number fact	4
Turners Syn.	6
Williams Syn.	5
Total Result	31

Conteggio di s_type	
s_type	
control	51.61%
number fact	12.90%
Turners Syn.	19.35%
Williams Syn.	16.13%
Total Result	100.00%

Tabelle pivot - procedura

1. Selezionare un intervallo compatto di dati.
2. Menu: **Dati > DataPilot > Avvia**

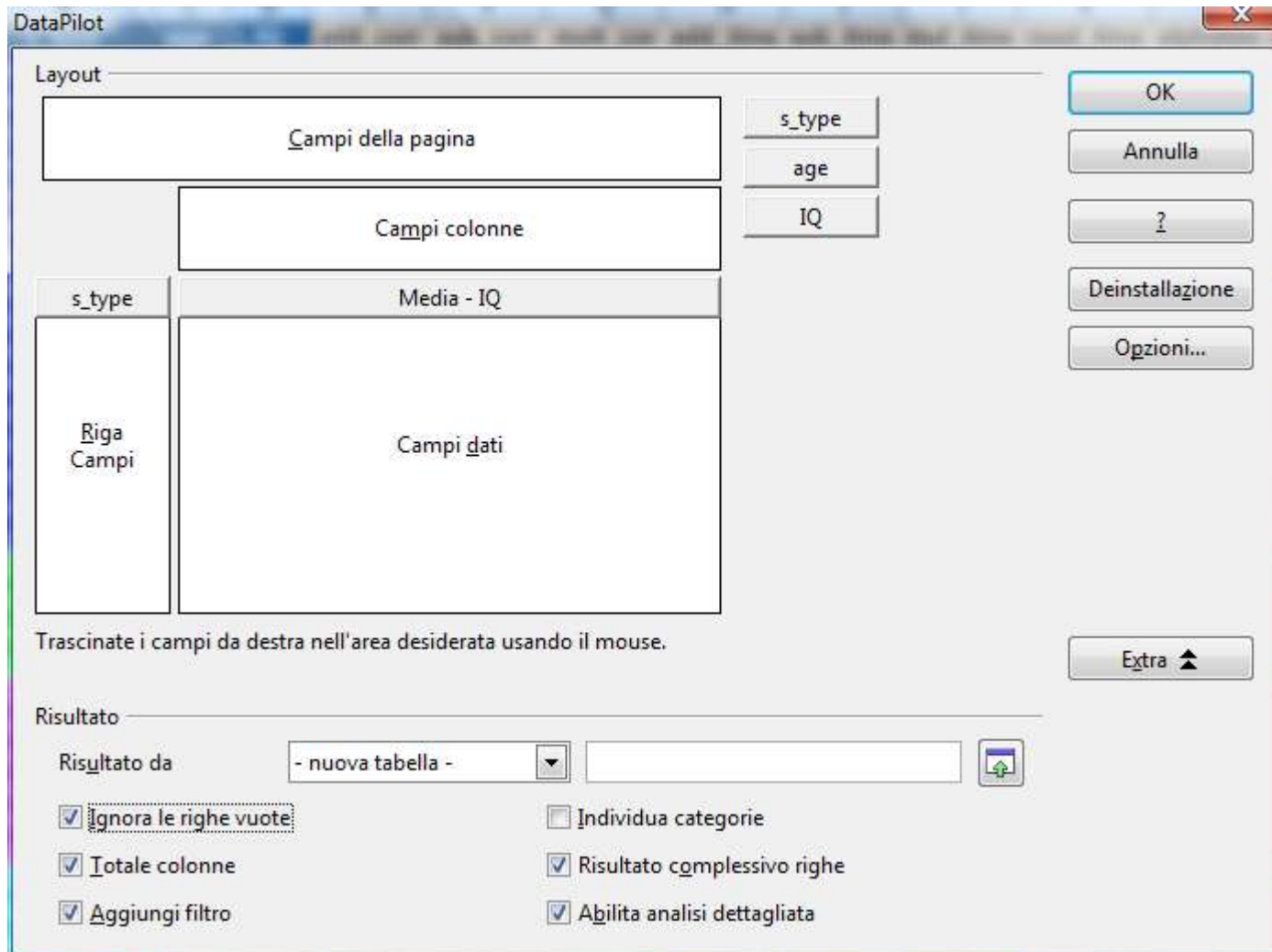


Tabelle pivot – procedura (2)

3) Disegnare la tabella pivot

- trascinare le categorie **causali** in: colonna, riga o pagina.
- trascinare le variabili **dipendenti** nel centro della tabella;
- selezionare la **misura di sintesi**, conteggio, media, ... per ciascuna variabile (con doppio click; si possono scegliere più di una sintesi)

4) Scegliere l'allocazione della tabella (Risultato da ..) (Ad esempio: *nuova tabella*)

Layout

Campi della pagina

Campi colonne

Media - IQ

Riga Campi

Campi dati

Trascinate i campi da destra nell'area desiderata usando il mouse.

Risultato

Risultato da - nuova tabella -

Ignora le righe vuote

Totale colonne

Aggiungi filtro

Individua categorie

Risultato complessivo righe

Abilita analisi dettagliata

Esercizio

File dati: **S3_Dan**

- 1) fare una tabella pivot ad una entrata **risp_verbale(livello)** dove:
 - la variabile dipendente è **risp_verbale**
 - la variabile indipendente è **livello**
 - esaminare la **media** della variabile dipendente
- 2) fare una tabella pivot a due entrate: **risp_verbale(luogo, livello)**
 - variabili dipendenti **risp_verbale** e **risp_verbale**,
 - variabile indipendente **livello**,
 - esaminare la **media** e **dev. standard** della variabile dipendente

Commento: i dati riguardano una serie di simulazioni di lesioni di un modello che simula la produzione verbale del participio passato. La variabile **livello** indica la estensione della lesione, mentre la variabile **logo** indica quale parte del modello è stata lesionata.