Calcoli con Fogli Elettronici

Lezione 3

Corso di Laurea in Biotecnologie
AA. 2010/2011

Docente del laboratorio:
Maria Silvia Pini
mpini@math.unipd.it

Frequenze

- Campione **X**: un insieme di ***N* osservazioni** $\{x_1, x_2 \dots x_N\}$ misurati con
- Scala di misura ***Y*** con ***κ* categorie** $[y_1 \dots y_K]$.

  Esempio: il *sesso* di *30* persone; categorie [M,F]

- Frequenza f_i di una categoria $y_i \in Y$ nel campione ***X***:
  il numero di osservazioni di y_i nel ***X***.
- Proporzione p_i di una categoria $y_i \in Y$ nel campione ***X***:
  la *frequenza* f_i divise per il numero totale ***N*** di osservazioni nel campione **X**.
- *Distribuzione di frequenze* ***F(y)*** nel campione ***X***:
  l'insieme di frequenze f_i di ciascuna categoria $y_i \in Y$ nel campione ***X***.
- *Distribuzione di probabilità empirica* ***P(y)*** nel campione ***X***:
  l'insieme di proporzioni p_i di ciascuna categoria $y_i \in Y$ nel campione ***X***.

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 abbinate

Se abbiamo dei **dati numerici con tanti livelli**, rischiamo di avere un osservazione per livello … Che cosa fare ?

Soluzione: definire una **scala S derivata** dalla scala originale **Y=[y_1 … y_K]**, con un limitato numero di livelli (**bin**) {s_1, s_2, …, s_M}.

Procedura:

- nella nuova scala S, ciascuno livello $\mathbf{s_i}$ raggruppa livelli $\mathbf{\{y_{i1}, y_{i2}, \ldots s_{ik}\}}$
- **Il numero di livelli raggruppati** in ciascun nuovo livello s_i deve essere **uguale**.
- Il numero di livelli in S dovrebbe essere **limitato** (per es., 10).

Categorie abbinate

Codice impiegato	Data di nascita
1	01/06/57
2	02/08/72
3	03/05/62
4	07/08/76
5	23/02/75
6	01/03/72
7	06/06/51
8	19/08/72
9	18/02/69
10	04/05/75

Decennio	
50	2
60	2
70	6
Total result	10

Distribuzioni di frequenze di dati numerici nei Fogli Elettronici

Se abbiamo dati ***X*** di tipo numerico (**scala ad intervalli**), possiamo utilizzare la funzione *frequenza()* 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 ***frequenza*(vettore_*osserv.; vettore_categorie*)**
   - *Selezionare* una colonna per il risultato (es., E2:E11)
   - Scrivere '=frequenza('
   - 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 calcolare la distribuzione dei dati numerici raggruppati. **La frequenza f_i di ciascuna categoria s_i** corrisponde al **numero delle osservazioni** con valori **$(s_{i-1} \ldots s_i]$.**

Distribuzioni: calcolo e rappresentazione grafico

Grafici:

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

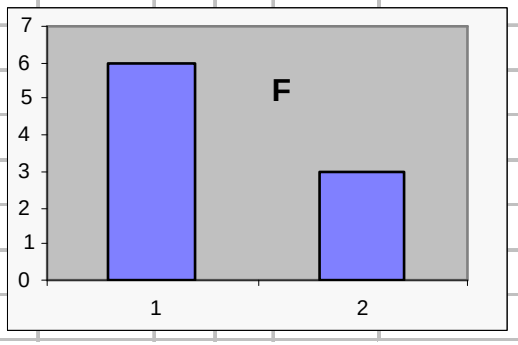
2. Tanti livelli: ***Grafico a linea***

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

Sesso	F	p()
1	6	0,667
2	3	0,375

=FREQUENZA(B2:B10;D2:D3)

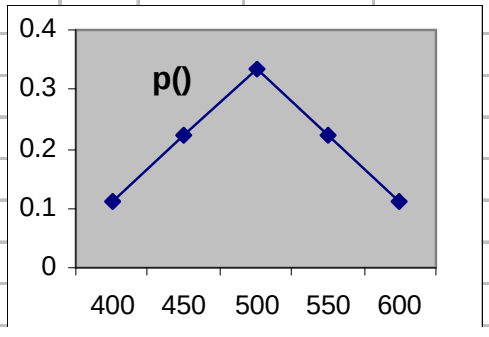
=E2/CONTA.NUMERI(B2:B10)



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	0.111
450	2	0.222
500	3	0.333
550	2	0.222
600	1	0.111

=FREQUENZA(B2:B10;D2:D6)



Esercizio: probabilità empirica

File dati: **S3_RT.ods**

- Calcolare le frequenze degli RT suddivisi nei seguenti livelli [400, 450, 500, 550,…]
- Calcolare la distribuzione delle probabilità
- Fare il grafico della distribuzione di probabilità

www.math.unipd.it/~mpini/Teaching-biotec.htm

Tendenze centrali

1. Se abbiamo dati in una colonna (es.: B2:B21)

2. **Media =** somma divisa per il numero dei punti.

   - selezionare una cella per il risultato (di solito sotto i dati) (es: B23)
   - scrivere '*=somma(*';
   - selezionare le osservazioni; ')'; Premere '*Invio*'
   - in un'altra cella, dividere la somma per il numero dei dati (*conta.numeri()*)

3. **Mediana o moda**: ordinare i dati e trovare:
   - la categoria centrale (mediana)
   - più-frequente (moda)

Variabilità

- Range / Intervallo (distanza tra i valori estremi): *max(*X) – *min*(X)
- Somma dei quadrati delle distanze dalla media (scarti) $SS_X = \sum \left(x_i - \overline{X}\right)^2$
- **Varianza** della **popolazione** (la media degli scarti**²**): $\sigma_X^2 = \frac{\sum \left(x_i - \overline{X}\right)^2}{n}$

$$\sigma_X^2 = \frac{\sum \left(x_i - \overline{X}\right)^2}{n} = \frac{\sum \left(x_i^2 - 2x_i \overline{X} + \overline{X}^2\right)}{n} = \frac{\sum x_i^2}{n} - \overline{X}^2 = \overline{X^2} - \overline{X}^2$$

- **Deviazione standard** (Scarto quadratico medio): $\sigma_X = \sqrt{\sigma_X^2}$

Calcolo della varianza / dev.st

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

$$Var_X = \frac{\sum (x_i - \overline{X})^2}{n}$$

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).

Algoritmo **B:** la differenza 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.

$$Var_X = \overline{X^2} - \overline{X}^2$$

II. La **dev. standard**: **dst**=*radq*(**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:

$$Cov(X,Y) = E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right]$$
$$= E(XY) - E(Y)E(X)$$

  - 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 Mx e My
2. Calcolare il prodotto degli scarti $D_i = (X_i - Mx)\ (Y_i - My)$
3. Calcolare la media Cov = media(D_i).

Algoritmo **B:**

1. Calcolare i prodotti $X_i Y_i$
2. Calcolare le medie Mx, My, Mxy
3. Calcolare la differenza **Cov**=Mxy - Mx My.

Funzioni per la statistica descrittiva

Funzione	Sintassi **(ad esempio, per le osservazioni in A2:A150)**
• MEDIA	*MEDIA*(A2:A150)
• MEDIANA	*MEDIANA*(A2:A150)
• Media distanza dalla Media	*MEDIA.DEV*(A2:A150)
• VARIANZA *della popolazione*	*VAR.POP*(A2:A150)
• SCARTO quadr.med. (*pop*.)	*DEV.ST.POP*(A2:A150)
• COVARIANZA tra 2 campioni	*COVARIANZA*(A2:A150; B2:B150)

Esercizio: medie e variabilità

- File dati: **S3_RT.ods**
  calcolare **con** e **senza** funzioni statistiche:
  la media, la varianza, e la dev.st. dei valori RT

- File dati: **S2_Math.ods**
  calcolare **con** e **senza** funzioni statistiche la **covarianza** tra *add_time* e *sub_time*

www.math.unipd.it/~mpini/Teaching-biotec.htm

Tabelle di contingenza (tabelle pivot)

- Sintetizzano 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, …) 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_iB_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

oggetti	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	Totale
control	**16**
number fact	**4**
Turners Syn.	**6**
Williams Syn.	**5**
Totale complessivo	31

Conteggio di s_type	
s_type	Totale
control	**51,61%**
number fact	**12,90%**
Turners Syn.	**19,35%**
Williams Syn.	**16,13%**
Totale complessivo	100,00%

Tabelle pivot – **procedura (1)**

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

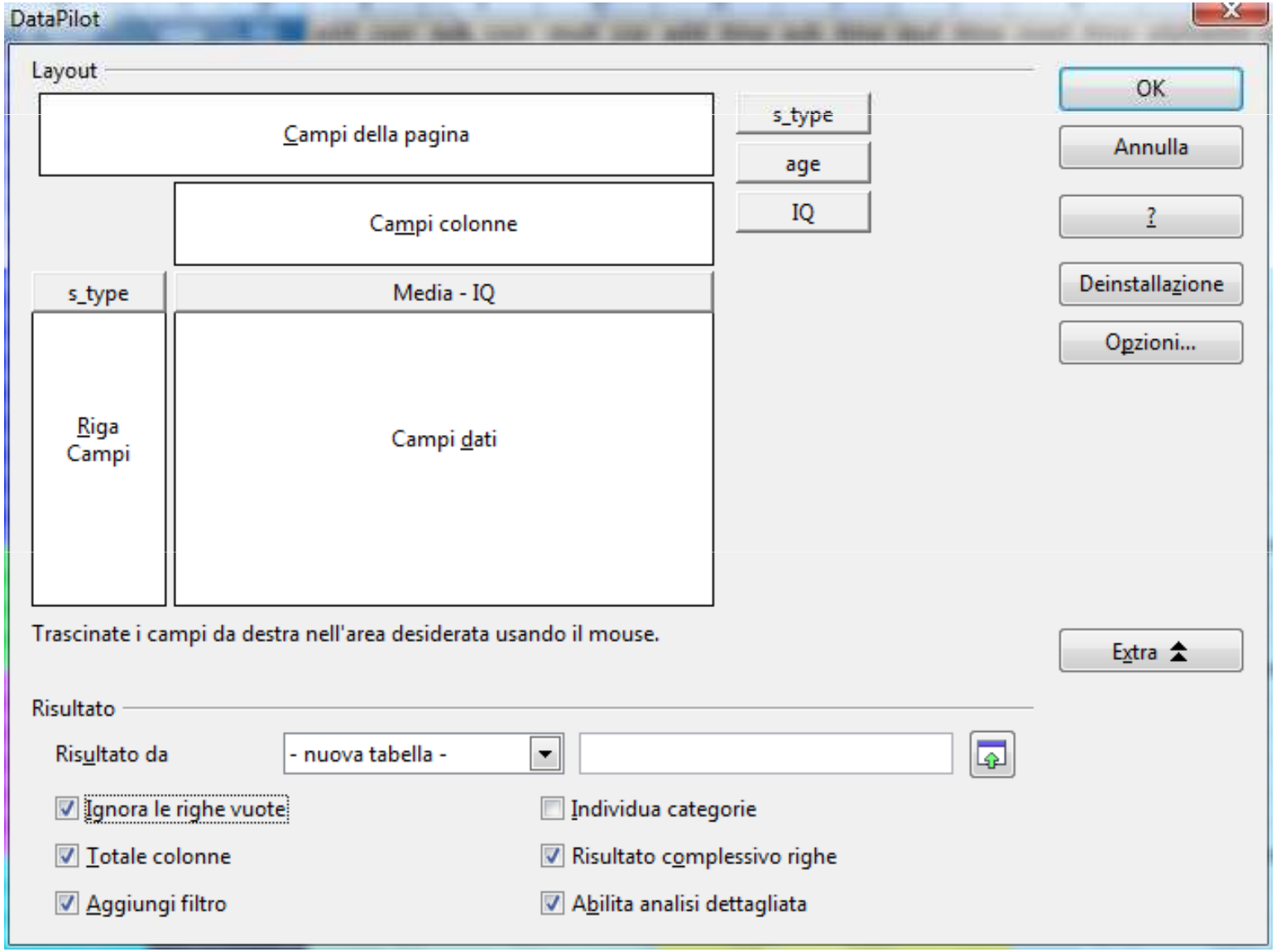


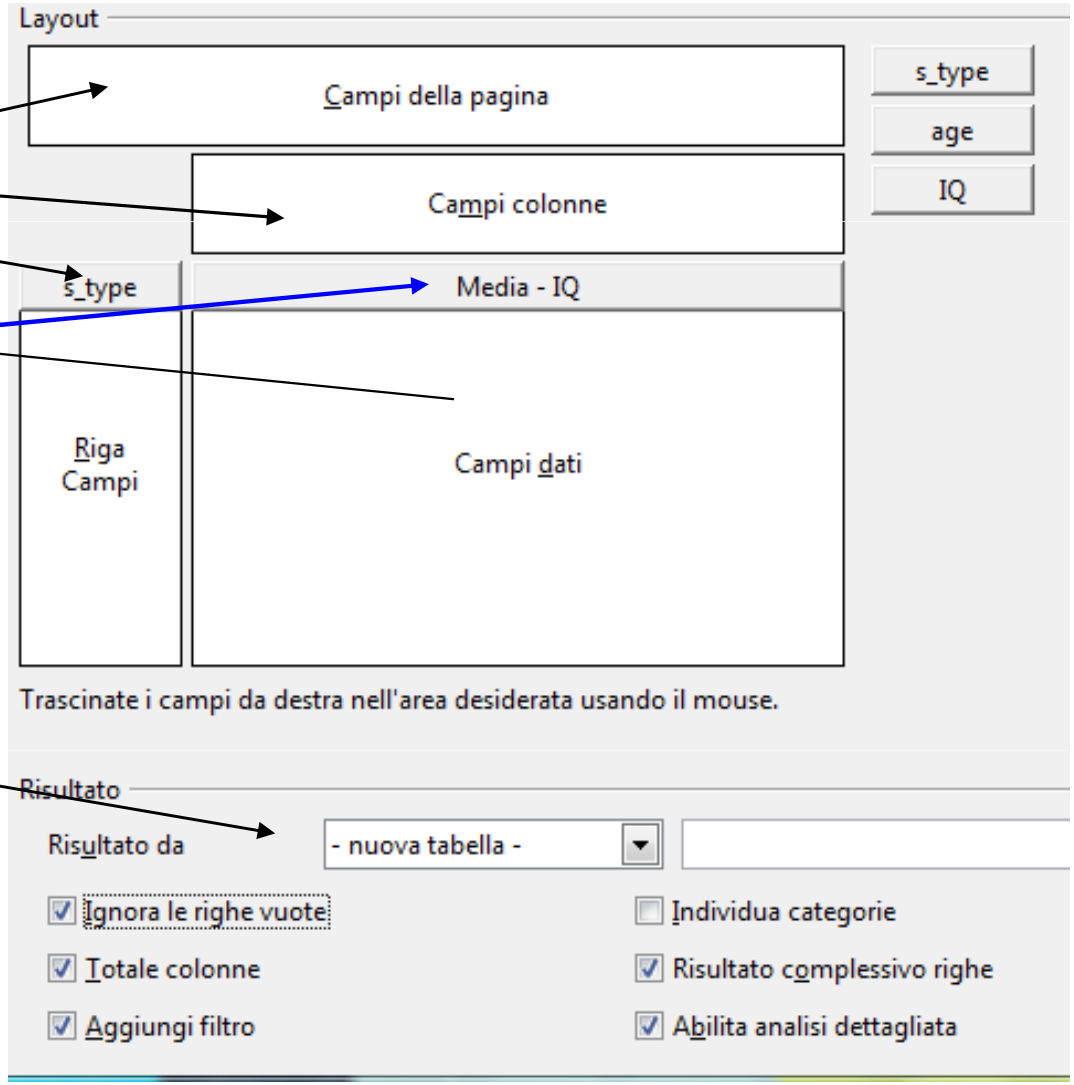
Tabelle pivot – **procedura (2)**

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*)



Esercizio

- File dati: **S3_Dan.ods**

  Creare in fogli diversi di S3_Dan.ods

  – una tabella pivot risp_verbale(luogo, livello)
    dove si esamina la media
  – una tabella pivot risp_verbale(luogo, livello)
    dove si esamina il conteggio
  – una tabella pivot risp_verbale(luogo, livello)
    dove si esamina la dev. st.
  – una tabella pivot risp_verbale(luogo, livello) e risp_sem(luogo, livello) dove si esaminano la media di risp_verbale e il massimo di risp_sem