Laboratorio in Informatica

Laurea in Biologia & Biologia Molecolare, Università di Padova

Statistica descrittiva con fogli di calcolo

- **distribuzione di probabilità empirica**
- **tendenze centrali e variabilità**
- **relazioni lineari**

Docente: Ivilin Stoianov
Assistenti: Michele De Filippo, Alberto Testolin

Lab P150 (Paolotti)
sito web: www.stoianov.it
Email: info@stoianov.it
info@stoianov.it

Distribuzioni di probabilità empirica

1) MISURAZIONI (osservazioni) di un fenomeno su una certa scala:

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) ENUMERARE ogni livello osservato

frequenza f_i del livello y_i nel campione ***X***: il **numero** di osservazioni di y_i nel ***X***

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

3) DISTRIBUZIONE del NUMERO di osservazioni di ogni livello

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***

Categorie abbinate

- 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**) $\{s_1, s_2, \ldots s_M\}$.

- Nella nuova scala S, ciascuno livello $\mathbf{s_i}$ raggruppa livelli $\{y_{i1}, y_{i2}, \ldots s_{ik}\}$

- 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 100 livelli, cioè, **0-100 cm**, dove ogni **cm** raggruppa **10 mm.**

info@stoianov.it

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 abbinati ***S*** in un'altra colonna (es: D2:D11)
3. Applicare la funzione ***frequency*(*osservazioni; livelli_abbinati*)**
   - *Selezionare* una colonna per il risultato (es., E2:E11)
   - Scrivere '=frequency('
   - Selezionare / riferire le osservazioni; scrivere '**;**'
   - Selezionare / riferire i livelli abbinati; scrivere ')'
   - Premere '***Ctrl-Maiusc-Invio***' (solo '*Invio*' calcola un solo valore!)

Si nota:

- I livelli indicati definiscono la nuova scala ***S***.
- La frequenza f_i di ciascun livello s_i riporta il numero di osservazioni con valori (s_{i-1} ... s_i].

Distribuzioni: calcolo e rappresentazione grafico

Grafici:

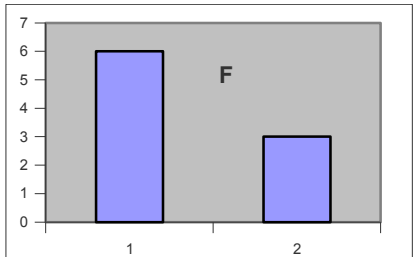
1. Pochi livelli: ***Istogramma*** (diagramma a barre)

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	0.6667
2	3	0.375

=FREQUENZA(B2:B10;D2:D3)

=E2/CONTA.NUMERI(B2:B10)

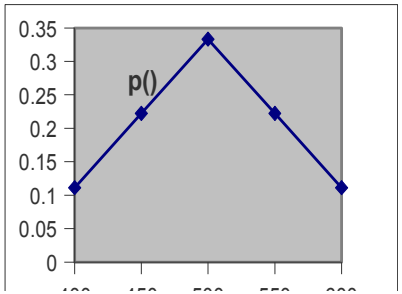


2. Tanti livelli: ***Grafico a linea***

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)



info@stoianov.it

Esercizio: probabilità empirica

- Il file ES_RT.xls contiene tempi misurati in **ms.**

Compito:

- calcolare le frequenze dei RT (in bin che raggruppano ogni 50 ms)
- calcolare la distribuzione di probabilità empirica
- fare un grafico della distribuzione di probabilità empirica.

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)

Funzioni: ***AVERAGE***(B2:B21), ***MEDIAN***(B2:B21), ***MODE***(B2:B21)

ESERCIZIO: usando i dati nel file **ES_Math.ods**,
calcolare la media, mediana, moda del **IQ** dei *soggetti di controllo*

Variabilità

- Range / Intervallo (distanza tra i valori estremi): *max*(X) – *min*(X)
- Somma dei quadrati delle distanze dalla media (scarti) $SS_X = \sum \left(x_i - \bar{X} \right)^2$
- **Varianza** della **popolazione** (la media degli scarti²): $\sigma_X^2 = \frac{\sum \left(x_i - \bar{X} \right)^2}{n}$

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

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

Funzioni statistiche:

VARIANZA *della popolazione* *VAR*(A2:A150)

SCARTO quadr.med. (*pop.*) *STDEV*(A2:A150) DEV.ST(...)

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).

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

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.

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

II. La **dev. standard**: **dst=*sqrt*(var)**

$$\sigma_X = \sqrt{Var_X}$$

ESERCIZIO: usando i dati nel file **ES_Math.ods**, calcolare la **varianza** e **dev.st.** del **IQ** dei *soggetti di controllo* con entrambi Algoritmi.

Covarianza

- Due variabili aleatorie **X** e **Y** possono essere relati per una loro relazione intrinseca (per.es.: voto in Statistica e voto in Informatica.)
- La covarianza – la media del prodotto dei scarti di **X** e **Y** rispetto alle loro medie – esprime il grado di dipendenza lineare tra X e Y:
  - se positivo: al **crescere** di x in media **cresce** anche y,
  - se negativo al **crescere** di x in media **decrescere** y.
- Nota: la covarianza di una variabile con se stessa = la varianza !

Algoritmo **A**: $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$

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

Algoritmo **B:** $= E(XY) - E(Y)E(X)$

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

Funzione: COVAR(A2:A150; B2:B150)

ESERCIZIO: calcolare la covarianza tra *add_time*, *sub_time*

Regressione Lineare

- Partendo da una serie di valori accoppiati (X_i, Y_i), cerchiamo la relazione **Y=b0+b1*X** (modello lineare con parametri **b0** e **b1**)
- Per un modello con parametri **b0** e **b1**, la previsione **Y' = b0+b1*X**
- **Metodo**: minimizzare l'errore *osservazione-predizione*: (**Y** – **Y'**)^2

Stima dei coefficienti:

- Le medie di **X** e **Y** — **mX** e **mY**
- I quadrati degli scarti di **X** e **Y** — $dX_i = (X_i - mX)^2$ e $dY_i = (Y_i - mY)^2$
- I prodotti degli scarti di **X** e **Y** — $dXY_i = (X_i - mX)*(Y_i - mY)$
- Le somme degli scarti — **DXY**=*sum*(**dXY**$_i$), **DX**=*sum*(**dX**$_i$), **DY**=*sum*(**dY**$_i$)
- Il coefficiente b1: — **b1** = **DXY** / **DX**
- Il coefficiente b0: — **b0** = **mY** – **b1***mX

Stima della precisione del modello:

- Le previsioni di **Y** su **X** — $Y_i' = b0 + b1*X_i$
- Errori delle previsioni — $eY_i = (Y_i - Y_i')^2$
- Somma degli errori — **EY**=*sum*(**eY**$_i$)
- La proporzione — **R2 = 1-EY/DY** indica la bonta del fit (valori 0-1)

ESERCIZIO: stimare la regressione ***add_time*** = ***b0*** + ***b1*** * ***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, …) 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

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**

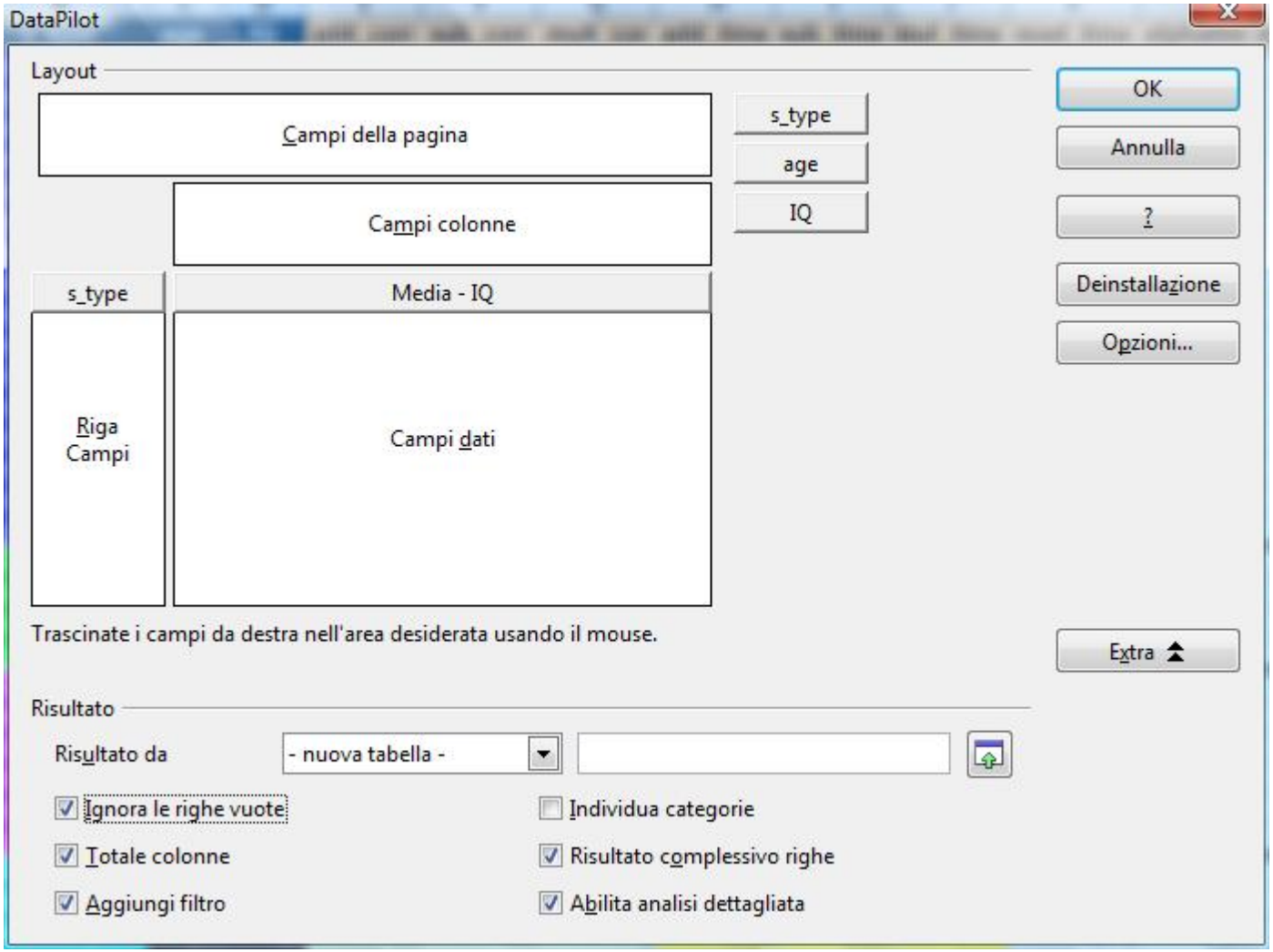


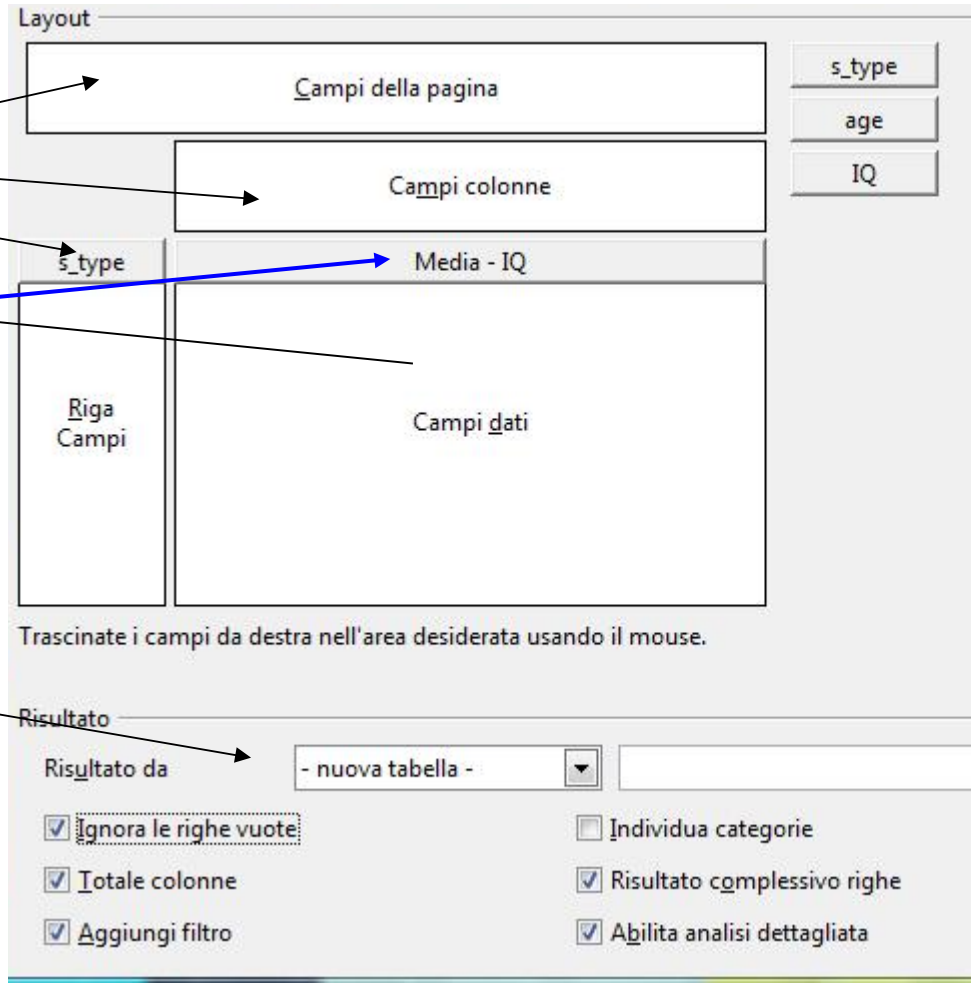
Tabelle pivot – **procedura (2)**

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*)



Esercizio

File dati: **ES_Les**

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.

Laboratorio Informatica

Laurea in Biologia & Biologia Molecolare, Università di Padova

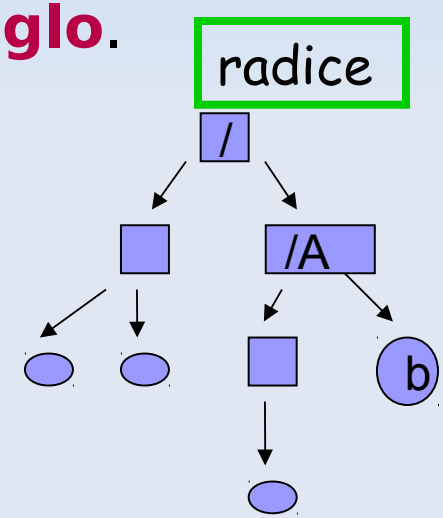
Linux

Docente: Ivilin Stoianov
Assistenti: Michele De Filippo, Alberto Testolin

Lab P150 (Paolotti)
sito web: www.stoianov.it
Email: info@stoianov.it

Organizzazione ad albero

- I dischi fissi sono divisi in partizioni.
- Ogni **partizione** è organizzata come un **albero**.
- Ogni nodo dell'albero e **file** o **directory**.
- I nodi sono organizzati con relazioni **padre-figlo**.
- Ogni figlo ha un solo padre.
- Un padre potrebbe avere più figli.
- **Directory:** un nodo con figli.
- **File**: un nodo senza figli.
- Il padre più in alto = **radice**



- Percorso assoluto: specifica la posizione di un file a partire dalla radice: "**/A**/B/d" (parte da "**/**")
- Percorso relativo: specifica la posizione a partire da una posizione interna: "B/d" (parte senza "**/**")

info@stoianov.it

Gestire Linux dal Terminal

Interfaccia grafica: accesso grafico agli applicazioni

Shell: accesso testuale agli applicazioni, tramite un programma chiata “Terminal” (“Shell”)

Sintassi: **nome_comando [opzioni] [argomento1…argomentoN]**

- in parentesi quadre: argomenti opzionali
- gli argomenti sono separati con spazio
- i commando sono scritti con lettere minuscole

info@stoianov.it

Commandi relativi al percorso

pwd indica la posizione attuale nel file system

ls visualizza il contenuto di una cartella o i file

ls [opzione] [nome_directory] [nome_file]

-a elenca tutti i file compresi quelli nascosti

-l fornisce informazioni aggiuntive

cd cambiare la cartella (la posizione nel file system)

cd [nome-directory]

cd .. sale di un livello nell'albero

cd ritorna all propria home

mkdir creare una cartella

mkdir [opzioni] nome_directory

rmdir eliminare una cartella

rmdir nome_directory

info@stoianov.it

Esercizio

- aprire un terminal
- individuare la posizione all'interno del albero dei file
- visualizzare I file che stanno nella cartella attuale
- andare in cartella “Documents”
- visualizzare il suo contenuto
- all'interno di tale cartella, creare una cartella ‘**biologia**’
- verificare che la cartella esiste
- spostarsi nella cartella “**biologia**”
- crerare una cartella ‘**biologia2**’ all'interno della cartella ‘biologia’
- verificare che la cartella esiste
- eliminare la cartella “**biologia2**”

Creare e visualizzare file (di testo)

pico semplice editore di testo (o "**nano**")

pico [nome_file]

Nell'editore: **Ctr-O** salva il file; **Ctr-X** esce dal editore

more visualizza il contenuto di un file

more nome_file

Tasti: **barra-*spazio*** per successiva pagina; "**Q**" per uscire.

less visualizza il contenuto di un file

Tasti: **frecce** per navigare; "**Q**" per uscire.

less nome_file

Esercizio

- spostarsi nella cartella "Documents"
- creare un file di testo e scrivere dentro qualcosa
- salvare il file
- visualizzare il suo contenuto utilizzando "more" e "less"

info@stoianov.it

Dupplicare ed eliminare dei file

cp copiare una cartella o file

- in una cartella

cp [opzioni] file_origine nome_directory

- con un nuovo nome

cp [opzioni] file_origine file_destinazione

Esempio: cp mydoc.txt cartella

rm eliminare una o più cartelle o file

rm [opzioni] nome-file

man manuale dei commandi di linux (o altre applicazioni)

Navigare con le **frecce** ed uscire con tasto "**Q**"

man nome-commando

Esercizio

- dopo aver creato un file di testo “prova.txt” (nella cartella “*Documents*”, duplicare il file con il nuovo nome ‘prova2.txt’

- copiare il file prova2.txt in nella cartella “biologia”, che è una sottocartella della cartella “Documents”

- eliminare il file “prova.txt”