Esercitazione I

Sommario della Esercitazione 1

- Statistica descrittiva con Matlab
  - Costruzione di istogrammi
  - Calcolo di frequenze
  - Caratterizzazione tendenza centrale di un campione
  - Caratterizzazione della variabilità di un campione

- Spazio Campionario
- Eventi e probabilità
- Probabilità condizionata

Statistica descrittiva

- Misuriamo, replicando molte volte, la viscosità di un polimero a valle di un reattore di polimerizzazione
- Si supponga, inoltre, che le condizioni sperimentali siano sempre le stesse (pressione, temperatura, ecc.)
- Si ottengono i dati in Pa×s riportati affianco
- Come ormai sappiamo bene i dati sono affetti da incertezze dovute agli errori di misura. (NON CI SONO ERRORI SISTEMATICI)
- Vogliamo caratterizzare le proprietà del campione:
  - Frequenze assolute, relative, cumulative, istogrammi, tendenza centrale, variabilità.

1.19	1.18
1.15	1.18
1.24	1.23
1.25	1.20
1.20	1.18
1.15	1.24
1.15	1.23
1.19	1.21
1.20	1.19
1.17	1.19
1.22	1.22
1.22	
1.18	
1.17	
1.18	
1.18	
1.21	
1.20	
1.16	

Statistica descrittiva

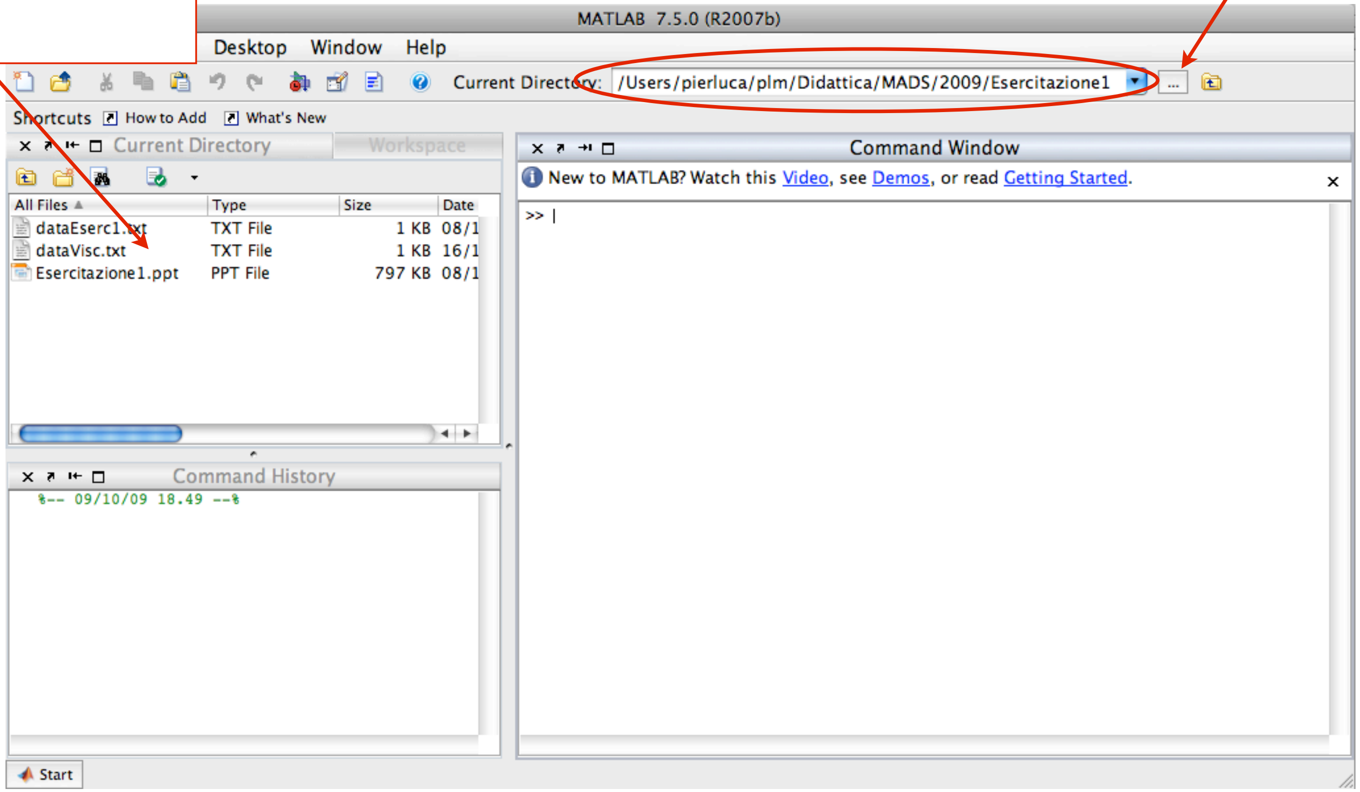
- Utilizzeremo MATLAB
- Dovremo:
  - Caricare i dati da un file
  - Ordinare i dati in senso crescente
  - Calcolare e diagrammare la FREQUENZA ASSOLUTA
  - Calcolare e diagrammare la FREQUENZA RELATIVA
  - Calcolare e diagrammare la FREQUENZA CUMULATIVA ASSOLUTA
  - Calcolare e diagrammare la FREQUENZA CUMULATIVA RELATIVA
  - Calcolare media, mediana, moda
  - Calcolare deviazione standard…

Uso di Matlab: Ingresso dati

- Caricamento dati

1 Cercare la directory dove sono i dati

2 Contenuto della directory



Uso di Matlab: Ingresso dati

- I dati sono nel file di testo: dataVisc.txt
- Caricamento:

```
var = load('filename.ext')
```

Variabile in cui memorizzare i dati

nome del file in cui sono presenti i dati

```
>> dati=load('dataVisc.txt')

dati =

    1.1900
    1.1500
    1.2400
    1.2500
    1.2000
    1.1500
    1.1500
    1.1900
    1.2000
    1.1700
    1.2200
    1.2200
    1.1800
    1.1700
    1.1800
    1.1800
    1.2100
    1.2000
    1.1600
    1.1800
    1.1800
    1.2300
    1.2000
    1.1800
    1.2400
    1.2300
    1.2100
    1.1900
    1.1900
    1.2200

>>
```

Uso di Matlab: dimensione del campione

- Possiamo determinare la dimensione del campione caratterizzare la struttura del vettore dati con i comandi:

```
length(dati)
size(dati)
```

```
>> length(dati)

ans =

    30

>> size(dati)

ans =

    30     1

>> |
```

- Ordinarli in senso crescente

```
sort(var)
```

```
>> datisort=sort(dati);
>>
```

```
length(dati)
load('filename.ext')
size(dati)
sort(dati)
```

Uso di Matlab: Istogrammi

- Determinazione dell'istogramma della frequenza assoluta:
  1. Determinazione dell'intervallo
  2. Suddivisione in classi
  3. Conteggio

- Determinazione dell'intervallo e suddivisione in classi

```
>> x=min(dati):0.01:max(dati)

x =

  Columns 1 through 8

    1.1500    1.1600    1.1700    1.1800    1.1900    1.2000    1.2100    1.2200

  Columns 9 through 11

    1.2300    1.2400    1.2500
```

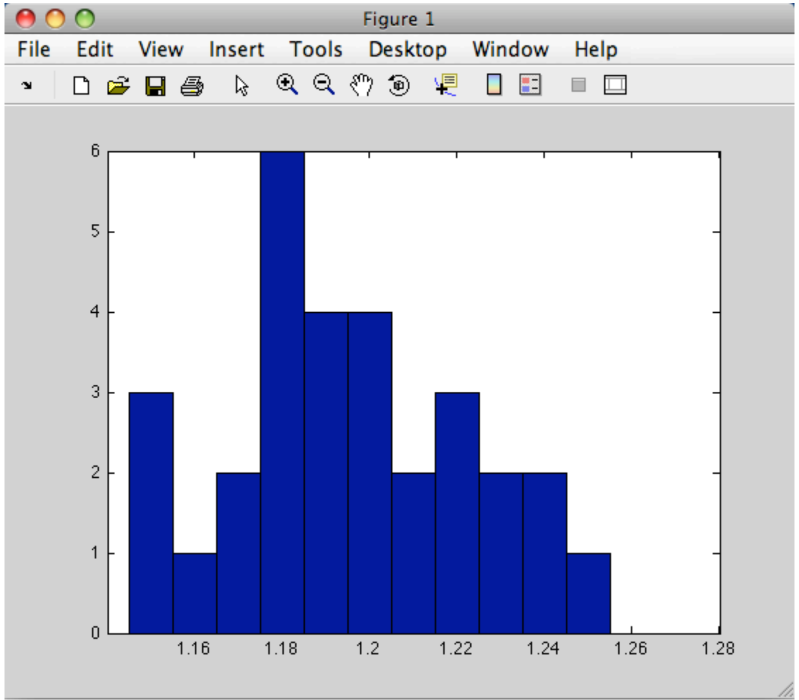
  - **Attenzione: abbiamo scelto intervallini di ampiezza 0.01**

Uso di Matlab: Istogrammi

- Costruzione dell'istogramma della frequenza assoluta: comando hist

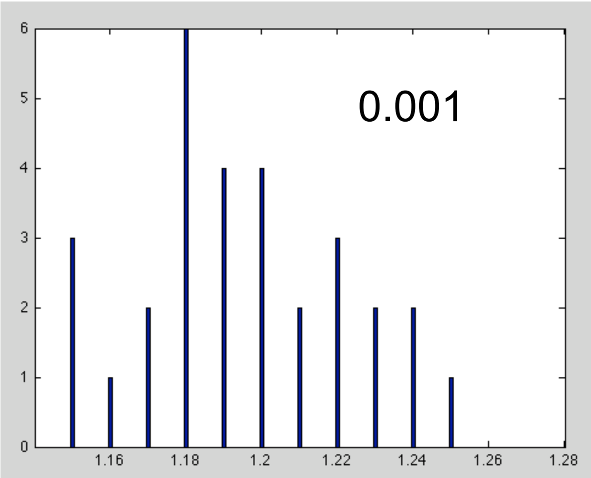
```
>> hist(dati,x)
```

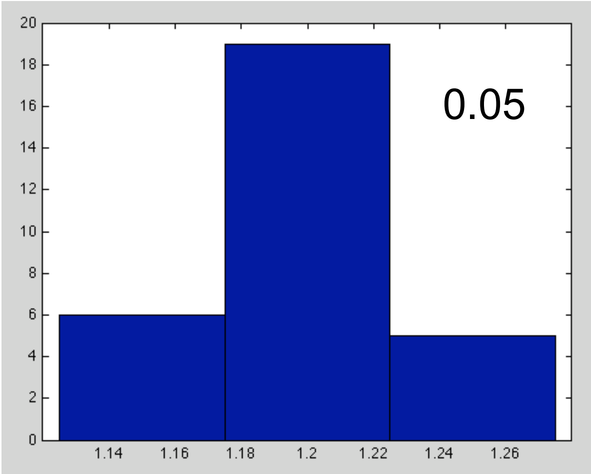
- Si genera una figura



Istogrammi: scelta del numero di classi

- Scelta dell'ampiezza delle classi

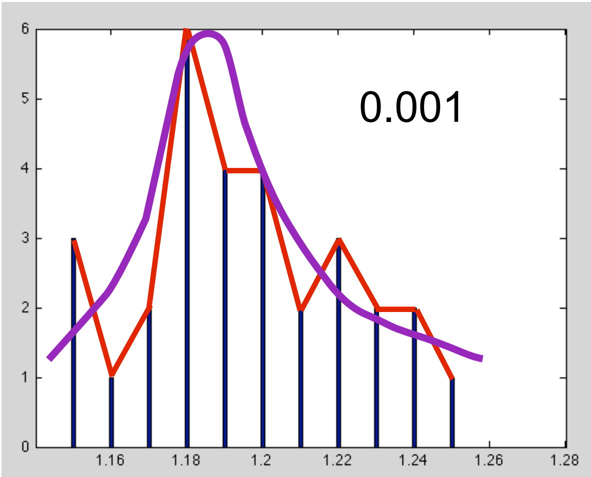


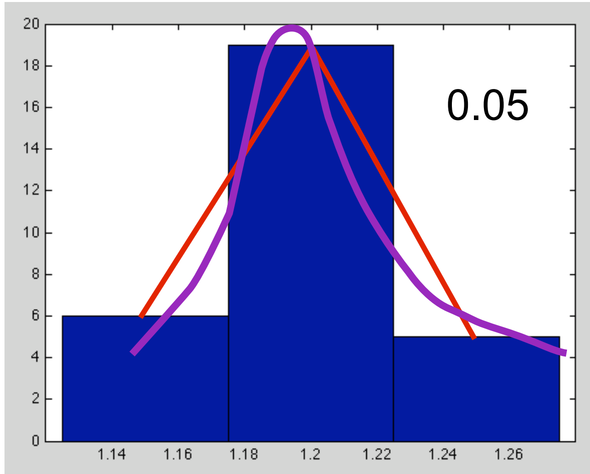


- Il numero di classi influenza fortemente l'istogramma
- Noi già sappiamo che l'istogramma ci fornisce indicazioni sulla sottostante distribuzione della popolazione.
- Costruiamo una stima della distribuzione facendo passare una poligonale per i punti medi delle barre dell'istogramma

Istogrammi: scelta del numero di classi

- "Stima" della distribuzione con la poligonale. (Immaginiamo di conoscere la verità in viola)





- Nel primo caso la poligonale è molto frastagliata e ci dà una idea ragionevole della distribuzione vera ma abbiamo varianze alte
- Nel secondo caso la poligonale è abbastanza regolare ma perde dettagli significativi della distribuzione vera

Istogrammi: scelta del numero di classi

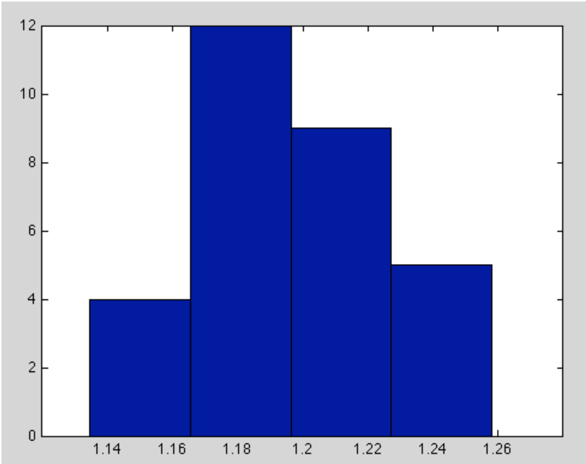
- Scelta ottimale difficile.
- Esistono regole empiriche per la scelta del numero di classi (ovviamente r deve essere un intero)

$$r = 1 + 3.3log(N) \qquad Sturges(1926)$$

$$r = 1 + 2.2log(N) \qquad Larson(1975)$$

- Criterio di ottimalità basato sulla variabilità del campione Scott (1979)

$$r = \frac{max - min}{3.49\sigma N^{-\frac{1}{3}}}$$



Uso di Matlab: Calcolo delle frequenze

- Come costruire frequenze assolute e relative
- Comando: `histc(dati,edges)`
  - dati: vettore dei dati
  - edges: vettore bordi delle classi (edges=min(dati):(max(dati)-min(dati))/r:max(dati);)
  - restituisce il conteggio delle frequenze assolute corrispondenti alle classi
  - se dividiamo il risultato per la dimensione del campione abbiamo il vettore delle frequenze relative
  - controllate i dettagli della funzione con help histc

- Costruzione alternativa dell'istogramma delle frequenze

```
bar(ascissa,ordinate)
```

Uso di Matlab: Calcolo delle frequenze cumulative

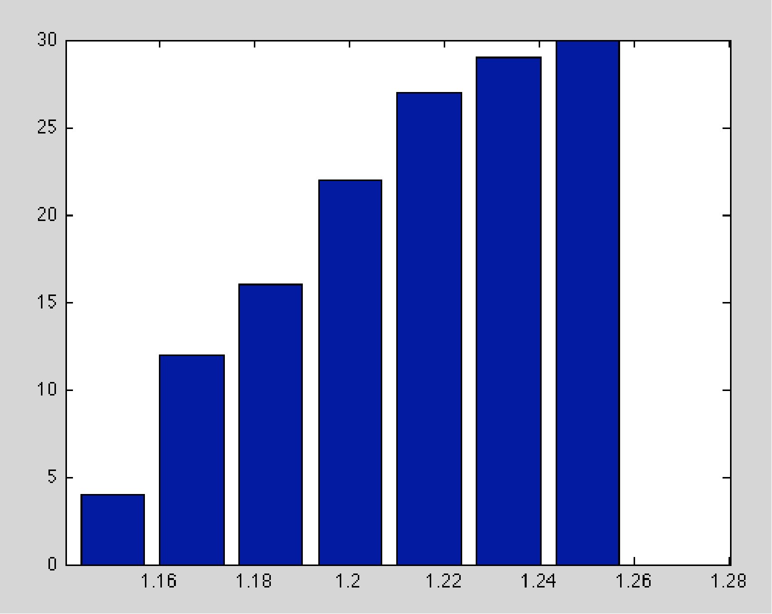
- Per calcolare le frequenze cumulative `cumsum(histc(x,edges))`
  - La funzione cumsum restituisce direttamente le frequenze cumulative (in questo caso assolute)

```
>> x=min(dati):(max(dati)-min(dati))/6:max(dati);
>> cumsum(histc(dati,x))

ans =

     4
    12
    16
    22
    27
    29
    30

>> bar(x,ans)
```



- Misure della tendenza centrale

```
mean(dati)
median(dati)
mode(dati)
```

```
>> mean(dati)

ans =

    1.1953

>> median(dati)

ans =

    1.1900

>> mode(dati)

ans =

    1.1800
```

- Misure di variabilità del campione (Matlab usa la definizione di deviazione standard e di varianza basata su N-1 gradi di libertà)

```
std(dati)
var(dati)
```

```
>> std(dati)

ans =

    0.0275

>> var(dati)

ans =

   7.5678e-04
```

Spazio campionario

- Una scatola contiene 10 transistor fabbricati da A e 10 da B
- Consideriamo i seguenti 4 esperimenti casuali
  - E1: 3 transistor sono estratti a caso con rimpiazzo e contiamo il numero di transistor A (tra quelli selezionati)
  - E2: 3 transistor sono estratti a caso con rimpiazzo e la marca di ciascun transistor è annotata
  - E3: I transistor sono estratti uno per volta a caso con rimpiazzo finchè non è estratto un transistor di marca A; il numero di B estratti fino al primo A viene annotato
  - E4: un transistor è estratto a caso e la sua durata viene misurata (in ore).
- Indicare lo spazio campionario dei 4 esperimenti
  - S1= {0,1,2,3}. Numero di eventi finito. Spazio campionario discreto
  - S2 = {AAA, AAB, ABA, BAA, ABB, BAB, BBA, BBB}. Numero di eventi finito. Spazio campionario discreto
  - S3 = {0,1,...}. Numero di eventi infinito ma numerabile. Spazio campionario discreto
  - S4 = $[0,\infty)$. Numero di eventi infinito non numerabile. Spazio campionario continuo

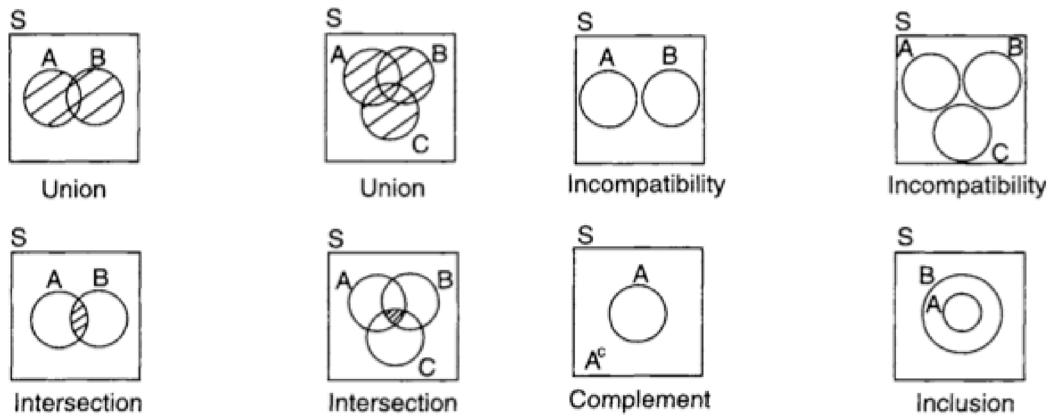
Esercizio 1

- Indicare gli eventi definiti rispetto agli spazi campionari appena visti:
  - A1: estraiamo un solo transistor A
  - A2: estraiamo un A e 2 B
  - A3: estraiamo 5 o 6 transistor B prima del primo A
  - A4: Il transistor estratto dura più di 200 ore

- A1={1};
- A2={ABB, BAB, BBA},
- A3={5,6},
- A4=(200, ∞)

Operazioni su eventi

- Operazioni con gli insiemi



1) $A \cup B = B \cup A$ and $A \cap B = B \cap A$ (Commutativity).

2) $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$ (Associativity).

3) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (Distributivity).

4) $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$ (De Morgan's laws).

Esercizio II

- 2 transistor sono estratti a caso senza rimpiazzo dalla scatola con 10 A e 10 B
- Quale è la probabilità di avere
  1. 2 A
  2. 2A o 2B
  3. un A ed un B
  - Ak= evento estrazione A alla k-sima estrazione
  - $B \cap C$: Risultati che appartengono sia a B che a C
  - $B \cup C$: Risultati che appartengono a B o a C

- SOLUZIONE

1. $P[A1 \cap A2] = P[A2|A1]P[A1] = 9/19 \times 10/20 = 9/38$
2. $P[(A1 \cap A2) \cup (B1 \cap B2)] = P[(A1 \cap A2)] + P[(B1 \cap B2)] = 9/38 + 9/38 = 9/19$
3. $P[((A1 \cap A2) \cup (B1 \cap B2))^C] = 1 - 9/19 = 10/19$

Esercizio III

- 1) Un esperimento casuale può avere esiti {a, b, c, d} con probabilità 0.1, 0.3, 0.5, e 0.1 rispettivamente. Si denoti con A l'evento {a, b}, B l'evento {b, c, d} e C l'evento {d}. Si determini: $P(A)$, $P(B)$, $P(C)$, $P(A \cap B)$, e $P(A \cup B)$.

- SOLUZIONE

- $P(A)=P(a)+P(b)=0.4$
- $P(B)=0.9$
- $P(C)=0.1$
- $P(A \cap B)=P(b)=0.3$
- $P(A \cup B)=1$

ESERCIZIO IV

- 2) In tabella è riportata la casistica di 940 wafer prodotti in un industria di semiconduttori. Si supponga di aver selezionato un wafer in modo casuale.
  1. Sia H l'evento che rappresenta una alta contaminazione. Quale è la probabilità P(H)?
  2. Si denoti con C l'evento che il wafer sia al centro dello strumento di sputtering. Quanto vale P(C)?
  3. Si determini la probabilità dell'evento $P(H \cup C)$ usando la regola dell'addizione.

Table 2-1 Wafers in Semiconductor Manufacturing Classified by Contamination and Location

	Location in Sputtering Tool		
Contamination	Center	Edge	Total
Low	514	68	582
High	112	246	358
Total	626	314	

- SOLUZIONE
  - 1) 358/940; 2) 626/940
  - 3) $P(H \cup C) = P(H) + P(C) - P(H \cap C) = 358/940 + 626/940 - 112/940 = 872/940$

Lista comandi Matlab utilizzati

```
bar(x,y)
cumsum(vett)
hist(dati,x)
histc(dati,edges)
length(dati)
load('filename.ext')
mean(dati)
median(dati)
size(dati)
sort(dati)
std(dati)
var(dati)
```