Statistica descrittiva

Sommario Lezione 2

- Molti di noi devono lavorare con dati. Dati prodotti con fatica da apparecchiature sperimentali. **I dati sono oggetti preziosi**

- Cominceremo a studiare la statistica analizzando come organizzare e descrivere dati (essenzialmente liste di numeri)

- Poi in modo naturale (si spera!) emergerà il concetto di distribuzione di probabilità.
  - Con la probabilità lavoreremo a lungo!

Prove ripetute

- Quando facciamo una misura y_1 di una quantità Y l'osservazione sperimentale approssima il valore vero di y. Se ripetiamo la misura $y_2 \neq y_1$ ma entrambe approssimano il valore vero.

- Facendo altre misure vediamo emergere una struttura nei dati che raccogliamo
  - Se siamo bravi e non commettiamo errori sistematici i valori tendono ad addensarsi attorno al valore vero

- Con una campagna sperimentale si ottengono sequenze di osservazioni che vengono riportate nell'ordine in cui si determinano.

- Per esempio in una sperimentazione abbiamo raccolto i trenta dati riportati nella tabella.

1.2	1.0
1.3	1.2
1.1	1.1
1.4	1.1
1.0	0.9
1.2	1.5
1.2	1.2
1.3	1.2
1.4	1.2
1.0	1.4
0.8	1.3
1.1	1.1
2.0	0.7
1.1	1.1
1.2	1.1

Campione

- INFORMAZIONI IMPORTANTI
  - Esiste un valore attorno a cui i dati si addensano?
  - In che intervallo si dispongono i dati?

- Il campione è caratterizzato da:
  - **Dimensione** (SIZE): il numero di dati
  - I valori dei dati

Dimensione del campione

- Dato un campione casuale, si chiama **statistica campionaria** qualunque funzione del campione.

Frequenza

- Il campo in cui variano i valori dell'esempio è: 0.7, 2.0

Range

- Suddividiamo il campo in **classi** (bin)
  - In genere non è elementare scegliere la dimensione delle classi.
    - Non troppo piccole
    - Né troppo grandi
  - Per questo esempio consideriamo le classi (0.7,0.8,…,2.0)

- Possiamo contare quanti dati appartengono a ciascuna classe ottenendo la **FREQUENZA ASSOLUTA**.

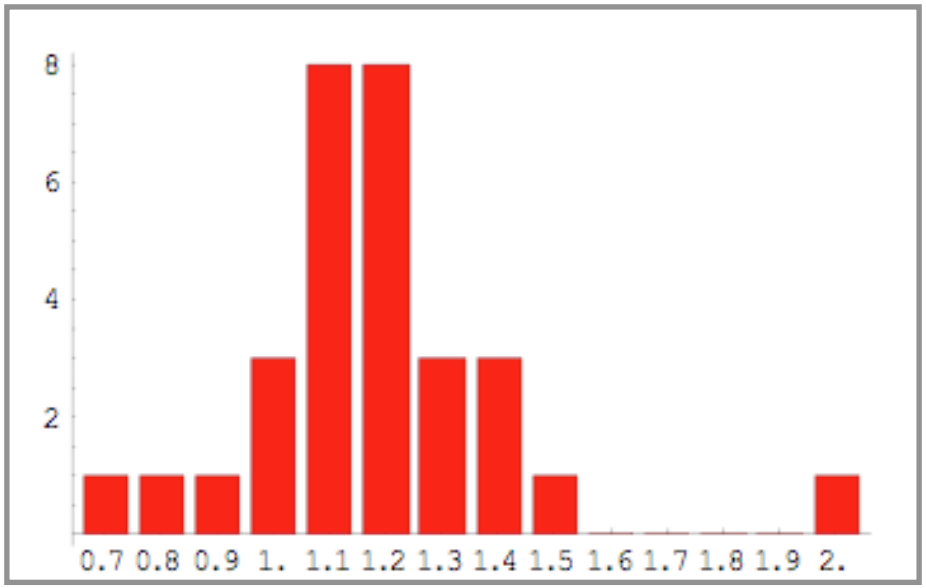
Frequenza assoluta

- La somma delle frequenze assolute è sempre pari alla dimensione del campione.

Ottimizzazione delle classi
http://statweb.calpoly.edu/chance/applets/Histogram.html

Istogramma

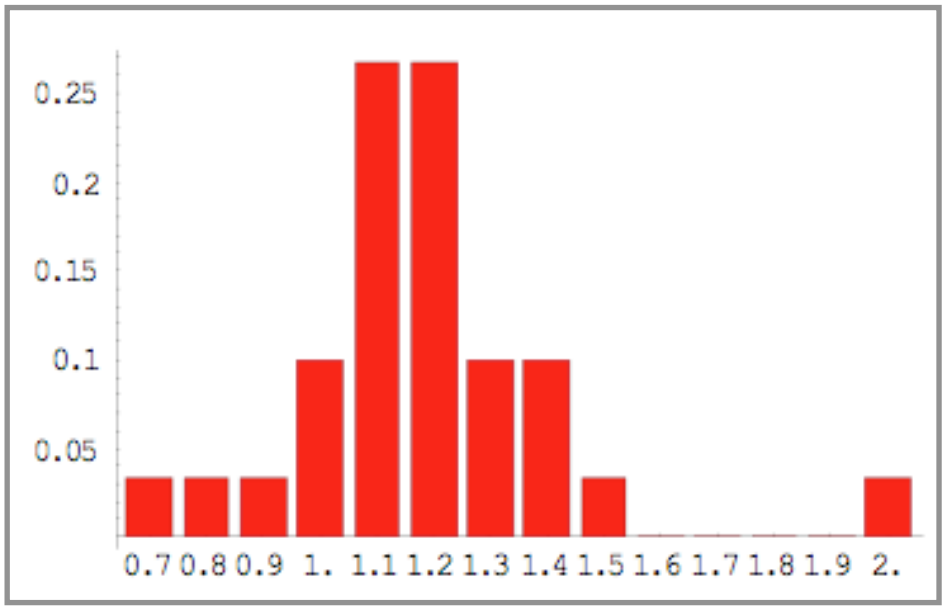
- Costruiamo un istogramma.
- **Frequenza assoluta**



Frequenza

- Alternativamente possiamo definire una **FREQUENZA RELATIVA** dividendo la frequenza assoluta per la dimensione del campione (in questo caso 30)

Frequenza relativa



Frequenza

- **Il valore della frequenza relativa è sempre compreso nell'intervallo [0,1], e la somma delle frequenze relative è sempre 1.**

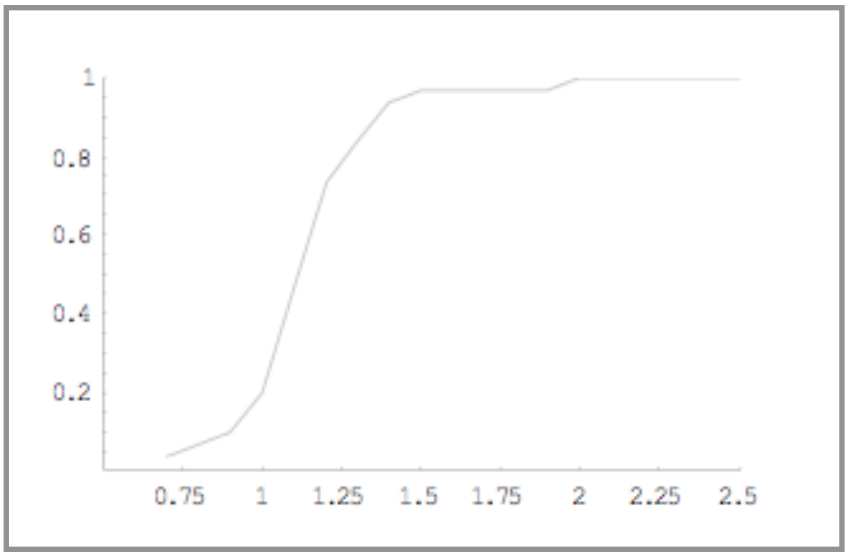
- Abbiamo anche altri modi per illustrare le frequenze.

- Per esempio si può descrivere il campione con le frequenze cumulative.

- La **FREQUENZA CUMULATIVA ASSOLUTA** è la sommatoria delle Frequenze Assolute per $x \leq x_0$

Frequenza cumulativa

- Analogamente si definisce una **FREQUENZA CUMULATIVA RELATIVA** dividendo quella assoluta per la dimensione del campione



Frequenza ed istogrammi con Excel

- Excel consente di creare una tabella di frequenze a partire da una tabella di dati mediante la funzione:

- FREQUENZA(Matrice_dati, Matrice_classi)
  - Matrice_dati è una tabella monodimensionale (array) che contiene l'insieme di valori di cui vogliamo calcolare le frequenze.
  - Matrice_classi è una tabella monodimensionale che contiene gli intervalli in cui vogliamo raggruppare i valori in Matrice_dati.

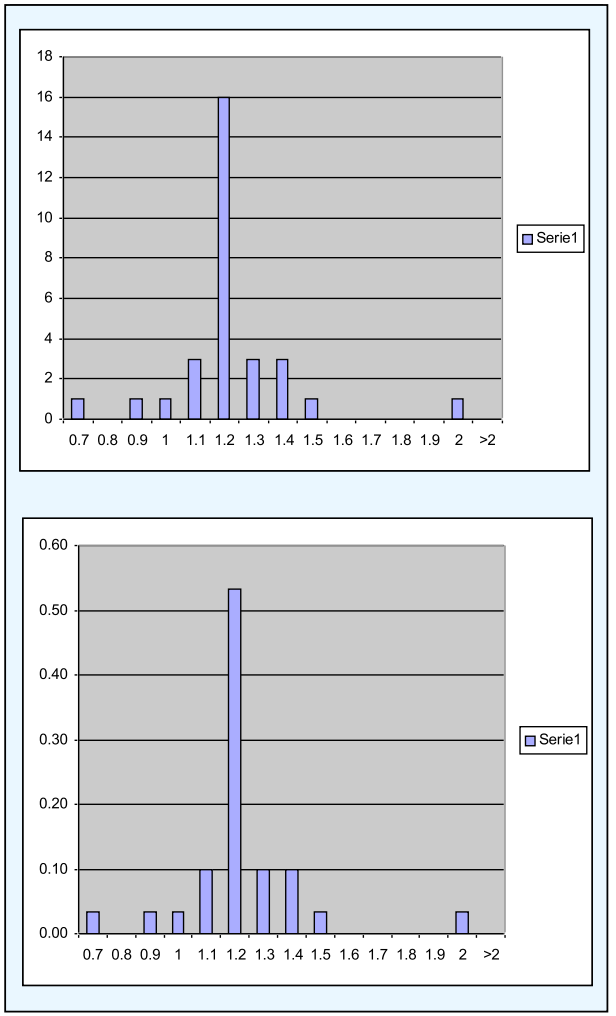
Frequenza ed istogrammi con Excel

- Step 1: in un'area dello stesso foglio che contiene la tabella dei dati o su un altro foglio costruire una tabella monodimensionale (un array) contenente il valore superiore per ciascuna classe

- Step 2: selezionare la colonna adiacente con un numero di celle pari a quello delle classi piu` 1
  - digitare la formula: FREQUENZA(Matrice_dati, Matrice_classi)
  - premere insieme i tasti CTRL-SHIFT-ENTER

- Step 3: Frequenze relative
  - Frequenze assolute / numero dei dati

- Step 4: costruzione diagrammi a barre



Frequenza e Probabilità

- Abbiamo già detto che: Quando facciamo una misura y_1 di una quantità Y l'osservazione sperimentale approssima il valore di Y. Se ripetiamo la misura $y_1 \neq y_2$ ma entrambe approssimano il valore vero.

- Se facessimo un numero infinito di misure allora potremmo descrivere il modo in cui sono distribuiti i dati osservati, e, quindi:
  - conosceremmo la funzione di frequenza.
  - conosceremmo la **probabilità** di ottenere un particolare valore in una misura.

Frequenza e Probabilità

- Per **probabilità** intendiamo numeri reali compresi tra 0 ed 1 collegati ad insiemi in un qualche spazio matematico.

Probabilità

- Le probabilità sono modelli delle frequenze con cui accadono **eventi**.

Evento

- L'idea è che abbiamo uno spazio di accadimenti che ci interessa: lo **spazio di probabilità**.
- In genere modelliamo questo spazio in modo da avere insiemi che chiamiamo eventi. Scegliamo *un evento A* (per esempio la viscosità misurata compresa tra 9 e 10 Poise) e vediamo quante volte si verifica tale evento rispetto al numero totale di accadimenti.
  - Questa è la frequenza relativa dell'evento *A*

Probabilità

- La probabilità può essere definita come il grado di fiducia che attribuiamo ad un fatto il cui verificarsi non è certo.
  - Se un consumatore entra in un negozio non è possibile sapere con certezza se acquisterà qualcosa.
  - Il responsabile di marketing di una azienda non può sapere con certezza se un nuovo prodotto avrà successo o meno.
  - Un investitore non sa con certezza se la quotazione di un certo titolo crescerà o meno
  - Lanciando una moneta non si può prevedere se comparirà testa o croce

- Associato al concetto di probabilità possiamo considerare quello di **esperimento aleatorio**
  - Data una prova, uno dei possibili risultati è caratterizzato da una certa probabilità di verificarsi

Frequenza e Probabilità

- La probabilità di un evento è il limite della frequenza relativa con cui si verifica tale evento in una lunghissima serie di prove ripetute in condizioni uguali.

- In un impressionante numero di casi questa frequenza obbedisce a regole che valgono per le **funzioni di probabilità**.
  - E questo è tanto più vero quanto più grande è il campione.

- Quindi possiamo dire che la probabilità che si verifichi l'evento A, $P(A)$, è il limite della frequenza di A per un campione infinito.

- Le probabilità sono numeri che ci dicono quanto spesso si verificano eventi.

Spazio campionario

- Il singolo risultato associato ad una prova è chiamato **evento elementare** (ω)

- L'insieme degli eventi elementari è chiamato **Spazio Campionario** (Ω)

- **Spazio campionario discreto**
  - Insieme finito (lanciare una volta una moneta)
  - Insieme infinito numerabile (lanciare ripetutamente una moneta fino a quando non compare la prima volta testa)

Spazio campionario discreto

- **Spazio campionario continuo**
  - Insieme infinito (durata di vita di un componente elettronico)

Spazio campionario infinito

Eventi

- Se Ω rappresenta lo **spazio campionario** ed E è un evento allora $E \subseteq \Omega$.
- **EVENTO CERTO**: $E = \Omega$.

Evento certo

- **EVENTO IMPOSSIBILE**: L'insieme vuoto, $\varnothing$, si definisce evento impossibile.

Evento impossibile

- **Complemento** A^C
- Dovrem(m!)o assegnare la probabilità ad eventi.
  - Astrazione della frequenza relativa di un evento
- La definizione di P necessita di Ω
  - P assume un valore per ogni sottoinsieme di Ω.

Scenari

Problema diretto

Conosciamo perfettamente la probabilità relativa ad un certo esperimento aleatorio

⇩

Possiamo calcolare la probabilità che si verifichi un evento

Facile
ma serve a poco

Problema inverso

Conosciamo un campione

⇩

Vogliamo calcolare le proprietà della probabilità che caratterizza l'esperimento

Difficile
È ciò che dovremo fare

Scenari

- Consideriamo la possibilità (irrealistica) di effettuare un infinito numero di repliche di uno stesso esperimento (**campione=spazio campionario**).

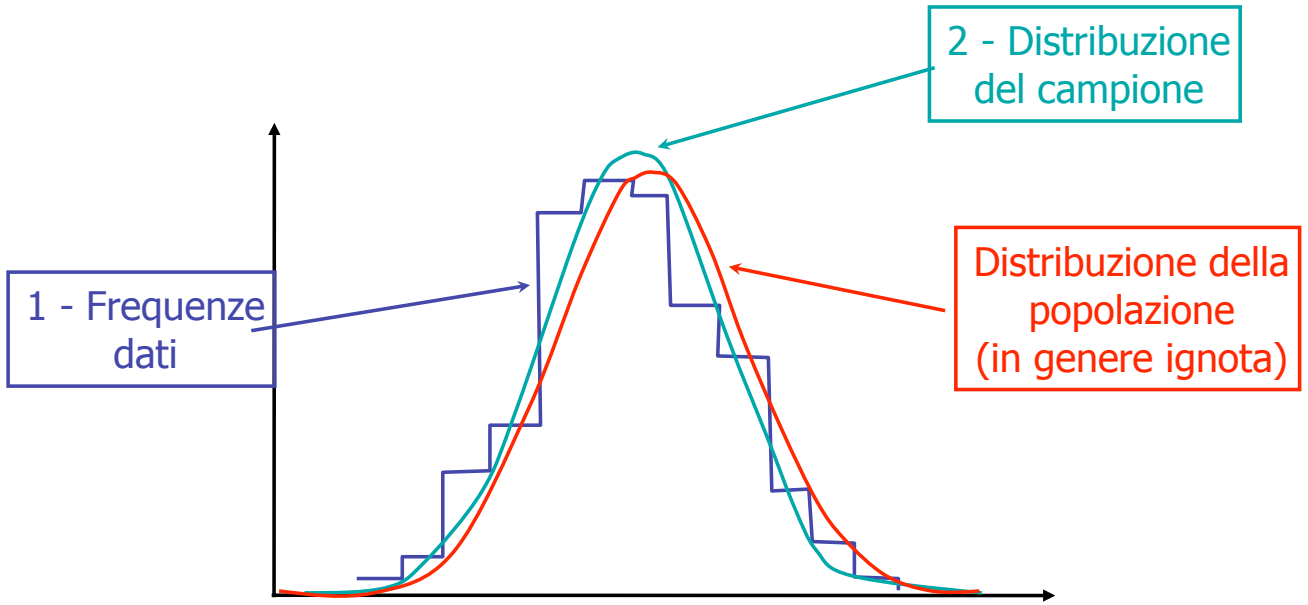
- Questa ipotesi di lavoro ci permetterebbe di arrivare a conoscere la probabilità di ottenere un particolare evento in una singola misura (**non il risultato DELL'ESPERIMENTO!!!**).
  - Conosciamo **la distribuzione che descrive lo spazio campionario**

- Al contrario quando (realisticamente) facciamo un numero finito di repliche possiamo ottenere solo la **DISTRIBUZIONE DEL CAMPIONE**.

- Nel limite di infinite repliche le due distribuzioni coincidono

Scenari

- A partire dai dati sperimentali impareremo a **IDENTIFICARE** la **distribuzione del campione** (curva verde), e questa approssimerà la **distribuzione della popolazione** (curva rossa).



Statistica descrittiva

Campione

- Sia K una proprietà osservata su N elementi estratti da uno spazio campionario.

- Date le N osservazioni $k_1, \ldots, k_N$ possiamo calcolare dei valori che caratterizzano la distribuzione della proprietà K nel campione.

- Tali valori che sono delle costanti sono solitamente detti **parametri**
  - **Media**
  - **Varianza**
  - **Range**
  - **Etc.**

Statistica descrittiva: centralità

Campione

- **Misure (in generale differenti) della tendenza centrale**
- **Media** del campione $\bar{k}$: $\bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i$
- **Mediana** del campione $\tilde{k}$: E' il valore che divide in due una distribuzione, ovvero è il valore che divide il campione in due parti ciascuna contenente la metà dei dati.
  - Se il numero dei dati è pari la mediana è a metà tra i due dati centrali
  - Se il numero dei dati è dispari la mediana è il valore centrale
- **Moda** del campione: il valore che compare con maggior frequenza ("il valore più probabile")

La lettera greca maiuscola Σ è un modo per dire in breve somma tutto

Statistica descrittiva: centralità

- Media e mediana coincidono in situazioni simmetriche
  - Nel caso simmetrico unimodale anche la moda coincide con media e mediana

- Quando le tre misure della tendenza centrale differiscono di solito vale

$$moda < mediana < media$$

- **La media non è un descrittore robusto**: asimmetrie significative o presenza di dati fuori scala hanno una grossa influenza sulla media. **La mediana è robusta.**

Statistica descrittiva: variabilità

MISURE DELLA DISPERSIONE

- Range
- Deviazione dalla media: $d_i \equiv k_i - \mu$.
- La media della deviazione dalla media è nulla (provate a dimostrarlo)
- Media del valore assoluto della deviazione $\alpha = \frac{1}{N-1}\sum |k_i - \bar{k}|$
  - Scomoda da usare
- **Deviazione standard** $\sigma = \sqrt{\frac{1}{N-1}\sum (k_i - \bar{k})^2}$
- **Varianza** $\sigma^2 = \frac{1}{N-1}\sum (k_i - \bar{k})^2$
  - Il denominatore è $N-1$ perché gli N dati del campione sono già stati usati per ottenere la media.
  - Con $N-1$ deviazioni si può determinare quella rimanente dato che le N deviazioni sommano a zero (GRADI DI LIBERTA')

Statistica descrittiva: variabilità

MISURE DELLA DISPERSIONE

- La deviazione standard misura la dispersione attorno alla media, va usata solo quando la media è utilizzata come misura della tendenza

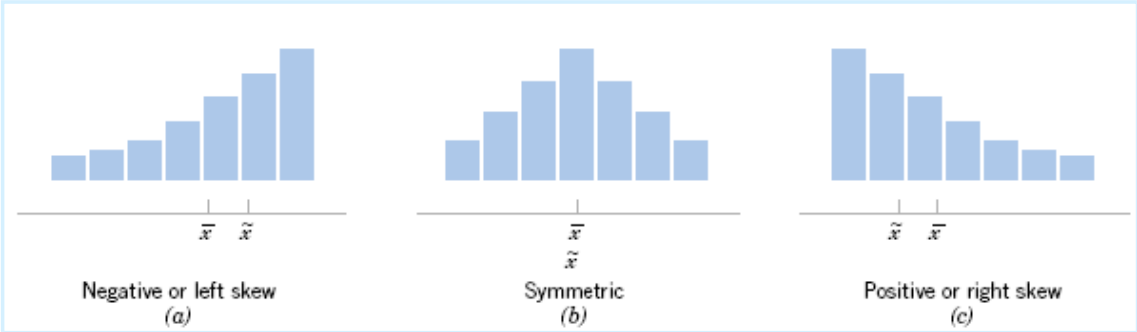
- $\sigma = 0$ implica che non c'è dispersione dei dati.

- In genere $\sigma > 0$

- La deviazione standard è, come la media, un descrittore non robusto.

Statistica descrittiva: variabilità

Forma della distribuzione



- Skewness (coeff. di asimmetria)

$$\frac{\sum\left(\frac{k_i-\bar{k}}{\sigma}\right)^3}{N}$$

>0 Coda a destra
=0 Simmetrica
<0 Coda a sinistra

- Curtosi (misura quanto la distribuzione è appuntita)

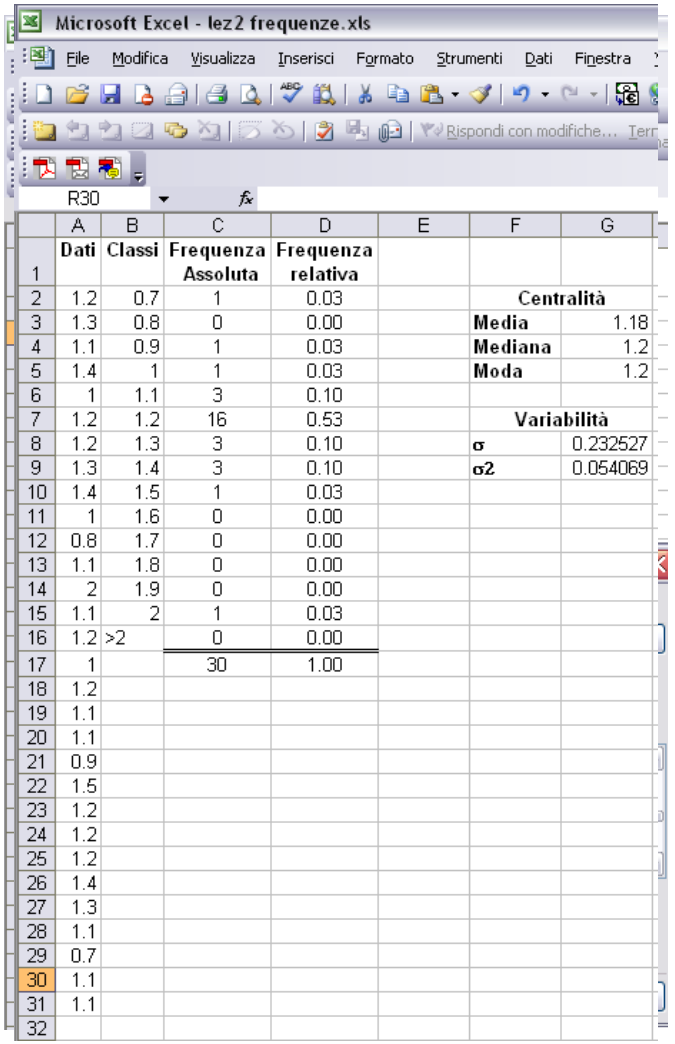
$$\frac{\sum\left(\frac{k_i-\bar{k}}{\sigma}\right)^4}{N}$$

>0 poco
<0 molto

Statistica descrittiva

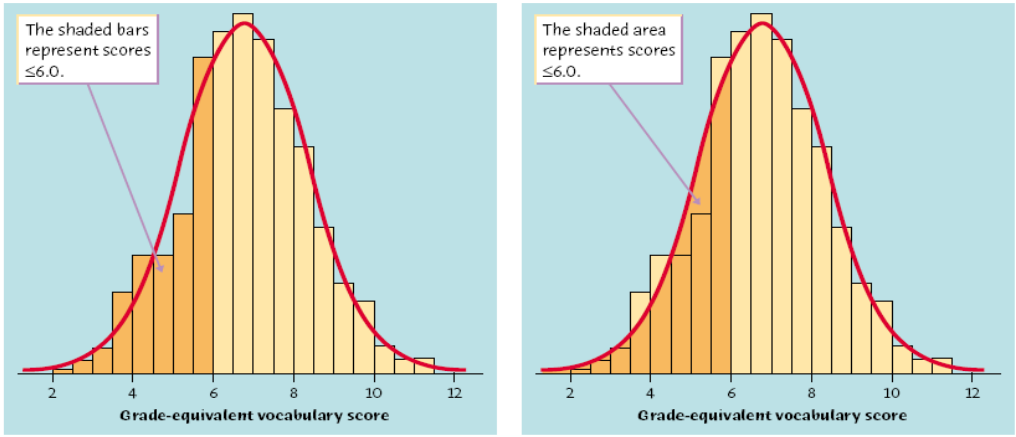
- **Con Excel**
- Scegliere una casella
- Selezionare inserisci dal menu
- Selezionare la funzione media dal menu statistico
  - Media
  - Mediana
  - Moda
- Selezionare i dati
- Premere OK
- Ripetere per altre funzioni
  - DEV.ST
  - Var

- **Inserire un outlier e commentare**



	A	B	C	D	E	F	G
1	Dati	Classi	Frequenza Assoluta	Frequenza relativa			
2	1.2	0.7	1	0.03		Centralità	
3	1.3	0.8	0	0.00		Media	1.18
4	1.1	0.9	1	0.03		Mediana	1.2
5	1.4	1	1	0.03		Moda	1.2
6	1	1.1	3	0.10			
7	1.2	1.2	16	0.53		Variabilità	
8	1.2	1.3	3	0.10		σ	0.232527
9	1.3	1.4	3	0.10		σ2	0.054069
10	1.4	1.5	1	0.03			
11	1	1.6	0	0.00			
12	0.8	1.7	0	0.00			
13	1.1	1.8	0	0.00			
14	2	1.9	0	0.00			
15	1.1	2	1	0.03			
16	1.2	>2	0	0.00			
17	1		30	1.00			
18	1.2						
19	1.1						
20	1.1						
21	0.9						
22	1.5						
23	1.2						
24	1.2						
25	1.2						
26	1.4						
27	1.3						
28	1.1						
29	0.7						
30	1.1						
31	1.1						
32							

Curva di densità



- Il nostro occhio risponde alle aree delle barre in un istogramma
- Abbiamo già visto che le barre rappresentano proporzioni tra le osservazioni.
- La curva disegnata tra le barre è fatta in modo che l'area ad essa sottesa sia unitaria.
  - Questa area rappresenta la porzione 1 ovvero tutte le osservazioni.
  - Aree sotto pezzi di curva rappresentano porzioni delle osservazioni
- La curva è detta curva di densità.

Curva di Densità

- Una **Curva di Densità**
  - E' sempre al di sopra dell'ascissa
  - Sottende un'area unitaria

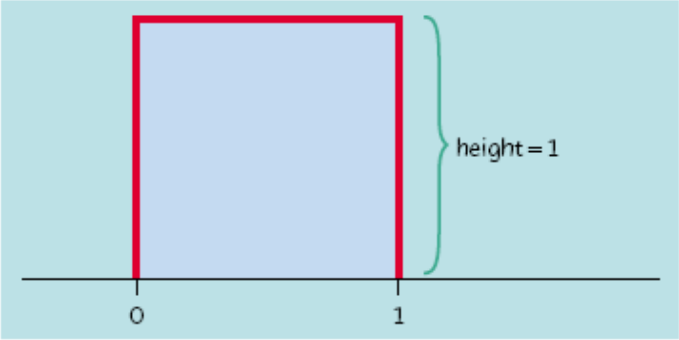
- Una curva di densità descrive la struttura di una **distribuzione**

- L'area sotto la parte di curva che sovrasta un certo intervallo fornisce la porzione di tutte le osservazioni che cadono in quell'intervallo.

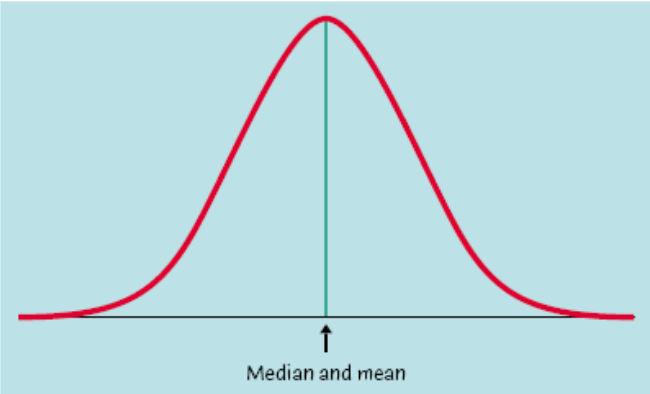
- Una curva di forma simile a quella illustrata nel grafico precedente si dice **Curva Normale**

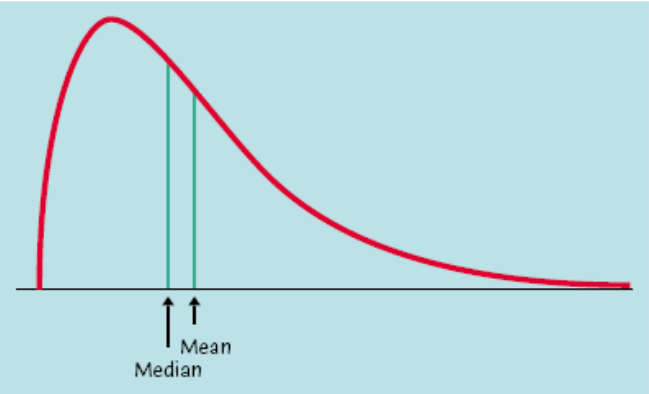
Parametri della distribuzione

- Curva di densità uniforme



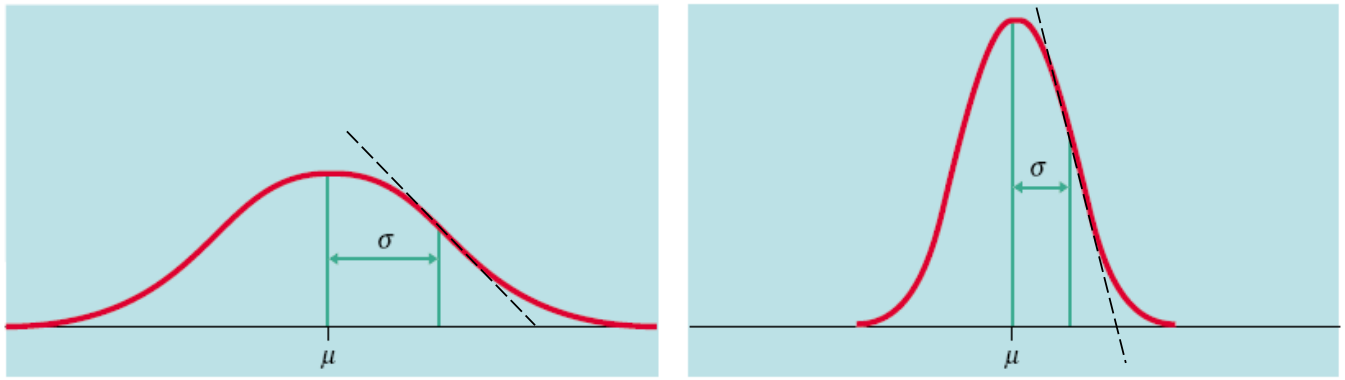
- Anche le curve di densità possono essere caratterizzate da descrittori.





Curva di Densità Normale

- Le **Curve di Densità Normale** giocano un ruolo fondamentale in statistica
  - Sono **simmetriche** ed unimodali
  - Tutte le curve di densità Normali hanno la stessa forma
  - Cambiano per la posizione della media (μ) e la "spanciatezza" (deviazione standard σ).



  - Cambiare media a σ costante vuol dire traslare orizzontalmente la curva.
  - Media e deviazione standard identificano univocamente la Curva di Densità Normale

Curva di Densità Normale

- Le CDN sono importanti per tre ragioni

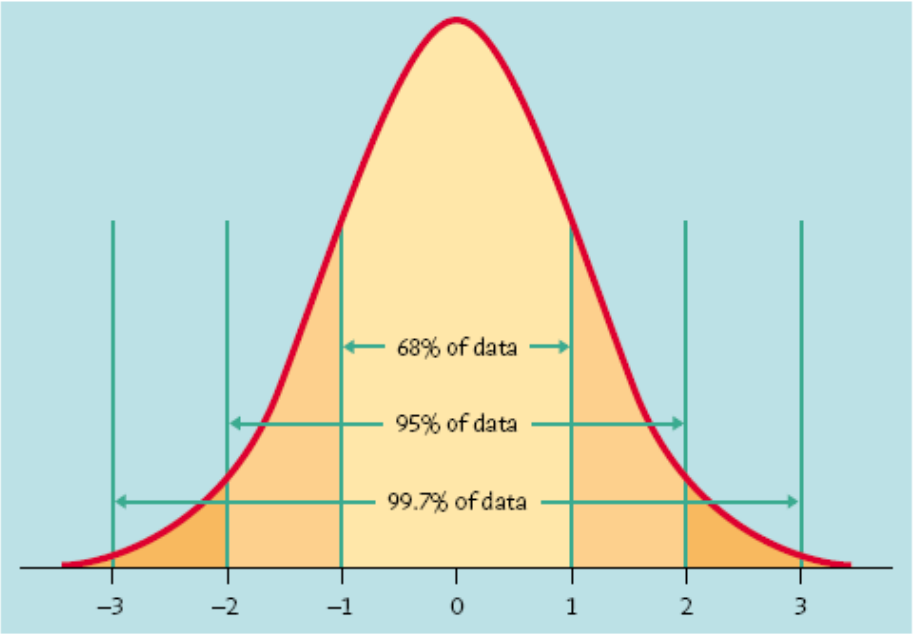
1. Sono buone descrizioni per distribuzioni di popolazioni reali.

2. Sono buone approssimazioni delle distribuzioni dei risultati di esperimenti aleatori

3. Molte procedure di inferenza statistica si basano su CDN.

Curva di densità Normale

- La regola **68-95.5-99.7**
  - Tutte le distribuzioni normali rispettano la
  - Il 68% delle osservazioni cade entro una distanza σ dalla media μ.
  - Il 95.5% delle osservazioni cade entro una distanza 2σ da μ.
  - Il 99.7% delle osservazioni cade entro una distanza 3σ da μ.

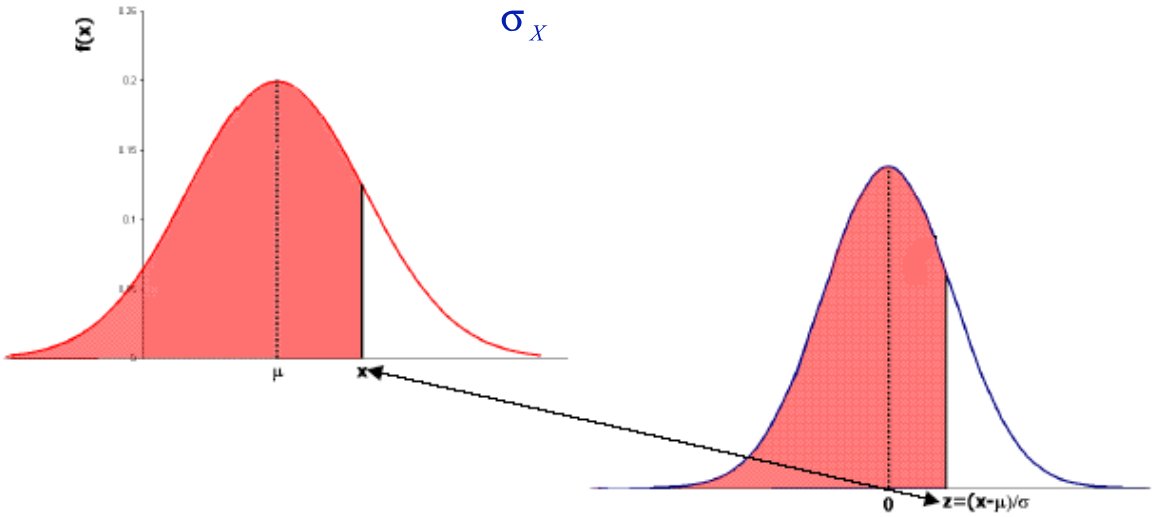


Curva di Densità Normale Standard

- La distribuzione **normale standard** ha media 0 e deviazione standard unitaria

- Si può standardizzare qualunque distribuzione normale: basta sottrarre la media e dividere per la deviazione standard

$$Z = \frac{X - \mu_X}{\sigma_X}$$

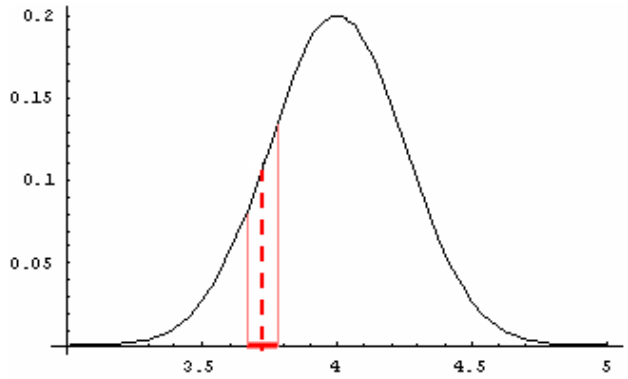


Distribuzioni di Probabilità

- La curva di densità è una **distribuzione** che rappresenta la probabilità di ottenere valori della variabile Y da un insieme di misure.

- L'area sottesa alla curva tra $y\text{-}dy/2$ e $y+dy/2$ dà il numero di eventi attesi in quella regione su un campione di dimensione 100

- Tale area divisa per l'area complessiva sottesa alla curva è la probabilità $P(y)dy$ che una misura dia un valore osservato compreso tra $y\text{-}dy/2$ ed $y+dy/2$.



Distribuzioni di Probabilità

- La funzione di probabilità è definita nel limite di infinite osservazioni.
- La frazione dN di osservazioni della variabile Y che fornisce valori compresi tra y e $y+dy$ è data da: $dN=P(y)dy$
- **Una distribuzione di probabilità è una funzione matematica che, per ogni valore della variabile, fornisce la probabilità che venga osservato quel valore.**
- Il processo di misura non è molto dissimile dal lancio dei dadi. Al lancio di un dado è associata una distribuzione di probabilità per il risultato che è particolarmente semplice: si tratta di una funzione che è diversa da zero solo in sei punti (corrispondenti ai sei "valori" delle facce, siano essi numerici o figurati), ed assume in quei punti lo stesso valore pari a 1/6.

Distribuzioni di Probabilità

- Questo è un esempio di una **distribuzione di probabilità discreta**, cioè diversa da zero solo in un **insieme numerabile** (non necessariamente finito) di punti.

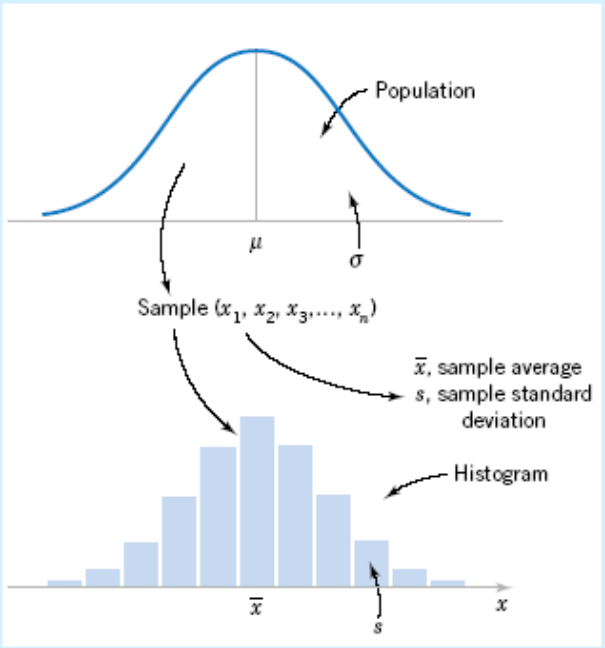
Probabilità discreta

- Esistono poi distribuzioni di probabilità **continue**, per le quali è possibile che si osservino valori compresi in un certo intervallo (eventualmente di ampiezza infinita) di numeri reali.
  - Qui le cose si complicano un poco, dal punto di vista matematico. Infatti, data l'infinità non numerabile dei numeri reali in un qualsiasi intervallo, dobbiamo concludere che non ha senso assegnare una probabilità finita a ciascuno di essi
  - Paradossalmente, ogni risultato, per quanto possibile, deve avere probabilità nulla. L'unica probabilità finita che ha senso definire è quella che il risultato cada in un certo intervallo finito di valori

Probabilità continua

Parametri della distribuzione

- Nella maggior parte dei problemi statistici si lavora su un campione di osservazioni estratto da una popolazione
- Spesso si usa una **Curva di Densità** (cioè una **distribuzione di probabilità)** come **modello** per una popolazione
- Di solito non si può osservare la popolazione, perciò dobbiamo usare il campione per prendere decisioni sulla popolazione.



Parametri della distribuzione

- La media, la mediana, la moda, la deviazione standard etc. sono parametri che caratterizzano l'informazione che stiamo cercando di determinare quando eseguiamo un esperimento.

- La media è collegabile al valore della grandezza che stiamo tentando di determinare (ha anche le sue dimensioni). **Potrebbe essere un buon candidato alla stima del valore "vero".**

- La varianza e la deviazione standard **caratterizzano l'incertezza** associata ai nostri tentativi sperimentali di determinare i valori "veri".
  - Per un fissato numero di osservazioni l'incertezza nella determinazione della media della distribuzione è proporzionale alla deviazione standard di quella distribuzione

Probabilità e parametri

- **Che connessione c'è tra la distribuzione di probabilità della popolazione e il campione sperimentale?**

- Le incertezze sperimentali precludono la possibilità di determinare valori "veri" dei parametri.

1. Dai dati sperimentali si descrive la distribuzione di frequenza del campione e se ne determinano media, varianza etc.

2. Dalla distribuzione del campione si stimano i parametri della distribuzione della popolazione (media, varianza etc.).

3. Dai parametri stimati della distribuzione della popolazione si ottengono i risultati.

Stime

Stima

- Il considerare i dati osservati come un campione della popolazione ci permette di **stimare** la forma e la dispersione della distribuzione della popolazione.

- Di conseguenza possiamo ottenere utili informazioni sulla precisione dei nostri risultati

- Le incertezze nei dati sperimentali sono di due tipi:
  - Qulle derivanti da fluttuazioni nella misura
    - Nell'esempio del viscosimetro le fluttuazioni nella misura dei tempi
  - Quelle associate alla descrizione teorica dei risultati
    - Nell'esempio del viscosimetro quelle dovute alla nostra ipotesi di Newtonianità del liquido.

- Lo studio della distribuzione dei risultati di misure ripetute della stessa quantità conduce alla comprensione delle incertezze di misura, e queste permettono **di stimare l'errore sperimentale**

Concetti importanti

- Frequenza e probabilità
- Spazio campionario
- Eventi
- Identificazione parametrica
- Campione
  - Misure di centralità del campione
  - Misure della variabilità del campione
  - Misure della forma del campione
- Curve di densità
  - La curva di sensità normale
- Distribuzione di probabilità