METODI STATISTICI E COMPUTAZIONALI

Stefania Spagnolo

Dipartimento di Matematica e Fisica, Univ. del Salento

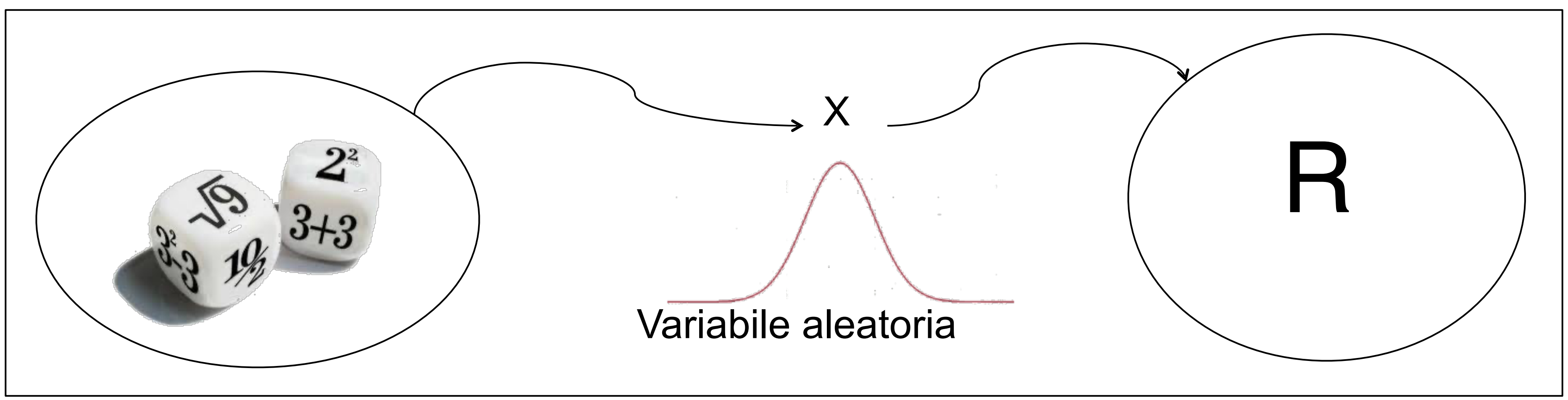
LEZIONE 9

TEORIA DEI CAMPIONI

[Dispense Prof. M. Garetto Univesità di Torino P. 167-216]

TEORIA DEI CAMPIONI

- Un ***insieme S contenente tutti i possibili risultati di un esperimento è detto spazio campione*** e ciascun risultato è detto elemento o punto di S. ***Un evento (A) in S è costituito da 1 o più elementi di S***. L'insieme di tutti gli eventi in S è anche detto campo di Borel
  - Spazio Campione/Spazio degli Eventi (Borel) = spazio che contiene tutti gli elementi/ eventi che caratterizzano un particolare processo fisico.

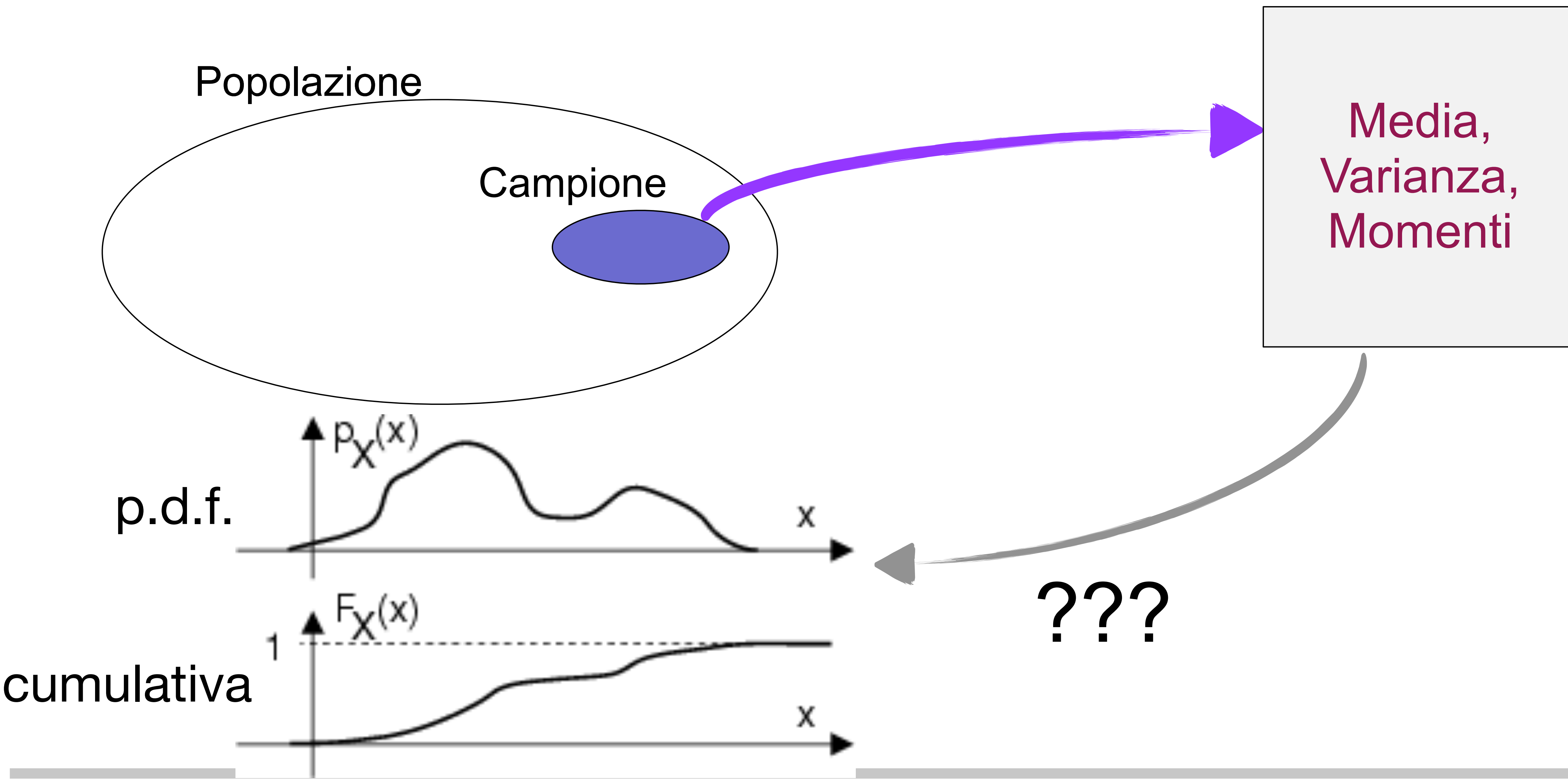


- ***Variabile aleatoria:*** un numero reale associato ad un evento.
- ***Probabilità*** che si verifichi un certo evento: probabilità che una variabile aleatoria assuma certi valori.
- In statistica lo ***Spazio Campione*** è anche indicato con il termine ***Popolazione***

TEORIA DEI CAMPIONI

Inferenza statistica

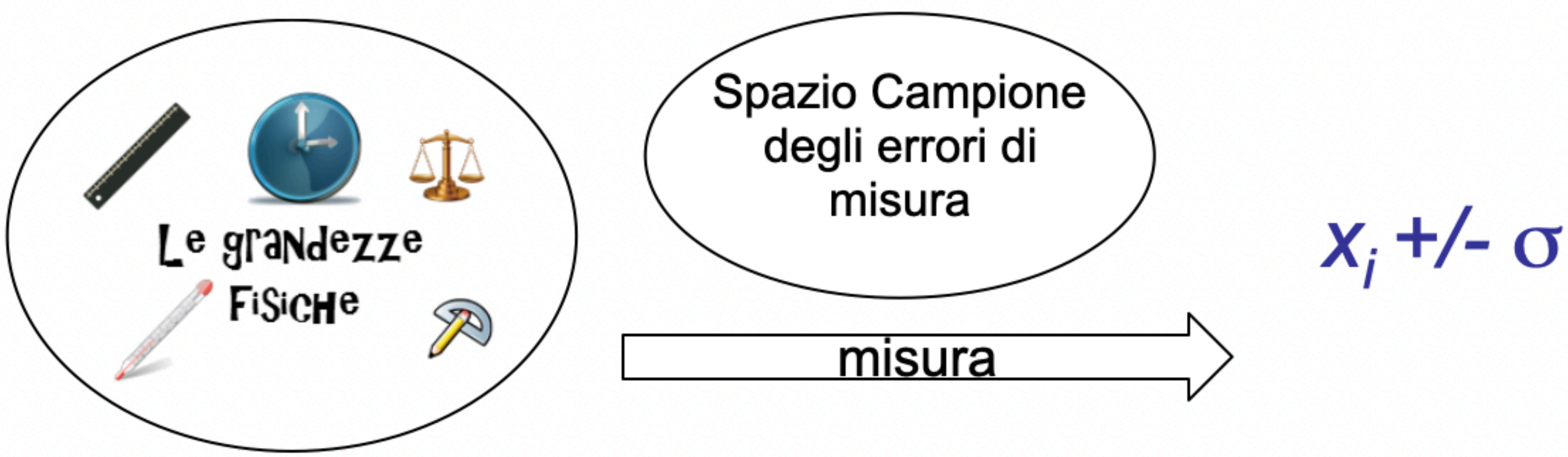
- determinare le caratteristiche di una *popolazione* analizzando solo un sottogruppo di essa (*campione*).



Teoria degli errori di misura

Quando dico che il valore misurato della grandezza fisica è x_i +/- σ con σ errore statistico sto facendo alcune assunzioni.

*Sto assumendo che **lo spazio campione degli errori di misura segue una distribuzione di probabilità nota** e che in un qualche modo **sono in grado di conoscerne la sigma**. Sto inoltre implicitamente assumendo che tale spazio campione abbia **valore di aspettazione nullo**.*



L'assumere che il valore di aspettazione della popolazione "*errori di misura*" associata al processo di misura in studio sia a valore di aspettazione nullo equivale a dire che la misura non sia affetta da "*errori sistematici*".

TEORIA DEI CAMPIONI & TEORIA DEGLI ERRORI DI MISURA

- Le domande a cui occorre sempre rispondere quando si fa una misura sono:

  - Se dispongo di un insieme (campione) di n misure x_i della grandezza fisica μ, quale è la miglior stima ($\hat{\mu}$) di μ?

  - Quanto $\hat{\mu}$ si discosta dal valore vero μ ?

- La ***teoria dei campioni***, alla base dell'inferenza statistica, studia le relazioni esistenti tra i parametri di una popolazione e i parametri di un campione.
  - Si applica, per esempio, per ***stimare la media μ e la varianza σ² =var(X) di una popolazione conoscendo la media <x> e la varianza s² di un campione.***

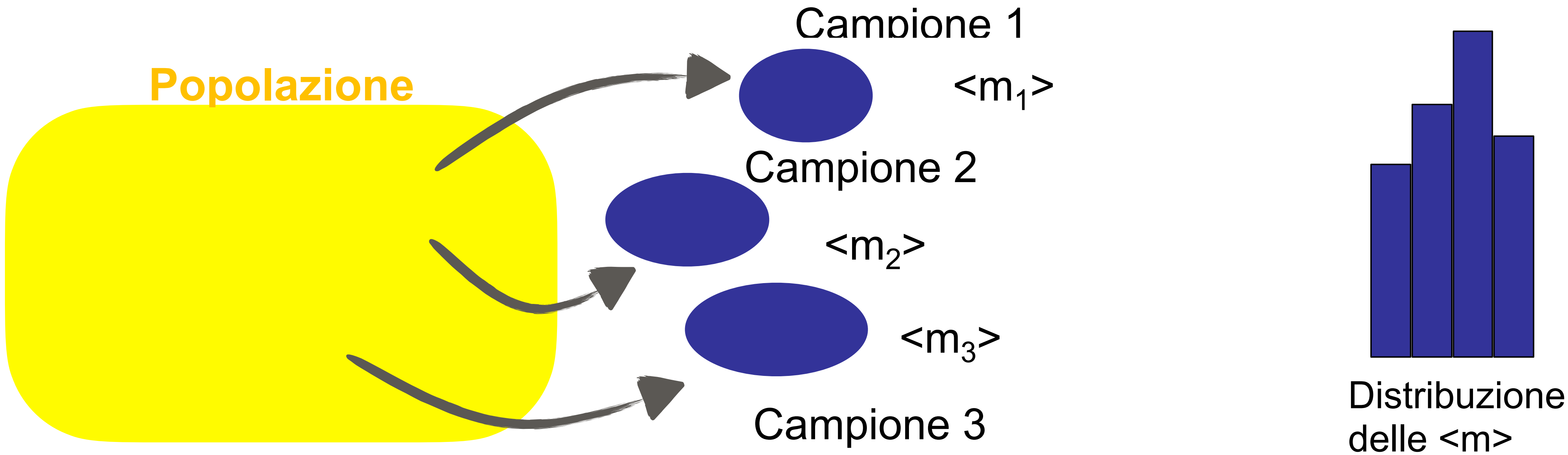
TEORIA DEI CAMPIONI

- Data una popolazione posso estrarre campioni con o senza re-immissione (del campione nella popolazione):
  - ***posso estrarre numerosi*** (infiniti) ***campioni da una*** stessa ***popolazione***,
  - ciascun campione avrà proprie grandezze statistiche (media, varianza, etc) diverse da campione a campione.
  - Possiamo quindi parlare di ***distribuzione delle grandezze campionarie***.

- ***Esempio:*** possiamo studiare la distribuzione di probabilità che segue la variabile aleatoria ***media dei campioni*** (detta anche ***media campionaria***).
  - Questa variabile aleatoria rappresenta la media che un campione di una data popolazione può assumere e fluttuerà in maniera casuale seguendo una qualche distribuzione.

TEORIA DEI CAMPIONI

- la ***distribuzione di campionamento***, la ***distribuzione***, cioè, ***di tutti i valori*** possibili ***assunti da un dato parametro statistico*** (denominato ***"statistica campionaria")*** calcolati da ***campioni delle stesse dimensioni*** estratti dalla stessa popolazione deve essere legata ai parametri della popolazione da cui i campioni sono stati estratti.



TEORIA DEI CAMPIONI

ESEMPIO: il dado

La popolazione del **“numero estratto da un dado”** è finita e costituita da 6 elementi tutti equi-probabili.
La media della popolazione è 3.5.

Un ***campione*** di dimensione 2 di questa popolazione può essere costituito dai numeri 3 e 5.
La media del campione è (3+5)/2=4

La ***distribuzione di campionamento*** in questo caso ***è la distribuzione di tutte le possibili medie ottenibili con due soli elementi della mia popolazione***

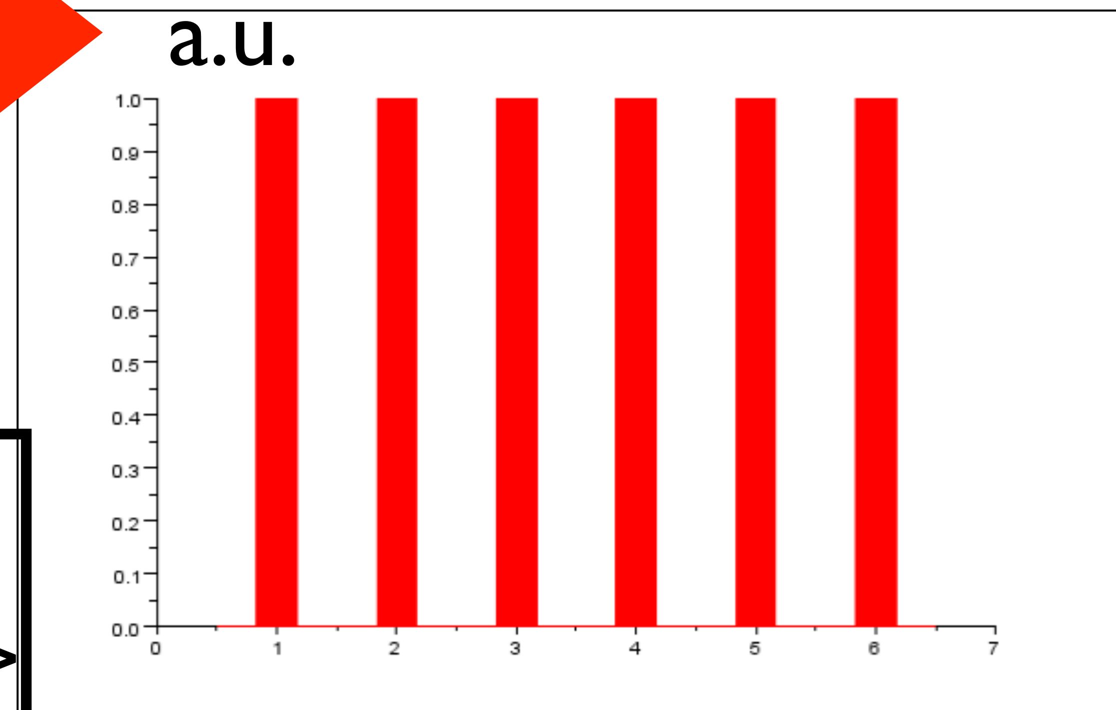
=> Qual è il legame tra la media di questo campione (4) e la media della popolazione (3.5) ?
=> Quale distribuzione di probabilità seguono le grandezze campionarie ?

TEORIA DEI CAMPIONI

La popolazione (1, 2, 3, 4, 5, 6)
(6 valori equiprobabili)

Media (1+2+3+4+5+6)/6. = 3.5
Varianza (1+4+9+16+25+36)/6-(3.5)^2= 2.92

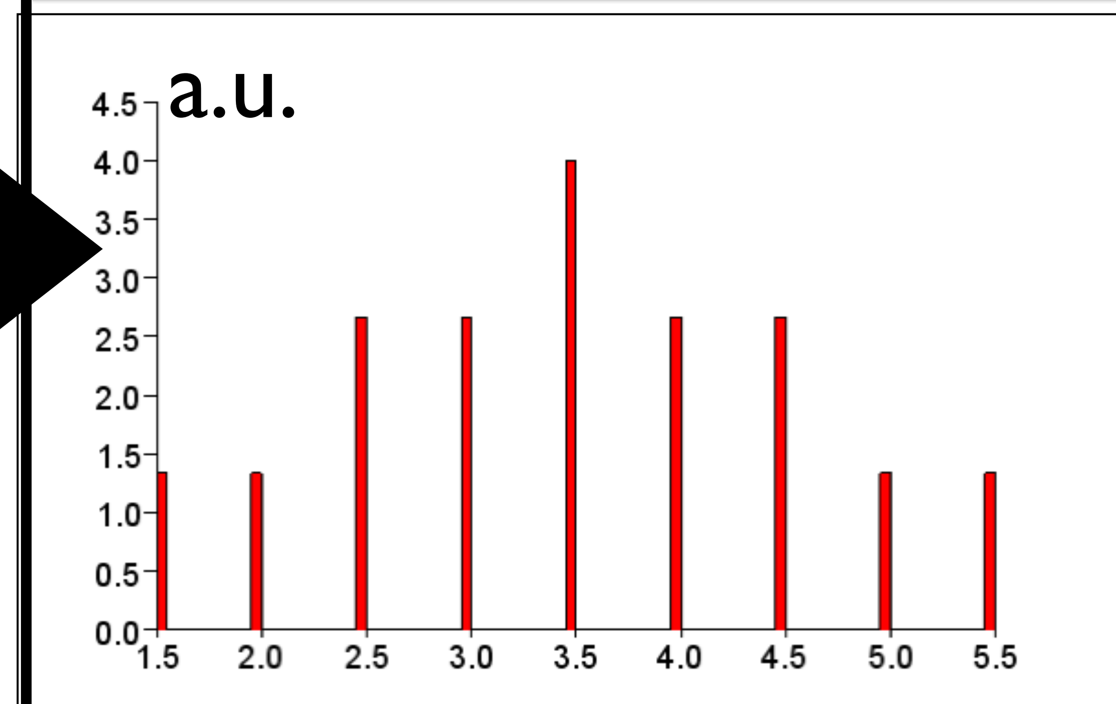
NOTA



Consideriamo ora tutti i possibili ***campioni*** di dimensione 2: (1,2) (1,3) (1,4) ……. (5,6) e calcoliamo la ***media per ciascun campione: <X>***

(1,2) 1.5
(1,3) 2
(1,4) (2,3) 2.5
(1,5) (2,4) 3
(1,6) (2,5) (3,4) 3.5
…

La distribuzione di frequenza della variabile aleatoria media campionaria ha una forma a campana simmetrica con media coincidente con la media della popolazione originale



TEORIA DEI CAMPIONI

- Osserviamo che da un campione non è possibile ricavare informazioni esatte sulla popolazione di provenienza. Occorre identificare grandezze che mi permettano di stimare i parametri della popolazione a partire dai parametri di un campione

- ***Definizioni:***
  - Ogni volta che una ***statistica campionaria*** è ***usata per stimare un parametro***, viene chiamata ***stimatore***. La sua realizzazione nel campione osservato costituisce la stima puntuale del parametro
  - Si chiama stimatore corretto (o non distorto) di un determinato parametro di una popolazione una "statistica campionaria" (variabile aleatoria) tale che il suo valore di aspettazione coincide con il parametro della popolazione stessa.

Uno stimatore *T* del parametro θ è detto *non distorto* se

$$E_\theta(T)=\theta, \text{ per tutti i valori } \theta \in \Theta$$

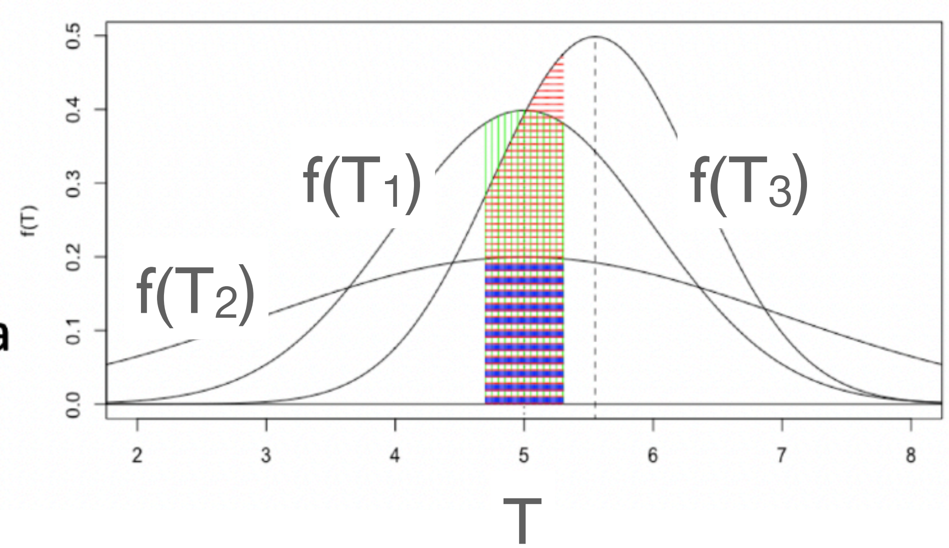
dove Θ è lo spazio parametrico del modello statistico considerato

TEORIA DEI CAMPIONI

- T stimatore (=statistica campionaria usata per stimare un parametro θ di una pdf) è una variabile aleatoria -> ha una sua pdf

Consideriamo la situazione rappresentata nella figura seguente, dove si ipotizzano tre stimatori alternativi per un dato parametro di popolazione, il cui vero valore è 5: T_1 e T_2 sono non distorti, mentreT_3 è distorto. Le aree colorate sottese a ciascuna curva rappresentano le probabilità che il corrispondente stimatore assuma valori in un dato intorno di 5.

T_3 è certamente migliore di T_2 (l'area blu è la più piccola di tutte) e forse è quasi altrettanto buono quanto T_1 (l'area rossa è quasi identica a quella verde).



La probabilita' che T3 cada nell'intervallino 5 +/- ε e' molto piu' alta della probabilita' che T2 cada nello stesso intervallo. Quindi T3 e' piu' efficace nel rappresentare il parametro che ha valore 5 della popolazione rispetto a T2

Tra gli stimatori non distorti alternativi si preferisce quello con minima varianza che viene chiamato stimatore piu' efficiente

TEORIA DEI CAMPIONI

Distribuzione della media campionaria

- Per studiare questa distribuzione si procede nel seguente modo:

- Si estrae un campione di dimensione ***n*** da una popolazione e se ne calcola la media **$<x_1>$**. Si estrae quindi un secondo campione sempre di dimensione ***n*** e sempre dalla stessa popolazione e se ne calcola la media **$<x_2>$**, etc.
- Le medie **$<x_i>$** dei vari campioni saranno diverse tra loro. ***La variabile aleatoria che rappresenta i valori della media dei vari campioni è detta <u>media campionaria</u> <X>***

- ***Teorema:***
  - Se si estraggono campioni casuali di ampiezza n da una popolazione con media μ e varianza σ^2 , la media campionaria cosi ottenuta avrà a sua volta
    - Media => $\mu_{<X>} = \mu$
    - Varianza => $\sigma^2_{<X>} = \sigma^2 / n$

Qualunque sia la distribuzione di probabilità che descrive la popolazione

TEORIA DEI CAMPIONI

Teorema:

- Se si estraggono campioni casuali di ampiezza n da una popolazione con media μ e varianza σ^2 , la media campionaria cosi ottenuta avrà a sua volta
  - Media => $\mu_{<X>} = \mu$
  - Varianza => $\sigma^2_{<X>} = \sigma^2 / n$

Dimostrazione

La dimostrazione si basa sull'osservazione che ognuna delle n misure (elementi estratte dalla popolazione) che costituiscono un campione possono essere interpretate come una variabile aleatoria con lo stesso valore e la stessa varianza della popolazione.

$$\mu_{<X>} = E[<X>] = E[\frac{1}{n}\sum_{i=1,n} x_i] = \frac{1}{n}\sum_{i=1,n} E[x_i] = \frac{1}{n} n\mu = \mu$$

La media campionaria è uno **stimatore corretto** del valore di aspettazione della popolazione in quanto il suo valore di aspettazione coincide con il valore di aspettazione della popolazione.

TEORIA DEI CAMPIONI

Teorema:

- Se si estraggono campioni casuali di ampiezza n da una popolazione con media μ e varianza σ^2 , la media campionaria cosi ottenuta avrà a sua volta
  - Media => $\mu_{<X>} = \mu$
  - Varianza => $\sigma^2_{<X>} = \sigma^2 / n$

<u>Dimostrazione</u>

$\mu_{<X>} = \mu$ Sost. Def <X>

$$\sigma^2_{<X>} = E[(<X> - \mu_{<X>})^2] = E[(<X> - \mu)^2] = E[(\frac{1}{n}\sum_{i=1,n} x_i - \mu)^2]$$

Molt e div μ per n

Porto fuori dall'int $1/n^2$

$$= E[(\frac{1}{n}\sum_{i=1,n}(x_i - \mu))^2] = \frac{1}{n^2}E[(\sum_{i=1,n}(x_i - \mu))^2] =$$

Quadrato di una somma = somma dei quadrati e doppi prod

$$= \frac{1}{n^2}\{\sum_{i=1,n} E[(x_i - \mu)^2] + \cancel{2\sum_{i \neq j} E[(x_i - \mu)(x_j - \mu)]}\} =$$

covarianza

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Ogni singola misura (elemento estratto dalla popolazione) è una variabile aleatoria indipendente. Questo è vero se la misura i-esima non è correlata con nessuna delle altre n-1 misure.

TEORIA DEI CAMPIONI

Con passaggi piu' espliciti

$$\sigma^2(< X >) = E\big[(< X > - \mu)^2\big] = E\Big[\big(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\big)^2\Big] \quad = \frac{\sigma^2}{N}$$

Infatti: $$\big(\frac{1}{N}\sum_{i=1}^{N} x_i - \frac{N\mu}{N}\big)^2 = \big(\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)\big)^2 =$$

$$\frac{1}{N^2}\Big(\sum_{i=1}^{N} (x_i - \mu)\Big)^2 = \frac{1}{N^2}\sum_{i=1}^{N} (x_i - \mu)^2 + \frac{2}{N^2}\sum_{i>j} (x_i - \mu)(x_j - \mu)$$

Ma il valore di aspettazione del primo termine ci da

$$\frac{1}{N^2}E\big[\sum_{i=1}^{N} (x_i - \mu)^2\big] = \frac{1}{N^2}\sum_{i=1}^{N} E\big[(x_i - \mu)^2\big] = \frac{\sigma^2}{N}$$

Il valore di aspettazione del secondo termine ci da

$\frac{2}{N^2}\sum_{i>j} E\Big[(x_i - \mu)(x_j - \mu)\Big] \rightarrow 0$ covarianza nulla, perche' due estrazioni dalla stessa popolazione sono indipendenti.

TEORIA DEI CAMPIONI

Teorema:

- Se si estraggono campioni casuali di ampiezza n da una popolazione con media μ e varianza σ^2 , la media campionaria cosi ottenuta avrà a sua volta
  - Media => $\mu_{<X>} = \mu$
  - Varianza => $\sigma^2_{<X>} = \sigma^2 / n$

- NOTA:
  - $\sigma_{<X>} = \sigma/\sqrt{n}$
    - E’ detta ***errore standard*** è rappresenta la variabilità della media dei campioni di dimensione n estratti da una popolazione.
  - da un campione non è possibile ricavare informazioni esatte sulla popolazione di provenienza
  - In base al ***teorema del limite centrale*** possiamo però affermare che per n grandi la distribuzione della variabile aleatoria media campionaria tende ad una distribuzione gaussiana con media proprio uguale alla media della popolazione e standard deviazione pari all’errore standard.
  - ***La legge dei grandi numeri***, inoltre mi dice che per n che tende ad infinito la media del mio campione tende alla media della popolazione.

TEORIA DEI CAMPIONI

- Riassumendo:
  - In base a quanto detto sino ad ora, la variabile aleatoria media campionaria è uno stimatore corretto della media della popolazione.
  - La varianza della media campionaria (da non confondere con la varianza del campione) mi da informazioni su come la variabile aleatoria media campionaria si distribuisce intorno al suo valore di aspettazione.
  - La sigma della distribuzione della variabile aleatoria media campionaria mi da informazioni sull'incertezza con cui si è identificata la media della popolazione tramite la media del campione. Per questo è anche detta errore standard.

Se dispongo di un insieme (campione) di n misure x_i della grandezza fisica μ, quale è la miglior stima ($\hat{\mu}$) di μ?

TEORIA DEI CAMPIONI: MISURA SINGOLA

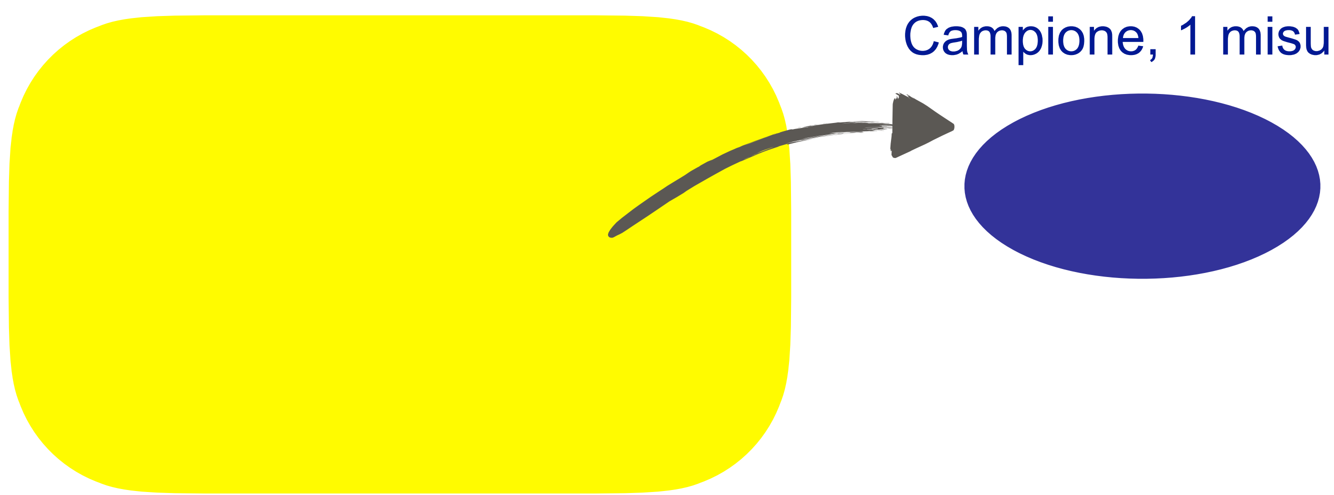
- Esempio: determinare il valore di una grandezza fisica mediante misura.

- ***Misura singola*** (campione costituito da un solo valore).
  - Il valore che ottengo è anche la miglior stima (unica) del valore di aspettazione della popolazione (valore vero). In pratica sto facendo la media di 1 misura.
  - Ricorro a questo metodo quando uno strumento è più preciso che sensibile.
  - La varianza della media campionaria coincide con la varianza della popolazione. Tipicamente in queste misure si assume di conoscere sia il tipo di distribuzione di probabilità che segue la popolazione (gaussiana, uniforme,…) che la varianza di questa (errore massimo). Tipicamente si assume come errore massimo la sensibilità dello strumento.
  - L'errore di misura consiste nella sigma della distribuzione delle medie campionarie, ma come detto, in questo caso, coincide con la sigma assunta per la distribuzione della popolazione.

TEORIA DEI CAMPIONI: MISURA SINGOLA

Popolazione

Campione, 1 misura



σ della popolazione nota

<m>=m stima di μ
Errore su stima

$$\sigma_{<m>} = \frac{\sigma}{\sqrt{n}} == \sigma$$

Esempio: determinare il valore di una grandezza fisica mediante misura.

1. ***Misura singola (campione costituito da un solo valore).***

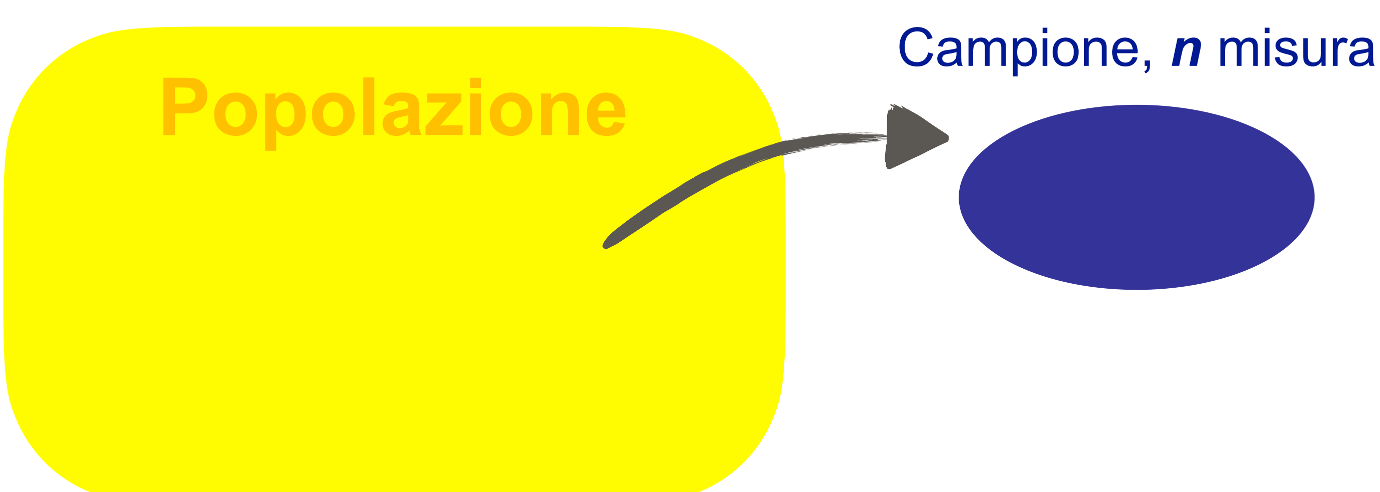
Misura di una lunghezza con un metro. Sensibilità = 0.5 mm = meta' della scansione minima

- Misura di una lunghezza: L=18.2 cm.
- Assumo nota la distribuzione di probabilità della popolazione: uniforme.
- Assumo nota la sigma della popolazione: (semi-ampiezza intervallo minimo)/radice(3) => σ=0.05/1.732=0.03 cm.
- Indico come errore di misura (errore standard) la sigma (nota) della popolazione in quanto in questo caso la sigma della distribuzione delle medie campionarie di un solo elemento coincide con la sigma della popolazione: L=18.20±0.03 cm

TEORIA DEI CAMPIONI: MISURA MULTIPLA

- Esempio: determinare il valore di una grandezza fisica mediante misura.

- ***Misura multipla*** (campione costituito da ***n*** valori).
  - E' conveniente ricorrere a questa procedura quando uno *strumento* è *poco preciso ma molto sensibile*. Questo succede in tutte quelle situazioni in cui ripetendo la misura si ottengono valori diversi. In questo caso si assume che le incertezze casuali nella procedura di misura siano dominanti rispetto alla sensibilità dello strumento.
  - Tipicamente si assume come errore statistico la **precisione** (non sensibilità) *[che determina la σ della popolazione]* dello strumento.



σ della popolazione nota

<m>=m stima di μ

Errore su stima

$$\sigma_{<m>} = \frac{\sigma}{\sqrt{n}}$$

TEORIA DEI CAMPIONI: MISURA MULTIPLA

Esempio: determinare il valore di una grandezza fisica mediante misura.

2. Misura multipla (campione costituito da n valori).

Misura del tempo impiegato da un pendolo a compiere N oscillazioni.

- la misura viene ripetuta 3 volte: T_1=8.2 s, T_2=8.4 s, T_3=8.1 s.
- Assumo nota la distribuzione di probabilità della popolazione: gaussiana.
- Assumo nota la sigma della popolazione: σ=0.2 s.
- Indico come migliore stima del tempo impiegato dal pendolo la media delle tre misure: $\langle T\rangle=(8.2+8.4+8.1)/3=8.2$
- Indico come errore di misura (errore standard) la sigma (nota) della popolazione diviso radice di 3: $\sigma_{\langle m\rangle}=0.2/\sqrt{3}=0.1$; $T=8.2\pm0.1$

NOTA: Questa situazione prevede la conoscenza della varianza "vera" dalla popolazione, cosa non sempre possibile.
Come stimare la varianza della popolazione?

TEORIA DEI CAMPIONI: MISURA MULTIPLA

Come stimare la varianza della popolazione?

- Partiamo dalla ***varianza del campione***

$$S_n^2 = \sum_{i=1}^{n} \frac{(X_i - \langle X \rangle)^2}{n}$$

- In realtà si dimostra facilmente che la varianza del campione non è uno stimatore corretto, mentre lo stimatore corretto della varianza della popolazione è

varianza campionaria

$$S_{n-1}^2 = \sum_{i=1}^{n} \frac{(X_i - \langle X \rangle)^2}{n-1}$$

E' una stima della varianza della popolazione migliore rispetto a S^2 perché non è distorta

TEORIA DEI CAMPIONI: MISURA MULTIPLA

ATTENZIONE: talvolta la varianza campionarie è indicata semplicemente con S^2 invece che S^2_{n-1}

zione?

- Partiamo dalla ***varianza del campione***

$$S^2_n = \sum_{i=1}^{n} \frac{(X_i - \langle X \rangle)^2}{n}$$

- In realtà si dimostra facilmente che la varianza del campione non è uno stimatore corretto, mentre lo stimatore corretto della varianza della popolazione è

varianza campionaria

$$S^2_{n-1} = \sum_{i=1}^{n} \frac{(X_i - \langle X \rangle)^2}{n-1}$$

E' una stima della varianza della popolazione migliore rispetto a S^2 perché non è distorta

TEORIA DEI CAMPIONI: MISURA MULTIPLA

- S^2 non è uno stimatore corretto infatti il suo valore di aspettazione ***non*** coincide con il parametro varianza della popolazione

$$E(S_n^2) = E\left(\sum_{i=1}^{n} \frac{(X_i - \mu + \mu - \overline{X})^2}{n}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} [(X_i - \mu) - (\overline{X} - \mu)]^2\right) =$$

$$= E\left(\frac{1}{n} \left[\sum_{i=1}^{n} (X_i - \mu)^2 - 2\,(\overline{X} - \mu)\left(\sum_{i=1}^{n} X_i - n\mu\right) + n(\overline{X} - \mu)^2\right]\right) =$$

$$= E\left(\frac{1}{n} \left[\sum_{i=1}^{n} (X_i - \mu)^2 - 2\,(\overline{X} - \mu)(n\overline{X} - n\mu) + n(\overline{X} - \mu)^2\right]\right) =$$

Si fattorizza n, e il termine diventa (uguale al terzo)

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2\right) - E((\overline{X} - \mu)^2) = \sigma^2 - \sigma^2_{\overline{X}} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

dal punto di vista pratico se il mio campione è sufficientemente grande (n>30) posso usare S^2 altrimenti è necessario usare S_{n-1}^2.

$$S_{n-1}^2 = \sum_{i=1}^{n} \frac{(X_i - \langle X \rangle)^2}{n-1}$$

Stimatore corretto della varianza della popolazione

$S_{n-1}^2 = \dfrac{n}{n-1} S_n^2$ ha valore di aspettazione σ^2

TEORIA DEI CAMPIONI: MISURA MULTIPLA

Esempio: determinare il valore di una grandezza fisica mediante misura.

2. Misura multipla (campione costituito da n valori).

Misura di una peso con una bilancia di precisione.

- Eseguo 4 misure e ne faccio la media: 15.7, 15.8, 15.8, 15.6

$$<p> = \frac{\sum p_i}{n} = 15.725 \text{ gr}$$

- Stimo la sigma della popolazione a partire dal campione:

$$s = \sqrt{\frac{\sum (p_i - <p>)^2}{n-1}} = 0.096 \text{ gr}$$

- Determino la sigma della distribuzione delle medie campionarie a partire dalla stima della sigma della popolazione :

$$\sigma_{<p>} = \frac{s}{\sqrt{n}} = 0.05 \text{ gr}$$

- Il risultato della misura è:

$$p = 15.72 \pm 0.05 \text{ gr}$$

TEORIA DEI CAMPIONI

Probabilità della media

Come faccio a valutare con che probabilità, dato un campione di dimensione nota *n*, la sua media si possa discostare dalla media della popolazione se la varianza della popolazione è nota?

E' utile introdurre la variabile aleatoria **media campionaria standardizzata**

$$Z = \frac{\langle X \rangle - \mu}{\frac{\sigma}{\sqrt{n}}}$$

σ nota !!

Si dimostra facilmente che essa ha valore di aspettazione nullo e varianza unitaria.

$$E\left[\frac{\langle X \rangle - \mu}{\sigma/\sqrt{n}}\right] = \frac{1}{\sigma/\sqrt{n}} E\left[\langle X \rangle - \mu\right] = \frac{1}{\sigma/\sqrt{n}}\left(E\left[\langle X \rangle\right] - \mu\right) = \frac{1}{\sigma/\sqrt{n}}(\mu - \mu) = 0$$

$$E\left[\left(\frac{\langle X \rangle - \mu}{\sigma/\sqrt{n}}\right)^2\right] = E\left[\frac{(\langle X \rangle - \mu)^2}{\sigma^2/n}\right] = \frac{E\left[(\langle X \rangle - \mu)^2\right]}{\sigma^2/n} = \frac{\sigma^2/n}{\sigma^2/n} = 1$$

TEORIA DEI CAMPIONI

La **densità di probabilità della variabile Z sarà normale** (media 0, varianza 1) **se la popolazione di partenza segue una distribuzione di probabilità normale.**

Nel caso in cui la popolazione iniziale non segua una distribuzione normale la Z per *n* grandi tenderà a una distribuzione normale.
Questo è una conseguenza del teorema del Limite Centrale.

n>=30

Z mi permette di rispondere alla domanda:

Come faccio a valutare con che probabilità, dato un campione di dimensione nota *n*, la sua media si possa discostare dalla media della popolazione se la varianza della popolazione è nota?

Il teorema del limite centrale (TLC) afferma che la somma (o la media) di un grande numero di variabili aleatorie indipendenti e dotate della stessa distribuzione è approssimativamente normale, indipendentemente dalla distribuzione soggiacente.

TEORIA DEI CAMPIONI

Vediamo "come fare una stima".

Esempio

Ipotizziamo che la variabile aleatoria X rappresenti una certa popolazione con media $\mu=5$ e varianza $\sigma^2=25$. Supponendo di estrarre un campione di 100 elementi quant'è la probabilità che la sua media sia maggiore di 5.4?

In base a quanto detto la variabile media campionaria essendo n grande si distribuisce in maniera gaussiana con media $\mu_{<X>}=\mu=5$ e varianza $\sigma^2_{<X>}=\sigma^2/100=0.25$

Per valutare, quindi, con che probabilità otterrei un campione con media maggiore di 5.4 devo valutare la frazione d'aria della distribuzione gaussiana con media 5 e varianza 0.25 tra 5.4 e infinito.

$$P(\langle X \rangle > 5.4) = \int_{5.4}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_{<X>})^2}{2\sigma^2}}$$

Più semplicemente posso ricorrere alla variabile media campionaria standardizzata

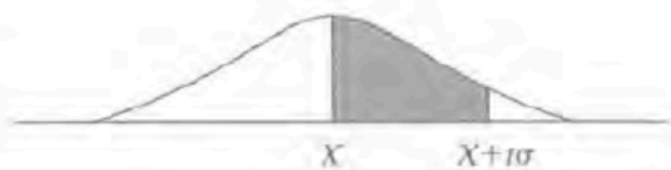
$$Z = \frac{5.4-5}{0.5} = 0.8$$

Dipartimento di Matematica e Fisica "Ennio De Giorgi"
INFN

Ho quindi l'identità: $P(\langle X \rangle > 5.4) = P(Z > 0.8)$

Posso quindi calcolarmi la probabilità richiesta da una qualunque tabella di percentili di una distribuzione normale

Table B. The percentage probability, $Q(t) = \int_X^{X+t\sigma} G_{X,\sigma}(x)dx$, as a function of t.



t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.40	0.80	1.20	1.60	1.99	2.39	2.79	3.19	3.59
0.1	3.98	4.38	4.78	5.17	5.57	5.96	6.36	6.75	7.14	7.53
0.2	7.93	8.32	8.71	9.10	9.48	9.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.88	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.89	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19

P(Z>0.8)=1-P(Z<0.8)=
1-(0.5+0.288)=0.212

TEORIA DEI CAMPIONI

Probabilità della media

σ NON nota !!

Con che probabilità, dato un campione di dimensione nota *n*, la sua media si discosta dalla media della popolazione ***se la varianza della popolazione non e' nota*** ?

Si dimostra che ***se la popolazione segue una distribuzione normale*** allora la variabile

$$T = \frac{\langle X \rangle - \mu}{\frac{S}{\sqrt{n}}}$$

S = *varianza campionaria*

È una variabile aleatoria che segue la distribuzione di probabilità detta: t di Student.

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-(\frac{\nu+1}{2})}$$

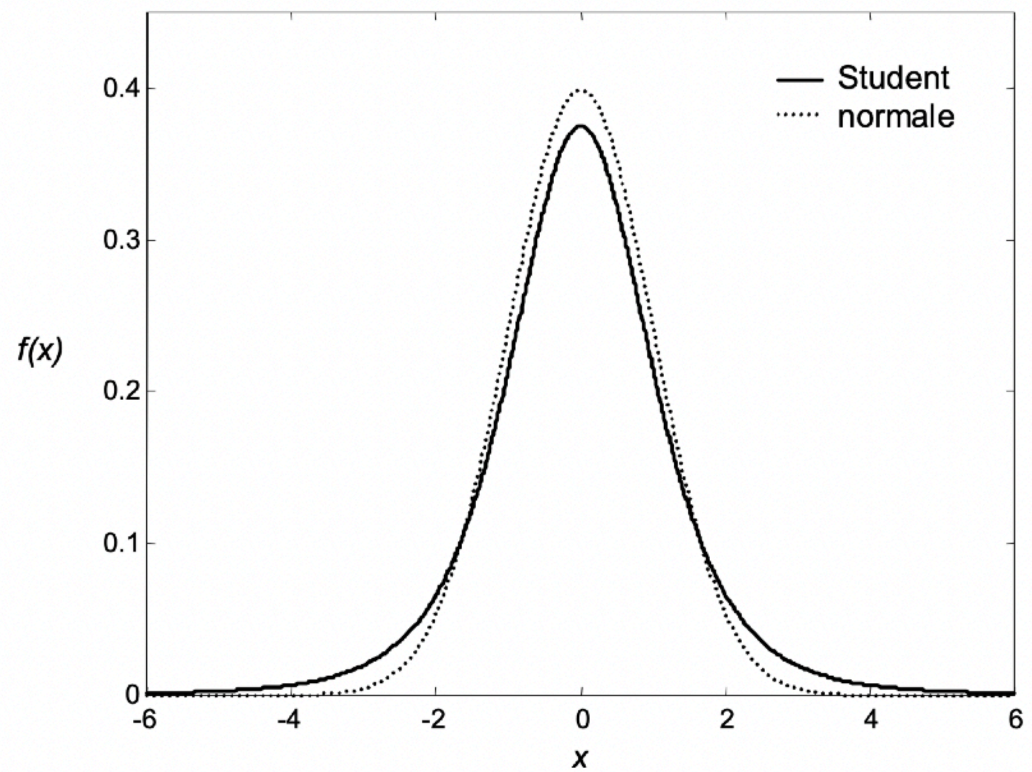
Dove la funzioni gamma sono definite come

$$\Gamma(\nu) = \int_0^\infty e^{-x}\, x^{\nu-1}\, dx$$

E ν (unico parametro) è detto numero di gradi di libertà e vale (nel nostra caso)

$$\boxed{\nu = n - 1}$$

TEORIA DEI CAMPIONI



ν=4

La distribuzione t di Student è simmetrica (come la gaussiana), ha media 0 ma varianza maggiore di 1 ($\nu/(\nu-2)$).

Come prevedibile (grazie anche al teorema del limite centrale) la t di Student tende alla distribuzione normale al tendere ad infinito del numero di gradi di libertà.

Per applicazioni pratiche, per $\nu>30$, si può approssimare la t di Student con una distribuzione normale.

TEORIA DEI CAMPIONI

Probabilità della varianza
quanto la stima dell'errore si discosta dal suo valore vero

Sino ad ora ci siamo limitati a valutare la distribuzione della media campionaria e a stimare la media della popolazione rispondendo alla domanda:
Come faccio a valutare con che probabilità, dato un campione di dimensione nota *n*, la sua media si possa discostare dalla media della popolazione?

In alcuni casi, però, possiamo trovarci di fronte alla necessità di fare delle stime della varianza della popolazione e quindi di dover rispondere alla domanda:
Come faccio a valutare con che probabilità, dato un campione di dimensione nota *n*, la sua varianza S^2 si possa discostare dalla varianza della popolazione?

Per rispondere a questa domanda mi devo chiedere che distribuzione di probabilità segue la variabile aleatoria **varianza campionaria.**

TEORIA DEI CAMPIONI

Probabilità della varianza

- La varianza di un campione altro non è che la somma di scarti quadratici per cui è definita positiva. Ne consegue che la sua distribuzione di probabilità non può essere simmetrica dato che è definita per X>0
- Ogni scarto, essendo ottenuto come differenza tra una variabile aleatoria e un valore di riferimento è esso stesso una variabile aleatoria
  - $s_i = (x_i - < X >)$
- nell'ipotesi che il valore ottenuto all'esperimento i-esimo non dipenda dall'esito dell'esperimento (i-1)-esimo, gli scarti sono variabili aleatorie indipendenti
  - Varianza campionaria = somma dei quadrati di variabili aleatorie indipendenti
- Si dimostra che ***se la popolazione è normale e da essa si estraggono campioni di ampiezza n***, se indico con S^2 la varianza campionaria allora la variabile

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

σ nota !!

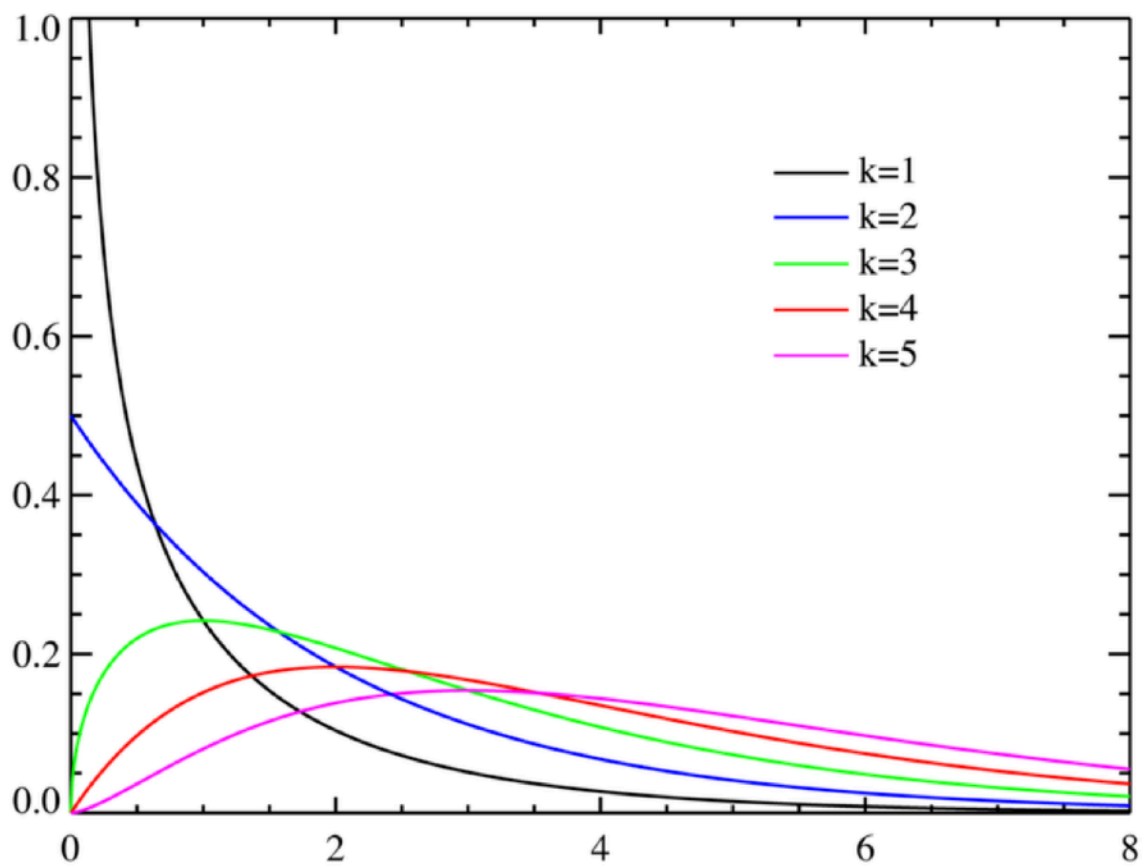
- segue una distribuzione di probabilita' di tipo χ^2 con n-1 gradi di libertà

TEORIA DEI CAMPIONI

Si può dimostrare che la distribuzione di probabilità che segue la variabile χ^2 ha la forma:

$$p_\nu(\chi^2) = \frac{1}{2^{\frac{\nu}{2}} \cdot (\frac{\nu}{2} - 1)!} (\chi^2)^{\frac{\nu}{2} - 1} e^{-\frac{\chi^2}{2}}$$

Dove ν è il numero di gradi di libertà (n-1 nel nostro caso).
La media della distribuzione è $\mu = \nu$ e la varianza è pari a 2ν.



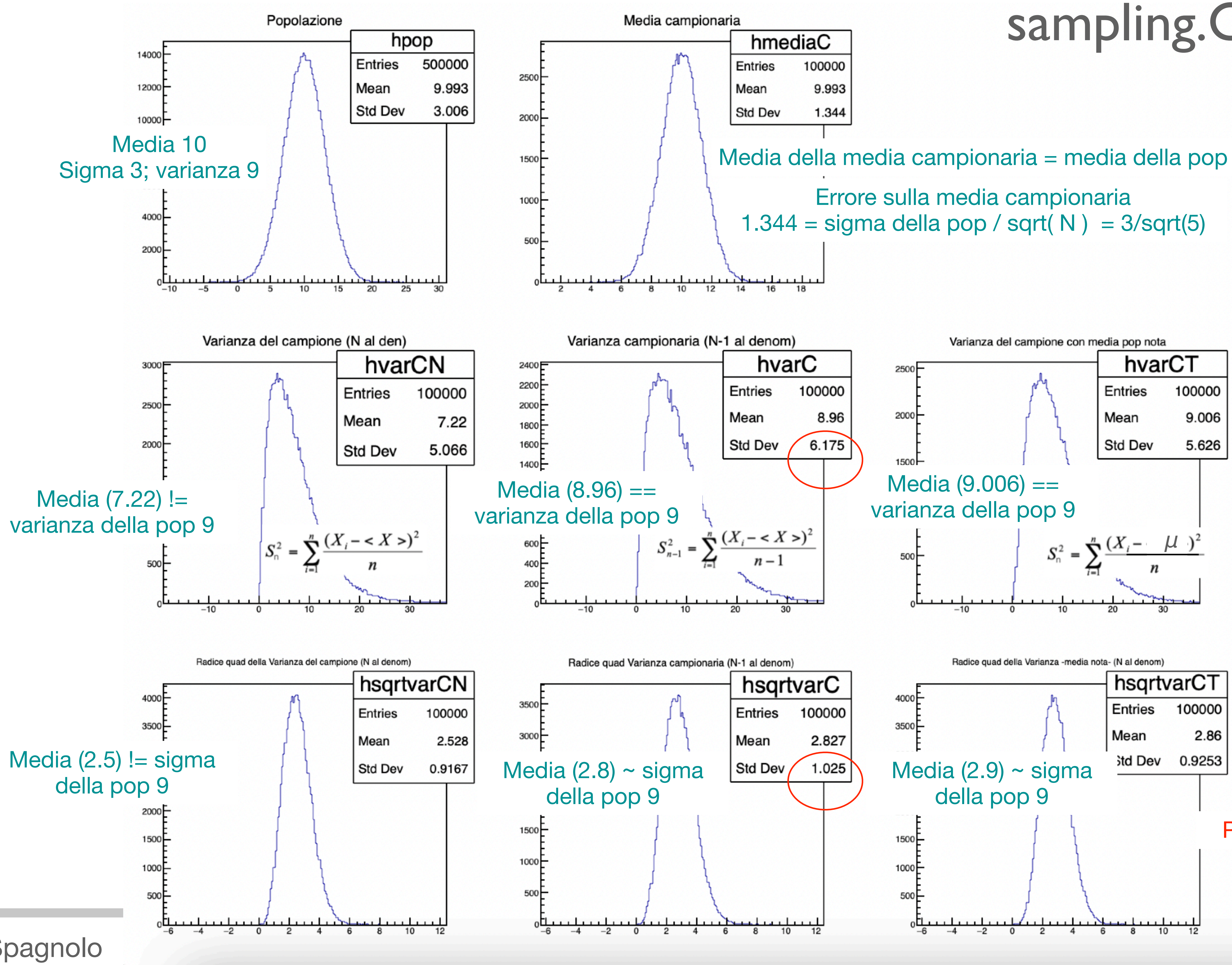
Anche la distribuzione del χ^2 tende ad una distribuzione normale se il numero di gradi di libertà tende ad infinito.

ESERCIZIO

- Popolazione: Gaussiana(10., 3.)
- Estrarre campioni di 5
  - Stimare e fare distribuzioni (100000 campioni) di
    - Media campionaria
    - Varianza (e sua radice quadrata) del campione
    - Varianza (e sua radice quadrata) campionaria
    - Varianza (e sua radice quadrata) assumendo di conoscere la media della popolazione

ESERCIZIO

sampling.C



?

Possiamo stimarli ?

Teoria dei campioni

Errore sulla varianza campionaria

Allora chi^2 con v=n-1 gradi di libertà ha varianza 2v, quindi

Siccome $\chi^2 = \frac{(n-1)S_{n-1}^2}{\sigma^2}$ è distribuita come una variabile χ^2 con n-1 ndf

S_{n-1}^2 ha varianza $var(S_{n-1}^2) = \frac{\sigma^4}{(n-1)^2} var(\chi^2) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{(n-1)}$

Quindi una stima dell'errore su $S_{n-1}^2 \rightarrow \sqrt{\frac{2}{n-1}}\sigma^2 \simeq \sqrt{\frac{2}{n-1}} S_{n-1}^2$

La radice quadrata della varianza della popolazione è stimata da $\sqrt{S_{n-1}^2}$

l'errore su questa stima e' (errore su una funzione di una variabile aleatoria):

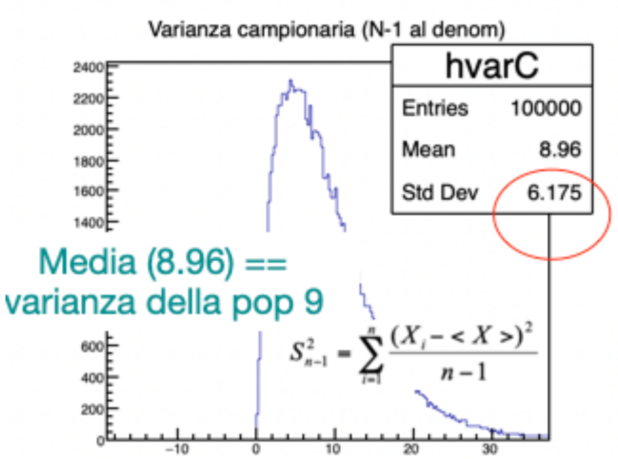
$$err(\sqrt{S_{n-1}^2}) = \frac{1}{2\sqrt{S_{n-1}^2}} err(S_{n-1}^2) = \frac{1}{2\sqrt{S_{n-1}^2}} \sqrt{\frac{2}{n-1}}\sigma^2 \simeq \frac{\sqrt{2}}{2}\sqrt{\frac{S_{n-1}^2}{n-1}}$$

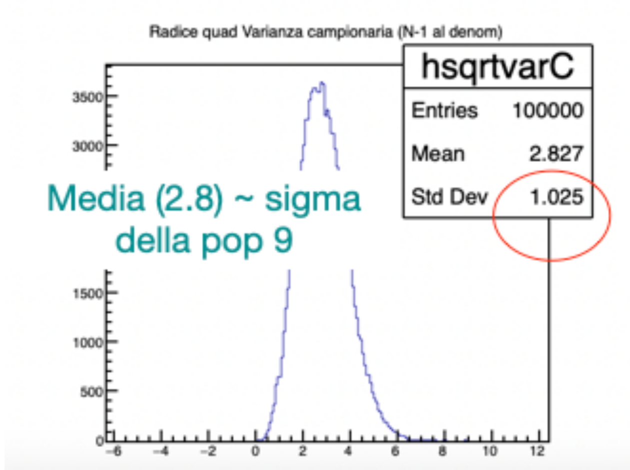
ESERCIZIO

sampling.C

- Popolazione: Gaussiana(10., 3.)
- Estrarre campioni di 5
  - Stimare e fare distribuzioni (100000 campioni) di
    - Media campionaria
    - Varianza (e sua radice quadrata) del campione

dia della






```
root [0] .x sampling.C
========
Popolazione con media      10
                varianza    9
                devStandard 3
========
Valore d'aspett. media campionaria: 10 da MC = 9.99264
Errore sulla media campionaria: 1.34164 da MC = 1.34368
========
Varianza del campione             [N]   da MC: 7.22012 +/- 5.06595
Varianza campionaria              [N-1] da MC: 8.95998 +/- 6.17491
Varianza del camp. media-pop nota [N]   da MC: 9.00591 +/- 5.62585
========
sqrt-Varianza del campione             [N]   da MC: 2.52838 +/- 0.916674
sqrt-Varianza campionaria              [N-1] da MC: 2.82682 +/- 1.02487
sqrt-Varianza del camp. media-pop nota [N]   da MC: 2.86047 +/- 0.925341
========
Errore sulla      varianza campionaria da Maximum Likelihood : 6.33566
Errore sulla sqrt-varianza campionaria da Maximum Likelihood : 0.999431
========
```

RIASSUMENDO

- Popolazione con pdf generica e parametri: α, β, … tra questi i momenti della distribuzione μ, σ^2
- Campione, dimensione n
  - Statistiche campionarie -> tra cui media campionaria <X>, varianza del campione, varianza campionaria S^2
  - Distribuzioni campionarie (distribuzioni delle statistiche campionarie); tra questi distribuzioni della media campionaria, della varianza campionaria, ecc
  - Una statistiche campionaria che stimi un parametro della popolazione si chiama stimatore
    - Uno stimatore e' corretto se il suo valore di aspettazione coincide con il parametro della popolazione stimato
- Teorema E[<X>] = μ, E[(<X>-μ)^2] = σ^2/n per qualunque pdf della popolazione
  - => il valore di aspettazione della media campionaria coincide con la media della popolazione μ, la varianza della media campionaria e' quale a σ^2/n, la media campionaria stima μ con errore pari a $\sigma/\sqrt{n} \simeq S/\sqrt{n}$
    - La media campionaria e' uno stimatore corretto
  - => Ossia la distribuzione campionaria della media campionaria ha media pari a μ e varianza pari a σ^2/n
- Dal teorema del limite centrale se n -> inf, la distribuzione campionaria della media campionaria e' gaussiana
- La varianza del campione ha valore di aspettazione pari a (n-1)σ^2/n => la varianza campionaria (non la varianza del campione) e' uno stimatore della varianza della popolazione non distorto
- Z = (<X> - μ)/σ/sqrt(n) => per n grande Z e' gaussiana normale (limite cen.) -> test di ipotesi con gaussiana
- T = (<X> - μ)/S/sqrt(n) => se X e' gaussiana -> T e' t-Student con ν=n-1 > faccio test di ipotesi con la t-Student
- S^2(n-1)/σ^2 -> segue una distribuzione χ^2 con ν=n-1 gradi di libertà => S^2 stima σ^2 con un errore che e' $\sigma^2\sqrt{2/(n-1)} \simeq S^2\sqrt{2/(n-1)}$

STIMA DI UN PARAMETRO DI UNA DISTRIBUZIONE DI PROBABILITÀ

Caso generale

Stima di un parametro della PDF della popolazione

- ***Stima di un parametro della distribuzione di probabilità della popolazione***
  - La trattazione sino ad ora fatta permette di stimare il valore di aspettazione di una popolazione a partire da un campione.
  - Permette, inoltre, di stimare la varianza di una popolazione a partire da un campione e fornisce strumenti per valutare con che probabilità la mia stima si discosta dal valore vero.

  - Consideriamo ora un caso più generale.
    - ***Supponiamo di conoscere la forma funzionale della distribuzione di probabilità di una popolazione e di sapere che tale forma funzionale dipende da un parametro incognito (λ).***

    - Ho un metodo in basato sulla teoria dei campioni che mi permette di stimare questo parametro?

METODO DELLA MASSIMA VEROSIMIGLIANZA

- Un metodo di carattere generale per stimare i parametri di una popolazione non nota a partire da un campione è il ***Metodo della Massima Verosimiglianza***

- Supponiamo che una certa popolazione abbia una distribuzione di probabilità che dipenda da un parametro sconosciuto λ e supponiamo di avere un campione di dimensione n di osservazioni {x1,x2,…,xn} e di conoscere la forma funzionale della distribuzione di probabilità f(x, λ).
  - Se ipotizzo che le singole osservazioni sono mutuamente indipendenti allora

$$L = f(x_1,\lambda)f(x_2,\lambda)f(x_3,\lambda)f(x_4,\lambda)....$$

  - E' detta funzione di Massima Verosimiglianza o Verosimiglianza.
  - Essa rappresenta la ***probabilità congiunta*** che le n misure si verifichino.
  - Il metodo della massima verosimiglianza stabilisce che la miglior stima del parametro λ è quella che rende massima la L.
  - Per trovare il massimo di L si calcola il logaritmo di ambo i membri in modo da passare da una produttoria ad una sommatoria, si deriva e si pone uguale a 0

METODO DELLA MASSIMA VEROSIMIGLIANZA

- Applicare il metodo di massima verosimiglianza equivale a risolvere l'equazione:

$$\frac{1}{f(x_1,\lambda)}\frac{\partial f(x_1,\lambda)}{\partial\lambda}+\frac{1}{f(x_2,\lambda)}\frac{\partial f(x_2,\lambda)}{\partial\lambda}+\ldots+\frac{1}{f(x_n,\lambda)}\frac{\partial f(x_n,\lambda)}{\partial\lambda}=0$$

- Il metodo della massima verosimiglianza è molto usato, richiede però la conoscenza della forma funzionale della densità di probabilità della popolazione incognita. Non sempre l'equazione precedente è risolvibile analiticamente. In molti casi occorre ricorrere a soluzioni numeriche.

METODO DELLA MASSIMA VEROSIMIGLIANZA

- Il caso a più parametri è un'estensione banale del caso a singolo parametro;

$$
\begin{cases}
\dfrac{1}{f(x_1,\lambda)}\dfrac{\partial f(x_1,\lambda)}{\partial\lambda_1}+\dfrac{1}{f(x_2,\lambda)}\dfrac{\partial f(x_2,\lambda)}{\partial\lambda_1}+\ldots+\dfrac{1}{f(x_n,\lambda)}\dfrac{\partial f(x_n,\lambda)}{\partial\lambda_1}=0\\
\dfrac{1}{f(x_1,\lambda)}\dfrac{\partial f(x_1,\lambda)}{\partial\lambda_2}+\dfrac{1}{f(x_2,\lambda)}\dfrac{\partial f(x_2,\lambda)}{\partial\lambda_2}+\ldots+\dfrac{1}{f(x_n,\lambda)}\dfrac{\partial f(x_n,\lambda)}{\partial\lambda_2}=0\\
\ldots\\
\dfrac{1}{f(x_1,\lambda)}\dfrac{\partial f(x_1,\lambda)}{\partial\lambda_n}+\dfrac{1}{f(x_2,\lambda)}\dfrac{\partial f(x_2,\lambda)}{\partial\lambda_n}+\ldots+\dfrac{1}{f(x_n,\lambda)}\dfrac{\partial f(x_n,\lambda)}{\partial\lambda_n}=0
\end{cases}
$$

Sistema di eq da risolvere

- Complicazione: le stime dei parametri $\boldsymbol{\lambda}_i$ non necessariamente saranno indipendenti le une dalle altre

METODO DELLA MASSIMA VEROSIMIGLIANZA

- Il parametro determinato con questo metodo, $\hat{\lambda}$, risulta una funzione di variabili casuali indipendenti, pertanto e' anche esso una variabile casuale

$$\frac{1}{f(x_1,\lambda)}\frac{\partial f(x_1,\lambda)}{\partial\lambda}+\frac{1}{f(x_2,\lambda)}\frac{\partial f(x_2,\lambda)}{\partial\lambda}+\ldots+\frac{1}{f(x_n,\lambda)}\frac{\partial f(x_n,\lambda)}{\partial\lambda}=0$$

- Qual e' la sua distribuzione di probabilita' ?
- Con quale errore $\hat{\lambda}$ stima il parametro della distribuzione λ ?

- Per definizione la varianza di $\hat{\lambda}$ e'

$$\sigma^2(\hat{\lambda}) = \int (\hat{\lambda}(x') - \hat{\lambda}_{medio})^2 L(x'_1, x'_2, \ldots, x'_n;\ \lambda)\ dx'_1 dx'_2 \ldots dx'_n$$

METODO DELLA MASSIMA VEROSIMIGLIANZA

- Per definizione la varianza di $\hat{\lambda}$ e'

$$\sigma^2(\hat{\lambda}) = \int (\hat{\lambda}(x') - \hat{\lambda}_{medio})^2 L(x'_1, x'_2, \ldots, x'_n;\ \lambda)\ dx'_1 dx'_2 \ldots dx'_n$$

- Difficile da calcolare, tipicamente si affronta numericamente
- Una relazione approssimata, valida per ***n*** grande, e'

$$\sigma^2(\hat{\lambda}) = -\left(\frac{d^2 \ln L}{d\lambda^2}\right)^{-1}$$

- Con lo stesso livello di approssimazione se la distribuzione di probabilita' dipende da più' parametri, l'inverso della matrice i cui elementi sono

$$U_{ij} = -\frac{d^2 \ln L}{d\lambda_i\ d\lambda_j}$$

rappresenta la matrice di covarianza, i cui elementi diagonali sono le varianze e quelli fuori diagonali le covarianze di coppie di variabili

METODO DELLA MASSIMA VEROSIMIGLIANZA

$$L^\star = -\ln L = A\lambda_1^2 + B\lambda_2^2$$

$$-\frac{d^2 \ln L}{d\lambda_1^2} = 2A$$

$$-\frac{d^2 \ln L}{d\lambda_2^2} = 2B$$

$$\begin{pmatrix} \frac{1}{2A} & 0 \\ 0 & \frac{1}{2B} \end{pmatrix} \begin{pmatrix} 2A & 0 \\ 0 & 2B \end{pmatrix} = \mathrm{I}$$

$$\hat{\lambda}_1 = 0 \quad \hat{\lambda}_2 = 0$$

$$\sigma(\lambda_1) = \sqrt{\frac{1}{2A}} \qquad \Delta L^\star = L^\star(\hat{\lambda}_1 + \sigma(\lambda_1), \hat{\lambda}_2) - L^\star(\hat{\lambda}_1, \hat{\lambda}_2) = \frac{1}{2}$$

$$\sigma(\lambda_2) = \sqrt{\frac{1}{2B}} \qquad \Delta L^\star = L^\star(\hat{\lambda}_1, \hat{\lambda}_2 + \sigma(\lambda_2)) - L^\star(\hat{\lambda}_1, \hat{\lambda}_2) = \frac{1}{2}$$

METODO DELLA MASSIMA VEROSIMIGLIANZA

Vediamone un' applicazione

- Supponiamo di avere n coppie di dati xi (note con errori trascurabili) e yi le quali hanno errori σi e supponiamo che le tra le variabili x e y siano legate dalla relazione: y=f(x,λ) dove λ è un parametro incognito.

- La probabilità che dato un certo x si ottenga un certo y è una probabilità condizionata ad una particolare scelta di λ. Si può valutare cioè per un fissato λ la probabilità di ottenere una particolare coppia di x e y.
  - $P(y_i|\lambda, x_i)$

- Se ipotizzo che le misure fluttuino in maniera gaussiana allora:

$$P(y_i|\lambda, x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - f(x_i,\lambda))^2}{2\sigma_i^2}}$$

- Applicando il principio di massima verosimiglianza ottengo

METODO DELLA MASSIMA VEROSIMIGLIANZA

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n}\left(\frac{1}{\sigma_i}\right)\prod_{i=1}^{n} e^{-\frac{(y_i - f(x_i,\lambda))^2}{2\sigma_i^2}}$$

$$\ln(L) = \cos t. + \sum_{i=1}^{n} -\frac{(y_i - f(x_i,\lambda))^2}{2\sigma_i^2}$$

- NOTA:
  - Dal metodo di massima verosimiglianza deriva il metodo dei minimi quadrati.
  - Se si applica il metodo della massima verosimiglianza per determinare i parametri di una Gaussiana (media e varianza) si ottiene lo stimatore corretto della media ma non quello della varianza.

METODO DELLA MASSIMA VEROSIMIGLIANZA PER DETERMINARE STIMATORI DELLA POPOLAZIONE

Consideriamo alcune distribuzioni note

Metodo della Massima Verosimiglianza

- Distribuzione di Poisson
  - Stimatore della media

$$L(\mu\,|\,x) = \prod_{i=1}^{N} \frac{\mu^{x_i}}{x_i!} e^{-\mu} = e^{-N\mu} \times \prod_{i=1}^{N} \frac{\mu^{x_i}}{x_i!}$$

$$L^\star = \ln L(\mu\,|\,x) = -N\mu + \sum_{i=1}^{N} x_i \ln \mu - \sum \ln x_i!$$

$$\frac{dL^\star}{d\mu} = -N + \frac{1}{\mu}\sum_{i=1}^{N} x_i = 0$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \,<X>$$

Nota: caso semplice, la Poissoniana dipende da un solo parametro

METODO DELLA MASSIMA VEROSIMIGLIANZA

- Distribuzione di Gauss
  - Stimatore della media,

Nota: due parametri

$$L(\mu, \sigma | x) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$L^\star = \ln L(\mu, \sigma | x) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2}$$

$$\begin{cases} \dfrac{dL^\star}{d\mu} = \displaystyle\sum_{i=1}^{N}\frac{x_i-\mu}{\sigma^2} = 0 \quad \rightarrow \quad \mu = \hat{\mu} = \sum_{i=1}^{N}\frac{x_i}{N} \\ \dfrac{dL^\star}{d\sigma^2} = -\dfrac{N}{2\sigma^2} + \dfrac{1}{2}\displaystyle\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^4} = 0 \rightarrow \quad \sigma^2 = \hat{\sigma}^2 = \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{N} \end{cases}$$

$\hat{\mu}$ = media campionaria

ma usando la prima eq $\mu=\hat{\mu}$ quindi $\hat{\sigma}^2 \rightarrow S^2_n$ (attenzione, estimatore distorto)

$$\frac{d^2L^\star}{d^2\mu} = -\frac{N}{\sigma^2} \quad \text{------------->> errore su } \hat{\mu} \rightarrow \sigma/\sqrt{N}$$

$$\frac{d^2L^\star}{d^2(\sigma^2)} = \frac{N}{2\sigma^4} - \frac{1}{2}\sum_{i=1}^{N}\frac{2(x_i-\mu)^2}{\sigma^6} \quad \text{------------->> errore su } \hat{\sigma}^2 \rightarrow \sigma^2\sqrt{2/N}$$

Con σ da valutare con il suo stimatore di massima verosimiglianza