Inferenza statistica

Cosa vuol dire distribuzione di una popolazione? Immaginiamo di avere una popolazione formata da individui che abbia una caratteristica misurabile (peso, altezza, età, reddito, opinione, abitudine di consumo, utilizzo di mezzi di trasporto) definibili in maniera certa. Modello la popolazione formata da individui rispetto a questa caratteristica, ciascuno con una possibilità di appartenere alla caratteristica, modellata come casuale dal punto di vista dell'osservatore esterno. Quando un individuo appartiene a una popolazione Ciascuno di noi ha un'altezza, ma vista come un'altezza di popolazione, possiamo pensarla come una variabile casuale distribuita in una certa maniera? Quando, rispetto alla sua caratteristica che studiamo, possiamo far entrare la variabile casuale che rispetta questa caratteristica in una serie di variabili casuali tutte identicamente distribuite. Ogni individuo è rappresentato da una variabile casuale corrispondente. Le variabili casuali sono tutte di è associata una distribuzione di probabilità, ma la distribuzione della probabilità non è una caratteristica globale, è associata alla variabile casuale di ogni individuo. Ciascun individuo porta con sé una variabile distribuite alla stessa maniera, tutte indipendenti. La distribuzione diventa una caratteristica della popolazione. L'insieme di tutti gli individui descritti dalla variabile casuale indipendente identicamente distribuita caratterizzati da questa distribuzione comune formano una popolazione. A quella popolazione casuale distribuita nella stessa maniera degli altri individui della popolazione.

Per esempio, per l'altezza relativa alla popolazione italiana, è chiaro che mischiare bambini con adulti, non avranno un'altezza distribuita nella stessa maniera, anche soltanto come media tra bambini e adulti è diversa. Non si può considerare come unica popolazione adulti e bambini, ma li devo distinguere in range di età, affinché la media possa essere ragionevolmente più o meno la stessa per tutte le variabili casuali di ciascun individuo della popolazione. Mischiando individui con caratteristiche diverse, ottengo un modello non corretto. Nel caso dell'altezza, essa dipende dalla genetica, alimentazione, sport… Tutte queste informazioni fanno sì che non è proprio vero che ogni individuo della popolazione adulta italiana abbia la stessa possibilità di avere una certa altezza, si tratta di un modello. Le semplificazioni sono la caratteristica matematica per fare previsioni, altrimenti avremmo tantissimi sottosistemi che non sono una popolazione.

L'ipotesi fondamentale introdotta è vera soltanto in maniera approssimata. Ogni popolazione è formata da sottoinsiemi di carattere relazionale, che possono condizionare le scelte di ciascuno. Nella descrizione della statistica, in un certo termine, questo aspetto viene trascurato. Quando si fa importante, allora si passa a una descrizione suddividendo la popolazione in più popolazioni caratterizzate da provenienze o esperienze diverse.

Cosa deve fare l'inferenza statistica? Considerare un campione.

CAMPIONE $(X_1, \dots, X_n)$
- indipendenti
- identicamente distribuite

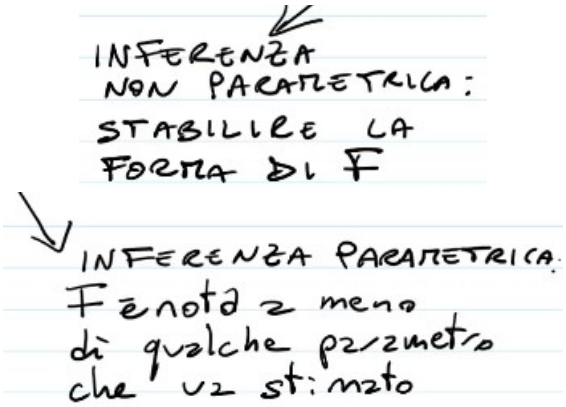
Sono ipotesi molto forti che ci consentono di lavorare in maniera semplice con tutti i calcoli che ne derivano.

- identicamente distribuite

(F è la distribuzione di prob. comune a tutti gli X_k)

Tutti gli Xk hanno la stessa funzione di distribuzione. Partendo dagli individui vogliamo o raccontare qualcosa sulla forma della distribuzione (modello normale, lognormale, uniforme, esponenziale…).

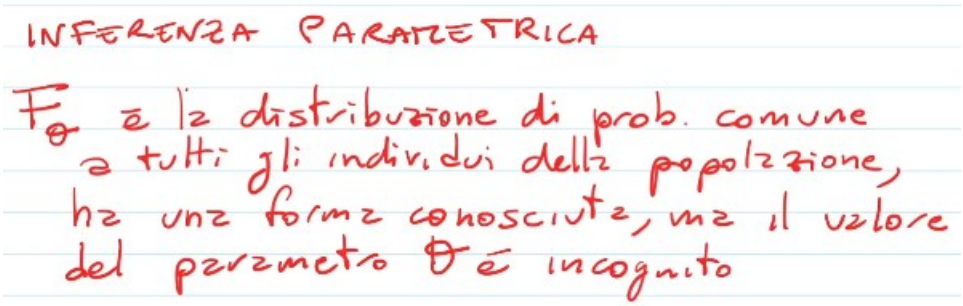
Oppure:



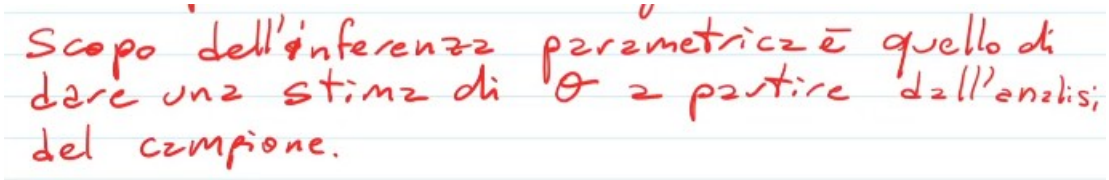
Per esempio, da studi precedenti so che l'altezza di una popolazione si distribuisce in maniera gaussiana, ma non conosco media e varianza. Oppure media conosciuta ma varianza no, oppure varianza sì e media no. In questi 3 casi, bisogna preoccuparsi di stabilire quali parametri mettere nelle costanti che caratterizzano le distribuzioni di probabilità.

È più ardua l'inferenza non parametrica, non conoscere la distribuzione di probabilità rende più difficile arrivare a risultati precisi.

Inferenza parametrica



La distribuzione è nota come forma, ma theta non è conosciuto.



Studio il campione, arrivo a una stima del parametro incognito. Conoscendone la stima e conoscendo la distribuzione, per tutti gli individui della popolazione posso dire che la loro caratteristica si comporterà secondo una funzione di distribuzione nota.

Per stimare θ usiamo uno STIMATORE $\hat{\theta}$ ovvero una STATISTICA (FUNZIONE DELLE VARIABILI DEL CAMPIONE) tale che:

Per essere uno stimatore per stimare theta, deve avere le seguenti caratteristiche.

1) sia corretto $E[\hat{\theta}] = \theta$

Se prendo una funzione delle variabili del campione con valor medio diverso dal parametro che voglio stimare, quello non può essere usato come stimatore del parametro.

2) sia efficiente: tra tutti gli stimatori corretti si considera quello con varianza più piccola $Var(\hat{\theta})$ minima

3) sia consistente: $\lim_{n \to +\infty} E[(\hat{\theta} - \theta)^2] = 0$

n è la numerosità del campione.

(se lo stimatore segue 1) $\Rightarrow$ $\lim_{n \to +\infty} Var(\hat{\theta}) = 0$)

Deve essere tale che se mando a infinito la numerosità del campione, la sua varianza tende a 0.

Un esempio di stimatore corretto, efficiente e consistente è la media campionaria.

es. 1 LA MEDIA CAMPIONARIA è lo stimatore del parametro μ di una popolazione gaussiana

Per popolazione gaussiana s'intende una popolazione i cui individui rispetto a una certa caratteristica hanno distribuzione gaussiana a tutti gli individui.

$(X_1, X_2 .. X_n)$ CAMPIONE

$\bar{X} = \sum_{k=1}^{n} \frac{X_k}{n}$ media campionaria

Media aritmetica delle variabili del campione.

Due proprietà ci danno correttezza e consistenza.

$$E[\bar{X}] = E\left[\sum_{k=1}^{n} \frac{X_k}{n}\right] = \frac{1}{n} \sum_{k=1}^{n} E[X_k] = \frac{1}{n} \sum_{k=1}^{n} \mu = \frac{n\mu}{n}$$

Sulla varianza entra in gioco l'ipotesi dell'indipendenza.

$$Var(\bar{X}) = Var\left(\sum_{k=1}^{n} \frac{X_k}{n}\right) = \frac{1}{n^2} Var\left(\sum_{k=1}^{n} X_k\right) = \underset{\text{INDIP.}}{=}$$

$$= \frac{1}{n^2} \sum_{k=1}^{n} Var(X_k) = \frac{1}{n^2} \sum_{k=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\lim_{n \to +\infty} Var(\bar{X}) = 0$$

Possiamo verificare la consistenza dello stimatore media campionaria quando ci interessa stimare il parametro mu.

Per l'efficienza, i calcoli sono più complessi.

Lo stimatore ottimale del sigma quadro è la varianza campionaria.

Es. 2 LA VARIANZA CAMPIONARIA è lo stimatore ottimale di σ^2 per una popolazione gaussiana

$$S^2 = \frac{\sum_{k=1}^{n} (X_k - \bar{X})^2}{n-1}$$

In fisica, ci si imbatte negli esperimenti con n al posto n – 1. Perché n – 1? Ci sono 2 risposte.

La più intuitiva è, se avessi il campione formato solo da un elemento, la media campionaria si riduce a X1. Quando io considero l'espressione a numeratore della varianza campionaria, è come dire 0. Allora, in questo caso particolare, se a denominatore ci fosse 1, avremmo varianza nulla. Invece così viene indeterminata ed è giusto, perché se la varianza misura quanto sono sparpagliati con un dato solo non posso sapere quanto sono sparpagliati i valori della variabile casuale.

Perché $(n-1)$? Se $n=1$ $\bar{X} = \frac{X_1}{1} = X_1$

$$\sum_{k=1}^{1} (X_k - \bar{X})^2 = (X_1 - X_1)^2 = 0$$

$$\Rightarrow S^2 = \frac{0}{0} \text{ forma indeterminata}$$

Risulta corretto dire che con un dato solo posso dare una stima della media, ma non della varianza. Per la varianza servono almeno 2 dati del campione.

infatti per stimare la varianza occorrono almeno 2 individui nel campione

La motivazione matematica è che se io lasciassi n lo stimatore non sarebbe corretto, cioè il suo valor medio non sarebbe sigma quadro.

$$2)\ E[S^2] = E\left[\frac{\sum_{k=1}^{n}(X_k - \bar{X})^2}{n-1}\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{k=1}^{n} X_k^2 + \sum_{k=1}^{n} \bar{X}^2 - 2\sum_{k=1}^{n}(\bar{X} X_k)\right] =$$

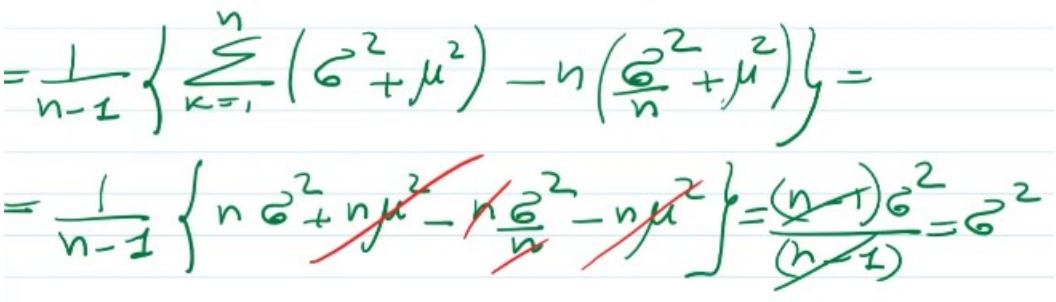
$$= \frac{1}{n-1}\left\{E\left[\sum_{k=1}^{n} X_k^2\right] + E\left[\sum_{k=1}^{n} \bar{X}^2\right] - 2E\left[\sum_{k=1}^{n} \bar{X} X_k\right]\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{k=1}^{n} E[X_k^2] + E[n\bar{X}^2] - 2E\left[\bar{X}\sum_{k=1}^{n} X_k\right]\right\} =$$

$$= \frac{1}{n-1}\left\{\sum_{k=1}^{n} E[X_k^2] + nE[\bar{X}^2] - 2E[n\bar{X}^2]\right\} =$$

$$= \frac{1}{n-1}\left\{\sum_{k=1}^{n} E[X_k^2] - nE[\bar{X}^2]\right\} =$$

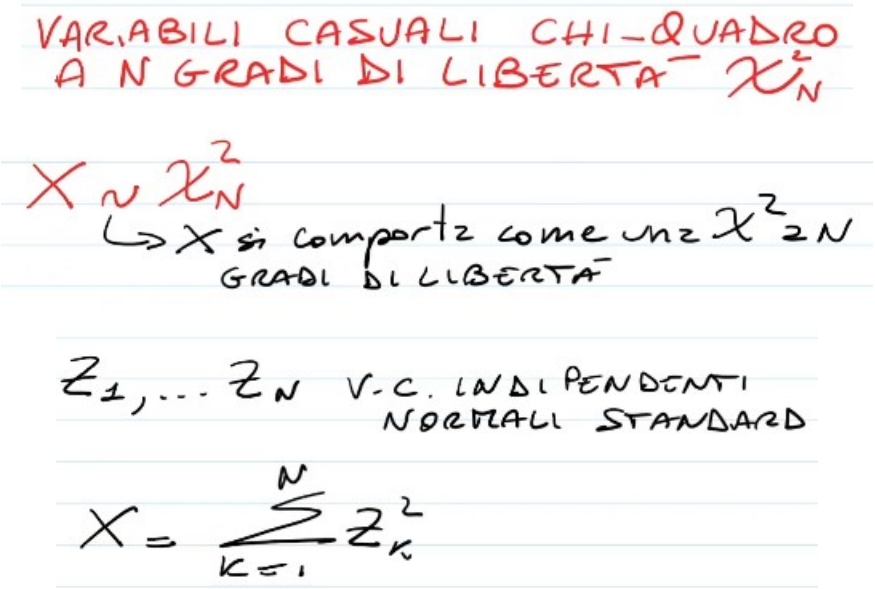
$$= \frac{1}{n-1}\left\{\sum_{k=1}^{n}\left(Var(X_k) + E^2[X_k]\right) - n\left(Var(\bar{X}) + E^2[\bar{X}]\right)\right\} =$$



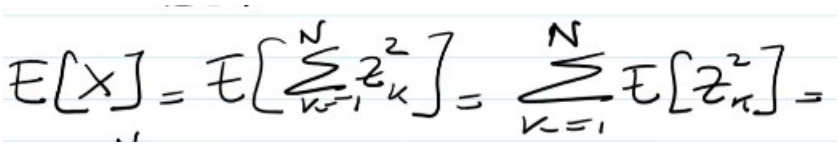
Siamo riusciti a verificare che il valor medio della varianza campionaria è sigma quadro.

Perché nel laboratorio di fisica si mette n? Perché quando n è molto grande, fare diviso n o n – 1 non cambia molto. In tante situazioni si divide per n per semplicità, ma da un punto di vista più rigoroso lo stimatore corretto è n – 1.

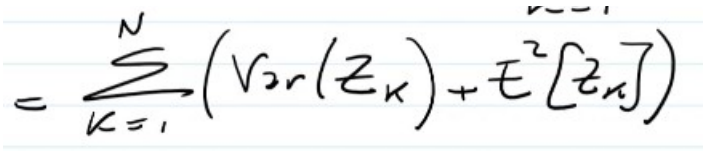
Variabili casuali chi-quadro a n gradi di libertà



La chi-quadro è una somma di normali standard al quadrato. Le sue funzioni di densità di probabilità e distribuzione non sono così semplici, ma da queste caratteristiche si può dire qualcosa: chi-quadro non può essere negativa (somma di quadrati).



Come prima, il valor medio del quadrato lo vediamo come il valor medio della varianza più il quadrato del valor medio.



Siccome le Z sono normali standard:

$$Z_k \sim N(0,1)$$

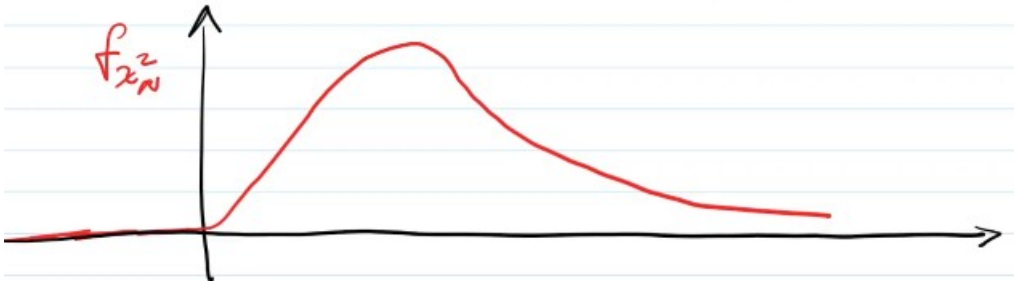
$$= \sum_{k=1}^{N} \left(\underbrace{Var(Z_k)}_{1} + \underbrace{E^2[Z_k]}_{0} \right) = \sum_{k=1}^{N} 1 = N$$

Il valor medio di una variabile casuale chi-quadro a N gradi di libertà è uguale al numero di gradi di libertà della variabile stessa.

PER CASA: QUANTO VALE $Var(X)$?

Si può lavorare con la varianza e la funzione generatrice dei momenti.

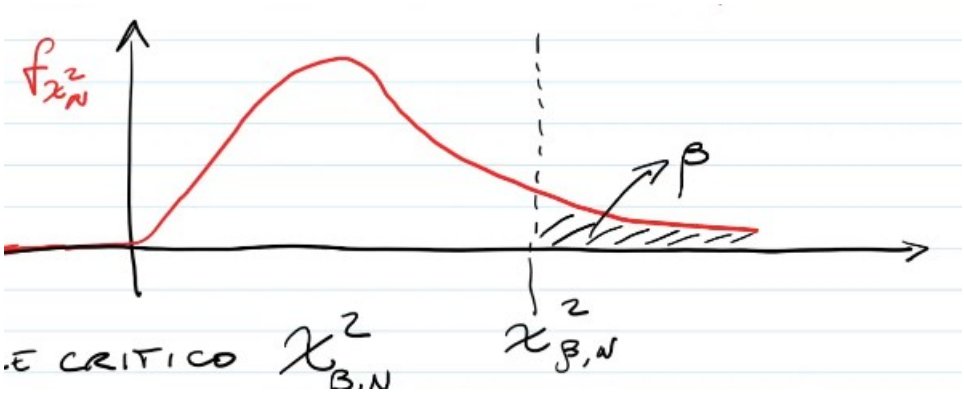
La funzione di densità è fatta così:



Ci sono tabelle della chi-quadro, che presentano una quantità particolare, il valore critico.

VALORE CRITICO $\chi^2_{\beta,N}$

È un valore per il quale succede:



$$P(X > \chi^2_{\beta,n}) = \beta$$

Questi valori si possono trovare sempre più lontani dall'asse Y se stiamo cercando un valore di beta molto piccolo e viceversa.

$$P(X \geq \chi^2_{\beta,n}) = \beta$$

T di Student a N gradi di libertà

T DI STUDENT A N GRADI DI LIBERTÀ

$Y \sim t_N$

↳ Y si comporta come una v.c. t di Student a N GRADI DI LIBERTÀ

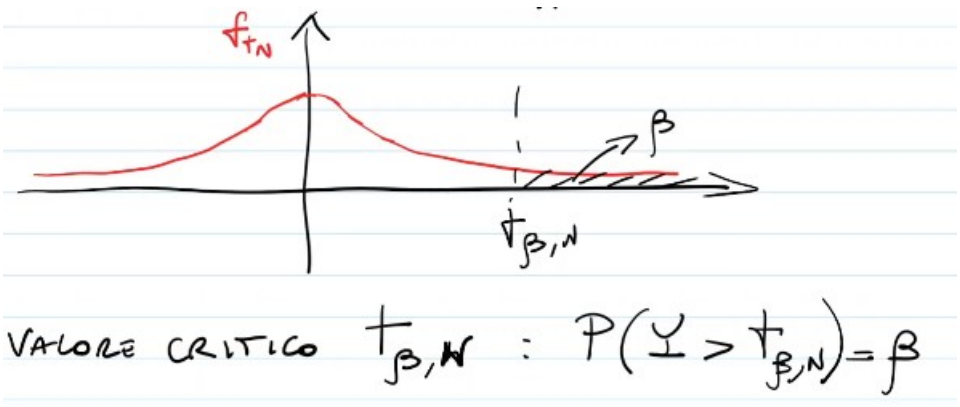
Servono 2 ingredienti:

$$Z \sim N(0,1)$$
$$C \sim \chi^2_N$$

$$Y = \frac{Z}{\sqrt{\frac{C}{n}}}$$

Teniamo conto che queste due distribuzioni di probabilità sono state create ad hoc per la statistica.

La funzione di densità risulta più schiacciata e anche qui c'è un valore critico.



VALORE CRITICO $t_{\beta,N}$: $P(Y > t_{\beta,N}) = \beta$

A differenza della chi-quadro, la T di Student è simmetrica.

Quindi:

$$t_{1-\beta,N} = -t_{\beta,N}$$

infatti: $P(Y > -t_{\beta,N}) = 1-\beta$

N.B.: l'indipendenza si introduce per semplificare i calcoli. A posteriori, va sempre fatta la verifica che tutto sia fatto ragionevolmente, prelevando un altro campione. Inoltre, l'indipendenza delle variabili del campione implica che gli individui scelti nel campione siano scelti in maniera casuale, cioè tutti gli individui hanno la stessa probabilità di far parte del campione, senza di questo l'ipotesi di indipendenza non si può fare, ci può essere un bias che mina l'indipendenza tra gli individui che formano un campione.

Stiamo cercando di definire degli stimatori per stimare dei parametri di una distribuzione di probabilità. Stiamo ragionando come in meccanica quantistica, pensando alle particelle come una distribuzione di probabilità, ma quando si fanno le misurazioni la distribuzione di probabilità non c'è più, c'è il valore della loro velocità o posizione. Lo stesso accade qui: finché si parla di stimatori come variabili casuali funzioni di variabili casuali del campione, si parla in astratto. Quando faccio un esperimento, ottengo dei valori, non più variabili casuali e lo stimatore diventa stima, un valore. Per esempio la media campionaria è il risultato dell'operazione "prendo un campione, misuro le variabili del campione, faccio la media aritmetica delle misure e ottengo un numero". La differenza tra stimatore e stima è: lo stimatore è la descrizione astratta di come si fa l'esperimento; la stima è il valore dello stimatore in un dato esperimento.

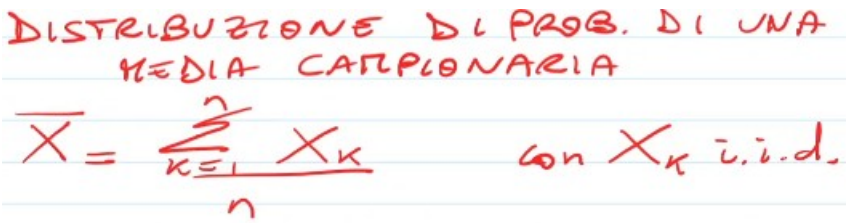
STIMATORE = VARIABILE CASUALE TRATTATA TEORICAMENTE PER ARRIVARE ALLA STIMA

STIMA = VALORE DELLO STIMATORE IN UN CERTO ESPERIMENTO

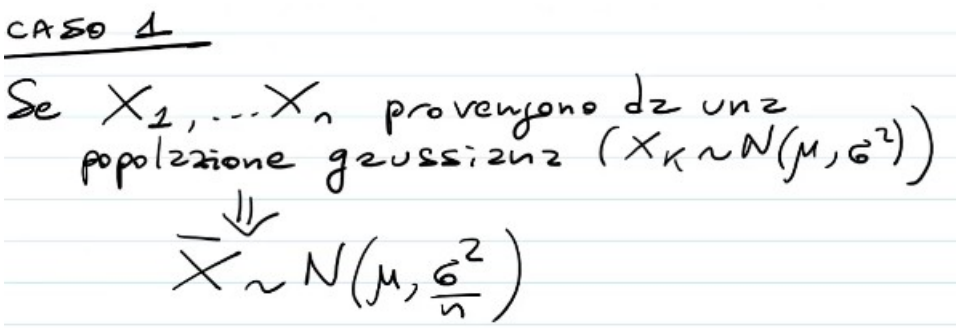
(RIPETENDO L'ESPERIMENTO LA STIMA PUÒ CAMBIARE PERCHÉ LO STIMATORE È UNA V.C.)

Come popolazione definiamo gli studenti di UNIBO. Ciascuno di noi preleva l'altezza di 50 studenti (il campione), misura le altezze e fa la media campionaria (aritmetica). Dalla media campionaria teorica si passa a una stima del valor medio, il valore che si ottiene sommando 50 altezze e dividendo per 50. Se ognuno di noi facesse così e si rifanno queste operazioni di stima, ognuno di noi otterrebbe un valore leggermente diverso da quelli altrui, perché a seconda degli individui presi si otterrà un valore non identico a quello degli altri. Lo stesso succedeva in fisica, tutti facevano le misure e si faceva la media, il valore era circa lo stesso ma ogni gruppo otteneva un valore diverso. Per questo lo stimatore è una variabile casuale: in ogni esperimento non è detto che il suo valore sia sempre lo stesso, sarà distribuito in una certa maniera. A proposito di questo, possiamo dire qualcosa sulla distribuzione della media campionaria e varianza campionaria.

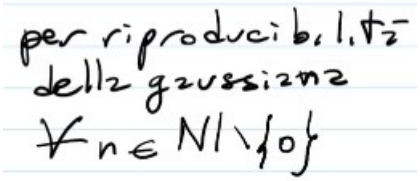
Distribuzione di probabilità di una media campionaria



Se la popolazione del campione fosse gaussiana (anche solo con 2 elementi), la media campionaria sarebbe gaussiana, una combinazione lineare di gaussiane.

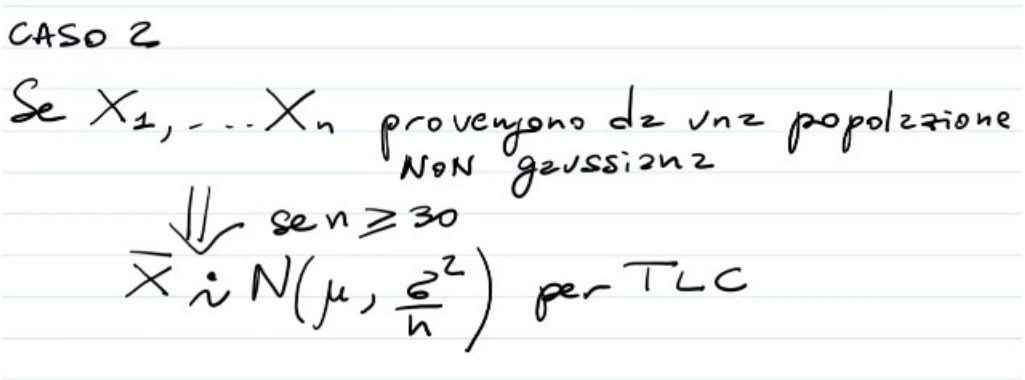


Questo a prescindere da n, la numerosità del campione.



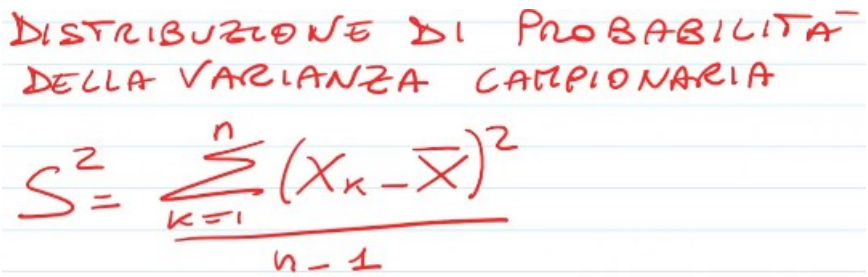
Nel caso in cui le variabili non appartengono a una gaussiana (tutti gli altri casi, distribuzioni strane).

Possiamo solo dire che con n vicino o superiore a 30, vale il teorema del limite centrale.

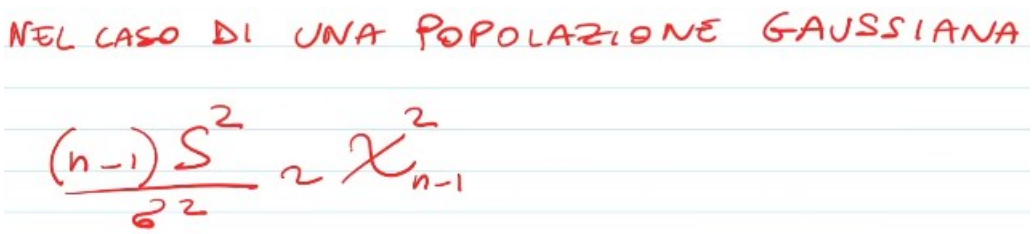


È un risultato teorico, con qualunque distribuzione, vale questo per il teorema del limite centrale.

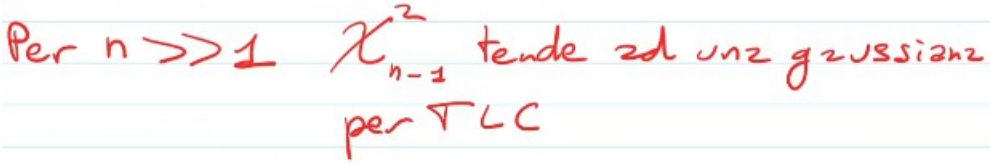
Distribuzione della probabilità della varianza campionaria



Se n è molto molto grande, anche S quadro si comporterà come una gaussiana. Però la questione delicata è che c'è la differenza tra la variabile e la media campionaria, che contiene in sé le stesse variabili. Quindi è vero che per n molto grande si comporta come una gaussiana, ma i risultati rigorosi li ricordiamo con una popolazione gaussiana.



Questo è il motivo per cui si è introdotta la chi-quadro. Moltissime situazioni sono descritte da una popolazione gaussiana, quindi è naturale cercare di codificare una distribuzione che descriva l'andamento della varianza campionaria di questa popolazione gaussiana. n – 1 perché c'è la media campionaria, toglierla fa abbassare di 1 il numero di gradi di libertà indipendenti.



Anche con popolazioni non gaussiane ma n molto grande S quadro si comporta circa come una gaussiana per il teorema del limite centrale.

Data una popolazione gaussiana

$X_1, \dots X_n$ CAMPIONE r.c. i.i.d.

$X_k \sim N(\mu, \sigma^2)$

Supponiamo di non conoscere sigma quadro, ci sarà una varianza ma non si sa quanto vale.

Prendiamo la media campionaria.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ per riproducibilità}$$

$$\Downarrow$$

$$Z = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

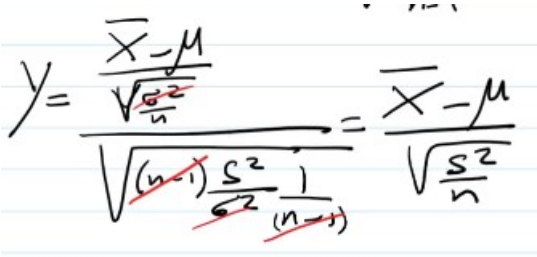
Otteniamo una normale standard.

Consideriamo la varianza campionaria.

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

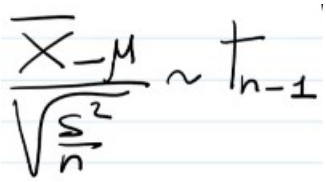
Creo una T di Student a n – 1 gradi di libertà dalle due variabili casuali introdotte.

$$\left.\begin{array}{l} Z = \dfrac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \\ C_{n-1} = (n-1)\dfrac{S^2}{\sigma^2} \sim \chi^2_{n-1} \end{array}\right\} \quad Y \sim t_{n-1} : Y = \frac{Z}{\sqrt{\frac{C_{n-1}}{n-1}}}$$

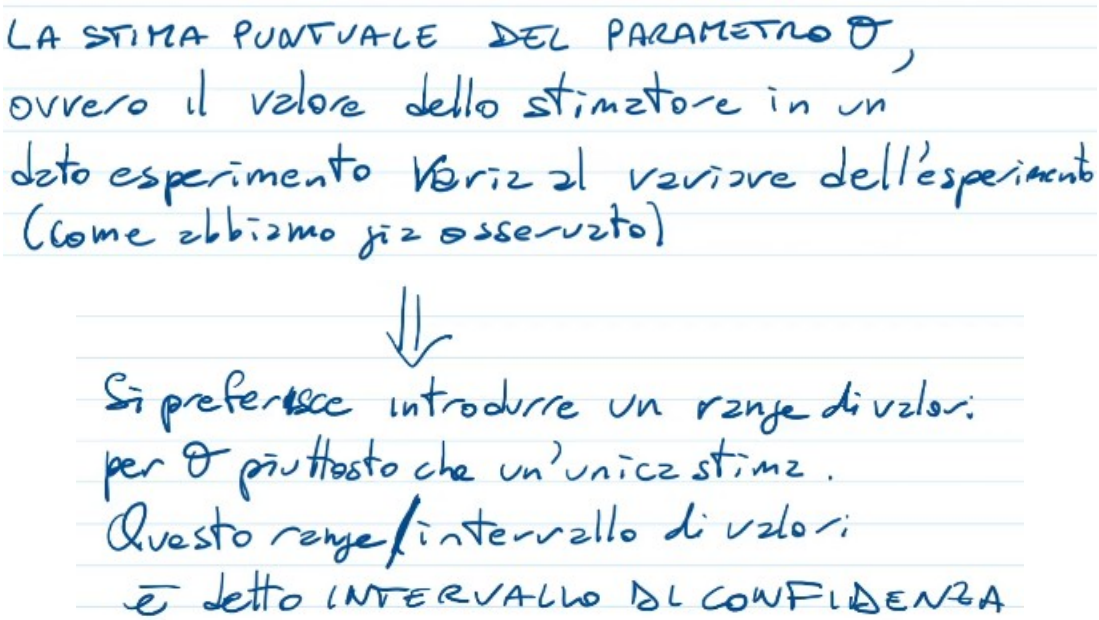


Non conoscendo sigma quadro e uso il suo stimatore (S quadro) ottengo una variabile casuale che si comporta come una T di Student. Ecco perché viene introdotta la T di Student. Student era lo pseudonimo di un chimico tedesco, "studente idealmente dei grandi matematici tedeschi".

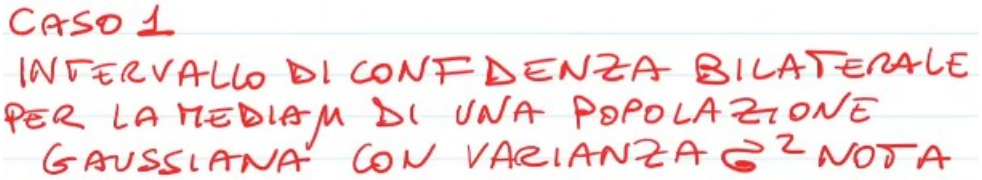
Tutte le volte che non conosciamo sigma quadro, usiamo S quadro per stimare sigma quadro, la media campionaria – mu diviso la radice quadrata di s quadro su n a cui sostituiamo S quadro a sigma quadro nella standardizzazione, si comporta come una T di Student a n – 1 gradi di libertà.



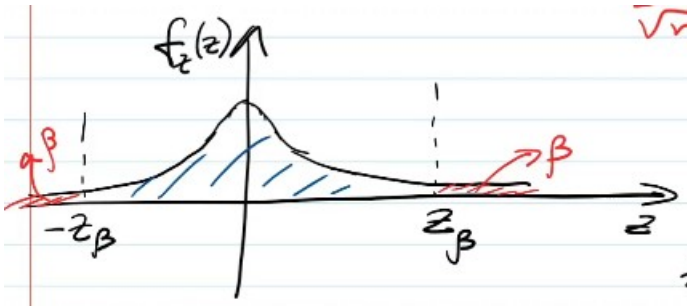
Tutto questo è importante perché purtroppo tutte le volte che facciamo un campionamento e le misurazioni otteniamo una stima diversa, così non sappiamo più quale valore attribuire al parametro che volevamo stimare. Si passa dalla stima secca ad un range di valori in cui ci aspettiamo che il parametro sia compreso, l'intervallo di confidenza.



Un intervallo di confidenza è quello per dare una stima del parametro mu di una popolazione gaussiana quando è nota sigma quadro.



$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$



z_β valore critico di $N(0,1)$

$Z \sim N(0,1)$

$$P(-z_\beta \leq Z \leq z_\beta) = 1-\beta-\beta = 1-2\beta$$

$$P\left(-z_\beta \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_\beta\right) = 1-2\beta$$

È un'affermazione della teoria della probabilità. Trasformiamola in un'affermazione di inferenza statistica, andando ad isolare al centro il parametro (da stimare) mu.

$$P\left(-z_\beta \frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z_\beta \frac{\sigma}{\sqrt{n}}\right) = 1-2\beta$$

$$P\left(-\bar{X} - z_\beta \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_\beta \frac{\sigma}{\sqrt{n}}\right) = 1-2\beta$$

$$P\left(\bar{X} - z_\beta \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_\beta \frac{\sigma}{\sqrt{n}}\right) = 1-2\beta$$

(moltiplicando per -1 scambio i termini e cambio di segno)

Questo è un intervallo di confidenza.