CAMPIONAMENTO

Gli statistici si basano sulle leggi fondamentali della probabilità e dell'inferenza statistica, per giungere a conclusioni sui sistemi scientifici studiati.
L'obiettivo è generalizzare l'esperimento singolo alla classe di tutti gli esperimenti simili, operando un'estensione dal particolare al generale, detta INFERENZA INDUTTIVA.
L'inferenza induttiva è perciò un processo d'azzardo: non si possono fare generalizzazioni assolutamente certe, si possono fare inferenze incerte e misurare il grado di incertezza in termini di probabilità.
DEFINIZIONE: La totalità delle osservazioni a cui siamo interessati è detta POPOLAZIONE OBIETTIVO.
(il numero delle osservazioni può essere finito o infinito).
Essendo poco pratico esaminare l'intera popolazione, si può esaminare una sua parte e fare inferenze sulla popolazione obiettivo.
DEFINIZIONE: Un sottoinsieme della popolazione è detto CAMPIONE.
Perché il campione sia rappresentativo della popolazione è necessario che il <u>campionamento</u> sia <u>casuale</u>
Nel campionamento casuale semplice ogni campione di una determinata dimensione ha la stessa proba= bilità di essere selezionato di qualsiasi altro campio= ne delle stesse dimensione.
(campionamenti indipendenti).

Supponiamo che la popolazione sia caratterizzata da una certa funzione di densità $f(x)$. Scegliendo un c.c. di dim = n dalla popolazione $f(x)$, definiamo la variabile casuale X_i, $i=1,\dots,n$ per rappresentare la i-esima misura del campione che si osserva.

$X_1,\dots,X_n$ sono un c.c. semplice ottenuto da $f(x)$ se le misure sono state ottenute ripetendo l'esperimento n volte in modo indipendente e alle stesse condizioni.

$\Rightarrow$ $X_1,\dots,X_n$ sono n v.c. indipendenti con la stessa densità di probabilità $f(x)$.

DEFINIZIONE: Siano $X_1,\dots,X_n$ n v.c. indipendenti con funzione di densità $f(x)$. $X_1,\dots,X_n$ è detto **CAMPIONE CASUALE** di dimensione n se

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n)=f(x_1)\cdots f(x_n)$$

(f. di densità congiunta = prodotto f. di densità marginali)

DEFINIZIONE: Il campione casuale è detto **POPOLAZIONE CAMPIONATA**.

La distribuzione congiunta

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n)=f(x_1)\cdots f(x_n)$$

è detto **DISTRIBUZIONE CAMPIONARIA** di $X_1,\dots,X_n$.

Lo scopo principale nel selezionare campioni casuali è quello di ottenere informazioni riguardo alcuni parametri sconosciuti della popolazione obiettivo.

Cioè è nota la forma di $f(\cdot,\theta)$, ma f contiene un parametro incognito θ.

Procedimento: si estrae un c.c. $X_1, \dots, X_n$ di dim $= n$ dalla densità $f(\cdot, \theta)$ e si stima il parametro incognito θ con il valore di una qualche funzione $t(x_1, \dots, x_n)$. Infine si determina quale tra queste funzioni sia la migliore per stimare il parametro θ.

DEFINIZIONE: una funzione t delle variabili casuali $X_1, \dots, X_n$ che costituiscono il campione casuale è detta **STATISTICA**.

La statistica $t(x_1, \dots, x_n)$ è a sua volta una variabile casuale che NON contiene alcun parametro incognito.

Esempi di statistiche utilizzate per misurare il centro di una serie di dati sono la media, la mediana e la moda:

Dato un c.c. di dim $= n$: $X_1, \dots, X_n$

a) MEDIA CAMPIONARIA

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

b) MEDIANA CAMPIONARIA

$$\tilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{se } n \text{ è dispari} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{se } n \text{ è pari} \end{cases}$$

c) MODA CAMPIONARIA

È il valore del campione che si presenta più frequentemente.

Altre importanti statistiche sono:

DEFINIZIONE: Dato $X_1, \dots, X_n$ c.c. di dim$=n$ estratto da una popolazione con densità $f(\cdot)$ si definisce MOMENTO CAMPIONARIO DI ORDINE r (ASSOLUTO) la quantità:

$$M'_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r$$

OSSERVAZIONE

Se $r=1$ $\quad M'_1 \equiv \bar{X}_n$.

DEFINIZIONE: Dato $X_1, \dots, X_n$ c.c. di dim$=n$ estratto da una popolazione con densità $f(\cdot)$ si definisce MOMENTO CAMPIONARIO DI ORDINE r RISPETTO A $\bar{X}_n$ la quantità:

$$M_r = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^r$$

N.B.: Se $r=1$ $\quad M_1 \equiv 0$.

OSSERVAZIONE

I momenti campionari assoluti rispecchiano i momenti della popolazione, cioè vale la seguente

TEOREMA 1: Dato $X_1, \dots, X_n$ c.c. di dim$=n$ estratto da una popolazione con densità $f(\cdot)$ si ha:

$$E[M'_r] = \mu'_r.$$

dove μ'_r sono i momenti di ordine r della popolazione.

Dim:

$$E[M'_r] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i^r\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i^r\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i^r]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mu'_r = \frac{1}{n} \cdot n \mu'_r = \mu'_r$$

perché X_i ha f. di densità f, $i=1, \dots, n$.

TEOREMA 1 bis : Dato $X_1, \dots, X_n$ c.c. di dim $= n$ estratto da una popolazione con densità $f(\cdot)$ si ha:

$$var[M'_r] = \frac{1}{n}\left[\mu'_{2r} - (\mu'_r)^2\right]$$

Dim :

$$var[M'_r] = var\left[\frac{1}{n}\sum_{i=1}^{n} X_i^r\right] = \left(\frac{1}{n}\right)^2 var\left[\sum_{i=1}^{n} X_i^r\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} var[X_i^r] \overset{(\circ)}{=}$$

perchè X_i sono v.c. indipendenti

In generale se W è una v.c.

$$var[W] = E[W^2] - (E[W])^2$$

$$\overset{(\circ)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\left\{E[X_i^{2r}] - (E[X_i^r])^2\right\}$$

$$= \frac{1}{n^2}\left\{n\,E[X^{2r}] - n(E[X^r])^2\right\}$$

perchè X_i hanno tutte la stessa funzione di densità f.

$$= \frac{1}{n}\left[\mu'_{2r} - (\mu'_r)^2\right]$$

OSSERVAZIONE

Se $r = 1$ $\quad E[M'_1] = E[\bar{X}_n] = \mu'_1 = \mu$

dove μ è la media della popolazione

$$var[M'_1] = var[\bar{X}_n] = \frac{1}{n}\left[\mu'_2 - (\mu'_1)^2\right] = \frac{\sigma^2}{n}$$

dove $\sigma^2 = \mu'_2 - (\mu'_1)^2$ è la varianza della popolazione.

$$E[\bar{X}_n] = \mu \qquad var[\bar{X}_n] = \frac{\sigma^2}{n}$$

Una misura di posizione o tendenza centrale, in un campione non fornisce da sola una chiara indicazione sulla natura del campione. Deve essere sempre considerato anche una misura di variabilità del campione.
Riguardo al momento campionario di ordine r rispetto alla media campionaria si ha

se $r=2$
$$M_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Anzichē utilizzare M_2 si preferisce usare:

DEFINIZIONE: Dato $X_1, \dots, X_n$ c.c. di dim $= n$ estratto da una popolazione con densità $f(\cdot)$ si definisce VARIANZA CAMPIONARIA la quantità

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

N.B.: se n è molto grande, non c'è differenza tra S^2 e M_2.

OSSERVAZIONE: si usa S^2 anzichè M_2 come misura della variabilità del campione poichè vale il seguente

TEOREMA 2: Dato $X_1, \dots, X_n$ c.c. di dim $= n$ estratto da una popolazione con funzione di densità $f(\cdot)$ si ha:

$$E[S^2] = \sigma^2$$

dove σ^2 è la varianza della popolazione.

Dim: (facoltativa).

Dimostrazione del TEOREMA 2.

$$\sum_{i=1}^{n}(X_i-\bar{X}_n)^2=\sum_{i=1}^{n}(X_i^2-2X_i\bar{X}_n+\bar{X}_n^2)=\sum_{i=1}^{n}X_i^2-2\bar{X}_n\underbrace{\sum_{i=1}^{n}X_i}_{n\bar{X}_n}+n\bar{X}_n^2$$

$$=\sum_{i=1}^{n}X_i^2-n\bar{X}_n^2$$

Perciò

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2-n\bar{X}_n^2\right]$$

Passando al valore atteso

$$(n-1)E[S^2]=E\left\{\sum_{i=1}^{n}X_i^2-n\bar{X}_n^2\right\}=E\left[\sum_{i=1}^{n}X_i^2\right]-nE[\bar{X}_n^2]$$

$$=\sum_{i=1}^{n}E[X_i^2]-nE[\bar{X}_n^2]\overset{(\bullet)}{=}$$

Dalla definizione di varianza.

$\forall$ v.c. W. $\quad var[W]=E[W^2]-(E[W])^2$

si ha

$$E[W^2]=var[W]+(E[W])^2$$

$$\overset{(\bullet)}{=}\sum_{i=1}^{n}\{var[X_i]+(E[X_i])^2\}-n\{var[\bar{X}_n]+(E[\bar{X}_n])^2\}$$

$$=n\,var[X]+n(E[X])^2-n\left\{\frac{\sigma^2}{n}+\mu^2\right\}$$

perchè le Xi hanno la stessa funzione di densità

$$=n\sigma^2+n\mu^2-\cancel{n}\frac{\sigma^2}{\cancel{n}}-n\mu^2=(n-1)\sigma^2$$

$$\Rightarrow E[S^2]=\sigma^2.$$

Dalle definizioni di M_2 e di S^2 si ha:

$$S^2 = \frac{m}{m-1} \cdot M_2 .$$

Infatti

$$S^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X}_m)^2 = \frac{m}{m-1} \left[\frac{1}{m} \sum_{i=1}^{m} (X_i - \bar{X}_n)^2 \right] = \frac{n}{n-1} M_2$$

$$M_2 = \frac{m-1}{m} S^2$$

Allora

$$E[M_2] = \frac{(m-1)}{m} E[S^2] = \frac{(m-1)}{m} \sigma^2 \neq \sigma^2$$

questo è il motivo per cui si usa la varianza campionaria al posto del momento di ordine 2 rispetto alla media campionaria come statistica per stimare la varianza della popolazione σ^2.

Riassumendo

M'_r stima μ'_r ; $\bar{X}_m$ stima μ ; S^2 stima σ^2

OSSERVAZIONE

Il teorema 1 per $r=1$ ci dice che la media campionaria $\bar{X}_m$ in media è uguale al parametro μ della popolazione ($E[\bar{X}_m] = \mu$), cioè la distribuzione di $\bar{X}_m$ è CENTRATA attorno a μ.

Mentre $var[\bar{X}_m] = \frac{\sigma^2}{m}$ prova che la dispersione dei valori di $\bar{X}_n$ intorno a μ è piccola se m, l'ampiezza del campione, è grande.

LEGGE DEI GRANDI NUMERI IN FORMA DEBOLE

La legge debole dei grandi numeri, che si dimostra usando la disuguaglianza di Chebychev, afferma che si possono fare inferenze attendibili per la media μ di una popolazione attraverso un numero finito di valori (campione casuale di dim=n) di X.

È possibile determinare un intero positivo n tale che, se si prende un c.c. di dimensione $\geq n$ da una popolazione di densità $f(\cdot)$ con media μ, la probabilità che la differenza tra la media campionaria $\bar{X}_n$ e la media μ della popolazione sia minore di una quantità fissata piccola a piacere, è vicina ad 1 quanto si vuole.

In formule:

$$\forall \varepsilon > 0 \quad e \quad 0 < \delta < 1 \quad \exists \, n > \frac{\sigma^2}{\varepsilon^2 \delta} :$$

$$P\left[\, |\bar{X}_n - \mu| < \varepsilon \,\right] \geq 1 - \delta$$

con μ e σ^2 rispettivamente media e varianza della densità $f(\cdot)$ della popolazione.

Dim:

Ricordiamo la disuguaglianza di Markov

$$P[g(X) \geq k] \leq \frac{E[g(x)]}{k}$$

$\forall k > 0$ e $g(x) \geq 0, \forall x \in \mathbb{R}$.

Formulazione analoga:

$$P[g(X) < r] \geq 1 - \frac{E[g(x)]}{r}$$

scelta $g(x) = (\bar{x}_m - \mu)^2$

scelto $r = \varepsilon^2$

$$P[|\bar{X}_m - \mu| < \varepsilon] = P[(\bar{X}_m - \mu)^2 < \varepsilon^2]$$

$$\geq 1 - \frac{E[(\bar{X}_m - \mu)^2]}{\varepsilon^2} \doteq$$

dalla definizione di varianza cioè:

X v.c. $\quad var[X] = E[(X - \mu_x)^2]$

poichè $E[\bar{X}_m] = \mu \Rightarrow E[(\bar{X}_m - \mu)^2] = var[\bar{X}_m]$

$$\doteq 1 - \frac{var[\bar{X}_m]}{\varepsilon^2} = 1 - \frac{\sigma^2}{m\varepsilon^2} \geq 1 - \delta,$$

$var[\bar{X}_m] = \frac{\sigma^2}{m}$

per $\delta > \frac{\sigma^2}{m\varepsilon^2}$ oppure $m > \frac{\sigma^2}{\delta\varepsilon^2}$.

Esempi

1) Data una popolazione con media μ incognita e varianza $\sigma^2 = 1$, calcolare la dimensione del c.c. estratto affinché sia almeno del 95% la probabilità che la media campionaria disti meno di 0.5 dalla media della popolazione

$$P[|\bar{X}_n - \mu| < \varepsilon] \geq 1 - \delta \Rightarrow P[|\bar{X}_n - \mu| < 0.5] \geq 0.95$$

$\Rightarrow \varepsilon = 0.5$

$\delta = 0.05$

$$n > \frac{\sigma^2}{\delta \varepsilon^2} = \frac{1}{(0.05)\cdot(0.5)^2} = 80$$

N.B. σ^2 è nota.

2) Quanto deve essere grande un campione casuale per essere sicuri al 99% che la media campionaria disti meno di (0.5σ) dalla media μ della popolazione?

$$P[|\bar{X}_n - \mu| < \varepsilon] = 0.99$$

$\Rightarrow \delta = 0.01$

$\varepsilon = 0.5\sigma$ $\quad$ σ è incognita

$$n > \frac{\sigma^2}{\delta \varepsilon^2} = \frac{\sigma^2}{(0.01)\cdot(0.5\sigma)^2} = \frac{1}{(0.01)\cdot(0.5)^2} = 400$$

IL TEOREMA DEL LIMITE CENTRALE

Sia $\bar{X}_n$ la media campionaria di un c.c. di dim $= n$ estratto da una popolazione avente funzione di densità $f(\cdot)$ INCOGNITA, con media μ e varianza finita σ^2.
Sia Z_n la v.c. definita da:

$$Z_n = \frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{var[\bar{X}_n]}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Allora la distribuzione di Z_n tende alla distribuzione normale standard $N(0,1)$ quando $n \to \infty$:

$Z_n \dot{\sim} N(0,1)$ $\quad \dot{\sim}$ = approssimativamente.

Problema del TLC: Quanto deve essere grande il campione affinché l'approssimazione sia valida?

Regola empirica $\rightarrow$ $n \geq 30$

OSSERVAZIONE

Se la densità della popolazione $f(\cdot)$ è NORMALE allora ogni elemento X_i di $\bar{X}_n$ è normale e quindi
$Z_n \sim N(0,1)$ $\quad \sim$ = esattamente
indipendentemente dalla numerosità n del campione.

$$\begin{cases} E[Z_n] = \frac{\sqrt{n}}{\sigma} E[\bar{X}_n - \mu] = \frac{\sqrt{n}}{\sigma}(\mu - \mu) = 0 \\ var[Z_n] = \frac{n}{\sigma^2} var[\bar{X}_n - \mu] = \frac{n}{\sigma^2} var[\bar{X}_n] = \frac{n}{\sigma^2} \cdot \frac{\sigma^2}{n} = 1. \end{cases}$$

valgono sempre.

Esempio

Si considerino delle sbarre di lunghezza data, caratterizzate da una $f(\cdot)$ incognita con $\sigma^2 = 0.04$ metri^2.

Scelto un c.c. di dim $= n$, calcolare n in modo che la media campionaria $\bar{X}_n$ disti dalla media della popolazione μ per meno di un centimetro, con una probabilità maggiore del 97%.

1° metodo: LGN

$1\ cm = 0{,}01\ m \Rightarrow \varepsilon = 0{,}01$

$\sigma^2 = 0.04\ m^2 \qquad \sigma = 0.02$

$$P[|\bar{X}_n - \mu| < \varepsilon] > 1 - \delta \qquad \delta > \frac{\sigma^2}{n\varepsilon^2} \quad \text{cioè} \quad n > \frac{\sigma^2}{\delta\varepsilon^2}$$

$\delta = 0.03$

$$\Rightarrow n > \frac{0.04}{(0.03)\cdot(0.01)^2} = 1.\bar{3} \cdot 10^4 \simeq 13333$$

2° metodo: TLC

$$|\bar{X}_n - \mu| < 0.01 \iff -0.01 < \bar{X}_n - \mu < 0.01$$

$$\iff \frac{-0.01}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{0.01}{\frac{\sigma}{\sqrt{n}}}$$

$$\iff -\frac{0.01}{\frac{0.02}{\sqrt{n}}} < Z_n < \frac{0.01}{\frac{0.02}{\sqrt{n}}}$$

$$\iff |Z_n| < \frac{\sqrt{n}}{20}$$

$$P[|X_n - \mu| < 0.01] > 0.97 \equiv P\left[|Z_n| < \frac{\sqrt{n}}{20}\right] > 0.97 \qquad \frac{\sqrt{n}}{20} = z_\alpha$$

$$P[|Z_n| < z_\alpha] = 2\,[P[Z_n < z_\alpha] - 0.5] = 2\,P[Z_n < z_\alpha] - 1$$

$$\Rightarrow 2\,P[Z_n < z_\alpha] - 1 > 0.97 \quad \text{cioè} \quad P[Z_n < z_\alpha] = 0.985$$

Dalla tabella della $N(0,1)$ $z_\alpha = 2.17$

$$\Rightarrow \frac{\sqrt{n}}{20} = 2.17 \Rightarrow n = 1883{,}56 \Rightarrow \boxed{n = 1.884}$$

CAMPIONAMENTO DA DISTRIBUZIONI NORMALI

Data una popolazione con funzione di densità normale $N(\mu, \sigma^2)$ abbiamo visto che la distribuzione della media campionaria $\bar{X}_n$ è ESATTAMENTE $N(\mu, \frac{\sigma^2}{n})$.
e quindi Z_n è ESATTAMENTE $N(0,1)$.
Per ogni X_i e c.c. di dim$=n$ si ha $X_i \sim N(\mu, \sigma^2)$
e quindi $Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$.

Definiamo la funzione:

$$U \doteq \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

(somma di quadrati di normali standard).
Si può provare

TEOREMA 3:

$$U \sim \chi^2_n$$ CHI-QUADRO con n gradi di libertà

Il "grado di libertà" è il n° di quadrati indipendenti nella sommatoria.
(χ^2 è una f. GAMMA con $\lambda = \frac{1}{2}$, $r = \frac{n}{2}$)
Poichè $Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, in base al TH. 3 si ha

$$Z_n^2 \sim \chi^2_1$$ CHI-QUADRO con $n=1$ gradi di libertà

Definiamo la funzione:

$$V \doteq \frac{(n-1)}{\sigma^2} S^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \overset{(\bullet)\text{ CONTI}}{=} \sum_{i=1}^{n} Z_i^2 - Z_n^2 = \underbrace{U}_{\sim \chi^2_n} - \underbrace{Z_n^2}_{\sim \chi^2_1}$$

Si può provare:

TEOREMA 4

$$V \sim \chi^2_{n-1}$$ CHI-QUADRO con $(n-1)$ gradi di libertà

In analogia a $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ è possibile definire

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right) = \frac{1}{n\sigma} (n\bar{X}_n - n\mu) = \frac{\bar{X}_n - \mu}{\sigma}$$

allora vale la relazione

$$\sum_{i=1}^{n} (Z_i - \bar{Z}_n) \equiv 0$$ VINCOLO che abbassa il grado di libertà.

Definiamo la funzione:

$$T \doteq \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{\bar{Z}_n}{\sqrt{\frac{V}{n-1}}}$$

Poiché $\bar{X}_n$ e S^2 sono statistiche indipendenti si può provare che $\bar{Z}_n$ e V sono indipendenti.

Si può provare:

TEOREMA 5

$$T \sim t_{n-1}$$ t di STUDENT con $(n-1)$ gradi di libertà

CONTI (•)

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{\sigma^2} \sum_i \left((X_i - \mu) - (\bar{X}_n - \mu) \right)^2 = \frac{1}{\sigma^2} \sum_i (X_i - \mu)^2$$

$$- \frac{2}{\sigma^2} (\bar{X}_n - \mu) \underbrace{\sum_i (X_i - \mu)}_{n\bar{X}_n - n\mu} + \frac{1}{\sigma^2} (\bar{X}_n - \mu)^2 \cdot n$$

$$= \sum_i \left(\frac{X_i - \mu}{\sigma} \right)^2 - \frac{n}{\sigma^2} (\bar{X}_n - \mu)^2 = \sum_i Z_i^2 - n \left(\frac{\bar{X}_n - \mu}{\sigma} \right)^2$$

$$= \left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \right)^2$$

$$= \sum_i Z_i^2 - n \bar{Z}_n^2 = \sum_i Z_i^2 - Z_n^2$$

TABELLA RIASSUNTIVA

Z_m è la statistica in grado di fare inferenza sulla media μ della popolazione quando σ^2 è nota

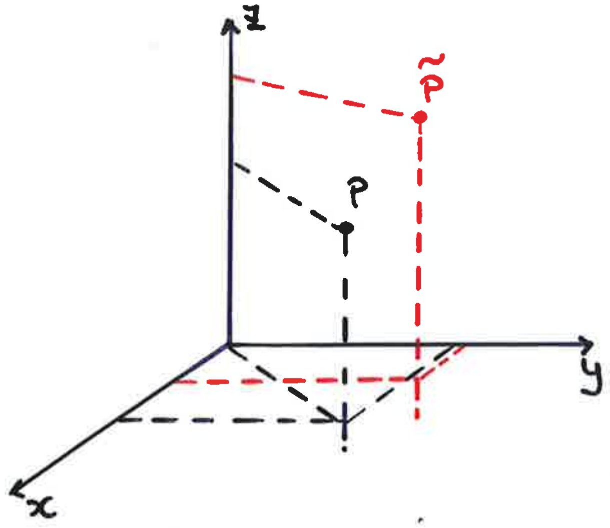
T è la statistica in grado di fare inferenza sulla media μ della popolazione quando σ^2 è incognita.
(in realtà vale per $m < 30$)

V è la statistica in grado di fare inferenza sulla varianza della popolazione σ^2 quando μ è incognita

U è la statistica in grado di fare inferenza sulla varianza σ^2 quando μ è nota.

Esempio

Si vuole localizzare un oggetto nello spazio, ma la misurazione porta un errore (in ognuna delle 3 direzioni x, y, z) che è una v.c. normale $N(\mu=0, \sigma=2m.)$. Supponendo i 3 errori indipendenti, calcolare la probabilità che la distanza tra posizione misurata e posizione reale sia maggiore di 3 metri.



$P(x, y, z)$ REALE

$\tilde{P}(x_1, y_1, z_1)$ MISURATA

$x_1 = x + \varepsilon_1$

$y_1 = y + \varepsilon_2$ $\quad$ ε_i errori

$z_1 = z + \varepsilon_3$ $\quad$ $i = 1, 2, 3$

$D =$ distanza tra P e $\tilde{P}$.

$$D^2 = \overline{P\tilde{P}}^2 = (x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2$$

$\varepsilon_i \sim N(0, 2)$

$$\hat{z}_i = \frac{\varepsilon_i - \mu}{\sigma} = \frac{\varepsilon_i}{2} \sim N(0, 1)$$

$Y = \sum_{i=1}^{3} \hat{z}_i^2$ somma di quadrati di normali standard

Per il TEOREMA 3 $\quad Y \sim \chi^2_{m=3}$

$$P[D > 3] = P[D^2 > 9] = P[\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 > 9] = P\left[\hat{z}_1^2 + \hat{z}_2^2 + \hat{z}_3^2 > \frac{9}{4}\right]$$

$$= P\left[Y > \frac{9}{4}\right] = 1 - \underbrace{P\left[Y \le \frac{9}{4}\right]}_{\approx 0.4778} \approx 0.5222$$

Se il problema della localizzazione avviene nel piano allora $P(x, y)$ $\quad \tilde{P}(x_1, y_1)$

$D^2 = \varepsilon_1^2 + \varepsilon_2^2 \quad \Rightarrow \quad Y \sim \chi^2_{m=2}$

Ma $\chi^2_{n=2} = \Gamma\left(\lambda = \frac{1}{2}, \kappa = \frac{n}{2} = 1\right) = \Gamma\left(\frac{1}{2}, 1\right) \equiv \exp\left(\lambda = \frac{1}{2}\right)$

Quindi $Y \sim \exp\left(\lambda = \frac{1}{2}\right)$ cioè:

$$f(y) = \lambda e^{-\lambda y}, \quad y \geq 0$$

$$F(y) = 1 - e^{-\lambda y}$$

Allora

$$P[D > 3] = P[D^2 > 9] = P\left[Y > \frac{9}{4}\right] = 1 - P\left[Y \leq \frac{9}{4}\right] = 1 - F(9/4)$$

$$= e^{-\lambda y}\Big|_{\lambda = \frac{1}{2},\, y = \frac{9}{4}} = e^{-9/8} \approx 0.3247$$