10 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 inferenza sulla popolazione obiettivo.
DEFINIZIONE: Un sottoinsieme della popolazione è detta CAMPIONE.
Perchè il campione sia rappresentativo della popola-

zione è necessario che il campionamento sia casuale. Nel campionamento casuale semplice ogni campione di una determinata dimensione ha la stessa probabilità di essere selezionato di qualsiasi altro campione della stessa dimensione (campionamenti indipendenti).
Supponiamo che la popolazione sia caratterizzata da una certa funzione di densità $f(x)$. Scegliendo un campione casuale di dimensione n dalla popolazione $f(x)$, definiamo la variabile casuale X_i, $i = 1, ..., n$ per rappresentare la $i-$esima misura del campione che si osserva.
$X_1, ..., X_n$ sono un campione casuale semplice ottenuto da $f(x)$ se le misure sono state ottenute ripetendo l'esperimento n volte in modo indipendente e alle stesse condizioni $\Rightarrow$ $X_1, ..., X_n$ sono n variabili casuali indipendenti con la stessa densità di probabilità $f(x)$.
DEFINIZIONE: Siano $X_1, ..., X_n$ n variabili casuali indipendenti con funzione di densità $f(x)$. $X_1, ..., X_n$ è detto CAMPIONE CASUALE di dimensione n se

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

(la funzione di densità congiunta è uguale al prodotto delle funzioni di densità marginali).
DEFINIZIONE: Il campione casuale viene chiamato POPOLAZIONE CAMPIONATA.
La distribuzione congiunta

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

è detta DISTRIBUZIONE CAMPIONARIA del campione $X_1, ..., X_n$.
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 campione casuale $X_1, ..., X_n$ di dimensione n dalla densità $f(\cdot, \theta)$ e si stima il parametro incognito θ con il valore di una qualche funzione $t(X_1, ..., X_n)$. Infine si determina quale tra queste funzioni sia la migliore per stimare il parametro θ.
DEFINIZIONE: Una funzione t delle variabili casuali $X_1, ..., X_n$ che costituiscono il campione casuale è detta STATISTICA.

La statistica $t(X_1, ..., X_n)$ 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, qui di seguito definite.
Dato un campione casuale $X_1, ..., X_n$ di dimensione n, si definiscono:

- MEDIA CAMPIONARIA

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

- MEDIANA CAMPIONARIA

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

- MODA CAMPIONARIA
  È il valore del campione che si presenta più frequentemente.

Altre importanti statistiche sono le seguenti:
DEFINIZIONE: Dato $X_1, ..., X_n$ campione casuale di dimensione 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 $r = 1 \qquad M'_1 = \overline{X}_n$.

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

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

Nota bene: se $r = 1 \qquad M_1 = 0$.

OSSERVAZIONE: I momenti campionari assoluti rispecchiano i momenti della popolazione, cioè vale il seguente

Teorema 1: Dato $X_1, ...X_n$ campione casuale di dimensione 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.

Dimostrazione

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

$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i^r] \underbrace{=}_{X_i \text{ hanno f.d. } f} \frac{1}{n}\sum_{i=1}^{n} \mu'_r = \frac{1}{n}\cdot n\mu'_r = \mu'_r$$

Teorema 1 bis: Dato $X_1, ..., X_n$ campione casuale di dimensione n estratto da una popolazione con densità $f(\cdot)$ si ha:

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

Dimostrazione

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

$$= \frac{1}{n^2}\,\text{var}\left[\sum_{i=1}^{n} X_i^r\right] \underbrace{=}_{X_i \text{ indip.nti}} \frac{1}{n^2}\sum_{i=1}^{n} \text{var}[X_i^r]. \qquad (1)$$

A questo punto notiamo che se W è una variabile casuale

$$\text{var}[W] = E[W^2] - E[W]^2,$$

quindi possiamo continuare l'equazione (1) con

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

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

OSSERVAZIONE

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

dove μ è la media della popolazione. Inoltre:

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

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

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

Una misura di posizione, o tendenza centrale, in un campione non fornisce da sola una chiara indicazione sulla natura del campione. Deve essere sempre

considerata anche una misura di variabilità del campione. Riguardo al momento campionario di ordine r rispetto alla media campionaria si ha

$$\text{se } r = 2 \qquad M_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

Anzichè utilizzare M_2 si preferisce usare la varianza campionaria che ora definiamo.

DEFINIZIONE: Dato $X_1, ..., X_n$ campione casuale di dimensione n estratto da una popolazione con densità $f(\cdot)$ si definisce VARIANZA CAMPIONARIA la quantità

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

Nota bene: 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 perchè vale il seguente

TEOREMA 2: Dato $X_1, ..., X_n$ campione casuale di dimensione 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.

Dimostrazione (facoltativa)

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

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

Perciò

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

Passando al valore atteso

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

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

Ma dalla definizione di varianza abbiamo che $\forall$ variabile casuale W vale

$$\text{var}[W] = E[W^2] - E[W]^2$$

e quindi

$$E[W^2] = \text{var}[W] + E[W]^2$$

e possiamo scrivere

$$\begin{aligned}(n-1)E[S^2] &= \sum_{i=1}^{n} \left\{\text{var}[X_i] + E[X_i]^2\right\} + \\ &- n\left\{\text{var}[\overline{X}_n] + E[\overline{X}_n]^2\right\} = \\ &= n\,\text{var}[X] + nE[X]^2 - n\left\{\frac{\sigma^2}{n} + \mu^2\right\}\end{aligned}$$

dove abbiamo utilizzato il fatto che tutte le X_i hanno la stessa funzione di densità. Quindi

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

da cui

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

Calcoliamo adesso $E[M_2]$. Dalla definizione di M_2 e

di S^2 si ha:

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

Infatti

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

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

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

da cui

$$E[M^2] = \frac{n-1}{n} E[S^2] = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

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

M'_r stima μ'_r; $\overline{X}_n$ stima μ; S^2 stima σ^2

OSSERVAZIONE

Il Teorema 1 per $r = 1$ ci dice che la media campionaria $\overline{X}_n$ in media è uguale al parametro μ della popolazione ($E[\overline{X}_n] = \mu$), cioè la distribuzione di $\overline{X}_n$ è CENTRATA attorno a μ.
Invece $\text{var}[\overline{X}_n] = \frac{\sigma^2}{n}$ prova che la dispersione dei valori di $\overline{X}_n$ intorno a μ è piccola se n, l'ampiezza del campione, è grande.

LA LEGGE DEI GRANDI NUMERI IN FORMA DEBOLE

La legge debole dei grandi numeri, che si dimostra usando la disuguaglianza di Chebyshev, afferma che si possono fare inferenze attendibili per la media μ di una popolazione attraverso un numero finito di valori (campione casuale di dimensione n) di X.
È possibile determinare un intero positivo n tale che, se si prende un campione casuale di dimensione $\geq n$ da una popolazione di densità $f(\cdot)$ con media μ, la probabilità che la differenza tra la media campionaria $\overline{X}_n$ e la media μ della popolazione sia minore di una quantità fissata piccola a piacere, è vicina ad 1 quanto si vuole.

In formule
$\forall \epsilon > 0$ e $0 < \delta < 1$ $\exists\, n > \frac{\sigma^2}{\epsilon^2 \delta}$:

$$P\left[|\overline{X}_n - \mu| < \epsilon\right] \geq 1 - \delta$$

con μ e σ^2 rispettivamente media e varianza della densità $f(\cdot)$ della popolazione.
Dimostrazione
Ricordiamo la disuguaglianza di Markov:

$$P[g(x) \geq r] \leq \frac{E[g(x)]}{r}, \quad \forall r > 0 \text{ e } g(x) \geq 0, \forall x \in \mathbb{R}$$

e la formulazione analoga

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

Scelti $g(x) = (\overline{x}_n - \mu)^2$ ed $r = \epsilon^2$

$$P\left[|\overline{X}_n - \mu| < \epsilon\right] = P\left[(\overline{X}_n - \mu)^2 < \epsilon^2\right]$$
$$\geq 1 - \frac{E\left[(\overline{X}_n - \mu)^2\right]}{\epsilon^2}$$

ma dalla definizione di varianza

$$\text{var}[X] = E[(X - \mu_X)^2]$$

poichè $E[\overline{X}_n] = \mu \Rightarrow E\left[(\overline{X}_n - \mu)^2\right] = \text{var}[\overline{X}_n]$, quindi

$$P\left[|\overline{X}_n - \mu| < \epsilon\right] \geq 1 - \frac{\text{var}[\overline{X}_n]}{\epsilon^2} = 1 - \frac{\sigma^2}{n\epsilon^2} \geq 1 - \delta$$

per $\delta > \frac{\sigma^2}{n\epsilon^2}$ oppure $n > \frac{\sigma^2}{\delta\epsilon^2}$.

Esempi

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

$$P[|\overline{X}_n - \mu| < \epsilon] \geq 1 - \delta \Rightarrow P[|\overline{X}_n - \mu| < 0.5] \geq 0.95$$
$$\epsilon = 0.5 \quad \delta = 0.05 \quad \Rightarrow$$
$$\Rightarrow n > \frac{\sigma^2}{\delta\epsilon^2} = \frac{1}{(0.05)\cdot(0.5)^2} = 80$$

Nota bene 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?

$$P[|\overline{X}_n - \mu| < \epsilon] = 0.99 \Rightarrow \delta = 0.01$$

$\epsilon = 0.5\sigma$ $\quad \sigma$ è incognita

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

IL TEOREMA DEL LIMITE CENTRALE

Sia $\overline{X}_n$ la media campionaria di un campione casuale di dimensione n estratto da una popolazione avente funzione di densità $f(\cdot)$ INCOGNITA, con media μ e varianza finita σ^2. Sia Z_n la variabile casuale definita da:

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

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

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

Problema del TLC: quanto deve essere grande il campione affinchè l'approssimazione sia valida?
Regola empirica $\to n \geq 30$

OSSERVAZIONE

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

Le uguaglianze

$$\begin{cases} E[Z_n] = \dfrac{\sqrt{n}}{\sigma} E[\overline{X}_n - \mu] = \dfrac{\sqrt{n}}{\sigma}(\mu - \mu) = 0 \\ \operatorname{var}[Z_n] = \dfrac{n}{\sigma^2} \operatorname{var}[\overline{X}_n - \mu] = \dfrac{n}{\sigma^2} \operatorname{var}[\overline{X}_n] \\ \qquad\qquad = \dfrac{n}{\sigma^2} \cdot \dfrac{\sigma^2}{n} = 1 \end{cases}$$

valgono sempre.

Esempio

Si considerino delle sbarre di lunghezza data, caratterizzate da una $f(\cdot)$ incognita con $\sigma^2 = 0.04\text{m}^2$. Scelto un campione casuale di dimensione n, calcolare n in modo che la media campionaria $\overline{X}_n$ disti dalla media della popolazione μ per meno di un centimetro, con una probabilità maggiore del 97%.

1° metodo: LGN

$1\text{cm} = 0.01\text{m} \Rightarrow \epsilon = 0.01$

$\sigma^2 = 0.04\text{m}^2 \Rightarrow \sigma = 0.2\text{m}.$

$$P[|\overline{X}_n - \mu| < \epsilon] > 1 - \delta \quad \delta > \frac{\sigma^2}{n\delta^2} \Rightarrow n > \frac{\sigma^2}{\delta\epsilon^2}$$

$$\delta = 0.03 \Rightarrow n > \frac{0.04}{(0.03)\cdot(0.01)^2} = 1.\overline{3} \cdot 10^4 \sim 13.333$$

2° metodo: TLC

$$|\overline{X}_n - \mu| < 0.01 \Leftrightarrow \frac{-0.01}{\frac{\sigma}{\sqrt{n}}} < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{0.01}{\frac{\sigma}{\sqrt{n}}}$$

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

$$P[|\overline{X}_n - \mu| < 0.01] > 0.97 \Rightarrow P[|Z_n| < \underbrace{\frac{\sqrt{n}}{20}}_{=z_\alpha}] > 0.97$$

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

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

$\Rightarrow \frac{\sqrt{n}}{20} = 2.17 \Rightarrow n = 1.883,56 \Rightarrow \boxed{n = 1.884}$

CAMPIONAMENTO DA DISTRIBUZIONI NORMALI

Da una popolazione con funzione di densità normale $N(\mu, \sigma^2)$ segue che la distribuzione della media campionaria $\overline{X}_n$ è ESATTAMENTE $N(\mu, \frac{\sigma^2}{n})$ e quindi Z_n è ESATTAMENTE $N(0,1)$.

Per ogni X_i elemento di un campione casuale di dimensione n si ha $X_i \sim N(\mu, \sigma^2)$ da cui segue che $Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$.

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 numero di quadrati indipendenti nella sommatoria (ricordiamo che χ^2 è una funzione GAMMA con $\lambda = 1/2$ ed $r = n/2$).

Poichè $Z_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, in base al Teorema 3 si ha

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

Definiamo la funzione:

$$V \doteq \frac{n-1}{\sigma^2} S^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)^2 =$$

$$= \sum_{i=1}^{n} Z_i^2 - Z_n^2 = \underbrace{U}_{\sim \chi^2_n} - \underbrace{Z_n^2}_{\sim \chi^2_1}$$

L'uguaglianza in verde verrà giustificata più avanti.
Si può provare:
TEOREMA 4

$$V \sim \chi^2_{n-1}$$

CHI QUADRO con $(n-1)$ gradi di libertà

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

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

$$= \frac{1}{n\sigma} (n\overline{X}_n - n\mu) = \frac{\overline{X}_n - \mu}{\sigma} \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{n} \left(Z_i - \overline{Z}_n \right) = 0$$

VINCOLO che abbassa il grado della libertà.

Definiamo la funzione:

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

Poichè $\overline{X}_n$ e S^2 sono statistiche indipendenti è possibile dimostrare che 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

Giustifichiamo adesso l'uguaglianza in verde introdotta prima del Teorema 4. Abbiamo:

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

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

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

Ma $\frac{n}{\sigma^2}(\overline{X}_n - \mu)^2 = \left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right)^2$, quindi

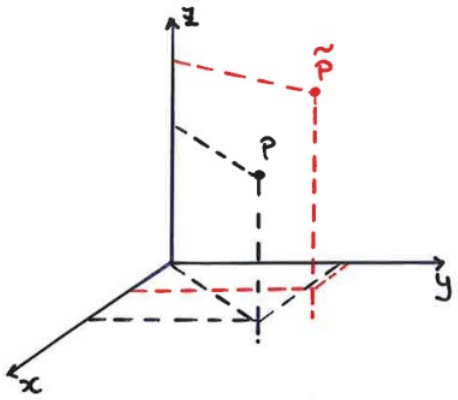
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}_n^2 = \sum_{i=1}^{n} Z_i - Z_n^2$$

TAVOLA RIASSUNTIVA

- Z_n 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.
- V è la statistica in grado di fare inferenza sulla varianza σ^2 della popolazione quando μ è incognita.
- U è la statistica in grado di fare inferenza sulla varianza σ^2 della popolazione quando μ è nota.

Esempio

Si vuole localizzare un oggetto nello spazio, ma il processo di misurazione porta un errore (in ognuna delle 3 dimensioni x, y, z) che si distribuisce come una variabile casuale normale $N(\mu = 0, \sigma = 2\text{m})$. 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

$x_1 = x + \epsilon_1$

$y_1 = y + \epsilon_2$

$z_1 = z + \epsilon_3$

ϵ_1, ϵ_2, ϵ_3 errori

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

$$D^2 = \overline{P\tilde{P}}^2 = (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2$$

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

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

$$Y = \sum_{i=1}^{3} \hat{Z}_i^2 = \text{somma di quadrati di normali stan-}$$

dard $\Rightarrow$ per il Teorema 3: $Y \sim \chi^2_{n=3}$

$$P[D > 3] = P[D^2 > 9] = P[\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 > 9] =$$
$$= P[\hat{Z}_1^2 + \hat{Z}_2^2 + \hat{Z}_3^2 > 9/4] = P[Y > 9/4] =$$
$$= 1 - \underbrace{P[Y \leq 9/4]}_{\simeq 0.4778} \simeq 0.5222$$

Se il problema della localizzazione avviene nel piano allora abbiamo la posizione reale $P(x, y)$ e quella misurata $\tilde{P}(x_1, y_1)$.
$D^2 = \epsilon_1^2 + \epsilon_2^2 \Rightarrow Y \sim \chi^2_{n=2}$
Ma $\chi^2_{n=2} = \Gamma(\lambda = 1/2, r = n/2 = 1) = \Gamma(1/2, 1) =$ $\exp(\lambda = 1/2)$
Quindi $Y \sim \exp(\lambda = 1/2)$ cioè:

$$f(y) = \lambda e^{-\lambda y} \quad \text{se} \quad y \geq 0,$$
$$F(y) = 1 - e^{-\lambda y}.$$

Allora
$P[D > 3] = P[D^2 > 9] = P[Y > 9/4] =$
$1 - P[Y \leq 9/4] = 1 - F(9/4) = e^{-\lambda y}\big|_{\lambda=2, y=9/4} =$
$e^{-9/8} \simeq 0.3247$