7 Campionamento

7.1 Introduzione

Gli statistici si basano sulle leggi fondamentali della probabilità e dell'inferenza statistica per giungere a conclusioni sui sistemi scientifici studiati. L'obiettivo è quello di 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 detto campione.

Perchè il campione sia rappresentativo della popolazione 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 di pari 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 $f(x)$, consideriamo la variabile casuale X_i, $i = 1, ..., n$ per rappresentare la $i-$esima misura del campione che si osserva. Le variabili casuali $X_1, ..., X_n$ 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

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

cioè 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

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

è detta distribuzione campionaria del c.c. $X_1, ..., X_n$.

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 c.c. $X_1, ..., X_n$ 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.

7.2 Statistiche

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

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

Nota bene: 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:

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

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

dove μ'_r, μ'_{2r} sono i momenti di ordine r, $2r$ della popolazione.

Osservazione

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

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.
Calcoliamo adesso $E[M_2]$. Dalla definizione di M_2 e di S^2 si ha:

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

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 σ^2 della popolazione.
Riassumendo

$$M'_r \text{ stima } \mu'_r; \quad \overline{X}_n \text{ stima } \mu; \quad S^2 \text{ stima } \sigma^2$$

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.

7.3 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.

La legge dei grandi numeri in forma debole

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

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

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

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

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

7.4 Il teorema del limite centrale

La formulazione della legge dei grandi numeri in forma debole stabilisce che i risultati delle singole prove influiscono poco sul risultato medio di un numero elevato di prove: le deviazioni dalla media, inevitabili in una prova singola, si livellano reciprocamente quando il numero delle prove è elevato.

Questo significa che quando il numero di prove è elevato il risultato medio diventa stabile e quindi può essere previsto.
Le possibilità di effettuare tali previsioni sono rese maggiori dal *teorema del limite centrale* che stabilisce quale distribuzione segua la somma di un numero sufficientemente grande di variabili casuali.
Tale teorema, detto *centrale* proprio per la sua importanza, permette di definire delle ipotesi e di stimare la loro probabilità di verificarsi.

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 $\rightarrow n \geq 30$

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 del campione.
Valgono sempre le uguaglianze

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

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

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

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

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

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

2° metodo: TLC

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

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

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

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

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

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

7.5 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

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

cioè somma di quadrati di normali standard.
Si può provare:
Teorema 3

$$U \sim \chi^2_n \qquad \text{CHI QUADRO con } n \text{ gradi di libertà}$$

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

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

Definiamo la funzione:

$$V \doteq \frac{n-1}{\sigma^2} S^2 = \underbrace{U}_{\sim \chi^2_n} - \underbrace{Z_n^2}_{\sim \chi^2_1}$$

Si può provare:
Teorema 4

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

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

Definiamo la funzione:

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

Si può provare:
Teorema 5

$$T \sim t_{n-1}$$

t di STUDENT con $(n-1)$ gradi di libertà

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.

7.6 La statistica proporzione campionaria*

Si consideri un campione casuale $X_1, ...X_n$ estratto da una popolazione Bernoulliana, rappresentata dalla variabile casuale $X \sim Ber(p)$.

La media campionaria $\overline{X}_n$ che rappresenta, in questo caso, la proporzione di successi nel campione casuale, dove il numero di successi è $Y = X_1 + ... + X_n$, viene chiamata proporzione campionaria e indicata con $\hat{p}$. La media e la varianza della proporzione campionaria sono:

$$E[\hat{p}] = E[\overline{X}_n] = p \ , var[\hat{p}] = var[\overline{X}_n] = \frac{pq}{n} \, .$$

Per il TLC, la distribuzione campionaria di $\hat{p}$ si approssima con una normale quando $n > 30$, $np > 5$, $nq > 5$:

$$\hat{p} \dot{\sim} N(p, \frac{pq}{n})$$

Esempio. Un medico ospedaliero rileva che il 40% dei suoi pazienti è stato ricoverato più di una volta nel suo reparto. Scelto un c.c. di 100 pazienti ricoverati nel reparto, determinare la probabilità che la proporzione del campione di pazienti con più di un ricovero sia compresa tra 0.4 e 0.5.

Si hanno 100 variabili Bernoulliane che valgono 1 se il paziente ha più di un ricovero e 0 altrimenti.

$X_i \sim Ber(0.4)$, $n = 100 \, , np = 40 \, , nq = 60$ quindi dal TLC

$\hat{p} \dot{\sim} N(p = 0.4, \frac{pq}{n} = 0.0024)$

Allora

$$P[0.4 < \hat{p} < 0.5] = P[0 < \frac{\hat{p} - p}{\sqrt{pq/n}} < \frac{0.5 - 0.4}{\sqrt{0.0024}}] =$$

$$= P[0 < Z < 2.04] = 0.9793 - 0.5 = 0.4793$$