Inferenza Statistica

- Consiste nel trarre delle conclusioni (*fare inferenza*) relativamente a una certa *popolazione* sulla base di un *campione*.
- *Popolazione*: totalità dei casi (unità statistiche) sui quali si manifesta il fenomeno oggetto di studio.
- *Campione*: sottoinsieme delle unità che compongono la popolazione estratte in modo casuale. L'estrazione può essere:
  - *con ripetizione* (o *con reimmissione*): una volta estratta, un'unità viene *reinserita* nella popolazione e quindi può essere selezionata più volte;
  - *senza ripetizione (o senza reimmissione)*: una volta estratta, una unità *non* viene reinserita nella popolazione e quindi può essere selezionata al più una volta.
- *Vantaggi:* in termini di tempo e costi rispetto a una rilevazione completa (censimento).

Esempio

- Si supponga che la *popolazione* sia composta di *N = 100* soggetti che indichiamo con le *etichette 1,2,....,100* e che siamo interessati a *stimare* l'altezza media dei soggetti che compongono questa popolazione sulla base di un *campione* di dimensione *n = 5*.
- *Estrazione con ripetizione*: per 5 volte si seleziona un numero casuale tra 1 e 100; il campione sarà formato dalle unità con le etichette corrispondenti:

  37, 25, **80**, 89, **80**
- *Estrazione senza ripetizione*: per 5 volte si seleziona un numero casuale tra 1 e 100 togliendo via via quelli già selezionati (se il primo numero selezionato è 8, questo non potrà essere più selezionato).

  8, 100, 88, 39, 29

- Per l'*estrazione* si può far ricorso a programmi informatici o alle tavole dei numeri casuali.
- L'*altezza media* della popolazione può essere stimata tramite l'altezza media dei soggetti che compongono il campione.
- Nel caso di *campionamento con ripetizione*:

Unità	Altezza
37	170,75
25	186,14
80	173,39
89	185,12
80	173,39
Media	**177,76**

Formalizzazione in termini statistici

- Si rappresenta il fenomeno di interesse tramite una *variabile aleatoria* (*X*), discreta o continua, che ha una certa distribuzione nella popolazione che viene indicata, in generale, con *f(x)*.
- Per semplicità si considera solo il campionamento con ripetizione o il campionamento da popolazione infinite.
- *Campione casuale*: variabile aleatoria multipla di dimensione *n*

  $(X_1, X_2, \ldots, X_n)$

  le cui componenti sono indipendenti e identicamente distribuite ognuna con distribuzione *f(x)*.
- *Campione osservato*: realizzazione di $(X_1, X_2, \ldots, X_n)$

  $(x_1, x_2, \ldots, x_n)$

- *Spazio campionario* (Ω): insieme di tutti i possibili campioni. Può essere discreto o continuo a seconda della natura (discreta o continua) della variabile *X*.
- *Distribuzione del campione*: funzione di densità o di probabilità del campione ($X_1, X_2, \ldots, X_n$)

$$f(x_1, x_2, \cdots, x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

- Es. : Estrazione di tre pezzi da un processo produttivo. La popolazione di partenza è di tipo Bernoulliano. Qual è la probabilità di estrarre un campione in cui i primi due pezzi siano sani e il terzo difettoso?

$$P(X_1 = 0, X_2 = 0, X_3 = 1) = f(0,0,1) = f(0)f(0)f(1) =$$

$$= \pi^0(1-\pi)^1 \cdot \pi^0(1-\pi)^1 \cdot \pi^1(1-\pi)^0 = \pi^1(1-\pi)^2$$

Esempio

- Si assuma che il *numero di figli per famiglia* (variabile discreta) abbia la seguente distribuzione nella popolazione

X	*f(x)*
0	0,1
1	0,3
2	0,6
	1

- Ci sono 9 campioni distinti di dimensione *n* = 2. Questi formano lo *spazio campionario* (Ω).

X_1	X_2	$f(x_1)$	$f(x_2)$	$f(x_1,x_2)$
0	0	0,1	0,1	0,01
0	1	0,1	0,3	0,03
0	2	0,1	0,6	0,06
…	…	…	…	…
2	1	0,6	0,3	0,18
2	2	0,6	0,6	0,36
				1

Esempio

- Si assuma che il numero di addetti per impresa (variabile discreta) abbia nella popolazione distribuzione di Poisson con parametro λ = 3,5.

$$X \sim Po(3{,}5)$$

- Lo spazio campionario (Ω) ha infiniti elementi.
- La probabilità del campione osservato

2, 4, 3, 3

è pari a

$$f(2,4,3,3) = f(2)\cdot f(4)\cdot f(3)\cdot f(3) =$$

$$= \frac{3{,}5^2}{2!}e^{-3,5} \cdot \frac{3{,}5^4}{4!}e^{-3,5} \cdot \frac{3{,}5^3}{3!}e^{-3,5} \cdot \frac{3{,}5^3}{3!}e^{-3,5} = 0{,}00163$$

- In pratica, la distribuzione della variabile aleatoria *X* non è completamente nota, ma nota a meno di uno o più *parametri* (θ) che la caratterizzano:

$$X \sim f(x;\theta)$$

- *Spazio dei parametri* (Θ): insieme di tutti i possibili valori che può assumere il parametro θ.
- Dato che il parametro θ non è noto, si vogliono trarre delle conclusioni su di esso sulla base di un campione estratto dalla popolazione. Cioè si vuole *fare inferenza* su θ.

- *Esempio 1*: si assume che l'*altezza* (variabile continua) abbia nella popolazione distribuzione normale con media μ e varianza σ^2 non noti:

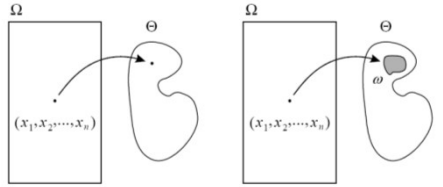
$$X \sim N(\mu, \sigma^2), \quad \mu \in (-\infty, \infty), \quad \sigma^2 \in (0, \infty),$$

- *Esempio 2*: si assume che il *numero di addetti per impresa* (variabile discreta) abbia nella popolazione distribuzione di Poisson con parametro λ non noto:

$$X \sim Po(\lambda), \quad \lambda \in (0, \infty)$$

Metodi di inferenza

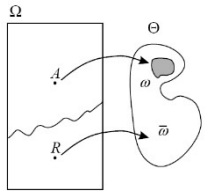
- ***Stima dei parametri***: sulla base del campione si assegna al parametro di interesse (θ) un valore (stima puntuale) o un insieme di valori (stima per intervallo).



- *Esempio 1*: sulla base di un campione di soggetti si stima che l'altezza media degli italiani è pari a 175,36cm o è compresa nell'intervallo (173; 178).

- *Esempio 2*: sulla base di un campione di famiglie si stima che il numero medio di figli per famiglia in Italia è 1,8 o che tale media è compresa nell'intervallo (1,5; 2,3).

- ***Verifica delle ipotesi***: si formula un'ipotesi, o congettura, sul parametro di interesse (θ) e si verifica, sulla base del campione, se tale ipotesi è o meno plausibile.



- *Esempio 1:* si formula l'ipotesi che l'altezza media degli italiani sia pari a 175cm e sulla base di 10 soggetti estratti casualmente si decide se tale ipotesi è plausibile o meno.

- *Esempio 2*: si formula l'ipotesi che il numero medio di figli delle famiglie lombarde sia inferiore di quello delle famiglie calabresi. Sulla base di due campioni di 50 famiglie ognuno, di cui il primo estratto in Lombardia e il secondo in Calabria, di decide se tale ipotesi è plausibile o meno.

Diversi approcci all'inferenza

- ***Classico*** (Fisher, Neyman, Pearson): concezione frequentista della probabilità; fa uso unicamente delle informazioni contenute nel campione.

- ***Bayesiano*** (Lindley, Savage, de Finetti): concezione soggettivista della probabilità; fa uso anche di informazioni a priori sui parametri espresse tramite distribuzioni di probabilità.

- ***Teoria delle decisioni*** (Wald): tiene conto delle conseguenze di decisioni alternative espresse tramite funzioni di perdita; può fare uso anche di informazioni a priori.

Statistiche campionarie

- Servono a sintetizzare l'informazione contenuta in un campione.
- Una ***statistica campionaria*** è una qualsiasi funzione del campione

$$t(X_1, X_2, \cdots, X_n)$$

- *Statistiche campionarie più comuni*:
  - Media campionaria: $\bar{X} = \sum_{i=1}^{n} X_i \Big/ n$
  - Varianza campionaria: $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 \Big/ (n-1)$
  - Statistiche d'ordine: $Y_1 \le Y_2 \le \cdots \le Y_n$
- *In particolare* saremo interessati alla media campionaria ($\bar{X}$) e alla varianza campionaria (S^2).

Esempio

- Si consideri nuovamente il campione di dimensione n = 5 soggetti per ognuno dei quali si riporta l'*altezza* (in cm.)

170,75; 186,14; 173,39; 185,12; 173,39

- La *media campionaria* risulta pari a

$$\bar{x} = \frac{170{,}75 + 186{,}14 + 173{,}39 + 185{,}12 + 173{,}39}{5} = 177{,}76$$

- La *varianza campionaria* risulta pari a

$$s^2 = \frac{(170{,}75 - 177{,}76)^2 + (186{,}14 - 177{,}76)^2 + \ldots + (173{,}39 - 177{,}76)^2}{4} = 52{,}93$$

Distribuzioni campionarie

- Ogni statistica campionaria ha una sua distribuzione, detta ***distribuzione campionaria***.
- Esempio: si riprendano i 9 campioni distinti di dimensione n = 2 sui quali si è osservato il numero di figli

X_1	X_2	$f(x_1,x_2)$	$\bar{X}$
0	0	0,01	0
0	1	0,03	0,5
0	2	0,06	1
1	0	0,03	0,5
1	1	0,09	1
1	2	0,18	1,5
2	0	0,06	1
2	1	0,18	1,5
2	2	0,36	2
		1	

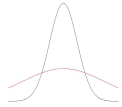
$\bar{X}$	$f(\bar{X})$
0	0,01
0,5	0,06
1	0,21
1,5	0,36
2	0,36
	1

Proprietà della distribuzione della media campionaria

- Per qualsiasi distribuzione $f(x)$ nella popolazione, il valore atteso e la varianza della *media campionaria* sono pari a:

$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

dove $\mu = E(X)$ è la media di X nella popolazione e $\sigma^2 = V(X)$ è la sua varianza.



Dimostrazione:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

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

Proprietà della distribuzione della varianza campionaria

- Per qualsiasi distribuzione $f(x)$ nella popolazione, il valore atteso della *varianza campionaria* è:

$$E\left(S^2\right)=\sigma^2$$

$$E(S^2)=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2\right)=$$
$$=\frac{1}{n-1}E\left(\sum_{i=1}^{n}[(X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2]\right)=$$
$$=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\sum_{i=1}^{n}(\bar{X}-\mu)^2\right)=$$
$$=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2-2n(\bar{X}-\mu)^2+n(\bar{X}-\mu)^2\right)=$$
$$=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}E(X_i-\mu)^2-nE(\bar{X}-\mu)^2\right)=$$
$$=\frac{1}{n-1}\left(\sum_{i=1}^{n}V(X_i)-nV(\bar{X})\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}V(X)-nV(\bar{X})\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma^2-n\frac{1}{n}\sigma^2\right)$$
$$=\frac{1}{n-1}(n\sigma^2-\sigma^2)=\frac{n-1}{n-1}\sigma^2=\sigma^2$$

- Esempio: si riprenda la variabile *numero di figli per famiglia*

X	***f(x)***	***X·f(x)***	***(X-μ)²· f(x)***
0	0,1	0	0,225
1	0,3	0,3	0,075
2	0,6	1,2	0,15
	1	**1,5**	**0,45**

da cui la *media* e la *varianza* di *X* sono μ =1,5 e σ^2 = 0,45.

- *Media* e *varianza campionaria* per ognuno dei 9 campioni

X_1	X_2	$f(x_1,x_2)$	$\bar{X}$	S^2
0	0	0,01	0	0
0	1	0,03	0,5	0,5
0	2	0,06	1	2
1	0	0,03	0,5	0,5
1	1	0,09	1	0
1	2	0,18	1,5	0, 5
2	0	0,06	1	2
2	1	0,18	1,5	0, 5
2	2	0,36	2	0
		1		

- Distribuzione della *media campionaria*:

$\bar{X}$	$f(\bar{x})$	$\bar{X}\cdot f(\bar{x})$	$[\bar{X}-E(\bar{X})]^2\cdot f(\bar{x})$
0	0,01	0	0,0225
0,5	0,06	0,03	0,06
1	0,21	0,21	0,0525
1,5	0,36	0,54	0
2	0,36	0,72	0,09
	1	**1,5**	**0,225**

- Si ha quindi

$$E(\bar{X})=1{,}5=\mu \quad e \quad Var(\bar{X})=0{,}225=\frac{\sigma^2}{n}$$

- Distribuzione della *varianza campionaria*:

S^2	$f(s^2)$	$S^2\cdot f(s^2)$
0	0,46	0
0, 5	0,42	0,21
2	0,12	0,24
	1	**0,45**

- Si ha quindi

$$E\left(S^2\right)=0{,}45=\sigma^2$$

Distribuzione della media campionaria in alcuni casi particolari

- Se *X* ha distribuzione *Bernoulliana*, *X* ~ Bin(1, *p*),

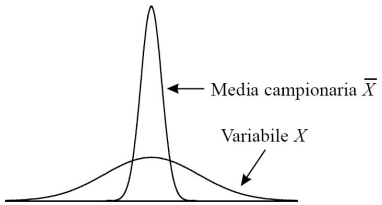
$$P(\overline{X}=\overline{x})=P(n\overline{X}=n\overline{x})=\binom{n}{n\overline{x}}p^{n\overline{x}}(1-p)^{n-n\overline{x}}$$

- Se *X* ha distribuzione *Normale*, *X* ~ N(μ,σ^2)

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

- Nel caso di *grandi campioni* (*n* ≥ 30), a prescindere dalla distribuzione di *X*, approssimativamente

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



Esempio

- Supponiamo che X abbia distribuzione Bernoulliana, $X\sim\text{Bin}(1,p)$, e che da tale popolazione si estraggano campioni di dimensione $n=5$.

- Se p fosse pari a 0,3, la distribuzione di $\overline{X}$ sarebbe

$\overline{x}$	$n\overline{x}$	$f(\overline{x})$
0,0	0	0,168
0,2	1	0,360
0,4	2	0,309
0,6	3	0,132
0,8	4	0,028
1,0	5	0,002
		1,000

$$P(\overline{X}=0)=P(5\overline{X}=0)=\binom{5}{0}0{,}3^0 0{,}7^5=0{,}168$$

$$P(\overline{X}=0{,}2)=P(5\overline{X}=1)=\binom{5}{1}0{,}3^1 0{,}7^4=0{,}360$$

Distribuzione della varianza campionaria in alcuni casi particolari

- Se *X* ha distribuzione *Normale*, *X* ~ N(μ,σ^2)

$$\frac{(n-1)S^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2}\approx\chi^2(n-1)$$

- La distribuzione χ^2(r) è chiamata *chi-quadro* con *r gradi di libertà* (parametro della distribuzione).

- La *media* e la *varianza* di una v.a. Y~ χ^2(r) sono pari a

$$E(Y)=r \quad e \quad V(Y)=2r$$

- Per il calcolo dei *centili* si fa ricorso ad opportune tavole che riportano i valori χ^2_α tali che

$$P(Y\geq\chi^2_\alpha)=\alpha$$

Esempio

- Si supponga che X abbia nella popolazione distribuzione *Normale*

$$X\sim\text{N}(10,20)$$

e che da tale popolazione vengano estratti campioni di dimensione $n=25$.

- La *media campionaria* ha distribuzione

$$\overline{X}\sim\text{N}(10,\,0{,}8)$$

- Per la *varianza campionaria* si ha

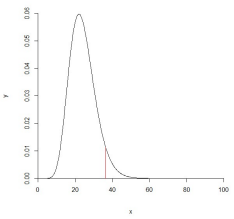
$$\frac{24 \cdot S^2}{20} \sim \chi^2(24)$$

da cui il centile corrispondente a $\alpha = 0{,}05$ è pari a 36,42, cioè

$$P\left(\frac{24 \cdot S^2}{20} \geq 36{,}42\right) = 0{,}05;$$

il centile corrispondente a $\alpha = 0{,}99$ è pari a 10,86, cioè

$$P\left(\frac{24 \cdot S^2}{20} \geq 10{,}86\right) = 0{,}99.$$



Media campionaria standardizzata

- In alcuni casi si farà uso della *media campionaria standardizzata*

$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}}$$

- Se *X* ha distribuzione *Normale, X* ~ N(μ, σ^2), si ha

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

- Usualmente la varianza della popolazione (σ^2) non è nota e quindi si utilizza al suo posto la varianza campionaria (S^2)

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}}$$

- Se *X* ha distribuzione *Normale, X* ~ N(μ, σ^2), si ha

$$T \approx t(n-1)$$

- La distribuzione *t(r)* è chiamata *t di Student* con *r gradi di libertà*.

- La ditribuzione *t di Student* è simmetrica intorno allo 0. La *media* e la *varianza* di una v.a. Y~ *t*(r) sono pari a

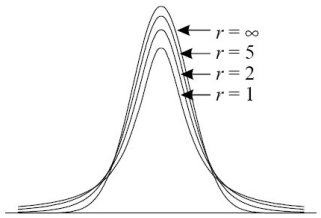
$$E(Y) = 0 \quad e \quad V(Y) = r/(r-2)$$

- Per il calcolo dei *centili* si fa ricorso ad opportune tavole che riportano i valori t_α tali che

$$P(Y \geq t_\alpha) = P(Y \leq -t_\alpha) = \alpha$$

- Quando *r* è elevato (≥100) la distribuzione *t di Student* è approssimabile con la Normale standardizzata.

Funzione di densità



Esempio

- Si supponga che *X* abbia nella popolazione distribuzione Normale

$$X \sim N(10\ ;\ 20)$$

e che da tale popolazione vengano estratti campioni di dimensione *n* = 25.

- La media campionaria standardizzata (con S^2) ha distribuzione

$$T \sim t(24)$$

- Il centile corrispondente a $\alpha = 0{,}05$ è $t_\alpha = 1{,}711$ mentre quello della normale standardizzata è 1,645.

- Se il campione ha dimensione *n* = 121

$$T \sim t(120)$$

- Il centile corrispondente a $\alpha = 0{,}05$ è $t_\alpha = 1{,}658$ molto più simile a quello della normale standardizzata.