**RP5\_BIS**

**INFERENZA STATISTICA: STIMATORI E STIME**

**Statistical inference**

Statistical inference is the process through which inferences about a population (population parameter) are made based on certain statistics calculated from a sample of data drawn from that population.



**Statistical inference**

* We make statistical inferences by using sample statistics to estimate population parameters…
* Using sample statistics we can make estimates about population parameters.
* There are two types of estimates of parameters: point estimates and interval estimates: a point estimate is a single number that is the best guess for the parameter; an interval estimate is an interval of numbers around the point estimate, within which the parameter value is believed to fall.

**Statistical inference: estimation**

* The term **estimator** refers to a particular type of statistic for estimating a parameter. This is conceptual. The term **estimate** is a noun that refers to the value of that particular statistic.
* Just as there are multiple possible sample statistics, there are many possible estimators. If we are interested, for example, in the population’s mean, then we could use the mean, the median, or the mode of the sample as our estimator.
* How do statisticians decide which estimator is best?

**Statistical inference: estimation**

**If** θ **is the unknown parameter of a given population and we want infer about it we need….**

A sampling statistics (g is a generic function)

and

which is the realization of the statistics based on a particular sample.

When the aim of the research is to infer on the unknown parameter of the distribution of the population, the sampling statistics is called ESTIMATOR of the unknown parameter (θ) and

is an ESTIMATE, so a realization or a value based on the sample values.

To understand the difference between T an t, the following detail on the simple random sample and i.i.d. distribution, are very useful. We will refer to the **Campionamento Casuale Semplice**(CCS) for simplicity but we could referer to a probabilistic sampling in general.

***Estrazioni i.i.d***

Y1, Y2, …Yn sono estratte casualmente dalla stessa popolazione

↓

La distribuzione marginale di ogni Yi è la stessa per ogni i

↓

Y1, Y2, …Yn si dicono *identicamente distribuite*

Dato lo schema di CCS, conoscere il valore di Y1, non fornisce indicazioni sulla probabilità di selezione di Y2

Prob (Y2|Y1)= Prob (Y2)

In generale

Prob (Yi|Yj)= Prob (Yi) per ogni i,j =1,…,n

Concludendo:

Nel c.c.s, *n* unità sono estratte casualmente da una popolazione e ogni unità ha la stessa probabilità di essere estratta. Il valore della v.c. Y per l’i-esima unità viene indicato con Yi. Siccome ogni unità ha la stessa probabilità delle altre di essere estratta, possiamo concludere che

Y1, Y2, …Yn sono estratte dalla stessa distribuzione e sono *indipendentemente distribuite* quindi sono ***indipendenti e identicamente distribuite (i.i.d)***

**Distribuzione campionaria della MEDIA CAMPIONARIA**

(Y1, Y2, …Yn) sono variabili casuali, la loro media è casuale.

Il valore della media campionaria cambia al variare degli elementi selezionati nel campione.

 Estimate

Quali valori può assumere e con quale probabilità? Dipende dallo schema di campionamento adottato (*sample design).*

Nel nostro caso CCS

Nell’ipotesi che le osservazioni campione (Y1, Y2, …Yn) siano i.i.d e che µY e σY2 siano rispettivamente media e varianza di Yi, allora

Valore atteso della media campionaria

Varianza della media campionaria

*..=*

N.B. tali risultati valgono indipendentemente dalla distribuzione di *Yi,* nessuna assunzione specifica è stata fatta.

***Distribuzione campionaria della MEDIA CAMPIONARIA per*** *Yi~N*

Dato che la media campionaria è una somma di v.c con Distribuzione Normale allora la media campionaria ha una Distribuzione Normale con media e varianza specificata sopra

In questo caso si parla di **distribuzione esatta** o **distribuzione per campioni finiti**.

***Distribuzione campionaria della MEDIA CAMPIONARIA per*** *Yi~*qualunque

**APPROSSIMAZIONE ALLA DISTRIBUZIONE CAMPIONARIA PER GRANDI CAMPIONI**

Se *Yi* non è Normale non è possibile determinare la distribuzione esatta e si determinerà la ***distribuzione asintotica.*** I due strumenti chiave sono

* La legge dei grandi numeri (permette di valutare il valore atteso di Yi)
* Il teorema del limite centrale (distribuzione della media campionaria)

***La legge dei grandi numeri e consistenza***

Nel 1713 Jakob Bernoulli, matematico svizzero, enunciò un teorema diventato famoso col nome di Legge dei Grandi Numeri. Questo teorema avvicina il concetto di probabilità statistica a quello di probabilità matematica fino a farli coincidere.

Infatti esso dice: **la frequenza relativa con cui un evento casuale si manifesta tende ad assumere il valore della sua probabilità matematica quanto più il numero delle osservazioni è alto**.

Cerchiamo di spiegare questo concetto con un semplice esempio.

La matematica afferma che lanciando una moneta esce Testa con probabilità p(Testa)=1/2; cioè nel 50% dei casi (probabilità matematica o teorica).

Ma siamo sicuri che lanciando una moneta 10 volte esca per 5 volte Testa?

Certamente no. Non c’è nessuna certezza che ciò avvenga; potrebbe venire Testa 6 volte su 10 oppure 4 su10 o anche solamente 2 su 10, come si fa a prevedere? Non si può!

Se però facciamo un esperimento che simuli il lancio casuale di una moneta per tantissime volte, e calcoliamo il rapporto fra il numero di volte che il risultato è stato Testa e il numero totale dei lanci, ci accorgiamo che al crescere del numero dei lanci, questo rapporto si avvicina sempre di più al **50%** stabilito dalla matematica.

Proviamo?!!!!!!!!

\*Programma STATA che simula il lancio di un dado ( per n volte: 1 T, 0 C)

set obs 10

 gen n= \_n

 \*assegna testa se il numero estratto>0.5

 gen x = uniform() > .5

gen heads = sum(x)

 gen pctheads = heads/n [valore della media campionaria]

 tab x

OUTPUT del programma che ripeto per diversi n (10, 100, 1000, 10000)

**n=10**

 x | Freq. Percent Cum.

------------+-----------------------------------

 0 | 4 40.00 40.00

 1 | 6 **60.00** 100.00

------------+-----------------------------------

 Total | 10 100.00

**n=100**

 x | Freq. Percent Cum.

------------+-----------------------------------

 0 | 52 52.00 52.00

 1 | 48 **48.00** 100.00

------------+-----------------------------------

 Total | 100 100.00

**n=1000**

 x | Freq. Percent Cum.

------------+-----------------------------------

 0 | 482 48.20 48.20

 1 | 518 **51.80** 100.00

------------+-----------------------------------

 Total | 1,000 100.00

**n=10000**

 x | Freq. Percent Cum.

------------+-----------------------------------

 0 | 5,023 50.23 50.23

 1 | 4,977 **49.77** 100.00

------------+-----------------------------------

 Total | 10,000 100.00

**Questo si verifica per qualunque**

**fenomeno casuale**

↓

“la media campionaria, che è una variabile casuale, è prossima alla media della popolazione con probabilità molto alta, quando n è grande”

Si dice che se Yi (i=1…n) sono i.i.d e la varianza di Yi è finita (la distribuzione non presenta outlier o sono improbabili).

* ***la media campionaria converge in probabilità alla media della popolazione al crescere di n***

o equivalentemente che

* ***la media campionaria è uno stimatore consistente della media della popolazione***

**Il teorema del limite centrale**

Con questo nome vengono presentati diversi enunciati che riguardano la **convergenza alla distribuzione normale di medie o somme di variabili aleatorie indipendenti**. Il teorema, nelle diverse formulazioni e generalizzazioni ha un ruolo importante, sia dal punto di vista teorico, sia, ancora di più, nelle applicazioni. Molti dei principali metodi utilizzati nell’inferenza statistica si basano su questo teorema.

Il teorema afferma che:

date *n* variabili aleatorie (Y1, Y2, …Yn) indipendenti con stessa media μY e stessa varianza σY2, la variabile aleatoria Z con



 “tende” ad avere una Distribuzione Normale di media 0 e varianza 1; dove “tende” significa che, all’aumentare di n, la distribuzione di Z assomiglia sempre più a quella di una variabile aleatoria N(0,1).

Il teorema, come già detto, è un risultato teorico e contiene, nel suo enunciato un limite per *n* che tende a infinito. Per renderlo operativo ci chiediamo quanto deve essere grande *n* perché il risultato valga?

La qualità dell’approssimazione dipende dalla distribuzione delle Yi che compongono la media, comunque in generale diciamo per n>30.

**Statistical inference: estimator properties**

* Just as there are multiple possible sample statistics, there are many possible estimators. If we are interested, for example, in the population’s mean, then we could use the mean, the median, or the mode of the sample as our estimator.
* How do statisticians decide which estimator is best?

Good estimators usually have two qualities:

they are **unbiased** and **efficient**.

**Statistical inference: estimator properties**

An unbiased estimator has its sampling distribution "centered" around the parameter; in other words, the expected value of the statistic is identical to the population parameter.

is unbiased is

This means that if we would draw all the possible samples, and for each of them we calculate t=g(x), the mean of all t would be exactly equal to .

If =B is a Biased estimator of

**Statistical inference: estimator properties**

An estimator is defined efficient if it has the minimum variance among the estimators that have the same sample size.

Unbiased and efficient estimators give us accurate and precise estimates of population parameters.

