UNIVERSITÀ
DEGLI STUDI
FIRENZE

AA 2018-19

INVENTARI FORESTALI

Dispensa 3

Inferenza statistica

Docente:

Prof. Gherardo CHIRICI
gherardo.chirici@unifi.it

Il campionamento

Scopo del campionamento è quantificare uno o più parametri statistici della popolazione
Il valore del parametro inferito dal campione è detto *stima*
La rappresentatività di un campione è data dalla sua capacità di fornire un'adeguata idea della popolazione dalla quale è stato estratto
Secondo l'approccio *design-based* tale obbiettivo dipende esclusivamente dalla modalità con cui un campione è estratto dalla popolazione
Per raggiungere questo obiettivo è necessario che l'estrazione sia casuale

Il campionamento soggettivo

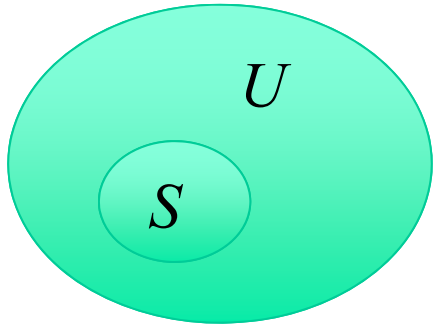
Nel campionamento soggettivo è l'operatore a selezionare le unità della popolazione da includere nel campione
Dato che in questo caso la probabilità di inclusione delle unità è incognita non è possibile il calcolo degli stimatori
Per questo i parametri statistici calcolati a partire da un campione estratto soggettivamente devono essere ritenuti intrinsecamente DISTORTI
Il campionamento soggettivo permette di ottenere informazioni esclusivamente sul campione e non sulla popolazione da cui è stato estratto
Non ha una minima numerosità del campione

Il campionamento probabilistico

Nel campionamento probabilistico le unità della popolazione da includere nel campione sono stabilite in modo oggettivo
L'oggettività è ottenuta attraverso una estrazione casuale
In questo caso la probabilità di inclusione delle unità della popolazione nel campione è nota ed è quindi possibile inferire dai parametri del campione ottenendo una stima dei parametri della popolazione
Il campionamento probabilistico deve ritenersi intrinsecamente corretto e deve essere applicato il più possibile
Richiede una numerosità campionaria spesso molto alta

Teoria del campionamento

$$U = \{u_1, u_2, \ldots, u_N\}$$



Universo o *popolazione*: insieme di N elementi omogenei rispetto a qualche caratteristica

Campione: sottoinsieme S di U di $n \leq N$ unità distinte scelte dall'universo tramite una procedura di selezione detta *piano di campionamento* o *disegno*.

$$S \subset U$$

Il campionamento è inevitabile quando non è nota la lista degli elementi che costituiscono la popolazione.
Si ricorre al campionamento anche quando N è noto per motivi economici.

$$\textit{Frazione di campionamento} = \frac{n}{N}$$

Variabili e parametri

Si dice *variabile* un aspetto quantitativo Y delle unità che deve essere analizzato.
Si indichino allora con $y_1, y_2, \ldots, y_N$ i valori assunti da tale variabile in corrispondenza delle N unità della popolazione.

Si dice *parametro* della popolazione una funzione dei valori assunti da Y nella popolazione,

$$\theta = \theta(y_1, y_2, \ldots, y_N)$$

Esempi di parametri calcolati sulla popolazione

Totale della popolazione:

$$T = \sum_{j=1}^{N} y_j$$

Media della popolazione:

$$\mu = \frac{T}{N} = \frac{1}{N} \sum_{j=1}^{N} y_j$$

Varianza della popolazione:

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{N} \left(y_j - \mu \right)^2$$

Stimatori campionari

Ogni elemento j della popolazione ha probabilità di inclusione π_j nel campione S

$$n = \sum_{j=1}^{N} \pi_j$$

Lo scopo di un'indagine campionaria è di valutare un parametro della popolazione sulla base dell'osservazione sul campione.
Si definisce *stimatore* del parametro θ una funzione che per ogni campione S associa un numero reale detto stima di θ:

$$\hat{\theta} = t(S)$$

Uno stimatore è *corretto* quando:

$$E(\hat{\theta}) = \sum_{\theta_k \in \Theta} \theta_k p(\theta_k) = \theta$$

Altrimenti lo stimatore è *distorto* della quantità:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

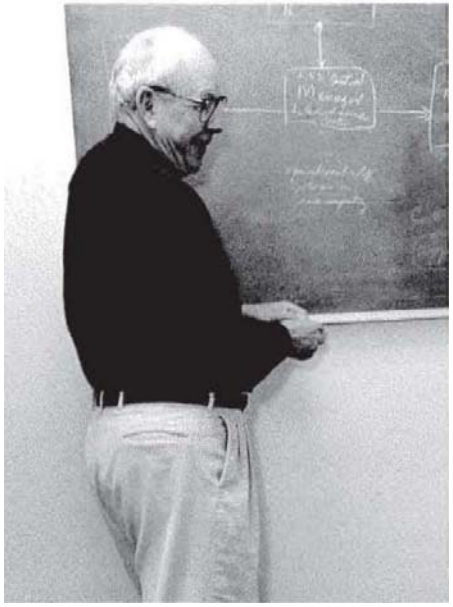
La precisione di uno stimatore è determinata dalla concentrazione della sua distribuzione di probabilità attorno al valore del parametro da stimare.
Se lo stimatore è corretto, la sua precisione è completamente determinata dalla sua varianza.

Proprietà degli stimatori

Le proprietà di uno stimatore si ottengono dalla sua distribuzione di probabilità, cha a sua volta può essere determinata solo conoscendo l'intera popolazione.
In pratica le proprietà di uno stimatore sono quindi incognite.
Stimatori non distorti per le quali è possibile l'espressione analitica della varianza sono definiti *stimatori lineari omogenei* (LO).
Gli stimatori LO permettono la stima corretta dei parametri e la quantificazione della precisione della stima.
Il più utilizzato degli stimatori LO è quello di Horvitz-Thompson (HT) proposto per la prima volta nel 1952.

Horvitz, D. G.; Thompson, D. J. (1952) "A generalization of sampling without replacement from a finite universe", Journal of the American Statistical Association, 47, 663–685, . JSTOR 2280784

Dr. Daniel Goodman Horvitz
1921 - 2008



Dr. Donovan J. Thompson
1919 - 1991

A GENERALIZATION OF SAMPLING WITHOUT REPLACEMENT FROM A FINITE UNIVERSE*

D. G. Horvitz† and D. J. Thompson
Iowa State College

This paper presents a general technique for the treatment of samples drawn without replacement from finite universes when unequal selection probabilities are used. Two sampling schemes are discussed in connection with the problem of determining optimum selection probabilities according to the information available in a supplementary variable. Admittedly, these two schemes have limited application. They should prove useful, however, for the first stage of sampling with multi-stage designs, since both permit unbiased estimation of the sampling variance without resorting to additional assumptions.

With the development of more efficient sampling systems, the system including both the sample design and the method of estimation. One sampling system is said to be more efficient than another if the variance or mean square error of the estimate with the first system is less than that of the second, provided the cost of obtaining the data and results is the same for both. The development of stratified, multi-stage, multiphase, cluster, systematic, and other sample designs beyond

The possibility of using unequal probabilities for selecting the sample elements from the universe as a means of increasing precision perhaps received its first impetus for applied sampling from Hansen and Hurwitz [2] in 1943. They introduced the selection of primary units (in a subsampling scheme) with probabilities proportionate to some measure of their size and presented the appropriate theory. Their sampling scheme was confined (when sampling without replacement) to samples of one primary unit per stratum, however, the theory not having been extended beyond this point. More recently, Midzuno [6] has generalized the Hansen and Hurwitz approach to sampling a combination of n elements of the universe with probability proportionate to some measure of size of the combination. Madow [5] has made some contributions to the theory of the systematic selection of several clusters with probability proportionate to a measure of size.

* Journal Paper No. J 2139 of the Iowa Agricultural Experiment Station, Ames, Iowa, Project 1005. Presented to the Institute of Mathematical Statistics, March 17, 1951.

† Now at the University of Pittsburgh.

Introduzione al teorema del limite centrale

Da una determinata popolazione possono essere estratti un certo numero di possibili campioni:

$$T = \frac{N!}{n!(N-n)!}$$

Ripetendo il campionamento sulla stessa popolazione per ciascun parametro statistico si avrà una distribuzione delle stime campionarie.
Per ogni distribuzione delle stime campionarie è possibile calcolare i relativi parametri statistici.
Esempio: della media campionaria può essere calcolata la media, la varianza, la deviazione standard, ecc.

Il teorema del limite centrale

Per popolazioni infinite la distribuzione della media campionaria $\hat{y}$ è approssimativamente una distribuzione normale con media $\mu_{\hat{y}}$ e varianza $\sigma^2_{\hat{y}}$ dove:

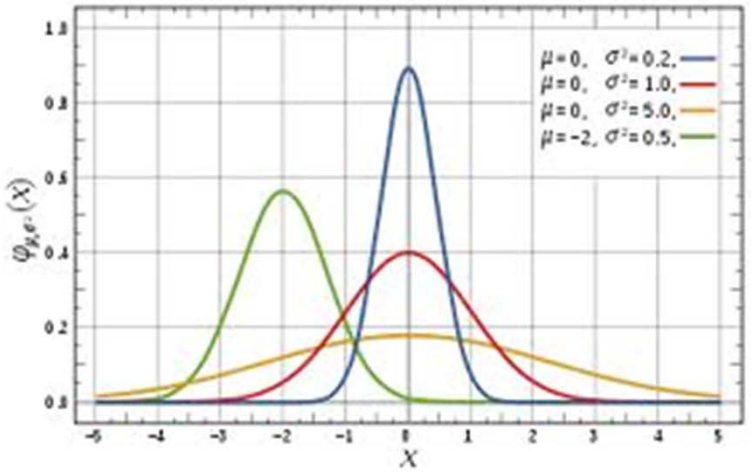
$$\mu_{\hat{y}} = \mu_y$$

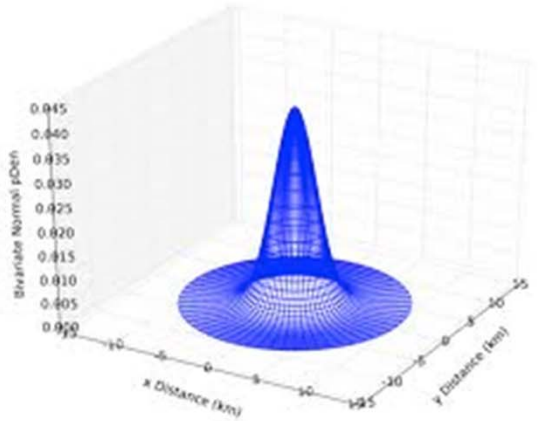
$$\sigma^2_{\hat{y}} = \frac{\sigma^2_y}{n}$$

La distribuzione normale (detta gaussiana) ha funzione di densità:

$$f(x) = ae^{-(x-b)^2/c^2}$$

$$n = \frac{1}{\sigma_y\sqrt{2\pi}} e^{\frac{-\frac{1}{2}(y-\mu_y)^2}{\sigma_y^2}}$$





Normal,
Bell-shaped Curve

Percentage of cases in 8 portions of the curve	.13%	2.14%	13.59%	34.13%	34.13%	13.59%	2.14%	.13%	
Standard Deviations	-4σ	-3σ	-2σ	-1σ	0	+1σ	+2σ	+3σ	+4σ
Cumulative Percentages		0.1%	2.3%	15.9%	50%	84.1%	97.7%	99.9%	
Percentiles		1	5	10	20 30 40	50	60 70 80	90	95 99
Z scores	-4.0	-3.0	-2.0	-1.0	0	+1.0	+2.0	+3.0	+4.0
T scores		20	30	40	50	60	70	80	
Standard Nine (Stanines)		1	2	3	4	5	6	7	8 9
Percentage in Stanine		4%	7%	12%	17%	20%	17%	12%	7% 4%

Stimatori HT

$$n = \sum_{j=1}^{N} \pi_j$$

Stimatore HT del totale:

$$\hat{T}_{HT} = \sum_{j \in S} \frac{y_j}{\pi_j}$$

$$\pi_j = \frac{n}{N}$$

se la popolazione è finita, ovvero se N è noto

con varianza:

$$V(\hat{T}_{HT}) = \sum_{j=1}^{N} \sum_{h>j} (\pi_j \pi_h - \pi_{jh}) \left(\frac{y_j}{\pi_j} - \frac{y_h}{\pi_h} \right)^2$$

Stimatore HT della media:

con varianza:

N noto

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{N}$$

$$V(\hat{\mu}_{HT}) = \frac{V(\hat{T}_{HT})}{N^2}$$

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{\hat{N}_{HT}} \qquad \hat{N}_{HT} = \sum_{J \in S} \frac{1}{\pi_j}$$

N ignoto

$$V(\hat{\mu}_{HT}) = \frac{V(\hat{T}_{HT})}{\hat{N}_{HT}^2}$$

Gli stimatori di HT sono sviluppati per il Campionamento Casuale Semplice (CCS)

Lo stimatore HT del totale è:

$$\hat{T}_{CCS} = \sum_{j \in S} \frac{y_j}{\pi_j} = \sum_{j \in S} y_j \frac{N}{n} = N \frac{1}{n} \sum_{j \in S} y_j = N\bar{y}$$

$$V(\hat{T}_{CCS}) = (N-n) N \frac{\delta^2}{n}$$

Dove $\bar{y}$ è la media calcolata nel campione

Lo stimatore HT della media:

$$\hat{\mu}_{CCS} = \frac{\hat{T}}{N} = \frac{N\bar{y}}{N} = \bar{y}$$

Nel CCS la stima della media della popolazione è uguale alla media del campione.

$$V(\hat{\mu}_{CCS}) = \frac{N-n}{N} \frac{\delta^2}{n}$$

La varianza degli stimatori diminuiscono all'aumentare della numerosità *n* del campione, annullandosi per *n* = *N*

Errore standard della stima

L'errore standard della stima (Standard Error, ES) definisce la variabilità delle stime e permette di calcolare l'intervallo di confidenza della stima.
L'errore standard è la radice quadrata della varianza dell'errore.
La varianza dell'errore e l'errore standard possono essere stimati per qualsivoglia parametro stimato (il totale, la media, ecc.). Qui facciamo il caso della media.
Se estraiamo tutti i possibili campioni da una popolazione, le stime di HT della media della popolazione che si ottengono da ogni campione, se l'estrazione del campione è casuale, si distribuiranno secondo una distribuzione normale. La media di tutte queste possibili medie è uguale alla media della popolazione (il valore vero).
La varianza di tutte queste possibili medie è la varianza dell'errore della media, la deviazione standard di tutte queste possibili medie è l'errore standard.
Dato però che dalla popolazione normalmente estraiamo solo un campione tra tutti i possibili potremo calcolare solo la stima dell'errore standard.

$$SE_{\hat{\bar{y}}} = \sqrt{\widehat{var}(\bar{y})} = \sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}}$$

Errore Standard della stima della media

Varianza della media calcolata sul campione, in quanto sappiamo che lo stimatore di HT della media della popolazione è la media del campione

Lo stimatore di HT della varianza è la varianza media sul campione

Si noti quindi che l'errore standard della stima dipende linearmente dalla variabilità nel campione (la deviazione standard *S* al numeratore) ma in forma quadratica con la numerosità campionaria (*n* al denominatore).
A parità di altri fattori per dimezzare SE bisogna quadruplicare *n*)

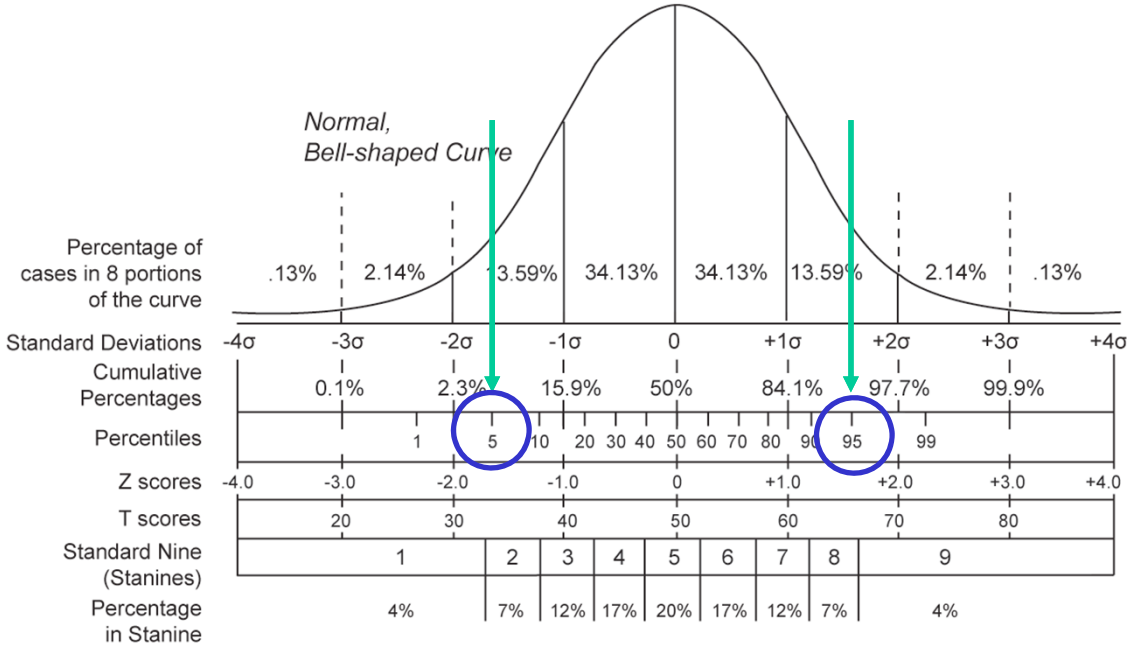
Intervallo di confidenza

L'intervallo di confidenza è un intervallo di valori in cui ci si aspetta ricada la stima con una certa probabilità. La probabilità (o sicurezza statistica) è in genere posta al 95% (comunque tra 90% e 99%). Quando la numerosità del campione è sufficientemente ampia (n>>30) allora gli estremi dell'intervallo di confidenza sono dati da:

$$\pm SE * t$$

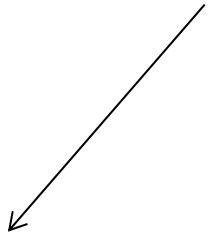
SE è l'errore standard della stima
t è il valore critico del t di Student, pari a 1,96 per la sicurezza statistica del 95%, e a 2,58 per il 99%

$$SE\% = SE / stima$$



Accuratezza: precisione, distorsione

Accuratezza: scostamento tra il valore di una stima campionaria di un dato parametro statistico e il suo valore vero incognito

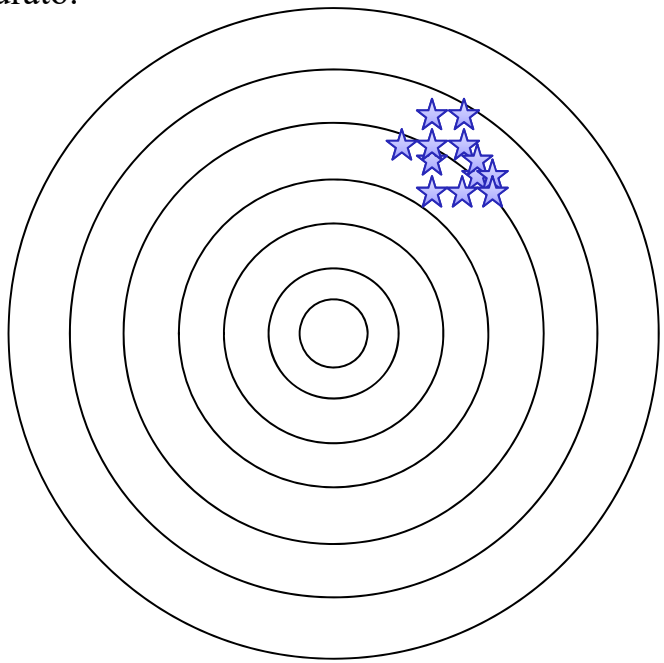


Precisione: dimensione degli scostamenti tra singole stime campionarie e il loro valore atteso

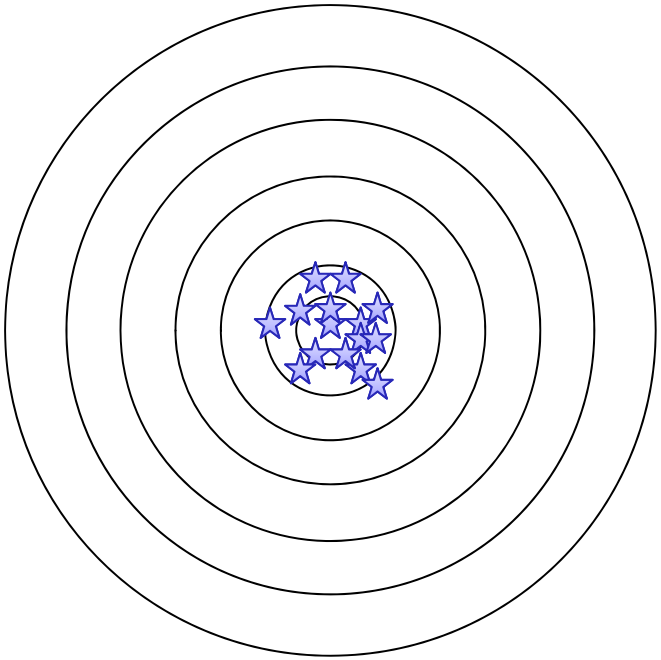
Distorsione: dimensione dello scostamento tra il valore atteso delle stime campionarie e il valore vero incognito del parametro statistico

Uno stimatore è accurato se è preciso e non distorto

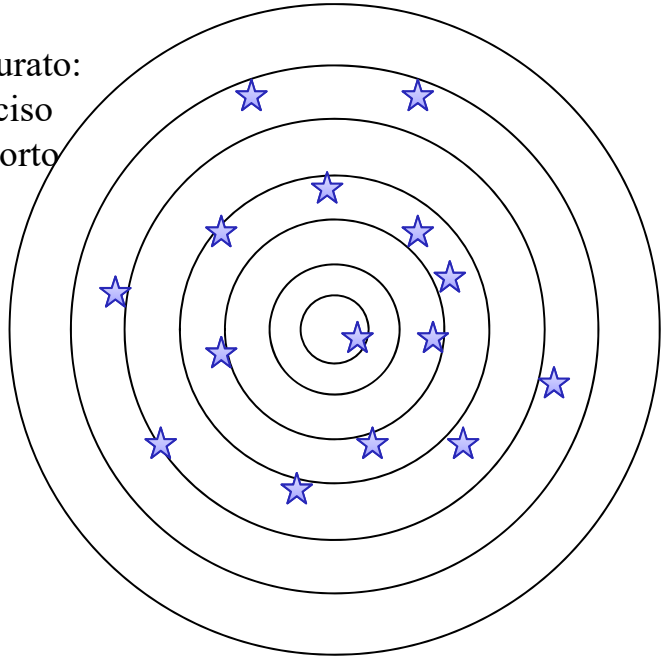
Non accurato:
Preciso
Distorto

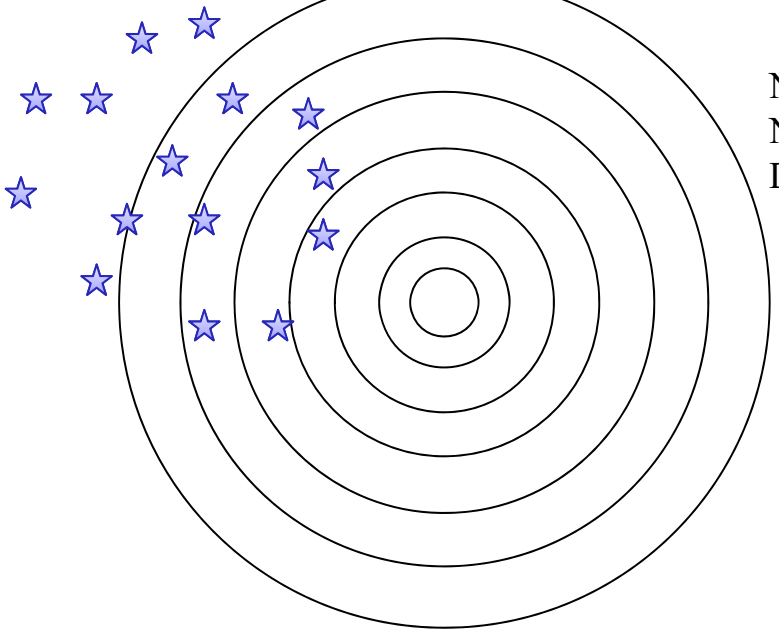


Accurato:
Preciso
Non distorto



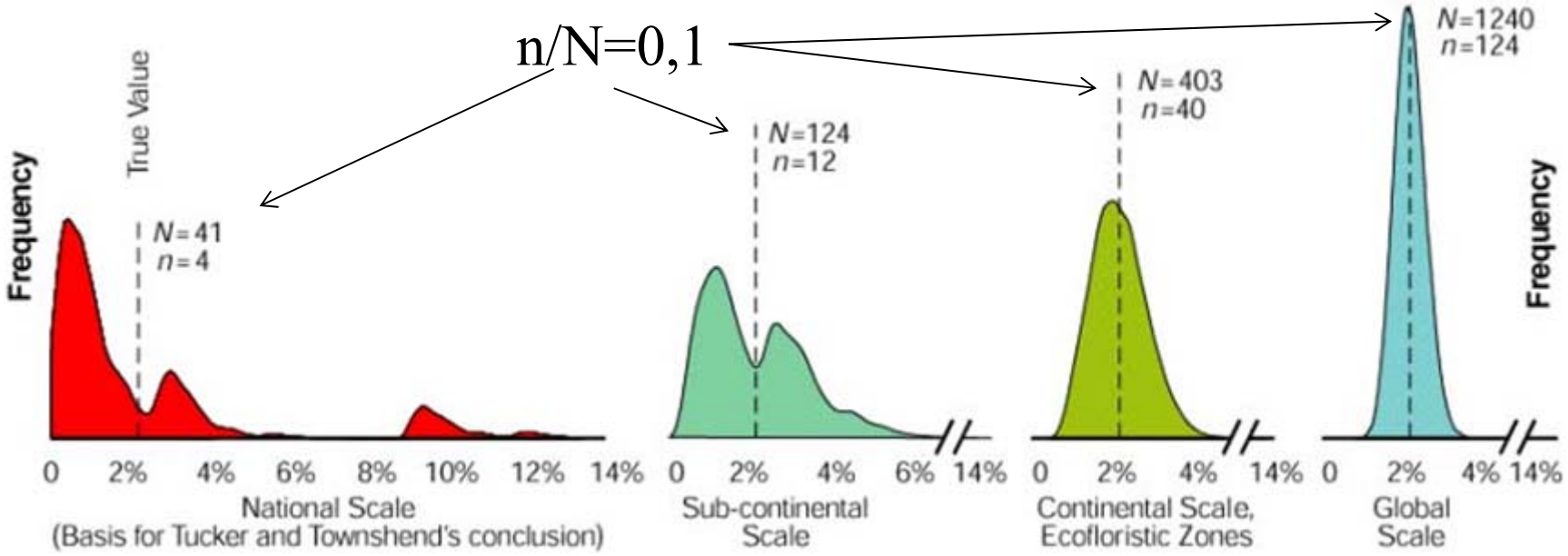
Non accurato:
Non preciso
Non distorto

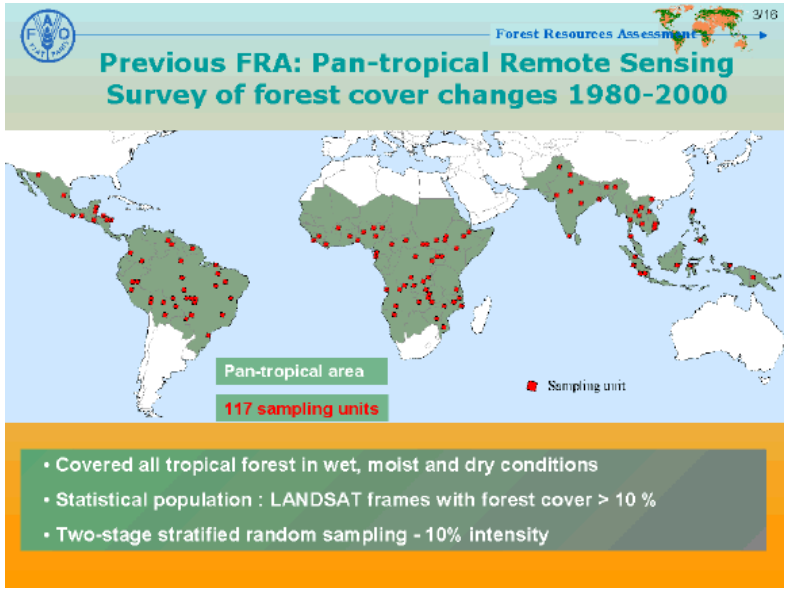




Non accurato:
Non preciso
Distorto

Dimensione del campione e intensità del campionamento





Sampling for estimating global deforestation
http://www.fao.org/docrep/007/ad058e/AD058E02.htm