UNIVERSITÀ DEGLI STUDI FIRENZE

AA 2018-19

INVENTARI FORESTALI

Dispensa 4

Disegni di Campionamento

Docente:

Prof. Gherardo CHIRICI
gherardo.chirici@unifi.it

Ogni variabile oggetto di stima ha le proprie unità statistiche che vengono distribuite secondo uno specifico disegno di campionamento.
Nel corso della progettazione dell'inventario si cercherà di far coincidere i disegni di campionamento delle diverse variabili.

Variabile		Unità di campionamento
Superficie Bosco Tipologie di bosco Altri uso del suolo	➡	Punto
Massa o volume Provvigione Biomassa Legno morto	➡	Area di saggio

STIMA DELLA SUPERFICIE FORESTALE

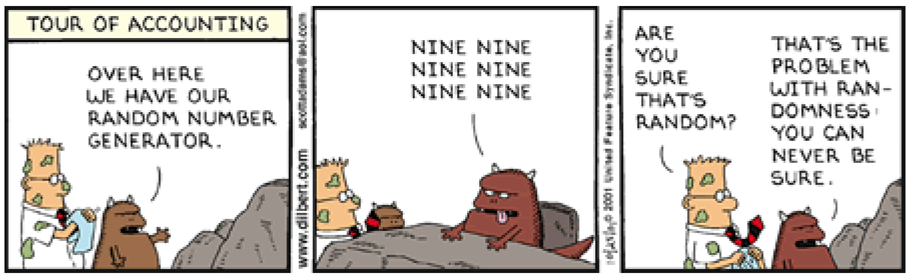
CAMPIONAMENTO CASUALE SEMPLICE

Nel campionamento casuale semplice (Simple Random Sampling – SRS) tutte le unità di campionamento hanno uguale probabilità di essere estratte, se l'estrazione è fatta con reinserimento o da popolazioni infinite.

E' presentato solo didascalicamente perché la base per lo sviluppo degli stimatori di HT

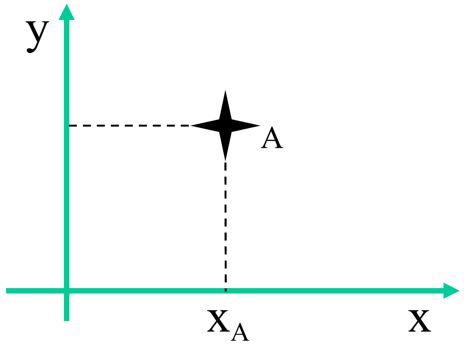
Nella pratica inventariale non viene utilizzato perchè poco vantaggioso in termini costi benefici.

Da infatti problemi di *clusterizzazione* della distribuzione delle unità di campionamento



In ambito inventariale, la modalità operativa per realizzare la stima della superficie forestale è un campionamento per punti

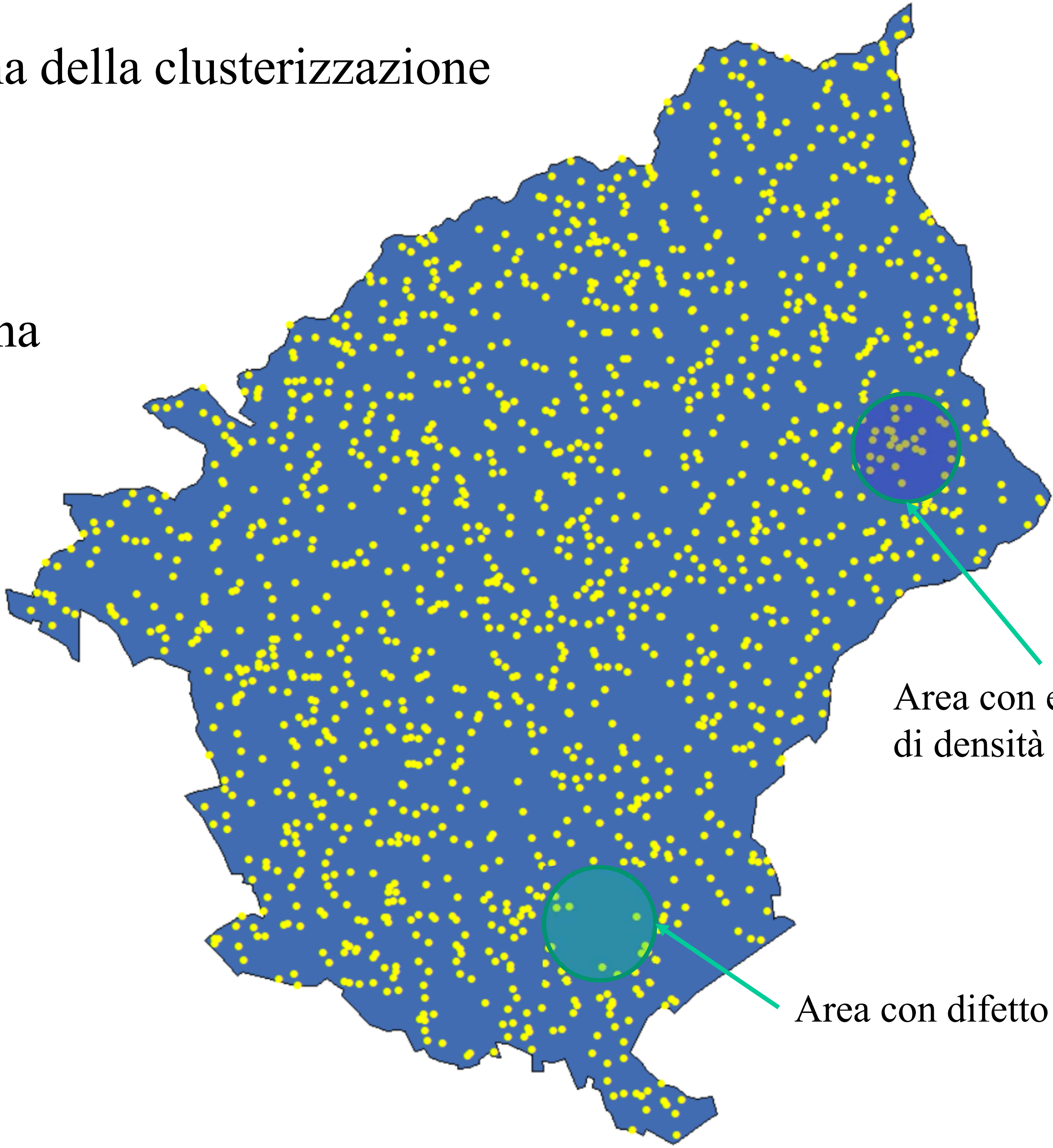
le unità statistiche sono identificate in cartografia da una coppia di coordinate; le unità statistiche con coordinate corrispondenti alle coppie di numeri casuali estratti sono incluse nel campione.



$\left.\begin{matrix} x_A \\ y_A \end{matrix}\right\}$ Coppia di coordinate casuali che identifica l'unità di campionamento A

Il problema della clusterizzazione

n=1435
A=1313 ha



Nel caso più semplice si dovrà stimare la superficie del bosco.
Data una certa definizione oggettiva di bosco.
La variabile y potrà quindi avere valore y=0 nel caso in cui l'unità di campionamento non sia in bosco, y=1 nel caso in cui l'unità sia in bosco.
Data una certa area oggetto dell'inventario di estensione nota A.
Essendo l'unità di campionamento (il punto) adimensionale, il numero delle possibili unità campionarie in A (N, la popolazione) è infinito.
Sarà quindi noto solo n, il numero delle unità del campione. Indichiamo n_f il numero di unità in bosco (in cui y=1) e n_{nf} le unità non in bosco (in cui y=0).
Dal generico stimatore di HT della media:

Stima del coefficiente di boscosità

$$\hat{\mu}_{HT} = \frac{\sum_{j=1}^{n} y_j}{n} = \frac{1}{n}\left(\left(n_f * 1\right) + \left(n_{nf} * 0\right)\right) = \frac{n_f}{n} = \hat{p}$$

$$var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$$

$$SE(\hat{p}) = \sqrt{var(\hat{p})}$$

$$\sigma^2 = p * q = p * (1 - p)$$

where $q = (1 - p)$. The variance, obviously, is a function of the mean *p* only!

The variance $\sigma^2 = p * q$ derives directly from the formula used for calculating the population variance in general:

$$\sigma_p{}^2 = \frac{\sum_i (y_i - \bar{y})^2}{N}$$

where $\bar{y}$ is the true proportion *p* and there are only two possibility of observations: (1) $y = 1$ for forest points which occur n_f times where

$$p = \frac{n_f}{N}$$

and $y = 0$ for non-forest points which occur $n_n = (n - n_f)$ times and

$$q = \frac{n_n}{N} = \frac{N - n_f}{N}$$

We rewrite then

$$\sigma^2 = \frac{\sum_i (y_i - \bar{y})^2}{N} = \frac{\sum_i (y_i - p)^2}{N} \frac{n_f(1-p)^2 + n_n(0-p)^2}{N}$$

$$= \frac{n_f}{N}(1-p)^2 + \frac{N - n_f}{N}(0-p)^2 = p(1-p)^2 + (1-p)^2 = p(1-p) = p * q$$

which is the parametric variance. The sample based estimation of that variance has $(n-1)$ degrees of freedom so that the sum of squares needs to be divided by $(n-1)$. We may use the same re-arrangement as before and care for the different denominator simply by introducing the factor $\mathrm{n}/(\mathrm{n}\text{-}1)$

$$s^2 = \frac{\sum_i (y_i - \hat{p})^2}{n} * \frac{n}{n-1} = \hat{p} * \hat{q} * \frac{n}{n-1}$$

The estimated error variance of $\hat{p}$ derives as usual from

$$v\hat{a}r(\hat{p}) = \frac{s^2}{n}$$

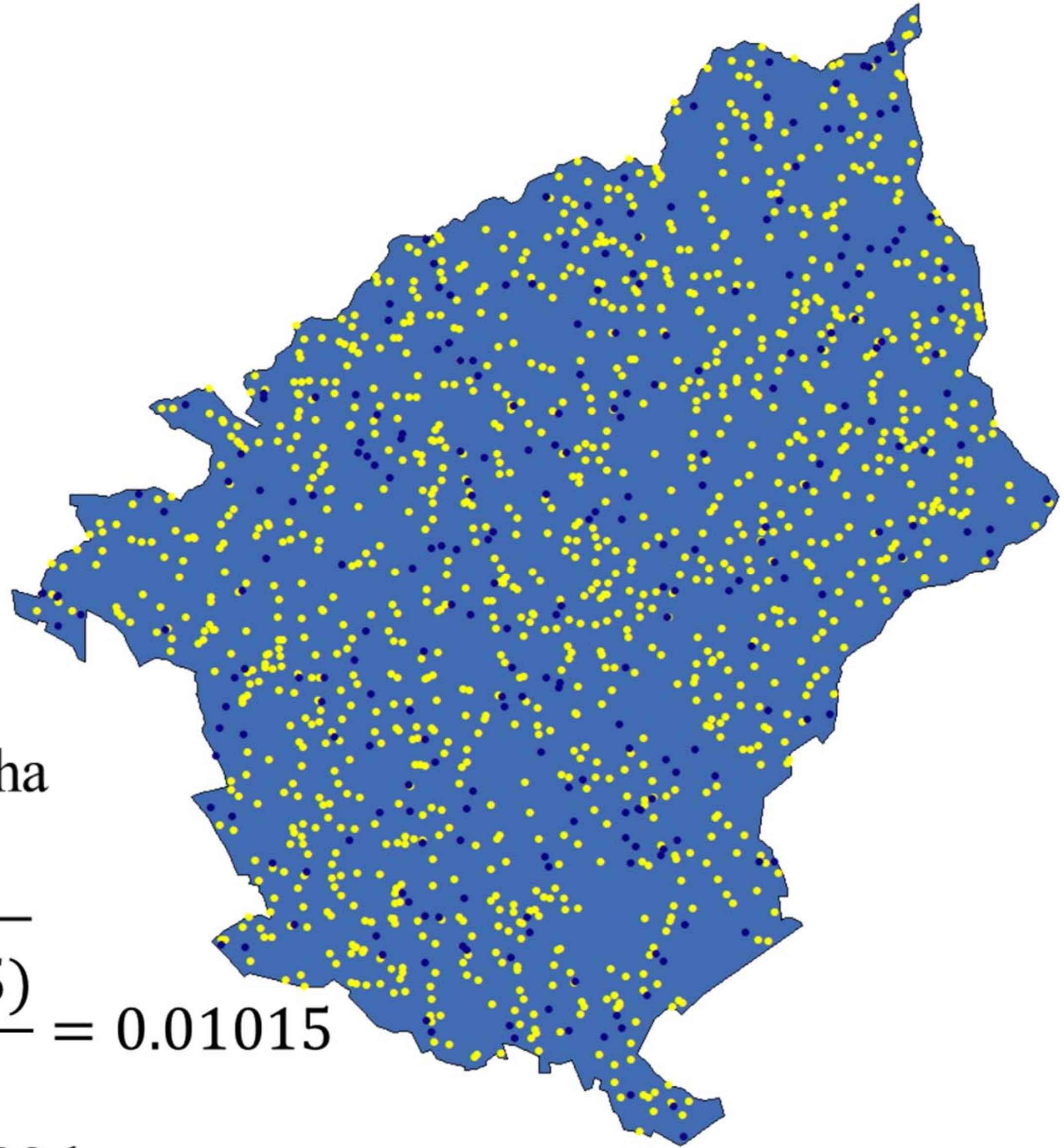
which in this case is

$$v\hat{a}r(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}$$

n=1435
A=1313 ha

n_f=1176

n_{nf}=259



$$\hat{p} = \frac{n_f}{n} = \frac{1176}{1435} = 0.8195$$

$$\hat{A}_f = 1313 * 0.8195 = 1076{,}02 \text{ ha}$$

$$SE(\hat{p}) = \sqrt{\frac{0.8195(1-0.8195)}{1434}} = 0.01015$$

$$SE(\hat{A}_f) = 0.01015 * 1313 = 13{,}33 \text{ ha}$$

$$SE\%(\hat{A}_f) = \frac{13{,}33}{1313} * 100 = 1{,}0156\%$$

STIMA DELLA PROVVIGIONE

Valore totale

Lo stimatore campionario del valore totale è pari a: $\hat{Y} = N\bar{y}$

Nel caso di inventari su ampie superfici con aree campione ordinarie di superficie unitaria pari ad *a*, essendo *N* circa pari a *A*/*a* (dove *A* è la superficie totale della popolazione considerata), la precedente formula può essere espressa nella formula:

$$\hat{Y} = \frac{A}{a}\bar{y}$$

Lo stimatore campionario della varianza del valore totale è pari a: $s_y^2 = N^2 s_{\bar{y}}^2$
o, se il campionamento è stato condotto su aree campione di superficie unitaria pari ad *a*:

$$s_y^2 = \left(\frac{A}{a}\right)^2 s_{\bar{y}}^2$$

posto cha *A* sia nota senza errore.

L'intervallo fiduciario della stima di $\hat{Y}$ è pari a $y \pm N\left(ts_{\bar{y}}\right)$ o, se il campionamento è stato condotto su aree campione di superficie unitaria pari ad a:

$$y \pm \frac{A}{a} ts_{\bar{y}}$$

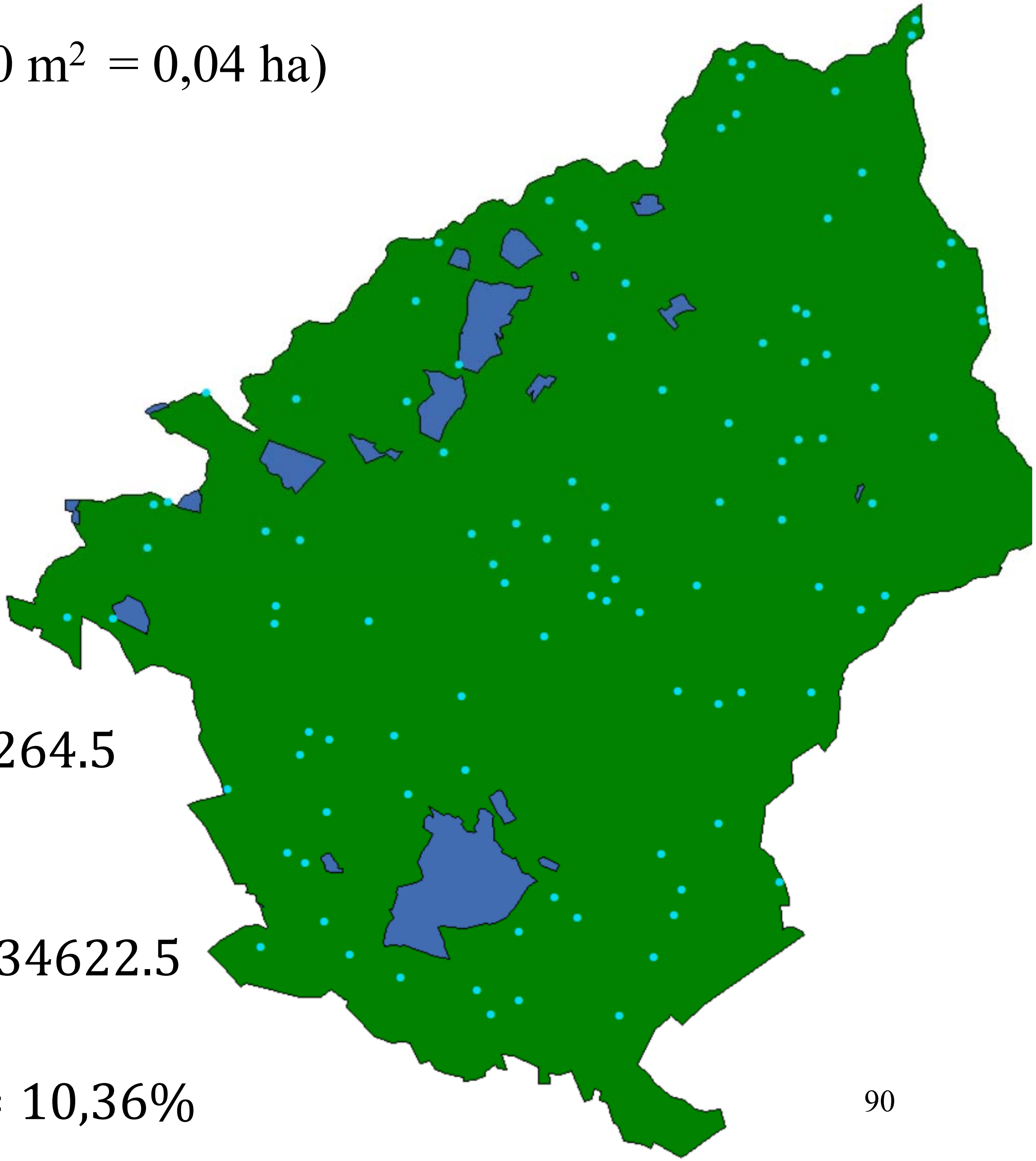
n=102 (ogni area di 400 m^2 = 0,04 ha)

a=0,04 ha

A=1259 ha

$\bar{y} = 10,62\, m^3$

$\sigma = 1,1\, m^3$



$$\hat{T} = \frac{1259}{0,04} 10,62 = 334264.5$$

$$\hat{\mu} = 10,62$$

$$SE(\hat{T}) = \frac{1259}{0,04} * 1,1 = 34622.5$$

$$SE\%(\hat{T}) = \frac{34622.5}{334264.5} = 10,36\%$$

Scelta della numerosità del campione

A parità di altre condizioni, tanto più numeroso è il campione, tanto maggiore è la precisione della stima. Dimensionare la numerosità del campione significa dunque prefissare la precisione delle stime che si vogliono ottenere in modo che siano caratterizzate da un errore di campionamento non superiore alla soglia massima tollerata ec_0:

$$n_0 = \left(\frac{ts_y}{ec_0} \right)^2$$

dove:

s_y= valore presunto della deviazione standard del parametro *y* nella popolazione considerata.

t = valore critico del *t di Student* corrispondente al prescelto livello di sicurezza statistica e agli appropriati gradi di libertà.

df	Area nella Coda di Destra sotto la Curva di Distribuzione *t*					
	0.10	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
∞	1.282	1.645	1.960	2.326	2.576	3.090

La suddetta formula può anche essere espressa secondo la notazione percentuale:

$$n_0 = \left(\frac{tCV\%}{ecp_0} \right)^2$$

dove:

CV% = valore presunto del coefficiente di variazione, espresso in termini percentuali, dell'attributo x nella popolazione considerata;

ecp_0 = errore percentuale massimo tollerato, dato dal rapporto percentuale tra il valore massimo tollerato dell'errore di campionamento e la media campionaria.

Dall'esame delle suddette formule, si rileva che il dimensionamento statistico del campione può essere operato solamente stabilendo a priori il valore di s_y, cioè avendo a disposizione informazioni preliminari sulla variabilità dell'attributo oggetto di stima. In altre parole, i valori di s_y vanno individuati tramite un campionamento preliminare, detto *campionamento pilota*.

Il valore di t da inserire nelle formule del dimensionamento campionario può essere desunto dalle tavole del *t di Student*, in prima approssimazione in corrispondenza di $n_{pil}-1$ gradi di libertà, dove n_{pil} è il numero di osservazioni del campione pilota. Per individuare più correttamente un valore di t esattamente commisurato alla dimensione numerica del campione definitivo, i gradi di libertà di t possono essere determinati con procedimento iterativo.

Osservazioni sul dimensionamento campionario

Il dimensionamento della numerosità del campione conduce solamente a una stima del numero minimo di unità campionarie necessario per contenere l'incertezza campionaria entro una data soglia massima tollerata. A esempio, definendo la numerosità del campione a un livello di sicurezza statistica del 95% esisterà sempre 1 possibilità su 20 che il campione estratto fornisca una stima del parametro di interesse al di fuori dell'intervallo fiduciario massimo tollerato, anche se la numerosità del campione sia stata, per ipotesi, esattamente quantificata sulla base del valore vero della varianza dell'attributo oggetto di interesse nella popolazione considerata.

Vantaggi e svantaggi del campionamento casuale

Vantaggi

- ❖ rispetta il requisito fondamentale degli schemi di campionamento probabilistico, ovvero l'estrazione delle unità campionarie è condotta con criteri di pura casualità;
- ❖ permette di variare in ogni momento, quando se ne presentino le necessità e le condizioni, la numerosità del campione;
- ❖ permette di dimensionare il campione in modo da ottenere la precisione di stima desiderata.

Svantaggi

- ❖ il campionamento casuale semplice può comportare il rischio di non tenere conto in modo adeguatamente uniforme di tutta la popolazione (alcune parti della popolazione possono risultare intensamente campionate, altre non campionate affatto). Per tale motivo il campionamento casuale semplice è affidabile soprattutto quando la popolazione è relativamente omogenea;
- ❖ nell'inventariazione delle risorse forestali la dislocazione delle unità campionarie risulta alquanto onerosa e disagevole, soprattutto se confrontata con quella di un campionamento di tipo sistematico.

CAMPIONAMENTO SISTEMATICO

Si consideri una popolazione le cui unità si presentano ordinate secondo un qualche criterio (ad esempio, secondo una seriazione spaziale o temporale).

Se, per ottenere da quella popolazione un campione di una data numerosità, si estrae la prima unità campionaria casualmente e le altre sono invece scelte a intervalli regolari a partire dalla prima unità estratta si realizza un campionamento cosiddetto *sistematico*.

Estrazione di un campione casuale

Assumendo che le unità statistiche di una popolazione siano disposte in sequenza e numerate da 1 a N e che sia definito un intervallo di campionamento Y ($<N$), scegliendo un numero k, compreso tra 1 e N, che individua la prima unità campionaria, il campione sistematico sarà costituito dalle unità numerate con:

$$k,\ k+Y,\ k+2Y,\ \ldots,\quad \text{e}\quad k-Y,\ k-2Y,\ \text{ecc.}$$

A rigore, il numero k deve essere scelto in modo completamente casuale, benché nella pratica del rilevamento delle risorse forestali viene spesso evitata l'estrazione casuale della prima unità campionaria, assumendo che esista una completa indipendenza tra la localizzazione delle unità campionarie e le variabili osservate.

Nelle applicazioni inventariali degli schemi di campionamento sistematico, le unità campionarie a terra sono in genere configurate come *strisce campione* e soprattutto come *aree campione*. Nei rilievi condotti su immagini telerilevate vengono in genere impiegati *fototransect* (omologhi alle strisce campione) e soprattutto *fotopunti*.

Strisce campione

La superficie da inventariare viene idealmente suddivisa in N strisce giustapposte. Le unità campionarie vengono prescelte a intervalli regolari, cioè una striscia ogni Y strisce. In pratica, si estrae un numero casuale compreso tra *1* e N e la striscia corrispondente viene scelta come unità campionaria iniziale: le altre strisce campione sono selezionate a intervalli regolari, una ogni Y strisce, nelle due direzioni a partire dalla striscia campione iniziale.

Nel caso in cui è sconosciuto il valore di N, ovvero del numero di strisce che compone la popolazione, si può operare selezionando direttamente la prima striscia campione, a partire da un dato margine, in corrispondenza di un numero casuale compreso tra *1* e Y (questo modo di operare può produrre un leggero sotto- o sovra-campionamento se N non risulta un multiplo esatto di Y).

Il campionamento sistematico a strisce è usato negli inventari delle foreste tropicali o comunque dove si hanno difficoltà di accesso all'interno della foresta.

Aree campione

La superficie da inventariare viene idealmente suddivisa in N aree giustapposte. Ciascuna area può essere rappresentata come una cella di intersezione delle righe e delle colonne di un reticolo. Le unità campionarie vengono prescelte a intervalli regolari, cioè un'area ogni Y aree secondo le due dimensioni del reticolo (righe, colonne).

L'estrazione delle n aree campione può essere condotta in maniera analoga a quanto descritto per le strisce campione, con la differenza che in questo caso si ha che fare con due dimensioni invece che con una. Si estrae un numero casuale compreso tra *1* e *Nr,* dove *Nr* rappresenta il numero delle righe, e un numero casuale compreso tra *1* e *Nc,* dove *Nc* rappresenta il numero delle colonne: questi due numeri casuali identificano l'area che viene scelta come unità campionaria iniziale. Le altre aree campione sono quindi selezionate a intervalli regolari, una ogni Y aree, in ambedue le direzioni (lungo le righe e lungo le colonne) a partire dall'area campione iniziale.

Nel caso in cui è sconosciuto il valore di *N*, ovvero del numero di aree che compone la popolazione, e di *Nr* e *Nc*, si può operare selezionando direttamente la riga della prima area campione, a partire da un dato margine, in corrispondenza di un numero casuale compreso tra *1* e *Y* e ripetendo la stessa operazione per le colonne (questo modo di operare può produrre un leggero sotto- o sovra-campionamento se *Nr* e/o *Nc* non sono multipli esatti di *Y*).

Stimatori campionari

Lo stimatore campionario della media dell'attributo *y* è:

$$\bar{y}_{sis} = \frac{\sum_{i=1}^{n} y_i}{n}$$

dove:

n = numerosità campionaria

y_i = valore dell'attributo della i-esima unità campionaria

Non esiste una procedura formalmente valida per la stima della varianza di $\bar{x}_{sis}$. Tuttavia, assumendo che l'estrazione sistematica abbia prodotto un ordine di selezione del campione sufficientemente casuale rispetto alla variabile di interesse, allora, in prima approssimazione, si possono applicare al campionamento sistematico gli stessi stimatori adottati per il campionamento casuale:

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

varianza dell'attributo *x*

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

deviazione standard dell'attributo *x*

$$s_{\bar{y}}^2 = \frac{s_y^2}{n} \frac{N-n}{N}$$

varianza della media campionaria

Lo stimatore del valore totale è pari a:

$$y_{sis} = N\bar{y}_{sis} = (Yn)\,\bar{y}_{sis}$$

E la sua varianza può essere stimata pari a:

$$s^2_{y_{sis}} = N^2 s^2_{\bar{y}_{sis}}$$

Se il campionamento è stato condotto su aree campione di superficie unitaria pari a *a*, essendo *N* circa pari a *A/a* (dove *A* è la superficie totale considerata, nota senza errore) si ha che:

$$y_{sis} = \frac{A}{a}\bar{y}_{sis}$$

$$s^2_{y_{sis}} = \left(\frac{A}{a}\right)^2 s^2_{\bar{y}_{sis}}$$

Dimensionamento di un campione sistematico

Non esistono metodi esatti per ottimizzare la scelta della numerosità di un campione sistematico al fine di non superare una data soglia dell'errore di campionamento. Tuttavia, nel caso di basse frazioni di campionamento[1], quali quelle che generalmente caratterizzano le applicazioni inventariali, tale numerosità può essere soddisfacentemente stabilita facendo riferimento alle procedure di dimensionamento dei campioni casuali:

$$n_0 = \left(\frac{t s_y}{e c_0} \right)^2$$

dove:

s_y = valore presunto della deviazione standard dell'attributo x nella popolazione considerata.

t = valore critico del *t di Student* corrispondente al prescelto livello di sicurezza statistica e agli appropriati gradi di libertà.

[1] Si definisce frazione di campionamento il rapporto tra la numerosità delle unità campionarie e la numerosità complessiva degli elementi (unità statistiche) della popolazione.

Vantaggi e svantaggi del campionamento sistematico

Vantaggi

- ❖ l'identificazione delle unità campionarie è molto più agevole rispetto a un campionamento casuale;
- ❖ nella gran parte delle situazioni e a parità di altre condizioni, il campionamento sistematico fornisce stime più accurate rispetto a quello casuale: non vi è alcuna possibilità che ampie porzioni omogenee della popolazione non vengano rappresentate da almeno qualche unità campionaria;
- ❖ la regolarità della distribuzione delle unità campionarie facilita l'impiego dei dati raccolti ai fini di una loro eventuale spazializzazione.

Svantaggi

- ❖ il campionamento sistematico può comportare il rischio che l'intervallo di campionamento coincida con eventuali fluttuazioni periodiche dei valori dell'attributo oggetto di stima;
- ❖ impossibilità di inferire in modo corretto la varianza della popolazione da quella del campione: si possono dunque ottenere solamente valori approssimati dell'errore di campionamento;
- ❖ una volta ultimato il campionamento non è possibile aggiungere nuove unità campionarie, a meno di non procedere ex novo al campionamento stesso;
- ❖ possono verificarsi, specialmente in popolazioni piccole, situazioni di sovra- o sotto-campionamento rispetto a quanto prestabilito.