Apunts de teoria de mostres

Llorenç Cerdà-Alabern
llorenc@ac.upc.edu
Barcelona, 3 de setembre de 2014

Índex

Prefaci

Vaig escriure aquests apunts mentre preparava l'assignatura *teoría de muestras* del curs 2013-14 de la carrera de matemàtiques de la UNED. Els apunts estan basats fonamentalment en el llibre de l'assignatura [1]. El contingut no és rigorós ni complert, i només hi ha demostracions senzilles orientades a entendre o ajudar a memoritzar les relacions que demostren. L'edició l'he fet amb LaTeX.

Capítol 1

Definicions

Població U (*population*) Conjunt d'elements a investigar. Només es consideraran poblacions finites de mida N. Farem servir la notació $U = \{U_1, \cdots, U_N\}$. Distingim:

- Població objectiu: La que es desitja investiga.
- Població marc (o simplement marc): Llista d'on es prenen les mostres. Idealment igual a la població objectiu.

Cens Quan s'agafa informació de tota la població, N.

Mostra de mida n**,** S_i (*sample*) col·lecció ordenada de n elements de la població $S_i = \{U_{i1}, \cdots, U_{in}\}$, $U_{ij} \in U$. També és possible que no importi l'ordre, i farem servir la notació $s_i = \{u_1, \cdots, u_n\}$, $u_i \in U$. Les mostres són variables aleatòries, en general dependents. Per tant:

$$P(S_i) = P\big(U_{i1}\big)\, P\big(U_{i2} \mid U_{i1}\big) \cdots P\big(U_{in} \mid U_{i1}, \cdots U_{in-1}\big) \quad (1.1)$$

Selecció de la mostra El mètode de mostreig és diferent si importa o no l'ordre, i si les mostres es prenen o no amb reposició. Segons això, la cardinalitat de l'espai mostral, S, amb mostres de mida n pot ser:

- Ordenada i amb reposició: $VR(N,n) = N^n$.
- Ordenada sense reposició: $V(N,n) = \binom{N}{n}\, n!$.
- No ordenada i amb reposició: $CR(N,n) = \binom{N+n-1}{n}$.
- No Ordenada sense reposició: $C(N,n) = \binom{N}{n}$.

A més, és possible que les mostres es puguin triar o no amb la mateixa probabilitat. Es diu mostreig amb *igual* o *desigual* probabilitat.

Característica poblacional X Assigna el valor numèric X_i per a cada unitat $U_i \in U$. Per exemple, el pes o altura d'una persona. Si la característica és de tipus qualitatiu, es pot associar l'indicador A_i que val 1 si la característica es presenta en U_i, i 0 altrament.

Paràmetre poblacional θ Funció $\theta(X_1, \cdots X_N)$ d'una característica poblacional. Exemples

- Mitjana poblacional:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{1.2}$$

- Total poblacional[1]:

$$X = \sum_{i=1}^{N} X_i = N\,\bar{X} \tag{1.3}$$

- Variancia poblacional:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \tag{1.4}$$

- Quasivariancia poblacional:

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2 = \frac{N}{N-1}\,\sigma^2 \tag{1.5}$$

- Mitjana quadràtica:

$$\bar{X}_Q = \sqrt{\frac{1}{N} \sum_{i=1}^{N} X_i^2} \tag{1.6}$$

- Mitjana geomètrica:

$$\bar{X}_G = \sqrt[N]{\prod_{i=1}^{N} X_i} \tag{1.7}$$

- Mitjana harmònica:

$$\bar{X}_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}} \tag{1.8}$$

Si la característica és qualitativa:

- Proporció poblacional:

$$P = \frac{1}{N} \sum_{i=1}^{N} A_i \tag{1.9}$$

- Total de classe poblacional:

$$A = \sum_{i=1}^{N} A_i = N\,P \tag{1.10}$$

Estimador (o estadístic) (*statistic*) Donada una mostra $s_i = \{u_1, \cdots, u_n\}$ farem servir la notació $s_i(X) = \{X_1, \cdots, X_n\}$ per identificar el conjunt de característiques poblacionals X_i que corresponen a cada element u_i de la mostra, i $S(X)$ per el conjunt de tots els valors possibles de les característiques poblacionals d'una mostra: $s_i(X) \in S(X)$. Un **estimador puntual** és una funció mesurable $\hat{\theta} : S(X) \in \mathbb{R}^n \to \mathbb{R}$, que estima la característica poblacional θ a partir d'una mostra: $\hat{\theta}(s_i(X)) = \hat{\theta}(X_1, \cdots X_n)$. Exemples:

- Mitjana mostral:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.11}$$

- Total mostral:

$$\hat{X} = \frac{N}{n} \sum_{i=1}^{n} X_i = N\,\bar{x} \tag{1.12}$$

- Proporció mostral:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} A_i \tag{1.13}$$

- Total de classe mostral:

$$\hat{A} = \frac{N}{n} \sum_{i=1}^{n} A_i = N\,\hat{p} \tag{1.14}$$

- En general, el paràmetre poblacional del tipus:

$$\theta = \sum_{i=1}^{N} Y_i$$

s'estima amb l'estimador mostral:

$$\hat{\theta} = \sum_{i=1}^{n} w_i\,Y_i$$

Procediment o mètode d'estimació Especifica el procediment de mostreig per a obtenir mostres s amb probabilitat $P(s)$, i estimador $\hat{\theta}$ per a determinar el paràmetre poblacional θ.

Distribució en el mostreig d'un estimador (*sampling distribution*) És la distribució d'un estimador. És a dir, sigui el conjunt $T = \{t \in \mathbb{R} \mid \exists (X_1, \cdots X_n) \in S(X), \hat{\theta}(X_1, \cdots X_n) = t\}$. Aleshores

$$P(T) = P\Big(\hat{\theta}(X_1, \cdots X_n) = t\Big) = \sum_{s_i \mid \hat{\theta}(s_i(X)) = t} P(s_i) \tag{1.15}$$

[1]Seguint la notació de [1, 2] es fa servir el mateix símbol que per una característica poblacional. Del context s'entén a què es refereix.

Esperança d'un estimador

$$\mathrm{E}\left(\hat{\theta}\right) = \sum_S \hat{\theta}(s_i)\, P(s_i) = \sum_{\mathbb{R}} t\, P\left(\hat{\theta} = t\right) \tag{1.16}$$

Variancia d'un estimador

$$\mathrm{V}\left(\hat{\theta}\right) = \mathrm{E}\left(\hat{\theta}^2\right) - \mathrm{E}\left(\hat{\theta}\right)^2 \tag{1.17}$$

Error en el mostreig (desviació típica) d'un estimador

$$\sigma\left(\hat{\theta}\right) = \sqrt{\mathrm{V}\left(\hat{\theta}\right)} \tag{1.18}$$

Error relatiu en el mostreig (coeficient de variació) d'un estimador

$$CV(\hat{\theta}) = \frac{\sigma\left(\hat{\theta}\right)}{\mathrm{E}\left(\hat{\theta}\right)} \tag{1.19}$$

Acuracitat (error quadràtic mitjà, EQM) d'un estimador

$$EQM(\hat{\theta}) = \mathrm{E}\left((\hat{\theta} - \theta)^2\right) \tag{1.20}$$

Biaix d'un estimador

$$B(\hat{\theta}) = \mathrm{E}\left(\hat{\theta}\right) - \theta \tag{1.21}$$

Es diu que un estimador és consistent quan $B(\hat{\theta}) \xrightarrow{n \to N} 0$.

1.1 Precisió d'un estimador

$$EQM(\hat{\theta}) = \mathrm{E}\left((\hat{\theta} - \theta)^2\right) = \mathrm{E}\left((\hat{\theta} - \mathrm{E}\left(\hat{\theta}\right) + \mathrm{E}\left(\hat{\theta}\right) - \theta)^2\right) = \mathrm{E}\left((\hat{\theta} - \mathrm{E}\left(\hat{\theta}\right) + B(\hat{\theta}))^2\right) = \sigma^2(\hat{\theta}) + B(\hat{\theta})^2 \tag{1.22}$$

És desitjable un estimador no biaixat (o centrat) $B(\hat{\theta}) = 0$, sempre i quan l'error (desviació típica) sigui petita. En la pràctica és acceptable un estimador biaixat sempre i quan:

$$\left|\frac{B(\hat{\theta})}{\sigma\left(\hat{\theta}\right)}\right| < \frac{1}{10} \tag{1.23}$$

Per a comparar la precisió de dos estimadors considerem el quocient entre els seus EQM: $EQM(\hat{\theta}_1)/EQM(\hat{\theta}_2)$. Si són centrats ($B(\hat{\theta}) = 0$), serà el quocient entre les seves variàncies: $\sigma^2(\hat{\theta}_1)/\sigma^2(\hat{\theta}_1)$.

1.2 Intervals de confiança

Si les estimacions $\hat{\theta}_1 \leq \hat{\theta}_2$ tals que:

$$P\left(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\right) \geq 1 - \alpha \tag{1.24}$$

aleshores, $[\hat{\theta}_1,\ \hat{\theta}_2]$ és un interval de confiança (IC) per θ de nivell de confiança $1 - \alpha$ (o significació α). Normalment es dóna la probabilitat en percentatge, és a dir, $100\,(1 - \alpha)$. Valors típics són intervals de confiança del 95% o 99%. La figura 1.1 mostra un IC per a una distribució simètrica (p.e. Gaussiana o Student).

Estimador no biaixat amb distribució normal Sigui

$$P(Z > z_{\alpha/2}) = \alpha/2, \tag{1.25}$$

és a dir, $z_{\alpha/2} = F^{-1}_{N(0,1)}\left(1 - \frac{\alpha}{2}\right)$ (veure la figura 1.1):

$$\hat{\theta} \sim N(\theta, \sqrt{\mathrm{V}\left(\hat{\theta}\right)}) \Rightarrow z = \frac{\hat{\theta} - \theta}{\sqrt{\mathrm{V}\left(\hat{\theta}\right)}} \sim N(0{,}1) \Rightarrow$$

$$P\left(-z_{\alpha/2} \leq z \leq z_{\alpha/2}\right) = 1 - \alpha \Rightarrow$$

$$\hat{\theta} - z_{\alpha/2}\sqrt{\mathrm{V}\left(\hat{\theta}\right)} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sqrt{\mathrm{V}\left(\hat{\theta}\right)} \tag{1.26}$$

i $\left[\hat{\theta} - z_{\alpha/2}\sqrt{\mathrm{V}\left(\hat{\theta}\right)},\ \hat{\theta} + z_{\alpha/2}\sqrt{\mathrm{V}\left(\hat{\theta}\right)}\right]$ és un IC.

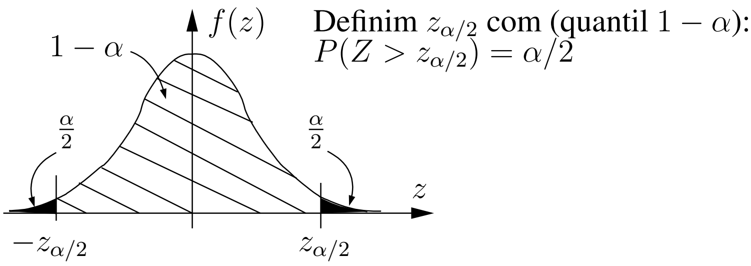


Figura 1.1: Interval de confiança en una distribució simètrica: IC $= \left[-z_{\alpha/2},\ z_{\alpha/2}\right]$.

Si estimem la variancia per la quasivariancia amb una mostra petita (n al voltant de 15 o menys), aleshores cal utilitzar una distribució de Student i l'IC és:

$$\hat{\theta} \pm t_{n-1;\alpha/2}\sqrt{\mathrm{V}\left(\hat{\theta}\right)} \tag{1.27}$$

Estimador no biaixat per a una distribució desconeguda
Amb la desigualtat de Chebyshev:

$$P\big(|X-\mu|\geq c\big)\leq\frac{\mathrm{V}\left(\hat\theta\right)}{c^2}\Rightarrow$$

$$P\left(|\hat\theta-\theta|\leq c\right)\geq 1-\frac{\mathrm{V}\left(\hat\theta\right)}{c^2}$$

$$P\left(\hat\theta-\frac{\sqrt{\mathrm{V}\left(\hat\theta\right)}}{\sqrt{\alpha}}\leq\theta\leq\hat\theta+\frac{\sqrt{\mathrm{V}\left(\hat\theta\right)}}{\sqrt{\alpha}}\right)\geq 1-\alpha \tag{1.28}$$

on s'ha posat $c=\frac{\sqrt{\mathrm{V}(\hat\theta)}}{\sqrt{\alpha}}$. Per tant, $\left[\hat\theta-\frac{\sqrt{\mathrm{V}(\hat\theta)}}{\sqrt{\alpha}},\hat\theta+\frac{\sqrt{\mathrm{V}(\hat\theta)}}{\sqrt{\alpha}}\right]$ és un IC.

Estimador biaixat amb distribució normal És fàcil comprovar que amb $B(\hat\theta)=\mathrm{E}\left(\hat\theta\right)-\theta$ es té que l'IC és $\left[\hat\theta-z_{\alpha/2}-B(\hat\theta)\sqrt{\mathrm{V}\left(\hat\theta\right)},\hat\theta+z_{\alpha/2}\sqrt{\mathrm{V}\left(\hat\theta\right)}-B(\hat\theta)\right]$.

Capítol 2
Mostreig amb probabilitats desiguals

Suposem una mostra sense reposició $\tilde{\boldsymbol{x}}$. Definim les VAs (indicadors):

$$e_i=\begin{cases}1, & \text{si } u_i\in\tilde{\boldsymbol{x}} \text{ amb probabilitat } \pi_i,\\ 0, & \text{si } u_i\notin\tilde{\boldsymbol{x}} \text{ amb probabilitat } 1-\pi_i.\end{cases} \tag{2.1}$$

$$e_i\,e_j=\begin{cases}1, & \text{si } u_i,u_j\in\tilde{\boldsymbol{x}} \text{ amb probabilitat } \pi_{ij},\\ 0, & \text{si } u_i,u_j\notin\tilde{\boldsymbol{x}} \text{ amb probabilitat } 1-\pi_{ij}.\end{cases} \tag{2.2}$$

Clarament:

$$\mathrm{E}(e_i)=\pi_i \tag{2.3}$$

$$\mathrm{E}\left(e_i^2\right)=\pi_i \tag{2.4}$$

$$\mathrm{V}(e_i)=\pi_i\,(1-\pi_i) \tag{2.5}$$

$$\mathrm{Cov}\left(e_i\,e_j\right)=\mathrm{E}\left(e_i\,e_j\right)-\mathrm{E}(e_i)\;E[e_j]=\pi_{ij}-\pi_i\,\pi_j \tag{2.6}$$

Propietats

$$\sum_{i=1}^{N}\pi_i=n \tag{2.7}$$

$$\sum_{\substack{i=1\\ i\neq j}}^{N}\pi_i=n-\pi_j \tag{2.8}$$

$$\sum_{\substack{i=1\\ i\neq j}}^{N}\pi_{ij}=(n-1)\,\pi_j \tag{2.9}$$

$$\sum_{\substack{i=1\\ i\neq j}}^{N}(\pi_{ij}-\pi_i\,\pi_j)=-\pi_j\,(1-\pi_j) \tag{2.10}$$

Demostració.
2.7: $\sum_{i=1}^N\pi_i=\sum_{i=1}^N\mathrm{E}(e_i)=E[\sum_{i=1}^N e_i]=n$ perquè per a qualsevol mostra $\tilde{\boldsymbol{x}}$ de mida n hi haurà n dels N indicadors e_i per els que $u_i\in\tilde{\boldsymbol{x}}$ serà cert. □

2.0.1 Estimador de Horvitz i Thompson

És un estimador no biaixat per mostreig sense reposició.

$$\text{Mitjana: } \bar x=\sum_{i=1}^{n}\frac{1}{N}\frac{X_i}{\pi_i} \tag{2.11}$$

$$\text{Total: } \hat X=\sum_{i=1}^{n}\frac{X_i}{\pi_i} \tag{2.12}$$

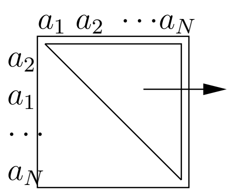
$$\text{Proporció: } \hat p=\sum_{i=1}^{n}\frac{1}{N}\frac{A_i}{\pi_i} \tag{2.13}$$

$$\text{Total classe: } \hat A=\sum_{i=1}^{n}\frac{A_i}{\pi_i} \tag{2.14}$$

Variancia del paràmetre $Y=\{X,\frac{X}{N},A,\frac{A}{N}\}$**:**
Tenint en compte que (veure la figura 2.1):

$$\mathrm{V}\left(\sum_{i=1}^{N}Y_i\right)=\mathrm{E}\left(\left(\sum_{i=1}^{N}Y_i-\mathrm{E}\left(\sum_{i=1}^{N}Y_i\right)\right)^2\right)=$$

$$\mathrm{E}\left(\left(\sum_{i=1}^{N}(Y_i-\mathrm{E}(Y_i))\right)^2\right)=\sum_{i=1}^{N}\sigma_i^2+2\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sigma_{ij} \tag{2.15}$$



$(\sum_{i=1}^N a_i)^2=\sum_{i=1}^N a_i^2+2\sum_{i=1}^N\sum_{j=i+1}^N a_i\,a_j$

Figura 2.1: Quadrat d'una suma.

Tenim que:

$$\hat{\theta} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i} = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} e_i \tag{2.16}$$

Substituint en 2.15:

$$\mathrm{V}\left(\hat{\theta}\right) = \sum_{i=1}^{N} \mathrm{V}\left(\frac{Y_i}{\pi_i} e_i\right) + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathrm{Cov}\left(\frac{Y_i}{\pi_i} e_i, \frac{Y_j}{\pi_j} e_j\right) = \sum_{i=1}^{N} \left(\frac{Y_i}{\pi_i}\right)^2 \mathrm{V}(e_i) + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \mathrm{Cov}(e_i, e_j)$$

D'on:

$$\mathrm{V}\left(\hat{\theta}\right) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} (1 - \pi_i) + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{Y_i\, Y_j}{\pi_i\, \pi_j} (\pi_{ij} - \pi_i\, \pi_j) \tag{2.17}$$

Estimadors no biaxials de la variancia del paràmetre Y:

$$\hat{\mathrm{V}}\left(\hat{\theta}\right) = \sum_{i=1}^{n} \frac{Y_i^2}{\pi_i} (1 - \pi_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{Y_i\, Y_j}{\pi_i\, \pi_j} (\pi_{ij} - \pi_i\, \pi_j) \tag{2.18}$$

$$\hat{\mathrm{V}}\left(\hat{\theta}\right) = \sum_{i=1}^{n} \sum_{j=1+1}^{n} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \frac{\pi_i\, \pi_j - \pi_{ij}}{\pi_{ij}} \tag{2.19}$$

L'equació 2.19 es coneix com estimador de la variancia de Yates i Grundy.

Capítol 3

Mostreig aleatori simple

Si no es diu el contrari, es sobreentén que és amb probabilitats de mostreig i sense reposició. En aquest cas,

$$\pi_i = \frac{n}{N} \tag{3.1}$$

$$\pi_{ij} = \frac{n\,(n-1)}{N\,(N-1)} \tag{3.2}$$

Substituint en 2.11, 2.12, 2.14, 2.13, s'obté respectivament 1.11, 1.12, 1.13, 1.14. Substituint en 2.17 i definint $f = n/N$, s'obtenen les variàncies:

$$\text{Mitjana: } \mathrm{V}(\bar{x}) = (1-f)\,\frac{S^2}{n} \tag{3.3}$$

$$\text{Total: } \mathrm{V}\left(\hat{X}\right) = N^2\,\mathrm{V}\left(\hat{\bar{x}}\right) \tag{3.4}$$

$$\text{Proporció: } \mathrm{V}(\hat{p}) = (1-f)\,\frac{S^2}{n} = \frac{N-n}{N-1}\,\frac{P\,Q}{n} \tag{3.5}$$

$$\text{Total classe: } \mathrm{V}\left(\hat{A}\right) = N^2\,\mathrm{V}(\hat{p}) \tag{3.6}$$

on $Q = 1 - P$. S^2 s'estima amb la quasivariancia mostral:

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2 \tag{3.7}$$

Per a variables qualitatives:

$$\hat{S}^2 = \frac{n\,\hat{p}\,\hat{q}}{n-1} \tag{3.8}$$

on $\hat{q} = 1 - \hat{p}$.

Formulari

Taula 3.1: **Estadístics**

	Mitjana	Total	Proporció	Classe	Variància	Q-Variancia
Poblacional	$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$	$X = \sum_{i=1}^{N} X_i$	$P = \frac{1}{N}\sum_{i=1}^{N} A_i$	$A = N\,P = \sum_{i=1}^{N} A_i$	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{x})^2$	$S^2 = \frac{N}{N-1}\,\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{x})^2$
Mostral	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\hat{X} = N\,\bar{x} = \frac{N}{n}\sum_{i=1}^{n} X_i$	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} A_i$	$\hat{A} = N\,\hat{p} = \frac{N}{n}\sum_{i=1}^{n} A_i$	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$	$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$Q = 1 - P$, $\hat{q} = 1 - \hat{p}$

$\sigma_A^2 = P\,Q$, $\hat{\sigma}_A^2 = \hat{p}\,\hat{q}$. $S_A^2 = \frac{N}{N-1}\,P\,Q$, $\hat{S}_A^2 = \frac{n}{n-1}\,\hat{p}\,\hat{q}$.

Taula 3.2: **Variàncies mostreig aleatori simple**

	Mostreig Simple (MS)		MS amb reposició	
Paràmetre	amb S^2	amb $\hat{S}^2$	amb S^2	amb $\hat{S}^2$
Mitjana $\mathrm{V}(\bar{x}) =$	$(1-f)\,\frac{S^2}{n}$	$(1-f)\,\frac{\hat{S}^2}{n}$	$\frac{\sigma^2}{n}$	$\frac{\hat{S}^2}{n}$
Total $\mathrm{V}\left(\hat{X}\right) =$	$N^2\,\mathrm{V}(\bar{x})$	$N^2\,\hat{\mathrm{V}}(\bar{x})$	$N^2\,\mathrm{V}(\bar{x})$	$N^2\,\hat{\mathrm{V}}(\bar{x})$
Proporció $\mathrm{V}(\hat{p}) =$	$(1-f)\,\frac{S_A^2}{n}$	$(1-f)\,\frac{\hat{S}_A^2}{n}$	$\frac{\sigma_A^2}{n}$	$\frac{\hat{S}_A^2}{n}$
Classe $\mathrm{V}\left(\hat{A}\right) =$	$N^2\,\mathrm{V}(\hat{p})$	$N^2\,\hat{\mathrm{V}}(\hat{p})$	$N^2\,\mathrm{V}(\hat{p})$	$N^2\,\hat{\mathrm{V}}(\hat{p})$

$f = n/N$. En MS $\mathrm{E}\left(\hat{S}^2\right) = S^2$. En MS amb reposició $\mathrm{E}\left(\hat{S}^2\right) = \sigma^2$. .

Capítol 4

Mostreig estratificat

Població N dividida en L estrats homogenis de mida N_h. $L = 1, \cdots L$. Mida relativa de l'estrat: $W_h = N_h/N$. Es prenen n_h mostres per estrat, $n = \sum_{h=1}^{L} n_h$.

Estadístics de l'estrat h: mitjana $\bar{X}_h$, total X_h, variancia σ_h^2, Quasivariancia S_h^2, proporció P_h i total de classe A_h.

Estadístics poblacionals (de la població estratificada): mitjana $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$, total $X = N\bar{X}$, variancia $\sigma^2 = 1/N \sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \bar{X})^2$, quasivariancia S_h^2, proporció $P_h = \sum_{h=1}^{L} W_h P_h$ i total de classe $A_h = N P$.

Estadístics mostrals de l'estrat h: mitjana $\bar{x}_h$, total $\hat{X}_h$, variancia $\hat{\sigma}_h^2$, quasivariancia $\hat{S}_h^2$, proporció $\hat{p}_h$ i total de classe $\hat{A}_h$. $\bar{x}_h$ és un estimador no biaixat de la mitjana de l'estrat h: $\bar{X}_h$, i $\hat{p}_h$ de P_h. Per tant, d'ells s'obtenen els estadístics mostrals, estimadors no baixats dels estadístics poblacionals: mitjana $\bar{x}_{st}$, total $\hat{X}_{st}$, proporció $\hat{P}_{st}$ i total de classe $\hat{A}_{st}$.

Taula 4.1: **Variàncies mostreig estratificat**

	Mostreig Simple (MS)		MS amb reposició	
Paràmetre	amb S_h^2	amb $\hat{S}_h^2$	amb S_h^2	amb $\hat{S}_h^2$
Mitjana				
$\mathrm{V}(\bar{x}_{st}) =$	$\sum_{h=1}^{L} W_h^2 (1-f_h) \frac{S_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 (1-f_h) \frac{\hat{S}_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 \frac{\hat{S}_h^2}{n_h}$
Total				
$\mathrm{V}\left(\hat{X}_{st}\right) =$	$N^2\,\mathrm{V}(\bar{x}_{st})$	$N^2\,\hat{\mathrm{V}}(\bar{x}_{st})$	$N^2\,\mathrm{V}(\bar{x}_{st})$	$N^2\,\hat{\mathrm{V}}(\bar{x}_{st})$
Proporció				
$\mathrm{V}\left(\hat{P}_{st}\right) =$	$\sum_{h=1}^{L} W_h^2 (1-f_h) \frac{S_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 (1-f_h) \frac{\hat{S}_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 \frac{\hat{S}_h^2}{n_h}$
Classe				
$\mathrm{V}\left(\hat{A}_{st}\right) =$	$N^2\,\mathrm{V}\left(\hat{P}_{st}\right)$	$N^2\,\hat{\mathrm{V}}\left(\hat{P}_{st}\right)$	$N^2\,\mathrm{V}\left(\hat{P}_{st}\right)$	$N^2\,\hat{\mathrm{V}}\left(\hat{P}_{st}\right)$

$S_h^2 = \frac{1}{N_h-1} \sum_{i=1}^{N} (X_{hi} - \bar{X}_h)^2$, $\hat{S}_h^2 = \frac{1}{n_h-1} \sum_{i=1}^{N} (X_{hi} - \bar{x}_h)^2$.
Per a la proporció: $S_h^2 = \frac{N_h}{N_h-1}\,\sigma_h^2 = \frac{N_h}{N_h-1}\,P_h\,Q_h$, $\hat{S}_h^2 = \frac{n_h}{n_h-1}\,\hat{p}_h\,\hat{q}_h$.
$f_h = n_h/N_h$. Per estrats de mida gran es pot aproximar $1 - f_h \approx 1$.
$S^2 = \frac{1}{N-1} \sum_h \sum_i (X_{ij} - \bar{X}) \approx \sum_h W_h\,S_h^2 + \sum_h W_h (\bar{X}_h - \bar{X})^2$.

Afixacions

- Igual: $n_h = \dfrac{n}{L}$
- Proporcional: $n_h = \dfrac{N_h}{N}\,n = W_h\,n$. També: $n_h = N_h\,f$, d'on $f_h = n_h/N_h = f$. Substituint per MS es té
  $\mathrm{V}(\bar{x}_{ap}) = \sum_{h=1}^{L} W_h\,(1-f)\,\dfrac{S_h^2}{n}$.
- Òptima amb cost c_h: $n_h = \dfrac{N_h\,S_h/\sqrt{c_h}}{\sum_h N_h\,S_h/\sqrt{c_h}}\,n$. Substituint per MS es té:
  $\mathrm{V}(\bar{x}_{ao}) = \dfrac{(\sum_h W_h\,S_h/\sqrt{c_h})(\sum_h W_h\,S_h\,\sqrt{c_h})}{n} - \dfrac{\sum_h W_h\,S_h^2}{N} \approx \dfrac{(\sum_h W_h\,S_h/\sqrt{c_h})(\sum_h W_h\,S_h\,\sqrt{c_h})}{n}$.
  Si el cost és el mateix, $n_h = \dfrac{N_h\,S_h}{\sum_h N_h\,S_h}\,n$, $\mathrm{V}(\bar{x}_{ao}) \approx (\sum_h W_h\,S_h)^2/n$.

Capítol 5
Estimador de raó

Quan es té informació Y_i relacionada amb la població estudiada X_i.

Raó poblacional : $$R = \frac{X}{Y} = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} Y_i} = \frac{\bar{X}}{\bar{Y}} \tag{5.1}$$

Raó mostral : $$\hat{R} = \frac{\hat{X}}{\hat{Y}} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{\bar{x}}{\bar{y}} \tag{5.2}$$

Estimador de la mitjana : $$\hat{\bar{X}}_R = \hat{R}\,\bar{Y} \tag{5.3}$$

Estimador del total : $$\hat{X}_R = N\,\hat{\bar{X}}_R = N\,\hat{R}\,\bar{Y} \tag{5.4}$$

Biaix : $$B(\hat{R}) = \mathrm{E}\left(\hat{R}\right) - R = \frac{R\,\mathrm{V}(\bar{y}) - \mathrm{Cov}(\bar{x},\bar{y})}{\bar{Y}^2} \tag{5.5}$$

Taula 5.1: **Variàncies estimador de raó**

	Mostreig Simple (MS)	MS amb reposició
Mitjana $\mathrm{V}\left(\hat{\bar{X}}_R\right) =$	$(1-f)\,\dfrac{S_x^2 - 2\,\hat{R}\,S_{xy} + \hat{R}^2\,S_y^2}{n}$	$\dfrac{S_x^2 - 2\,\hat{R}\,S_{xy} + \hat{R}^2\,S_y^2}{n}$
Total $\mathrm{V}\left(\hat{X}_R\right) =$	$N^2\,\mathrm{V}\left(\hat{\bar{X}}_R\right)$	$N^2\,\mathrm{V}\left(\hat{\bar{X}}_R\right)$
Raó $\mathrm{V}\left(\hat{R}_R\right) =$	$\mathrm{V}\left(\hat{\bar{X}}_R\right)/\bar{Y}^2$	$\mathrm{V}\left(\hat{\bar{X}}_R\right)/\bar{Y}^2$
Biaix $B(\hat{R}) =$	$(1-f)\,\dfrac{R\,S_y^2 - S_{xy}}{n\,\bar{Y}^2}$	$\dfrac{R\,\sigma_y^2 - \sigma_{xy}}{n\,\bar{Y}^2}$

Normalment farem servir els estimadors de les quasivariancies en les fórmules anteriors:
$\hat{S}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{x})^2$, $\hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{y})^2$, $\hat{S}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{x})\,(Y_i-\bar{y})$.
Notar que es pot calcular: $S_x^2 - 2\,\hat{R}\,S_{xy} + \hat{R}^2\,S_y^2 = (\sum X_i^2 + R^2\sum Y_i^2 - 2\,R\sum X_i\,Y_i)$/(n-1)

Nota: per el càlcul de la variancia es fa l'aproximació $\hat{R} - R = \frac{\bar{x}-R\,\bar{y}}{\bar{y}} \approx \frac{\bar{x}-R\,\bar{y}}{\bar{Y}}$, d'on:

$$\mathrm{V}\left(\hat{R}\right) = \mathrm{V}\left(\hat{R}-R\right) \approx \mathrm{V}\left(\frac{\bar{x}-R\,\bar{y}}{\bar{Y}}\right) = (1-f)\,\frac{S_x^2 - 2\,\hat{R}\,S_{xy} + \hat{R}^2\,S_y^2}{n\,\bar{Y}^2}$$

Capítol 6
Estimador per regressió

Quan es té informació Y_i relacionada amb la població estudiada X_i.

Estimador de la mitjana : $$\bar{x}_{rg} = \hat{x} + b\,(\hat{Y} - \hat{y}) \tag{6.1}$$

- El valor de b que minimitza la variancia és la pendent de la recta de regressió de X sobre Y: $b_o = S_{xy}/S_y^2$, que s'estima per $b_o = \hat{S}_{xy}/\hat{S}_y^2$.
- Si b és una constant, l'estimador $\bar{x}_{rg}$ és no biaixat.

Taula 6.1: **Variàncies estimador per regressió**

	Mostreig Simple (MS)	MS amb reposició
Mitjana $\mathrm{V}\big(\bar{x}_{rg}\big) =$	$(1-f)\,\dfrac{S_x^2 - 2\,b\,S_{xy} + b^2\,S_y^2}{n}$	$\dfrac{S_x^2 - 2\,b\,S_{xy} + b^2\,S_y^2}{n}$
Total $\mathrm{V}\Big(\hat{X}_{rg}\Big) =$	$N^2\,\mathrm{V}\Big(\hat{\bar{X}}_{rg}\Big)$	$N^2\,\mathrm{V}\Big(\hat{\bar{X}}_{rg}\Big)$

Taula 6.2: **Variàncies mínimes (amb $b_o = S_{xy}/S_y^2$)**

	Mostreig Simple (MS)	MS amb reposició
Mitjana $\mathrm{V}\big(\bar{x}_{rg}\big) =$	$(1-f)\,\dfrac{S_x^2\,(1-\rho^2)}{n}$	$\dfrac{\sigma_x^2\,(1-\rho^2)}{n}$
Total $\mathrm{V}\Big(\hat{X}_{rg}\Big) =$	$N^2\,\mathrm{V}\Big(\hat{\bar{X}}_{rg}\Big)$	$N^2\,\mathrm{V}\Big(\hat{\bar{X}}_{rg}\Big)$

Coeficient de correlació: $\rho = \mathrm{Cov}(X,Y)\,/(\sigma_x\,\sigma_y) = S_{xy}/(S_x\,S_y)$

Notar que:

$$S_{xy} = \frac{1}{n-1}\sum_i((X_i-\bar{x})(Y_i-\bar{y})) = \frac{1}{n-1}\left(\sum_i(X_i\,Y_i) - n(\bar{x}\,\bar{y})^2\right) = \frac{1}{n-1}\left(\sum(X_i\,Y_i) - \frac{\sum X_i \sum Y_i}{n}\right)$$

D'on:

$$S_x^2 = \frac{1}{n-1}\left(\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right)$$

i

$$\rho = \frac{S_{xy}}{S_x\,S_y} = \frac{\sum(X_i\,Y_i) - \frac{\sum X_i \sum Y_i}{n}}{\sqrt{\left(\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right)\left(\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right)}}$$

Capítol 7
Mostreig sistemàtic

S'organitza la població $N = n\,k$ amb n zones de mida k:

	1	2	$\cdots$	k
1	n_{11}	n_{12}	$\cdots$	n_{1k}
2	n_{21}	n_{22}	$\cdots$	n_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	n_{n1}	n_{n2}	$\cdots$	n_{nk}

Característica poblacional $\theta = \sum_{i=1}^{n}\sum_{j=1}^{k} Y_{ij}$, estimada per $\hat{\theta} = \frac{N}{n}\sum_{i=1}^{n} Y_{ij}$. És a dir, l'espai mostral són les columnes, que es trien amb probabilitat $1/k$. El mostreig és sense reposició. Si es tria la columna j els estimadors són:

Mitjana : $$\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_{ij} = \bar{x}_j \tag{7.1}$$

Total : $$\hat{X} = N\,\bar{x}_j \tag{7.2}$$

Proporció : $$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} A_{ij} = \hat{P}_j \tag{7.3}$$

Classe : $$\hat{A} = N\,\hat{P}_j \tag{7.4}$$

Quasivariancia inter-mostral:
$$S_{bs}^2 = \frac{1}{k-1}\sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{x}_j - \bar{X})^2 = \frac{n}{k-1}\sum_{j=1}^{k}(\bar{x}_j - \bar{X})^2 \tag{7.5}$$

Quasivariancia intra-mostral:
$$S_{ws}^2 = \frac{1}{N-k}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \bar{x}_j)^2 \tag{7.6}$$

Relació fonamental:
$$\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X})^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{x}_j+\bar{x}_j-\bar{X})^2 =$$
$$\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{x}_j)^2 + \sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{x}_j-\bar{X})^2 +$$
$$\underbrace{2\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{x}_j)(\bar{x}_j-\bar{X})}_{=0} =$$
$$\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{x}_j)^2 + \sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{x}_j-\bar{X})^2$$

D'on es té:
$$(N-1)\,S^2 = (N-k)\,S_{ws}^2 + (k-1)\,S_{bs}^2 \tag{7.7}$$

d'on:
$$\mathrm{V}\Big(\hat{\bar{X}}\Big) = (1-f)\,\frac{S_{bs}^2}{n} = \sigma^2 - \frac{n-1}{n}\,S_{ws}^2 \tag{7.8}$$

Per a les proporcions, amb $1/n\sum_{i=1}^{n}(X_{ij}-\bar{x}_j)^2 = \hat{P}_j\,\hat{Q}_j$, s'obté:
$$\mathrm{V}\Big(\hat{P}\Big) = P\,Q - \frac{1}{k}\sum_{j=1}^{k}\hat{P}_j\,\hat{Q}_j \tag{7.9}$$

Capítol 8
Mostreig per conglomerats

És el cas complementari del mostreig estratificat. Notació:

- N: nombre de conglomerats.
- n: nombre de conglomerats en la mostra.
- M: unitats per conglomerat (els suposem tots de la mateixa mida).

Característica poblacional $\theta = \sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij}$, estimada per $\hat{\theta} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} Y_{ij}$. Els estimadors són:

Mitjana : $$\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{j=1}^{M} X_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{X}_i = \bar{\bar{x}}$$

Total : $$\hat{X} = N\;M\,\bar{\bar{x}}$$

Proporció : $$\hat{P} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{j=1}^{M} A_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{P}_i$$

Classe : $$\hat{A} = N\;M\,\hat{P}$$

Quasivariancia inter-conglomerats:
$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{X}_i-\bar{X})^2 = \frac{M}{N-1}\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2 \tag{8.1}$$

Quasivariancia intra-conglomerats:
$$S_w^2 = \frac{1}{N\,M-N}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\bar{X}_i)^2 \tag{8.2}$$

Relació fonamental:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\bar{X})^2=\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X})^2=$$
$$\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{X}_i-\bar{X})^2+$$
$$\underbrace{2\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\bar{X}_i)(\bar{X}_i-\bar{X})}_{=0}=$$
$$\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{X}_i-\bar{X})^2$$

D'on es té:

$$(N\,M-1)\,S^2=(N\,M-N)\,S_w^2+(N-1)\,S_b^2 \quad (8.3)$$

D'on la variancia de la mitjana ($f=(n\,M)/(N\,M)=n/N$):

$$\mathrm{V}(\bar{\bar{x}})=(1-f)\,\frac{S_b^2}{n\,M} \quad (8.4)$$

Per a la proporció l'equació 8.3 té la forma:

$$\sigma^2=\sigma_w^2+\sigma_b^2 \quad (8.5)$$

on:

$$\sigma^2=P\,Q \quad (8.6)$$

$$\sigma_b^2=\frac{1}{N}\sum_{i=1}^{N}(P_i-P)^2 \quad (8.7)$$

$$\sigma_w^2=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}(P_{ij}-\bar{P}_i)^2=\frac{1}{N}\sum_{i=1}^{N}P_i\,Q_i \quad (8.8)$$

on P_i és la proporció en cada conglomerat. Com que la proporció és un mostreig aleatori simple dels conglomerats:

$$\mathrm{V}\left(\hat{P}\right)=(1-f)\,\frac{S_b^2}{n}=\frac{N-n}{N-1}\,\frac{\sigma_b^2}{n} \quad (8.9)$$

Coeficient de correlació intra-conglomerats: Sigui (X_{ij},X_{iz}) un parell de mostres del conglomerat i, amb $j\neq z$ (en total $N\binom{M}{2}$ parelles):

$$\delta=\frac{\mathrm{Cov}\left(X_{ij},X_{iz}\right)}{\sigma(X_{ij})\,\sigma(X_{iz})}=$$
$$\frac{\mathrm{E}\Big((X_{ij}-\mathrm{E}(X_{ij}))(X_{iz}-\mathrm{E}(X_{iz}))\Big)}{\sigma^2}=$$
$$\frac{\sum_{i=1}^{N}\sum_{j\neq z}^{M}(X_{ij}-\bar{X})(X_{iz}-\bar{X})}{N\binom{M}{2}\sigma^2}=$$
$$\frac{2\sum_{i=1}^{N}\sum_{j\neq z}^{M}(X_{ij}-\bar{X})(X_{iz}-\bar{X})}{(N\,M-1)\,(M-1)\,S^2} \quad (8.10)$$

Es té:

$$\mathrm{V}(\bar{\bar{x}})=(1-f)\,\frac{S^2}{n\,M}\,(1+(M-1)\,\delta) \quad (8.11)$$

Per tant, si $\delta=0$ té la mateixa precisió que el mostreig simple, i és millor si $\delta<0$.

8.1 Conglomerats de mides diferents

Si les mides no són molt diferents, definim

$$\bar{M}=\sum_i M_i/N \quad (8.12)$$

i les fórmules són les mateixes que abans substituint M per $\bar{M}$.

Si els conglomerats es seleccionen amb probabilitats π_i (normalment proporcionals a la mida dels conglomerats: $\pi_i=M_i/\sum_i M_i$). Suposem que volem estimar el total X si els totals de cada conglomerat són X_i, $X=\sum_i X_i$. L'estimació del total és

$$\hat{X}=\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\pi_i}=\frac{1}{n}\sum_{i=1}^{n}\frac{M_i\,\bar{X}_i}{\pi_i} \quad (8.13)$$

amb variancia:

$$\mathrm{V}\left(\hat{X}\right)=n\,\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\pi_i}\right)=$$
$$\frac{1}{n}\sum_{i=1}^{n}\pi_i\left(\frac{X_i}{\pi_i}-X\right)^2=\frac{1}{n}\left(\sum_{i=1}^{n}\frac{X_i^2}{\pi_i}-X^2\right) \quad (8.14)$$

Si X és desconegut, un estimador no biaixat de $\mathrm{V}\left(\hat{X}\right)$ és:

$$\hat{\mathrm{V}}\left(\hat{X}\right)=\frac{1}{n\,(n-1)}\sum_{i=1}^{n}\left(\frac{X_i}{\pi_i}-\hat{X}\right)^2 \quad (8.15)$$

Per a la mitjana $\hat{\bar{X}}=\hat{X}/M$, on ara $M=\sum_i M_i$, l'estimació de la variancia és $\hat{\mathrm{V}}\left(\hat{\bar{X}}\right)=\hat{\mathrm{V}}\left(\hat{X}\right)/M^2$.

Capítol 9
Mostreig doble

Població N estratificada en L estrats.

Primera fase: mostra $n' = \sum n'_h$. Definim $W_h = N_h/N$, estimat per n_h/n: proporció de la població en l'estrat h.

Segona fase: mostra $n = \sum n_h$, on les mostres n_h s'agafen d'entre les n'_h.

Estimador de la mitjana:

$$\hat{\bar{X}} = \sum_h W_h \, \bar{x}_h$$

on W_h s'estima per n'_h/n de la primera fase, i $\bar{x} = x_h/n_h$ de la segona fase. Es té que:

$$\mathrm{V}\left(\hat{\bar{X}}\right) = (1-f)\,\frac{S^2}{n'} + \sum_h (\frac{1}{v_h} - 1)\, W_h \frac{S_h^2}{n'} \tag{9.1}$$

On $f = n'/N$, $v_h = n_h/n'_h$.

9.1 Falta de resposta

La població està dividida en 2 estrats $N = N_1 + N_2$: N_1 individus que responen, N_2 individus que no responen.

Primera fase: mostra de n' d'entre N: n'_1 contesten, n'_2 no.

Segona fase: mostra de $n_1 = n'_1$ (s'agafen totes les respostes) i n_2 d'entre n'_2. És d'esperar que la mostra n_2 serà més cara, doncs als que no contesten s'haurà de fer un tractament especial, per exemple, una entrevista personalitzada.

Definim $f_{21} = n_2/n'_2$. De 9.1:

$$\mathrm{V}\left(\hat{\bar{X}}\right) = (1-f)\,\frac{S^2}{n'} + (\frac{1}{f_{21}} - 1)\, W_2 \frac{S_2^2}{n'} \tag{9.2}$$

Nota: El terme en 9.1 $(\frac{1}{v_1} - 1)\, W_1 \frac{S_1^2}{n'} = 0$, doncs $v_1 = n_1/n'_1 = 1$.

Apèndixs

A. Recordatori probabilitat

Canvis de variable Si $\boldsymbol{X}$ és una VA k-dimensional absolutament continua i $\boldsymbol{Y} = g(\boldsymbol{X})$, on $g : \mathbb{R} \to \mathbb{R}^k$ és injectiva i diferenciable amb Jacobià J_g, i inversa h amb Jacobià J_h:

$$\begin{aligned} f_Y(\boldsymbol{y}) &= f_X(g^{-1}(\boldsymbol{y}))\,|J_g(g^{-1}(\boldsymbol{y}))|^{-1} \\ &= f_X(h(\boldsymbol{y}))\,|J_h(\boldsymbol{y})| \end{aligned}$$

Desigualtat de Markov Si $g : [0,\infty) \to [0,\infty)$ és una funció creixent, per qualsevol $c > 0$:

$$P\big(|X| \geq c\big) \leq \frac{\mathrm{E}\big(g(|X|)\big)}{g(c)}.$$

En particular:

$$P\big(|X - a| \geq c\big) \leq \frac{\mathrm{E}\big(|X - a|^r\big)}{c^r}.$$

i la **desigualtat de Chebyshev**:

$$P\big(|X - \mu| \geq c\big) \leq \frac{\sigma^2}{c^2}.$$

Teorema central del límit Sigui $\{X_j\}$ una successió de VA iid amb variància finita σ^2 i mitjana μ. Definim la suma parcial: $S_n = X_1 + \cdots X_n$, $\mathrm{E}(S_n) = n\,\mu$, $\sigma^2(S_n) = n\,\sigma^2$. Aleshores, la variable tipificada

$$Z_n = \frac{S_n - n\,\mu}{\sigma\sqrt{n}} \xrightarrow{\mathrm{d}} Z \sim N(0,1)$$

De forma que per n gran es pot aproximar S_n per $N(n\,\mu,\sigma\sqrt{n})$. Quan S_n és discreta és millor l'aproximació:

$$P(S_n = a) \approx F_{N(0,1)}\left(\frac{a + 0.5 - n\,\mu}{\sigma\sqrt{n}}\right) - F_{N(0,1)}\left(\frac{a - 0.5 - n\,\mu}{\sigma\sqrt{n}}\right)$$

on $F_{N(0,1)}(x)$ és la funció de distribució de $N(0,1)$.

B. Poblacions Normals

B.1 Distribució gamma i χ^2

$$\gamma(n,\theta) = \frac{\theta^n\,x^{n-1}\,\mathrm{e}^{-\theta\,x}}{\Gamma(n)},\, x \geq 0 \tag{9.3}$$

Mitjana: n/θ, variança: n/θ^2.

$$\chi^2_n = \gamma(n/2, 1/2) = \frac{1/2^n\,x^{n/2-1}\,\mathrm{e}^{-x/2}}{\Gamma(1/2)},\, x \geq 0 \tag{9.4}$$

Mitjana: n, variança: $2n$.

Propietats:

$$X_i \sim \gamma(n_i,\theta) \Rightarrow \Sigma\,X_i \sim \gamma(\Sigma\,n_i,\theta) \tag{9.5}$$

$$X \sim \gamma(n,\theta) \Rightarrow 2\theta\,X \sim \chi^2_{2n} \tag{9.6}$$

$$\chi^2_n + \chi^2_m \sim \chi^2_{n+m} \tag{9.7}$$

$$X_i \sim N(0,1) \Rightarrow \sum_{i=1}^{n} X_i^2 \sim \chi^2_n \tag{9.8}$$

$$X_i \sim N(0,\sigma) \Rightarrow \frac{n\,s^2}{\sigma^2} \sim \chi^2_n \tag{9.9}$$

B.2 Teorema de Fisher

Si $X_i \sim N(\mu,\sigma)$:

$$s^2 \text{ i } \bar{X} \text{son independents} \tag{9.10}$$

$$\bar{X} \sim N(\mu,\sigma/\sqrt{n}) \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{9.11}$$

$$\frac{n\,s^2}{\sigma^2} \sim \chi^2_{n-1} \tag{9.12}$$

B.3 Teorema de Student

$$\frac{N(0,1)}{\sqrt{\frac{1}{n}\chi^2_n}} \sim t_n = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\,\pi}\,\Gamma(\frac{n}{2})}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}},\, t \in (-\infty,\infty) \tag{9.13}$$

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\frac{(n-1)S^2}{\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \sim t_{n-1} \tag{9.14}$$

B.4 Distribució F de Snedecor

Si: $X_i \sim N(0,1)$ i $Y_i \sim N(0,1)$:

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\frac{1}{m}\sum_{i=1}^{m} Y_i^2} \sim \frac{\frac{1}{n}\chi^2_n}{\frac{1}{m}\chi^2_m} \sim F_{n,m} = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\left(\frac{n}{m}\right)^{-\frac{n}{2}}\left(1 + \frac{n}{m}t\right)^{-\frac{n+m}{2}},\, t \geq 0 \tag{9.15}$$

Si: $X_i \sim N(\mu_1,\sigma_1)$ i $Y_i \sim N(\mu_2,\sigma_2)$:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \frac{\frac{1}{n-1}\chi^2_{n-1}}{\frac{1}{m-1}\chi^2_{m-1}} \sim F_{n-1,m-1} \tag{9.16}$$

Nota: Com que $T \sim F_{n,m} \Rightarrow \frac{1}{T} \sim F_{m,n}$

$$p = P\big(T > x_p\big) = P\big(1/T < 1/x_p\big) = P\big(T' < 1/x_p\big) = 1 - P\big(T' > 1/x_p\big) \Rightarrow x_{n,m;p} = 1/x_{m,n;1-p} \quad (9.17)$$

B.5 Altres relacions

Si μ i σ són desconeguts:

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < t_{n-1;\alpha/2}\right) = 1 - \alpha \quad (9.18)$$

Si σ és coneguda:

$$P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha \quad (9.19)$$

Si: $X_i \sim N(\mu_1, \sigma_1)$ i $Y_i \sim N(\mu_2, \sigma_2)$:

$$\left.\begin{array}{l} \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}} \sim N(0,1) \\ \frac{(n-1)\,S_1^2}{\sigma_1^2} + \frac{(m-1)\,S_2^2}{\sigma_2^2} \sim \chi^2_{n+m-2} \end{array}\right\} \Rightarrow$$

$$\frac{\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}}}{\sqrt{\frac{1}{n+m-2}\left[\frac{(n-1)\,S_1^2}{\sigma_1^2} + \frac{(m-1)\,S_2^2}{\sigma_2^2}\right]}} \sim t_{n+m-2} \quad (9.20)$$

Si $n, m \geq 15 \Rightarrow S_1^2 \approx \sigma_1^2$, $S_2^2 \approx \sigma_2^2$:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim N(0,1) \quad (9.21)$$

Referències

[1] César Pérez López. *Técnicas de muestreo estadístico*. Ibergarceta Publicaciones, S.L., 2010.

[2] Alfonso García Pérez. *Problemas resuleteos de Teoría de Muestras en Poblaciones Finitas*. UNED, 1994.

Quadre amb algunes distribucions

Distribució	Paràmetres	Funció de densitat	Mitjana	Variança	Funció característica
Bernoulli	$0 \leq p \leq 1$ $q = 1 - p$	$p^k\,(1-p)^{1-k}$ $k = 0{,}1$	p	$p\,(1-p)$	$q + p\,\mathrm{e}^{i\,t}$
Binomial	$0 \leq p \leq 1$ $q = 1 - p$	$\binom{n}{k} p^k\,(1-p)^{n-k}$ $k = 0{,}1,\ldots n$	$n\,p$	$n\,p\,(1-p)$	$(q + p\,\mathrm{e}^{i\,t})^n$
Geomètrica	$0 \leq p \leq 1$ $q = 1 - p$	$p\,(1-p)^k$ $k \geq 0$	$\dfrac{1-p}{p}$	$\dfrac{1-p}{p^2}$	$\dfrac{p}{1 - q\,\mathrm{e}^{i\,t}}$
Binomial negativa	$r > 0$ $0 \leq p \leq 1$ $q = 1 - p$	$\binom{k+r-1}{k} p^r\,q^k$ $k \geq 0$	$r\,\dfrac{1-p}{p}$	$r\,\dfrac{1-p}{p^2}$	$\left(\dfrac{p}{1 - q\,\mathrm{e}^{i\,t}}\right)^r$
Poisson	$\lambda > 0$	$\dfrac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}, \quad k \geq 0$	λ	λ	$\exp\left\{\lambda\,(\mathrm{e}^{i\,t} - 1)\right\}$
Normal $N(\mu,\sigma)$	$\mu \in \mathbb{R}$ $\sigma > 0$	$\dfrac{1}{\sqrt{2\,\pi\,\sigma^2}}\,\exp\left\{\dfrac{-(x-\mu)^2}{2\,\sigma^2}\right\}$ $x \in \mathbb{R}$	μ	σ^2	$\exp\left\{\mu\,i\,t - \dfrac{t^2\,\sigma^2}{2}\right\}$
Uniforme	$a < b$	$\dfrac{1}{b-a}, \quad a < x < b$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$	$\dfrac{\mathrm{e}^{i\,t\,b} - \mathrm{e}^{i\,t\,a}}{i\,t\,(b-a)}$
Exponencial	α	$\alpha\,\mathrm{e}^{-\alpha\,x}, \quad x > 0$	$\dfrac{1}{\alpha}$	$\dfrac{1}{\alpha^2}$	$\left(1 - \dfrac{i\,t}{\alpha}\right)^{-1}$
Gamma $\gamma(n,\alpha)$	$\alpha > 0,$ $n > 0$	$\dfrac{\alpha^n\,x^{n-1}\,\mathrm{e}^{-\alpha\,x}}{\Gamma(n)}, \quad x > 0$	$\dfrac{n}{\alpha}$	$\dfrac{n}{\alpha^2}$	$\left(1 - \dfrac{i\,t}{\alpha}\right)^{-n}$
Beta $\beta(p,q)$	$p > 0,$ $q > 0$	$\dfrac{x^{p-1}\,(1-x)^{q-1}}{\mathrm{B}(p{,}q)},$ $0 < x < 1$	$\dfrac{p}{p+q}$	$\dfrac{p\,q}{(p+q)^2(p+q+1)}$	
Cauchy	$a > 0,$ $b \in \mathbb{R}$	$\dfrac{a}{\pi\,\left[a^2 + (x-b)^2\right]}$ $x \in \mathbb{R}$	No existeix	No existeix	$\mathrm{e}^{i\,t\,b - a\,\|t\|}$

$$\Gamma(x) = \int_0^\infty e^{-t}\,t^{x-1}\,dt,\ \Gamma(n) = (n-1)! \quad \mathrm{B}(p{,}q) = \int_0^1 t^{p-1}\,(1-t)^{q-1}\,dt,\ \mathrm{B}(p{,}q) = \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}$$