

Quadre amb algunes distribucions

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

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

Apunts de teoria de mostres

Llorenç Cerdà-Alabern

llorencc@ac.upc.edu

Barcelona, 3 de setembre de 2014

ritzar les relacions que demostren. L'edició l'he fet amb L^AT_EX.

Índex

1	Definicions	1
1.1	Precisió d'un estimador	3
1.2	Intervals de confiança	3
2	Mostreig amb probabilitats desiguals	4
2.0.1	Estimador de Horvitz i Thompson	4
3	Mostreig aleatori simple	5
4	Mostreig estratificat	7
5	Estimador de raó	8
6	Estimador per regressió	9
7	Mostreig sistemàtic	10
8	Mostreig per conglomerats	10
8.1	Conglomerats de mides diferents	11
9	Mostreig doble	12
9.1	Falta de resposta	12
Apèndixs		
A.	Recordatori probabilitat	13
B.	Poblacions Normals	13
B.1	Distribució gamma i χ^2	13
B.2	Teorema de Fisher	13
B.3	Teorema de Student	13
B.4	Distribució F de Snedecor	13
B.5	Altres relacions	14

Prefaci

Vaig escriure aquests apunts mentre preparava l'assignatura *teoria de mostres* 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 complet, i només hi ha demostracions senzilles orientades a entendre o ajudar a memo-

Capítol 1

Definicions

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

- **Població objectiu:** La que es desitja investigar.
- **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 $S_i = \{U_{i_1}, \dots, U_{i_n}\}$, $U_{ij} \in U$. També és possible que no importi l'ordre, i farem servir la notació $s_i = \{u_1, \dots, u_n\}$, $u_j \in U$. Les mostres són variables aleatòries, en general dependents. Per tant:

$$P(S_i) = P(U_{i_1})P(U_{i_2} | U_{i_1}) \dots P(U_{i_n} | U_{i_1}, \dots, U_{i_{n-1}}) \quad (1.1)$$

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 $\theta(X_1, \dots, X_N)$ d'una característica poblacional. Exemples

- Mitjana poblacional:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (1.2)$$

- Total poblacional¹:

$$X = \sum_{i=1}^N X_i = N \bar{X} \quad (1.3)$$

- Variancia poblacional:

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

- Quasivariancia poblacional:

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

- Mitjana quadràtica:

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

- Mitjana geomètrica:

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

- Mitjana harmònica:

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

Si la característica és qualitativa:

- Proporció poblacional:

$$P = \frac{1}{N} \sum_{i=1}^N A_i \quad (1.9)$$

- Total de classe poblacional:

$$A = \sum_{i=1}^N A_i = NP \quad (1.10)$$

Estimador (o estadístic) (*statistic*) Donada una mostra $s_i = \{u_1, \dots, u_n\}$ farem servir la notació $s_i(X) = \{X_1, \dots, X_n\}$ 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 $\hat{\theta} : S(X) \in \mathbb{R}^n \rightarrow \mathbb{R}$, que estima la característica poblacional θ a partir d'una mostra: $\hat{\theta}(s_i(X)) = \theta(X_1, \dots, X_n)$. Exemples:

- Mitjana mostral:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i \quad (1.11)$$

- Total mostral:

$$\hat{X} = \frac{N}{n} \sum_{i=1}^n X_i = N \bar{x} \quad (1.12)$$

- Proporció mostral:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n A_i \quad (1.13)$$

- Total de classe mostral:

$$\hat{A} = \frac{N}{n} \sum_{i=1}^n A_i = N \hat{p} \quad (1.14)$$

- 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 $T = \{t \in \mathbb{R} \mid \exists (X_1, \dots, X_n) \in S(X), \theta(X_1, \dots, X_n) = t\}$. Aleshores

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

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

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

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

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 $Y_i \sim N(\mu_2, \sigma_2)$:

1.1. Precisió d'un estimador

Esperança d'un estimador

$$E(\hat{\theta}) = \sum_S \hat{\theta}(s_i) P(s_i) = \sum_{\mathbb{R}} t P(\hat{\theta} = t) \quad (1.16)$$

Variància d'un estimador

$$V(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2 \quad (1.17)$$

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

$$\sigma(\hat{\theta}) = \sqrt{V(\hat{\theta})} \quad (1.18)$$

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

$$CV(\hat{\theta}) = \frac{\sigma(\hat{\theta})}{E(\hat{\theta})} \quad (1.19)$$

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

$$EQM(\hat{\theta}) = E((\hat{\theta} - \theta)^2) \quad (1.20)$$

Biaix d'un estimador

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta \quad (1.21)$$

Es diu que un estimador és consistent quan

$$B(\hat{\theta}) \xrightarrow{n \rightarrow \infty} 0.$$

1.1 Precisió d'un estimador

$$EQM(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = E((\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2) = E((\hat{\theta} - E(\hat{\theta})) + B(\hat{\theta}))^2 = \sigma^2(\hat{\theta}) + B(\hat{\theta})^2 \quad (1.22)$$

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:

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

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: $\sigma^2(\hat{\theta}_1)/\sigma^2(\hat{\theta}_2)$.

1.2 Intervalls de confiança

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

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

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 dona 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

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

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

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

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

$$P(-z_{\alpha/2} \sqrt{V(\hat{\theta})} \leq \hat{\theta} - \theta \leq z_{\alpha/2} \sqrt{V(\hat{\theta})}) \quad (1.26)$$

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

Definim $z_{\alpha/2}$ com (quantil $1 - \alpha$):

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

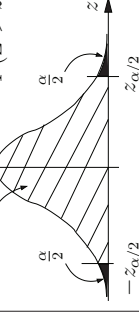


Figura 1.1: Interval de confiança en una distribució simètrica: IC = $[\hat{\theta} - z_{\alpha/2} \sqrt{V(\hat{\theta})}, \hat{\theta} + z_{\alpha/2} \sqrt{V(\hat{\theta})}]$.

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

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

Estimador no biaixat per a una distribució desconeuda

Amb la desigualtat de Chebyshev:

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

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

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

on s'ha posat $c = \sqrt{\frac{V(\hat{\theta})}{\alpha}}$.

Per tant,

$$\left[\hat{\theta} - \sqrt{\frac{V(\hat{\theta})}{\alpha}}, \hat{\theta} + \sqrt{\frac{V(\hat{\theta})}{\alpha}}\right] \text{ és un IC.}$$

Estimador biaixat amb distribució normal És fàcil comprovar que amb $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ és té que l'IC és

$$\left[\hat{\theta} - z_{\alpha/2} - B(\hat{\theta}), \hat{\theta} + z_{\alpha/2} - B(\hat{\theta})\right].$$

Capítol 2

Mostreig amb probabilitats desiguals

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

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

$$e_i e_j = \begin{cases} 1, & \text{si } u_i, u_j \in \tilde{x} \text{ amb probabilitat } \pi_i \pi_j, \\ 0, & \text{si } u_i, u_j \notin \tilde{x} \text{ amb probabilitat } (1 - \pi_i)(1 - \pi_j). \end{cases} \quad (2.2)$$

Clarament:

$$E(e_i) = \pi_i \quad (2.3)$$

$$E(e_i^2) = \pi_i \quad (2.4)$$

$$V(e_i) = \pi_i(1 - \pi_i) \quad (2.5)$$

$$\text{Cov}(e_i e_j) = E(e_i e_j) - E(e_i) E(e_j) = \pi_i \pi_j - \pi_i \pi_j \quad (2.6)$$

Propietats

$$\sum_{i=1}^N \pi_i = n \quad (2.7)$$

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

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

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

Demostració.

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

2.0.1 Estimador de Horvitz i Thompson

És un estimador no biaixat per mostreig sense reposició.

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

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

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

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

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

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

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

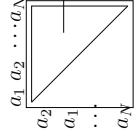


Figura 2.1: Quadrat d'una suma.

Apèndix

A. Recordatori probabilitat

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

$$f_Y(\mathbf{y}) = f_X(g^{-1}(\mathbf{y})) |J_g(g^{-1}(\mathbf{y}))|^{-1}$$

$$= f_X(h(\mathbf{y})) |J_h(\mathbf{y})|$$

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

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

En particular:

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

i la desigualtat de Chebyshev:

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

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

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

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

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

Mitjana: n/θ , variància: n/θ^2 .

$$\chi_n^2 = \gamma(n/2, 1/2) = \frac{1/2^n x^{n/2-1} e^{-x/2}}{\Gamma(1/2)}, x \geq 0 \quad (9.4)$$

Mitjana: n , variància: $2n$.

Propietats:

$$X_i \sim \gamma(n_i, \theta) \Rightarrow \sum X_i \sim \gamma(\sum n_i, \theta) \quad (9.5)$$

$$X \sim \gamma(n, \theta) \Rightarrow 2\theta X \sim \chi_{2n}^2 \quad (9.6)$$

$$\chi_n^2 + \chi_m^2 \sim \chi_{n+m}^2 \quad (9.7)$$

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

$$X_i \sim N(0, \sigma) \Rightarrow \frac{n s^2}{\sigma^2} \sim \chi_n^2 \quad (9.9)$$

B.2 Teorema de Fisher

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

$$s^2 \text{ i } \bar{X} \text{ son independents} \quad (9.10)$$

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

$$\frac{n s^2}{\sigma^2} \sim \chi_{n-1}^2 \quad (9.12)$$

B.3 Teorema de Student

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

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

B.4 Distribució F de Snedecor

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

$$\frac{1}{n} \sum_{i=1}^n X_i^2 \sim \frac{1}{n} \chi_n^2 \sim F_{n, m} =$$

$$\frac{1}{m} \sum_{i=1}^m Y_i^2 \sim \frac{1}{m} \chi_m^2$$

$$\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{n}{m}\right)^{-\frac{n}{2}} \left(1 + \frac{n}{m}t\right)^{-\frac{n+m}{2}}, t \geq 0 \quad (9.15)$$

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

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \frac{1}{m-1} \chi_{m-1}^2 \sim F_{n-1, m-1} \quad (9.16)$$

Capítol 9

Mostreig doble

Població N estratificada en L estrats.

Primera fase: mostra $n'_h = \sum n'_{hk}$. 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'_{hk} .

Estimador de la mitjana:

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

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:

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

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:

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

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

Tenim que:

$$\hat{\theta} = \sum_{i=1}^n \frac{Y_i}{\pi_i} = \sum_{i=1}^n \frac{Y_i}{\pi_i} e_i \quad (2.16)$$

Substituint en 2.15:

$$\begin{aligned} \mathbf{V}(\hat{\theta}) &= \sum_{i=1}^n \mathbf{V} \left(\frac{Y_i}{\pi_i} e_i \right) + \\ & 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{Cov} \left(\frac{Y_i}{\pi_i} e_i, \frac{Y_j}{\pi_j} e_j \right) = \\ & \sum_{i=1}^n \left(\frac{Y_i}{\pi_i} \right)^2 \mathbf{V}(e_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n \frac{Y_i Y_j}{\pi_i \pi_j} \text{Cov}(e_i, e_j) \end{aligned}$$

D'on:

$$\begin{aligned} \mathbf{V}(\hat{\theta}) &= \sum_{i=1}^n \frac{Y_i^2}{\pi_i} (1 - \pi_i) + \\ & 2 \sum_{i=1}^n \sum_{j=i+1}^n \frac{Y_i Y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j) \quad (2.17) \end{aligned}$$

Estimadors no biaxials de la variancia del paràmetre Y :

$$\begin{aligned} \hat{\mathbf{V}}(\hat{\theta}) &= \sum_{i=1}^n \frac{Y_i^2}{\pi_i} (1 - \pi_i) + \\ & 2 \sum_{i=1}^n \sum_{j=i+1}^n \frac{Y_i Y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j) \quad (2.18) \\ \hat{\mathbf{V}}(\hat{\theta}) &= \sum_{i=1}^n \sum_{j=1+1}^n \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \quad (2.19) \end{aligned}$$

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,

$$\pi_i = \frac{n}{N} \quad (3.1)$$

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

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:

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

$$\text{Total: } \mathbf{V}(\hat{X}) = N^2 \mathbf{V}(\hat{\hat{x}}) \quad (3.4)$$

$$\text{Proporció: } \mathbf{V}(\hat{p}) = (1-f) \frac{S^2}{n} = \frac{N-n}{N-1} \frac{PQ}{n} \quad (3.5)$$

$$\text{Total classe: } \mathbf{V}(\hat{A}) = N^2 \mathbf{V}(\hat{p}) \quad (3.6)$$

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

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

Per a variables qualitatives:

$$\hat{S}^2 = \frac{n \hat{p} \hat{q}}{n-1} \quad (3.8)$$

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

6

Capítol 3. Mostreig aleatori simple

8.1. Conglomerats de mides diferents

11

Formulari

Taula 3.1: Estadístics

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

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

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

Taula 3.2: Variàncies mostreig aleatori simple

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

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

Relació fonamental:

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

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

D'on es té:

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

D'on la variància de la mitjana ($f = (nM)/(NM) = n/N$):

$$V(\bar{x}) = (1-f) \frac{S_b^2}{nM} \quad (8.4)$$

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

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

on:

$$\sigma^2 = PQ \quad (8.6)$$

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

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

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

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

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

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

Es té:

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

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

$$\bar{M} = \sum_{i=1}^n M_i / N \quad (8.12)$$

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 variància:

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

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

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

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

Capítol 7

Mostreig sistemàtic

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

	1	2	...	k
1	n_{11}	n_{12}	...	n_{1k}
2	n_{21}	n_{22}	...	n_{2k}
...
n	n_{n1}	n_{n2}	...	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 mostrat són les columnes, que es trien amb probabilitat $1/k$. El mostreig és sense reposició. Si es tria la columna j els estimadors són:

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

$$\text{Total : } \hat{X} = N \bar{x}_j \quad (7.2)$$

$$\text{Proporció : } \hat{P} = \frac{1}{n} \sum_{i=1}^n A_{ij} = \hat{P}_j \quad (7.3)$$

$$\text{Classe : } \hat{A} = N \hat{P}_j \quad (7.4)$$

Quasivariància inter-mostrat:

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

Quasivariància intra-mostrat:

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

Relació fonamental:

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{X})^2 &= \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{x}_j + \bar{x}_j - \bar{X})^2 = \\ &= \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{x}_j)^2 + \sum_{i=1}^n \sum_{j=1}^k (\bar{x}_j - \bar{X})^2 + \\ &= 2 \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{x}_j)(\bar{x}_j - \bar{X}) = \\ &= \underbrace{\sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{x}_j)^2 + \sum_{i=1}^n \sum_{j=1}^k (\bar{x}_j - \bar{X})^2}_{=0} \end{aligned}$$

D'on es té:

$$(N-1)S^2 = (N-k)S_{ws}^2 + (k-1)S_{bs}^2 \quad (7.7)$$

d'on:

$$V(\hat{X}) = (1-f) \frac{S_{bs}^2}{n} = \sigma^2 - \frac{n-1}{n} S_{ws}^2 \quad (7.8)$$

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:

$$V(\hat{P}) = PQ - \frac{1}{k} \sum_{j=1}^k \hat{P}_j \hat{Q}_j \quad (7.9)$$

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:

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

$$\text{Total : } \hat{X} = NM \bar{\bar{X}}$$

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

$$\text{Classe : } \hat{A} = NM \hat{P}$$

Quasivariància inter-conglomerats:

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^N \sum_{j=1}^M (\bar{X}_i - \bar{X})^2 = \frac{M}{N-1} \sum_{i=1}^N (\bar{X}_i - \bar{X})^2 \quad (8.1)$$

Quasivariància intra-conglomerats:

$$S_w^2 = \frac{1}{NM-N} \sum_{i=1}^N \sum_{j=1}^M (X_{ij} - \bar{X}_i)^2 \quad (8.2)$$

Capítol 4

Mostreig estratificat

Població N dividida en L estrats homogenis de mida N_h , $L = 1, \dots, L$. 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 , variància σ_h^2 . Quasivariància S_h^2 , proporció P_h i total de classe A_h .

Taula 4.1: Variàncies mostreig estratificat

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

$$S_h^2 = \frac{1}{N_h-1} \sum_{i=1}^N (X_{hi} - \bar{X}_h)^2, \hat{S}_h^2 = \frac{1}{n_h-1} \sum_{i=1}^n (X_{hi} - \bar{x}_h)^2.$$

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

Afixacions

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

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

Estadístics mostrals de l'estrat h : mitjana \bar{x}_h , total X_h , variància σ_h^2 , quasivariància 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 biaixats dels estadístics poblacionals: mitjana \bar{x}_{st} , total \hat{X}_{st} , proporció \hat{P}_{st} i total de classe \hat{A}_{st} .

Capítol 5

Estimador de raó

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

Raó poblacional :

$$R = \frac{\bar{X}}{\bar{Y}} = \frac{\sum_{i=1}^N X_i}{\sum_{i=1}^N Y_i} = \frac{\bar{X}}{\bar{Y}} \quad (5.1)$$

Raó mostral :

$$\hat{R} = \frac{\hat{X}}{\hat{Y}} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i} = \frac{\bar{x}}{\bar{y}} \quad (5.2)$$

Estimador de la mitjana :

$$\hat{X}_R = \hat{R} \bar{Y} \quad (5.3)$$

Estimador del total :

$$\hat{X}_R = N \hat{X}_R = N \hat{R} \bar{Y} \quad (5.4)$$

Biaix :

$$B(\hat{R}) = E(\hat{R}) - R = \frac{R \text{V}(\bar{y}) - \text{Cov}(\bar{x}, \bar{y})}{\bar{Y}^2} \quad (5.5)$$

Taula 5.1: Variàncies estimador de raó

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

Normalment farem servir els estimadors de les quasivariàncies 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: $S_x^2 - 2\hat{R}S_{xy} + \hat{R}^2 S_y^2 = (\sum X_i^2 + R^2 \sum Y_i^2 - 2R \sum X_i Y_i) / (n-1)$

Nota: per el càlcul de la variància 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:

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

Capítol 6

Estimador per regressió

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

Estimador de la mitjana :

$$\bar{x}_{r,y} = \hat{x} + b(\hat{Y} - \hat{y}) \quad (6.1)$$

- El valor de b que minimitza la variància és la pendent de la recta de regressió de X sobre Y ; $b_0 = S_{xy}/S_y^2$, que s'estima per $b_0 = \hat{S}_{xy}/\hat{S}_y^2$.

- Si b és una constant, l'estimador $\bar{x}_{r,y}$ és no biaixat.

Taula 6.1: Variàncies estimador per regressió

	Mostreig Simple (MS)	MS amb reposició
Mitjana		
$V(\bar{x}_{r,y}) =$	$(1-f) \frac{S_x^2 - 2bS_{xy} + b^2 S_y^2}{n}$	$\frac{S_x^2 - 2bS_{xy} + b^2 S_y^2}{n}$
Total		
$V(\hat{X}_{r,y}) =$	$N^2 V(\hat{X}_{r,y})$	$N^2 V(\hat{X}_{r,y})$

Taula 6.2: Variàncies mínimes (amb $b_0 = S_{xy}/S_y^2$)

	Mostreig Simple (MS)	MS amb reposició
Mitjana		
$V(\bar{x}_{r,y}) =$	$(1-f) \frac{S_x^2 (1-\rho^2)}{n}$	$\frac{\sigma_x^2 (1-\rho^2)}{n}$
Total		
$V(\hat{X}_{r,y}) =$	$N^2 V(\hat{X}_{r,y})$	$N^2 V(\hat{X}_{r,y})$

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

Notar que:

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

D'on:

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

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

i