OBSERVATÓRIO ASTRONÓMICO
UNIVERSIDADE DE COIMBRA

Meeting SPE/CIM
Methodological Issues in Official Statistics

Centro Internacional de Matemática

» Estimação de Variância

Casos Práticos «

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Rita Sousa
(Com a colaboração do DMSI/ME)

(17-11-2007)

»

Estimação de Variância

- Contextualização
- Principais Estimadores
- Calibração
- Plano Complexo de Amostragem
- Método de Linearização
- Método dos Grupos Aleatórios
- Métodos de Reamostragem
- Casos Práticos

Inquéritos por Amostragem

- Nos inquéritos por amostragem pretende-se analisar as características de uma dada **população** de **dimensão** N, com base numa **amostra** de n **unidades** extraídas dessa mesma população.

- No INE, IP, os **métodos** de amostragem são, em geral, **probabilísticos** (aleatórios), pelo que é quase sempre possível associar a cada elemento da população a respectiva **probabilidade de inclusão** na amostra.

Bases de Amostragem

Amostra-Mãe

- A **Amostra-Mãe** é uma **base de amostragem de alojamentos familiares**, a partir da qual são seleccionadas as amostras para os inquéritos a realizar pelo INE, IP, junto das **famílias**.

- Foi **criada em 2001** através dos **Censos da população** e, passados estes anos, é necessário proceder à actualização da informação, atribuindo-lhe melhor qualidade, que se reflecte também nos inquéritos efectuados pelo INE, IP.

Bases de Amostragem

Amostra-Mãe

- A **actualização da Amostra-Mãe** processa-se a partir de um **Inquérito próprio** de actualização e de **Inquéritos às Famílias** (IE, IDEF, ICOR, INS, IPTR, IUTICF)*;

- Em 2006 foi feito um inquérito piloto a 14 áreas mas no **1º trimestre de 2007 teve início o** efectivo **processo de actualização**. Prevê-se que tenha a **duração de 3 anos**, já que se selecciona cerca de 1/12 do nº total de áreas, dando-se prioridade às **áreas com maior índice de esgotamento** (com menor nº de alojamentos de residência principal).

---

* IE – Inquérito ao Emprego;
IDEF – Inquérito às Despesas das Famílias;
ICOR – Inquérito às Condições de Vida e Rendimento;
INS – Inquérito Nacional de Saúde;
IPTR – Inquérito à Procura Turística de Residentes;
IUTICF – Inquérito à Utilização das Tecnologias de Informação e Comunicação nas Famílias.

Bases de Amostragem

Ficheiro de Unidades Estatísticas (FUE)

- O **FUE** é uma base de amostragem das **Empresas** (Sociedades ou Empresários em Nome Individual), **Instituições sem Fins Lucrativos** e **Administração Pública**, a partir da qual são seleccionadas as amostras para os inquéritos a realizar pelo INE, IP, junto das **empresas**.

- A **actualização** do FUE é feita com base em inquéritos do INE, IP, (**fontes internas**) e com base em ficheiros administrativos (**fontes externas**), como é o caso da IES*, que permite melhorar a consistência e actualização da informação de periodicidade anual.

---

* IES – Informação Empresarial Simplificada

Estimação da Variância

- De um modo geral, nos inquéritos por amostragem, pretende-se estimar características da população como **totais**, **médias** ou **proporções** (de indivíduos com uma determinada restrição).

- No cálculo das **estimativas populacionais** é **conveniente avaliar** a **precisão** dessas mesmas estimativas, através das variâncias e respectivos coeficientes de variação.

Total Populacional

Seja Y a variável aleatória em estudo e $\{Y_1, Y_2, \ldots, Y_n\}$ a respectiva amostra aleatória de dimensão n, para uma população de N indivíduos.

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

- Seja σ^2 a variância da população e S^2 o respectivo estimador centrado da variância:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \text{, com } \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Total/Média Populacionais

Amostragem Aleatória Simples

- Para uma **amostra aleatória simples** (sem reposição) o **estimador do total** populacional é:

$$\hat{Y}_T = N\overline{Y} = N\frac{\sum_{i=1}^{n} Y_i}{n}$$

Estimador da variância: $S_T^2 = \hat{\sigma}_T^2 = N^2\hat{\sigma}_{\overline{Y}}^2 = N^2\left[(1-f)\frac{S^2}{n}\right]$

Em que $f = \frac{n}{N}$ corresponde à fracção de amostragem.

Total/Média Populacionais

Amostragem Aleatória Estratificada

Na **amostragem aleatória estratificada** a população é dividida em **grupos** (estratos) considerados **mais homogéneos** em relação a determinadas características (variáveis de estratificação).

- Neste caso, para k estratos, temos:

$N = \sum_{h=1}^{k} N_h$, que representa a **dimensão** total da **população**;

$n = \sum_{h=1}^{k} n_h$, que representa a **dimensão** total da **amostra**;

Em que N_h e n_h correspondem ao número de unidades no estrato h para a população e para a amostra, respectivamente.

Total/Média Populacionais

Amostragem Aleatória Estratificada

- Para uma **amostra aleatória estratificada**, considerando a população dividida em k estratos, o **estimador do total** populacional é:

$$\hat{Y}_T = N\overline{Y} = N\sum_{h=1}^{k}\frac{N_h}{N}\overline{Y}_h = \sum_{h=1}^{k}N_h\overline{Y}_h$$

Estimador da variância: $S_T^2 = \hat{\sigma}_T^2 = N^2\hat{\sigma}_{\overline{Y}}^2 = N^2\sum_{h=1}^{k}W_h^2\left(1-f_h\right)\frac{S_h^2}{n_h}$

Em que $W_h = \frac{N_h}{N}$ corresponde ao peso do estrato h na população e $f_h = \frac{n_h}{N_h}$ corresponde à fracção de amostragem no estrato h.

Proporção Populacional

Para uma dada população de dimensão N, a proporção de indivíduos que verifica uma dada característica é dada por:

$$p_T = \frac{R}{N}$$

- Sendo R a soma de N variáveis com distribuição binomial, de parâmetros N e p, a variância da proporção populacional é dada por:

$$\sigma_p^2 = \frac{p(1-p)}{N}$$

Proporção Populacional

Amostragem Aleatória Simples

- Para uma **amostra aleatória simples** (sem reposição) o **estimador da proporção** populacional é:

$$\hat{p}_T = \frac{r}{n}$$

Estimador da variância: $S_p^2 = \hat{\sigma}_p^2 = (1 - f)\frac{\hat{p}(1-\hat{p})}{n-1}$

Em que $f = \frac{n}{N}$ corresponde à fracção de amostragem.

Proporção Populacional

Amostragem Aleatória Estratificada

- Para uma **amostra aleatória estratificada**, considerando a população dividida em k estratos, o **estimador da proporção** populacional é:

$$\hat{p}_T = \sum_{h=1}^{k} \frac{N_h}{N} \hat{p}_h = \sum_{h=1}^{k} \frac{N_h}{N} \frac{r_h}{n_h}$$

Estimador da variância:

$$S_p^2 = \hat{\sigma}_p^2 = \sum_{h=1}^{k} W_h^2 S_{p_h}^2 = \sum_{h=1}^{k} W_h^2 \left(1 - f_h\right) \frac{\hat{p}_h (1 - \hat{p}_h)}{n_h - 1}$$

Em que $W_h = \frac{N_h}{N}$ corresponde ao peso do estrato h na população e $f_h = \frac{n_h}{N_h}$ corresponde à fracção de amostragem no estrato h.

Estimação da Variância

O cálculo da variância é uma **tarefa mais complexa** quando:

- Temos **estimadores** populacionais **não lineares** ou que não correspondem a nenhuma combinação linear de médias ou totais;
- Se recorre a **planos de amostragem complexos**;
- Se faz **imputação** (de não respostas);
- Se recorre à **calibração** da amostra.

Calibração

- A **calibração** do desenho amostral consiste em recorrer a informação auxiliar para fazer o **ajustamento dos ponderadores iniciais**, obtendo-se totais marginais idênticos aos conhecidos da população.

- Este processo pretende **corrigir algumas distorções da amostra**, em relação à população de referência, que se podem dever por exemplo ao **carácter aleatório** ou a **não respostas**.

- A calibração de ponderadores iniciais pode ser feita, por exemplo, por um **processo de ajustamento por margens**, que cria novos ponderadores tão próximo quanto possível dos iniciais, por **minimização de uma dada função distância.**

Método de Ajustamento por Margens

- **Sejam:**

d_i – Pesos iniciais, de amostragem;

w_i – Pesos finais ajustados, ou seja, as ponderações procuradas;

$\mathbf{G}$ – Função distância, de argumento w_i/d_i, que mede a distância entre os ponderadores;

$\boldsymbol{s}$ – Amostra seleccionada;

X_i – vector linha com os valores da observação i para as variáveis auxiliares;

$\mathbf{X}$ – vector linha das margens de ajustamento.

O problema consiste em encontrar os pesos w_k que são **solução** de:

$$Min \sum_{i \in s} d_i G(w_i / d_i)$$, sob a restrição $$\sum_{i \in s} w_i X_i = \mathrm{X}$$

Calibração

Exemplo

Um **exemplo** do recurso às metodologias de calibração, no INE, IP, é o **Inquérito ao Emprego** (IE):

- O **ponderador inicial** é calculado por estrato de acordo com o desenho amostral; $$d_i = \frac{1}{probsel_i}$$
- Aplica-se um factor de correcção para as não-respostas;

$$w_i = \frac{1}{probsel_i} \cdot \frac{PopNUTSIII_r}{\sum_{i=1}^{n_r} \frac{1}{probsel_i}}$$

- Calibram-se os ponderadores iniciais para que os totais marginais coincidam com os da população, recorrendo-se a informação externa* ao inquérito.

* Projecções mensais da população.

Calibração

Exemplo

As **margens** utilizadas no IE baseiam-se nas **estimativas mensais independentes da população**, para as seguintes variáveis de desagregação:

- Por região **NUTS II**, para o **escalão etário quinquenal** (sendo o último de 75 e + anos) e **sexo** e para o **escalão etário 18-24 anos** e **sexo**;
- Por região **NUTS III**, para **seis escalões etários** (0-14; 15-24; 25-34; 35-44; 45-64; 65 e + anos) e **sexo**.

Função de distância utilizada é a do "**método *logit*** (método *ranking ratio* com limites)"

Calibração

Exemplo

Efeitos da Calibração

região	area	aloj	nfam	nind	sexo	idade	sexee	peso inicial	peso final	quope
norte	22	228	1	1	1	31	7	294,87	812,19	2,75
norte	1	107	1	2	2	68	30	297,46	200,10	0,67
centro	448	195	1	1	2	54	27	381,95	135,97	0,36
centro	518	883	1	3	1	4	1	402,94	1457,98	3,62
centro	641	90	1	2	2	21	21	387,60	888,35	2,29
centro	653	185	1	3	1	14	3	394,33	606,48	1,54
lisboa	635	29	1	3	2	7	18	352,17	196,45	0,56
lisboa	929	292	1	1	1	58	12	417,85	343,63	0,82
alentejo	1076	310	1	3	1	1	1	253,46	931,31	3,67
alentejo	1077	280	1	2	2	49	26	225,82	279,23	1,24
alentejo	1131	164	1	2	2	38	24	111,07	117,26	0,95
alentejo	1135	360	1	1	2	41	25	136,11	132,06	1,03
algarve	1271	401	1	3	1	6	2	97,37	87,66	0,90
algarve	1259	629	1	2	1	38	8	98,55	183,18	1,86
açores	1312	237	1	3	1	19	4	60,30	30,31	0,50
açores	1301	427	1	2	2	64	29	60,08	59,26	0,99
madeira	1363	183	1	1	2	56	28	63,92	45,88	0,72
madeira	1401	238	1	3	2	13	19	63,75	117,40	1,84

Plano Amostral Complexo

- Um **plano amostral complexo** é aquele cujo desenho incorpora alguns **níveis de complexidade**, tais como **estratificação**, **conglomeração** e **probabilidades desiguais de selecção**.

- As **estimativas pontuais** dos parâmetros são **influenciadas pela ponderação da amostra**, enquanto que as **estimativas da variância** dos estimadores dos parâmetros do modelo são **influenciadas pelos efeitos de estratificação e conglomeração**.

Conglomeração

- Um **exemplo** do recurso à amostragem por conglomerados é a **amostragem multi-etápica** em que numa primeira fase se agrupam as unidades em **subpopulações**, facilmente identificáveis (por exemplo: escolas, hospitais, áreas geográficas, etc.).

- Estes subgrupos são considerados heterogéneos e a **amostragem é feita sobre os conglomerados** e não sobre os indivíduos da população.

Amostragem Multi-etápica

Quando existe apenas uma **única etapa**, **observam-se todos os elementos dos conglomerados seleccionados**.

Na amostragem **multi-etápica** a amostra é seleccionada em **duas ou mais etapas** consecutivas:

- **PSU** – "primary sampling unit" (unidade de amostragem primária);
- **SSU** – "secondary sampling unit" (unidade de amostragem secundária);
- **TSU** – "tertiary sampling unit" (unidade de amostragem terciária);
- …

Amostragem Multi-etápica

Exemplo

No **Inquérito Nacional de Saúde (INS)** pretende-se obter estimativas sobre o estado de saúde/doença da população portuguesa, os principais factores e a utilização de cuidados de saúde. Recorre-se a uma **amostra probabilística multi-etápica**, com **estratificação a nível geográfico**. A selecção é feita em **duas etapas**, a partir da Amostra-Mãe (AM):

- As **unidades primárias** correspondem às **áreas** da AM;
- As **unidades secundárias** correspondem aos **alojamentos familiares de residência principal**, existentes em cada uma das áreas. Nestas unidades não se realiza qualquer amostragem, dado que se recolhe informação de todos os indivíduos que aí residem.

Estimação da Variância

Métodos

Para **estimação da variância** dos estimadores, em inquéritos complexos, existem vários **métodos**:

- Método da **Linearização**;
- Método dos **Grupos Aleatórios**;
- Métodos de **Reamostragem**.

Linearização

- Quando o estimador não é linear pode recorrer-se a uma **aproximação** por um linear, com o propósito de estimar a variância, por exemplo pelo **Método das Séries de *Taylor*.**

- Geralmente este método é aplicado quando dispomos de uma **função** (designada por "***smooth***") expressa pelas **médias** e/ou **totais populacionais**.

- Apesar deste método já estar teoricamente bem desenvolvido, pode tornar-se **complexo na dedução das derivadas parciais**.

Linearização

- Por exemplo, para um dado **parâmetro populacional θ**, função não linear de k médias populacionais $(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_k)$:

$$\theta = h\left(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_k\right) \Rightarrow \hat{\theta} = h\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2, ..., \hat{\overline{Y}}_k\right)$$

Linearizando a função ***Smooth***, pelo **método de *Taylor***, temos:

$$h\left(\hat{\overline{Y}}_1, \hat{\overline{Y}}_2, ..., \hat{\overline{Y}}_k\right) = h\left(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_k\right) + \sum_{i=1}^{k} \frac{\partial h\left(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_k\right)}{\partial \overline{Y}_i} \left(\hat{\overline{Y}}_i - \overline{Y}_i\right)$$

Variância do Estimador:

$$Var\left(\hat{\theta}\right) = Var\left[\sum_{i=1}^{k} \frac{\partial h}{\partial \overline{Y}_i} \hat{\overline{Y}}_i\right]$$

Método de *Taylor*

Exemplo

- Para uma **amostra aleatória simples** $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ do par de variáveis (X, Y) o estimador da **razão** populacional é:

$$\hat{R} = h\left(\hat{\bar{X}}, \hat{\bar{Y}}\right) = \frac{\hat{\bar{Y}}}{\hat{\bar{X}}}$$

Linearizando, pelo **método de *Taylor***, temos:

$$\hat{R} = h\left(\hat{\bar{X}}, \hat{\bar{Y}}\right) = h\left(\bar{X}, \bar{Y}\right) + \frac{\partial h\left(\bar{X}, \bar{Y}\right)}{\partial \bar{Y}}\left(\hat{\bar{Y}} - \bar{Y}\right) + \frac{\partial h\left(\bar{X}, \bar{Y}\right)}{\partial \bar{X}}\left(\hat{\bar{X}} - \bar{X}\right)$$

$$= \frac{\bar{Y}}{\bar{X}} + \frac{1}{\bar{X}}\left(\hat{\bar{Y}} - \bar{Y}\right) - \frac{\bar{Y}}{\bar{X}^2}\left(\hat{\bar{X}} - \bar{X}\right)$$

Método de *Taylor*

Exemplo

- Para o **estimador da razão** populacional:

$$\hat{R} = \frac{\bar{Y}}{\bar{X}} + \frac{1}{\bar{X}}\left(\hat{\bar{Y}} - \bar{Y}\right) - \frac{\bar{Y}}{\bar{X}^2}\left(\hat{\bar{X}} - \bar{X}\right)$$

Temos

$$Var\left(\hat{R}\right) = \left(\frac{1}{\bar{X}}\right)^2 Var\left(\hat{\bar{Y}}\right) + \left(\frac{\bar{Y}}{\bar{X}^2}\right)^2 Var\left(\hat{\bar{X}}\right) - 2\left(\frac{1}{\bar{X}}\right)\left(\frac{\bar{Y}}{\bar{X}^2}\right) Cov\left(\hat{\bar{X}}, \hat{\bar{Y}}\right)$$

Estimador da variância:

$$\hat{Var}\left(\hat{R}\right) = (1-f)\left[\left(\frac{1}{\bar{X}}\right)^2 \frac{S_Y^2}{n} + \left(\frac{\bar{Y}}{\bar{X}^2}\right)^2 \frac{S_X^2}{n} - 2\left(\frac{1}{\bar{X}}\right)\left(\frac{\bar{Y}}{\bar{X}^2}\right)\frac{S_{XY}}{n}\right]$$

Em que $f = \frac{n}{N}$ corresponde à fracção de amostragem.

Estimação da Variância

Critérios de Selecção dos Métodos

	Diferenciável	Não Diferenciável
Estimador Linear	**Dedução algébrica da variância do estimador ou Métodos de Reamostragem, no caso de se recorrer a imputação, calibração ou do plano de amostragem ser complexo**	
Estimador Não Linear	**Métodos de Linearização do Estimador pelo Método de Taylor**	**Estimadores "*plug-in*" que são uma extensão do método de Taylor, desenvolvida por *Deville****

* Macros em SAS cedidas pelo EUROSTAT, que são usadas por exemplo nos principais indicadores do Inquérito às Condições de Vida e Rendimento (ICOR).

Método dos Grupos Aleatórios

- Baseia-se no conceito de **replicação do desenho amostral**, dividindo a amostra em G **grupos disjuntos**, de tal forma que **cada grupo** constitua uma **versão mais reduzida da amostra inicial**.

- Este método é simples de ser aplicado mas o pressuposto da independência dos grupos **nem sempre garante a consistência das estimativas**:
  - Quando a amostra só pode ser dividida num **número reduzido de grupos**, por exemplo se a amostragem for estratificada e alguns estratos dispuserem de poucas observações.

- É necessário **replicar o plano amostral em cada um dos grupos** aleatórios.

Métodos de Reamostragem

- Os métodos de reamostragem utilizam a **amostra** aleatória original **como** se se tratasse de uma **população**, **extraindo** dessa **várias amostras aleatórias**.

- A **variância** de um estimador populacional é **estimada com base na variabilidade das várias subamostras**.

- Nos métodos de reamostragem podemos optar por uma das **técnicas**:
  - ***Bootstrap***;
  - ***Jackknife***.

Método *Bootstrap*

Historial

- O método de reamostragem ***Bootstrap*** foi introduzido por ***Efron*** em **1979** e foi considerado uma das primeiras técnicas estatísticas mais complexas em termos computacionais. Esta técnica veio **substituir** as habituais **deduções algébricas** pelas **simulações em computador**.

- O recurso ao *Bootstraping* na estimação da variância, no contexto da **amostragem**, foi estudado nos **anos 80** e começou a ser aplicada também na estratificada mas apresentou algumas limitações em estratos com restrições de dimensão.

- Em *Shao* & *Tu* (1996) podem ser consultados alguns desenvolvimentos e sugestões ao método inicial, como foi o caso do método ***bootstrap*** **com reposição em 1985** (*McCarthy* and *Snowden*).

Método *Bootstrap*

- Método estatístico desenvolvido para **estimar a distribuição amostral de um estimador** por amostragem da própria amostra original.

- Método de reamostragem que se baseia na extracção de B **amostras *bootstrap***: amostras aleatórias com reposição, de dimensão igual à amostra inicial, ou seja, com possibilidade de repetição de elementos.

- A **variabilidade** da estimativa populacional é **estimada com base na variabilidade das estimativas** calculadas a partir **das amostras *bootstrap***.

Método *Bootstrap*

Seja Y a variável aleatória em estudo e $\{Y_1, Y_2, \ldots, Y_n\}$ a respectiva amostra aleatória de dimensão n.

Amostras *Bootstrap*: $\left(Y_1^b, Y_2^b, \ldots, Y_n^b\right)$, com $b = 1, \ldots, B$.

- O **estimador *Bootstrap* da variância um dado estimador** $\hat{\theta}$ é a variância amostral dos B pseudo-estimadores:

$$\hat{Var}_B\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \hat{\theta}\right)^2$$

Método *Bootstrap*

Exemplo

Para uma amostra de **815 empresas** da **secção de CAE 55** (Alojamento e Restauração) foi feita uma simulação da técnica *Bootstrap* para avaliar a convergência dos resultados em **10**, **100** e **1000 réplicas**. A tabela que se segue mostra-nos os resultados para a amostra original.

ECAE	EPS	ESTRATO	SECÇÃO	UNIVERSO	AMOSTRA	k=N/n	TOT_VVN	VAR_ESTRATO	CV_ESTRATO
551	1	551/1	H	1156	68	17	101228918	812038298239662.00	28.15%
551	2	551/2	H	494	38	13	337147960	1923399132477150.00	13.01%
551	3	551/3	H	130	65	2	753919150	6204232989311930.00	10.45%
552	1	552/1	H	258	43	6	17491319	30225666772858.50	31.43%
552	2	552/2	H	36	18	2	18211796	3686517547138.22	10.54%
552	3	552/3	H	6	6	1	26469610	0.00	0.00%
553	1	553/1	H	9750	195	50	883152372	3277565088615650.00	6.48%
553	2	553/2	H	1264	79	16	636533816	6081381979865940.00	12.25%
553	3	553/3	H	56	28	2	322903018	9702986631667580.00	30.51%
554	1	554/1	H	8960	140	64	682958035	2841735964615300.00	7.81%
554	2	554/2	H	672	42	16	267685069	839517129971771.00	10.82%
554	3	554/3	H	18	18	1	47605255	0.00	0.00%
555	1	555/1	H	192	48	4	22351993	15771095444570.20	17.77%
555	2	555/2	H	36	9	4	34220090	116484613083822.00	31.54%
555	3	555/3	H	18	18	1	316402197	0.00	0.00%
			TOTAL	**23046**	**815**		**4468280598**	**31849025107613400.00**	**3.99%**

Método *Bootstrap*

Exemplo

» **10 Réplicas**

ECAE	EPS	ESTRATO	SECÇÃO	UNIVERSO	AMOSTRA	k=N/n	TETA_BS	VAR_BS	CV_ESTRATO
551	1	551/1	H	1156	68	17	86333886.3	222909979421797.00	17.29%
551	2	551/2	H	494	38	13	381253028	778483119946529.00	7.32%
551	3	551/3	H	130	65	2	873656405.8	1920024159755740.00	5.02%
552	1	552/1	H	258	43	6	22405068.6	82109681338835.60	40.44%
552	2	552/2	H	36	18	2	19599120	5220737876033.78	11.66%
552	3	552/3	H	6	6	1	26469610	0.00	0.00%
553	1	553/1	H	9750	195	50	941126050	1866608116659440.00	4.59%
553	2	553/2	H	1264	79	16	1017081387	77216858609864400.00	27.32%
553	3	553/3	H	56	28	2	289707752.8	9878560114973540.00	34.31%
554	1	554/1	H	8960	140	64	750718873.6	2300588875452140.00	6.39%
554	2	554/2	H	672	42	16	262881867.2	773897295460184.00	10.58%
554	3	554/3	H	18	18	1	47605255	0.00	0.00%
555	1	555/1	H	192	48	4	20128135.6	7776591076254.04	13.85%
555	2	555/2	H	36	9	4	26380674.8	13201771199321.10	13.77%
555	3	555/3	H	18	18	1	316402197	0.00	0.00%
			TOTAL	**23046**	**815**		**5081749312**	**95066239053024200.00**	**6.07%**

Método *Bootstrap*

Exemplo

» **100 Réplicas**

ECAE	EPS	ESTRATO	SECÇÃO	UNIVERSO	AMOSTRA	k=N/n	TETA_BS	VAR_BS	CV_ESTRATO
551	1	551/1	H	1156	68	17	86247766.85	235121991190490.00	17.78%
551	2	551/2	H	494	38	13	352550370.1	1707157056796020.00	11.72%
551	3	551/3	H	130	65	2	870408052.7	7678484140469740.00	10.07%
552	1	552/1	H	258	43	6	22446676.32	58983591930000.80	34.21%
552	2	552/2	H	36	18	2	19781531.02	4320242624898.06	10.51%
552	3	552/3	H	6	6	1	26469610	0.00	0.00%
553	1	553/1	H	9750	195	50	964200135.5	2431297063730670.00	5.11%
553	2	553/2	H	1264	79	16	910308163.5	54838916535252700.00	25.73%
553	3	553/3	H	56	28	2	280779794.2	4870016209488630.00	24.85%
554	1	554/1	H	8960	140	64	749035952	5290529015004800.00	9.71%
554	2	554/2	H	672	42	16	270778168.5	963117019732187.00	11.46%
554	3	554/3	H	18	18	1	47605255	0.00	0.00%
555	1	555/1	H	192	48	4	20180817.52	4115417776624.09	10.05%
555	2	555/2	H	36	9	4	25690963.96	23030604733558.80	18.68%
555	3	555/3	H	18	18	1	316402197	0.00	0.00%
			TOTAL	**23046**	**815**		**4962885454**	**781050888888730300.00**	**5.63%**

Método *Bootstrap*

Exemplo

» **1000 Réplicas**

ECAE	EPS	ESTRATO	SECÇÃO	UNIVERSO	AMOSTRA	k=N/n	TETA_BS	VAR_BS	CV_ESTRATO
551	1	551/1	H	1156	68	17	85795091.73	220285078201609.00	17.30%
551	2	551/2	H	494	38	13	356782390.1	1436174121497780.00	10.62%
551	3	551/3	H	130	65	2	867613356.6	7918541949442880.00	10.26%
552	1	552/1	H	258	43	6	21151366.09	45469449831007.80	31.88%
552	2	552/2	H	36	18	2	20053726.23	4340876506585.03	10.39%
552	3	552/3	H	6	6	1	26469610	0.00	0.00%
553	1	553/1	H	9750	195	50	965582474.4	2816450575927030.00	5.50%
553	2	553/2	H	1264	79	16	914297237.5	54342960839019700.00	25.50%
553	3	553/3	H	56	28	2	279204147.8	5969639937128260.00	27.67%
554	1	554/1	H	8960	140	64	751974461.4	4424804660495200.00	8.85%
554	2	554/2	H	672	42	16	273797750.5	1000608358740050.00	11.55%
554	3	554/3	H	18	18	1	47605255	0.00	0.00%
555	1	555/1	H	192	48	4	20157909.31	5686934014987.88	11.83%
555	2	555/2	H	36	9	4	25209734.62	27197456973075.60	20.69%
555	3	555/3	H	18	18	1	316402197	0.00	0.00%
			TOTAL	**23046**	**815**		**4972096708**	**78212160237778100.00**	**5.62%**

Método *Jackknife*

Historial

- Esta técnica foi inicialmente desenvolvida por ***Quenouille***, em **1949**, como o objectivo de reduzir e **estimar o enviesamento de estimadores**, num contexto de **populações infinitas**.

- Mais tarde, em **1958**, ***Turkey*** sugeriu a implementação do método *Jackknife* na **estimação de variâncias**.

- Para **populações finitas** a técnica *Jackknife* foi introduzida por ***Durbin***, em **1959**, sendo posteriormente **desenvolvida por *Wolter*** (**1985**).

Método *Jackknife*

- Este método pressupõe a **criação de várias subamostras** que se obtêm **retirando uma ou mais observações da amostra inicial**.

- A **variância é estimada** com base na **variabilidade entre as estimativas obtidas** (a partir das subamostras constituídas) **e a calculada pela totalidade da amostra**.

Método *Jackknife*

Retirando, à amostra inicial, uma observação de cada vez, obtemos n subamostras de dimensão n-1.

- O **estimador *Jackknife* da variância um dado estimador** $\hat{\theta}$ é dado por:

$$\hat{Var}_{JK}\left(\hat{\theta}\right)=\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i-\hat{\theta}\right)^2$$

Método *Jackknife*

Aplicações em R

- Na metodologia dos apuramentos de diversos inquéritos*, no INE, IP, é possível recorrer ao ***package survey***, em **linguagem R**, que está em constante actualização e que foi desenvolvido por ***Thomas Lumley*** (Universidade de *Washington*).

- **Contactos frequentes com o autor permitiram adicionar determinadas especificidades dos inquéritos do INE, IP**, às funcionalidades já existentes.

---

* Por exemplo, o Inquérito ao Emprego (IE) e o Inquérito Nacional de Saúde (INS).

Método *Jackknife*

No **Inquérito Nacional de Saúde (INS)**

1. **Definição do desenho amostral;**

desenho <- **svydesign** (~area, pond_iniciais, dados_norte)

- **area** — PSU´S do desenho amostral (no caso da região Norte são 372 áreas geográficas).
- **pond_iniciais** — Ponderadores (inverso da probabilidade de selecção ao qual é aplicado um factor de correcção de não respostas).
- **dados_norte** — Ficheiro de dados.

Método *Jackknife*

2. Criação de réplicas *Jackknife* do desenho amostral;

```
desenho_replicas <- as.svrepdesign (desenho, type="JK1")
```

- **desenho** — desenho amostral definido no ponto 1.
- **type="JK1"** — Tipo das réplicas a criar, neste caso *Jackknife*. Neste exemplo são criadas 372 réplicas, cada uma delas obtida retirando uma unidade primária (área) da amostra inicial.

Método *Jackknife*

3. Ajustamento por Margens;

Calibra<- **calibrate** (desenho_replicas, ~sexoe, pop, bounds=c(0.25,4), calfun="logit")

- **desenho_replicas** — réplicas criadas no ponto 2.
- **sexoe** — variável de ajustamento, de 38 categorias, construída a partir dos 19 escalões etários e dos 2 géneros, para as quais são conhecidos os totais populacionais.
- **pop** — Total populacional da região norte.
- **calfun="logit" —** Define a função de distância utilizada, que neste caso é a *logit*;
- **bounds=c(0.25,4) —** Limite inferior e superior para a distância entre os pesos ajustados e pesos iniciais.

Método *Jackknife*

Exemplo da base de dados (2005/2006) – Região Norte

regiao	area	naloj	nfam	nind	idade	sexo	sexoe	asma	diabetes	tensao	pond_inicial	pond_final	quope
1	1	447	1	1	49	1	11	0	1	1	621.28	1316.60	2.12
1	1	448	1	2	67	2	34	0	0	1	612.75	1016.98	1.66
1	17	353	1	2	14	2	23	1	0	0	619.14	1477.66	2.39
1	65	131	1	3	88	1	19	0	1	0	608.77	1168.65	1.92
1	85	88	1	1	34	1	8	0	0	1	625.92	1793.89	2.87
1	111	113	1	3	18	1	5	0	0	1	597.52	1337.90	2.24
1	129	544	1	3	7	1	3	1	0	0	633.72	1656.88	2.61
1	153	35	1	1	75	2	36	0	1	1	625.92	1216.61	1.94
1	255	327	1	2	48	2	30	1	1	0	602.82	1243.70	2.06
1	255	328	1	5	0	2	20	1	0	0	600.52	1268.14	2.11
1	305	383	2	2	20	2	25	1	0	0	602.63	2185.52	3.63
1	311	273	1	1	55	1	13	1	0	1	604.03	1202.17	1.99
1	332	38	1	1	66	2	34	1	1	1	519.58	862.33	1.66
1	339	272	1	4	33	2	27	1	1	1	512.76	1329.59	2.59
1	350	508	1	1	39	1	9	0	0	1	630.42	1852.87	2.94
1	394	151	1	1	50	1	12	0	1	1	615.33	1260.57	2.05
1	394	152	1	2	65	2	34	1	1	0	615.33	1021.24	1.66
1	409	229	1	1	70	1	16	1	1	0	610.70	1301.34	2.13

Método *Jackknife*

4. Apuramento das variáveis em análise e cálculo das respectivas variâncias.

Apura <- **svyby** (~popT, ~sexo +asma, calibra, vartype=c("var,cvpct"))

- **popT** — variável de *breakdown* para a análise em questão (população total).
- **sexo+asma** — variáveis para as quais se deseja efectuar o apuramento.
- **calibra** — réplicas do desenho amostral devidamente calibradas, calculadas no ponto 3.
- **vartype=c("var","cvpct")** — estatísticas a serem disponibilizados adicionalmente além da estimativa (neste caso, variância e coeficiente de variação).

Método *Jackknife*

» **Quadro de Apuramento**

População residente que tem ou já teve asma, diabetes ou tensão alta, por sexo, na região Norte (NUTS II), segundo dados de 2005/2006 (4º INS).

doenca	sexo	estimativa	var	cvpct
asma	1	82241	118427581	13.23
	2	131490	168451649	9.87
diabetes	1	116818	107973632	8.90
	2	157714	108757686	6.61
tensao	1	258716	197773515	5.44
	2	418317	256562038	3.83

Importância da Estimação de Variância

- O **desvio padrão**, o **coeficiente de variação** e os **intervalos de confiança** são as principais **medidas** utilizadas para **avaliar a precisão** de um dado estimador, medindo o seu **erro amostral** que resulta da aleatoriedade das estimativas.

- A **variância de um estimador** é o indicador essencial para podermos avaliar a **qualidade** e **fiabilidade** das **estimativas**.

- Toda a **metodologia** associada a qualquer **operação estatística** é definida **com base em indicadores de qualidade** e **critérios de precisão**, para os quais o **cálculo da variância** se torna **imprescindível**.

Conclusões

- Nos inquéritos por amostragem é possível deduzir, de forma relativamente simples, a variância dos estimadores mais usuais, no contexto da amostragem aleatória, simples ou estratificada.
- Torna-se necessário recorrer a métodos de linearização quando os estimadores não são combinações lineares de totais ou médias.
- É aconselhável aplicar métodos de reamostragem quando o plano amostral é complexo, como em diversos inquéritos do INE, IP.
- Quando se procede à imputação de não respostas ou quando se recorre à calibração dos dados os métodos usuais, de dedução algébrica, não nos permitem obter resultados fidedignos.

Referências Bibliográficas

- Deville, J.C. (1999). *Variance estimation for complex statistics and estimators*: *linearization and residual techniques*. Survey Methodology, vol. 25 no. 02.
- Cochran, W. G. (1977). *Sampling techniques*. 3rd Edition, New York: Springer.
- Effron, B. (1979). *Bootstrap methods: Another look at the jaccknife*. Annals of Statistics 7, 1-26.
- Lumley, T. (2004). *Analysis of complex survey samples*. Journal of Statistical Software 9(1): 1-19.
- Lumley, T. (2006). *Survey: analysis of complex survey samples*. R package version 3.6-5.
- Särdnal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, Springer.
- Shao, J. and Tu (1995). *The jackknife and bootstrap*. New York: Springer-Verlag.
- Turkey, J.W. (1958). *Bias and Confidence in not-quite large samples*. Annals of Mathematical Statistics, 29-614.
- Wolter, K.M. (1985). *Introduction to variande estimation*. New York: Springer-Verlag.