Inferência com probabilidades desiguais

Prof. Caio Azevedo

Introdução

- Todos os métodos inferenciais, vistos até o momento, foram baseados em esquemas probabilísticos onde todas as amostras tinham a mesma probabilidade de serem selecionadas.
- Apresentaremos, aqui, técnicas inferenciais baseadas em esquemas probabilísticos mais gerais.
- Um problema que surge é a obtenção de expressões para o vício e para a variância dos estimadores.
- As notações e definições são basicamente aquelas definidas para AAS_c e AAS_s, com as devidas adaptações.
- Embora as probabilidades de seleção não sejam iguais para as diferentes amostras, utilizaremos A_1 e A_2 para identificar que aquelas foram selecionadas com e sem reposição, respectivamente.

Motivação

- Os esquemas de seleção sob probabilidades desiguais buscam levar em consideração características (do problema/dados) que permitam obter amostras mais "representativas" (distribuição amostral da resposta, próxima da respectiva distribuição populacional).
- Tais informações podem ser usada para criar "grupos" (de modo que, por exemplo, as probabilidades de seleção dependam dos seus tamanhos) e/ou como covariáveis a serem usadas na construção das probabilidades de seleção (dentre outras possibilidades).

Exemplo

- Considere uma população dividida em grupos ou conglomerados de tamanhos $N_\alpha, \alpha = 1, ..., A$.
- Desenvolve-se um esquema probabilístico com reposição, onde as probabilidades de inclusão são proporcionais aos tamanhos dos grupos $N_\alpha, \alpha = 1, ..., A$.
- Considere uma população com A = 6 grupos dados na Tabela do próximo slide.
- Para selecionar uma unidade, escolhe-se um número aleatório entre 1 e 25. Suponha que seja o número 11. Como o número 11 cai no intervalo correspondente à unidade 3, que vai de 6 a 13, a unidade 3 é selecionada. As unidades seguintes que farão parte da amostra serão selecionadas com reposição. Portanto, a unidade 3 pode novamente fazer parte da amostra.

Tabela do exemplo anterior

Unidade	N_α	$\sum_{\alpha=1}^{A} N_\alpha$	Intervalo
1	3	3	1-3
2	2	5	4-5
3	8	13	6-13
4	4	17	14-17
5	1	18	18
6	7	25	19-25

Exemplo

- O exemplo que apresentamos a seguir considera o caso em que um único conglomerado é selecionado. As probabilidades de seleção neste caso são estabelecidas pelo pesquisador como sendo proporcionais aos tamanhos dos conglomerados.
- Considere novamente a população $\mathcal{U}$, com N $= 6$ elementos onde $\boldsymbol{d} = (2, 6, 10, 8, 10, 12)'$.
- Para esta população, $\mu = 8$. A população está dividida nos três conglomerados: $C_1 = \{1, 2\}$, com $\mu_1 = 4$; $C_2 = \{3\}$, com $\mu_2 = 10$; $C_3 = \{4, 5, 6\}$, com $\mu_3 = 10$.

Cont.

- Procedendo como no Exemplo anterior, as probabilidades de inclusão dos grupos 1, 2 e 3 são iguais a 2/6, 1/6 e 3/6, respectivamente.
- Selecionando um conglomerado de acordo com as probabilidades acima, tem-se a distribuição do estimador $\widehat{\mu}_c$ (veja aqui), dada na Tabela seguir.

$\widetilde{\mu}_c$	4	10
$P(\widehat{\mu}_c = \widetilde{\mu}_c)$	2/6	4/6

Então $\mathcal{E}_A(\widehat{\mu}_c) = 4\frac{2}{6} + 10\frac{4}{6} = 9$ e

$\mathcal{V}_A(\widehat{\mu}_c) = \frac{2}{6}(4-8)^2 + \frac{4}{6}(10-8)^2 = 8.$

Caso geral

- Considere uma população com N unidades que podem ser inclusive grupos (estratos) ou conglomerados.
- Suponha que associada à unidade i da população tem-se uma medida m_i, obtida segundo algum critério estabelecido previamente.
  - Por exemplo, amostrando hospitais, essa medida poderia ser o número de leitos.
  - Por outro lado, em levantamentos de indústrias, uma medida do tamanho pode ser o número de empregados ou o faturamento em um determinado período.

Cont.

- Definida a medida do tamanho da unidade i por m_i, a probabilidade de seleção associada ao elemento i será $z_i = \frac{m_i}{m_0}$, $i = 1, \ldots, N$, onde $m_0 = \sum_{i=1}^{N} m_i$.
- Seleciona-se então, com reposição e probabilidade de seleção z_i para cada unidade, uma amostra $\boldsymbol{S}$ de tamanho n da população.
- Suponha o interesse em estimar o total populacional (τ). Como estimador podemos considerar:

$$\widehat{\tau}_{ppz} = \frac{1}{n} \sum_{i \in \boldsymbol{S}} \frac{y_i}{z_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{z_i}$$

Cont.

- Seja F_i (link) o número de vezes que a unidade i fora selecionada, $i = 1, \ldots, N$. Temos que $\boldsymbol{F} = (F_1, \ldots, F_N) \sim$ Multinomial $(n, \boldsymbol{z})$ $(\boldsymbol{z} = (z_1, ..., z_N)')$ ou seja:

$$P(\boldsymbol{F} = \boldsymbol{f}) = \frac{n!}{f_1! f_2! \ldots f_N!} \prod_{i=1}^{N} z_i^{f_i} \tag{1}$$

em que $\sum_{i=1}^{N} f_i = n$ e $\sum_{i=1}^{N} z_i = 1$.

- Logo $\mathcal{E}_A(F_i) = nz_i$, $\mathcal{V}_A(F_i) = nz_i(1 - z_i)$ e $\text{Cov}(F_i, F_j) = -nz_i z_j$.
- Note ainda que

$$\widehat{\tau}_{ppz} = \frac{1}{n} \sum_{i=1}^{N} F_i \frac{y_i}{z_i} \tag{2}$$

Cont.

- Caso todas as amostras tenham a mesma probabilidade de serem selecionadas, ou seja, se $m_i = 1, \forall i$, então $m_0 = N$ e $z_i = \frac{1}{N}, \forall i$. Assim:

$$\widehat{\tau}_{ppz} = \frac{1}{n}\sum_{i \in \boldsymbol{s}} \frac{y_i}{1/N} = N\frac{1}{n}\sum_{i \in \boldsymbol{s}} y_i = N = N\widehat{\mu} = \widehat{\tau}$$

- Além disso, temos que:

$$\mathcal{E}_{A_1}(\widehat{\tau}_{ppz}) = \tau; \mathcal{V}_{A_1}(\widehat{\tau}_{ppz}) = \frac{1}{n}\sum_{i=1}^{N} z_i \left(\frac{y_i}{z_i} - \tau\right)^2$$

- Com efeito, de (1) em (2), temos que:

$$\mathcal{E}_{A_1}\left(\widehat{\tau}_{ppz}\right) = \frac{1}{n}\sum_{i=1}^{N} \mathcal{E}(F_i)\frac{y_i}{z_i} = \frac{1}{n}\sum_{i=1}^{N} nz_i\frac{y_i}{z_i} = \sum_{i=1}^{N} y_i = \tau$$

Cont.

- Além disso, temos que:

$$\begin{aligned}
\mathcal{V}_{A_1}(\widehat{\tau}_{ppz}) &= \frac{1}{n^2}\left\{\sum_{i=1}^{N}\left(\frac{y_i}{z_i}\right)^2 \mathcal{V}_{A_1}(F_i) + 2\sum_{i<j}\frac{y_i}{z_i}\frac{y_j}{z_j}Cov(F_i, F_j)\right\} \\
&= \frac{1}{n}\left\{\sum_{i=1}^{N}\left(\frac{y_i}{z_i}\right)^2 z_i(1-z_i) - 2\sum_{i<j}\frac{y_i}{z_i}\frac{y_j}{z_j}z_i z_j\right\} \\
&= \frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2}{z_i} - \tau^2\right) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - \tau\right)^2 \qquad (3)
\end{aligned}$$

- Um estimador não viciado de (3) é dado por:

$$\widehat{\mathcal{V}}_{A_1}(\widehat{\tau}_{ppz}) = \frac{1}{n(n-1)}\sum_{i\in \boldsymbol{s}}\left(\frac{y_i}{z_i} - \widehat{\tau}_{ppz}\right)^2$$

Cont.

- A média populacional pode ser estimada por : $\widehat{\mu}_{ppz} = \frac{1}{N}\widehat{\tau}_{ppz}$.
- Para os planos AE (link 1, link 2), AC e AC2E, bem como para os estimadores razão e regressão, os resultados podem ser adaptados (veja a literatura sugerida no programa).
- Discuteremos um pouco sobre como fazer inferência quando a amostra é selecionada sem reposição (via estimador Horvitz-Thompson).

Estimador de Horvitz-Thompson

- Assume-se que as unidades que compõem a amostra são selecionadas sem reposição.
- A população é constituída por A unidades (podem ser, por exemplo, conglomerados ou grupos na amostragem estratificada) e dessas A unidades, a são selecionadas sem reposição.
- Definimos:
  - π_i: a probabilidade de que a unidade i faça parte da amostra.
  - π_{ij}: a probabilidade de que as unidades i e j façam parte da amostra. $i, j = 1, \ldots, A$.

Cont.

- Como visto anteriormente, $\pi_i = \sum_{i \in \boldsymbol{s}} P(\boldsymbol{s})$ e $\pi_{ij} = \sum_{i,j \in \boldsymbol{s}} P(\boldsymbol{s})$.
- Assim, é possível provar que (exercício):

$$\sum_{i=1}^{A} \pi_i = a; \sum_{i \neq j} \pi_{ij} = (a-1)\pi_i; \sum_{i=1}^{A} \sum_{j>i} \pi_{ij} = \frac{a(a-1)}{2}$$

- O estimador de Horvitz-Thompson (HT) para o total populacional é dado por:

$$\widehat{\tau}_{HT} = \sum_{i \in \boldsymbol{s}} \frac{y_i}{\pi_i} = \sum_{i=1}^{a} \frac{Y_i}{\pi_i} = \sum_{i=1}^{A} F_i \frac{y_i}{\pi_i}$$

Cont.

- Lembre que $F_i \sim \text{Bernoulli}(\pi_i)$, assim $\mathcal{E}_{A_2}(F_i) = \pi_i$ e $\mathcal{V}_{A_2}(F_i) = \pi_i(1-\pi_i)$.
- Além disso, $Cov(F_i, F_j) = \pi_{ij} - \pi_i\pi_j$.
- Portanto, vem que:

$$\mathcal{E}(\widehat{\tau}_{HT}) = \sum_{i=1}^{A} \mathcal{E}(F_i)\frac{y_i}{\pi_i} = \sum_{i=1}^{A} \pi_i \frac{y_i}{\pi_i} = \sum_{i=1}^{A} y_i = \tau, \text{ e}$$

$$\begin{aligned}
\mathcal{V}(\widehat{\tau}_{HT}) &= \sum_{i=1}^{A} \left(\frac{y_i}{\pi_i}\right)^2 \mathcal{V}_{A_2}(F_i) + 2\sum_{i=1}^{A}\sum_{j>i} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} Cov(F_i, F_j) \\
&= \sum_{i=1}^{A} y_i^2 \left(\frac{1-\pi_i}{\pi_i}\right) + 2\sum_{i=1}^{A}\sum_{j>i} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} y_i y_j
\end{aligned}$$