Introdução aos pacotes sampling e survey

Prof. Caio Azevedo

Introdução

- Três etapas fundamentais na resolução de problemas :
  1. Definição do plano amostral (PA).
  2. Seleção da amostra segundo o PA do item 1): pacote "sampling" (ou a função "sample", no caso do PA *AAS* (com ou sem reposição)).
  3. Análise dos dados (amostra) considerando o PA do item 1): pacote "survey".
- Aqui pode-se encontrar diversos pacotes/funções que lidam com metodologias de amostragem no R.

Pacote sampling

- Já vimos que a função "sample" seleciona amostras segundo os planos AAS_c e AAS_s.
- Permite a seleção de amostra segundo planos amostrais (complexos) (incluindo o AAS). Ferramentas:
  - Amostragem: Estratificação, dois estágios, probabilidades desiguais, amostragem balanceada.
  - Estimativa: estimador de calibração e regressão.
  - Ferramentas: cálculo de probabilidades de inclusão, cruzamento de estratos.
  - Contem algumas bases de dados reais
- Manual e página.
- Apresentação.

AAS no sampling

- srswr: amostragem aleatória simples com reposição. Sintaxe:

  srswr(n,N)

- srswor: amostragem aleatória simples sem reposição.Sintaxe:

  srswr(n,N)

  em que n é o tamanho da amostra e N é o tamanho da população.

- srswr: Retorna um vetor de tamanho N, em que cada elemento indica o número de repetições para a unidade k na amostra.

- srswr: Retorna um vetor (com elementos 0 e 1) de tamanho N, em que cada elemento deste vetor indica o status da unidade k (1, se a unidade k é selecionada na amostra; 0, caso contrário).

Amostragem com probabilidades desiguais no sampling

- UPbrewer.
- UPmaxentropy.
- UPmidzuno, UPmidzunopi2.
- UPmultinomial.
- UPpivotal, UPrandompivotal.
- UPpoisson.
- UPsampford.
- UPsystematic, UPrandomsystematic, UPsystematicpi2.
- UPtille, UPtillepi2.

Pacote survey

- O pacote survey permite analisar dados sob amostragem complexa.
- O esquema de amostragem pode ser explicitamente descrito ou representado por pesos de replicação.
- Instalação: install.packages("survey")/load("survey").
- Página oficial, manual, página no CRAN, artigo, apresentação e livro.
- A seleção da amostra pode ser utilizando algum outro recurso computacional ou o pacote sampling do R, por exemplo.

Opções de metodologias de análise

- Cálculo de médias, totais, razões, quantis, tabelas de contingência.
- Modelos de regressão, modelos loglineares, curvas de sobrevivência, testes de classificação, para toda a amostra e para domínios (sub - populações).
- Cáculo de variâncias por linearização de Taylor ou por pesos replicados (BRR, jackknife, bootstrap, bootstrap multiestágio ou fornecidos pelo usuário)
- Incorporação de amostragem multi-estágio com ou sem reposição, na análise.

Cont.

- Análise multivariada: componentes principais, análise fatorial (experimental).
- Razão de verossimilhança (Rao-Scott) teste para mlg, modelos de Cox, modelos loglineares.
- Amostragem PPS (probability proportional to size) com ou sem reposição: estimadores de Horvitz-Thompson e Yates-Grundy e uma variedade de outras aproximações.
- Pós-estratificação, ranking / calibração generalizada, estimativa de GREG, corte ("aparamento") de pesos.
- Planejamento de duas fases. Pesos estimados para estimadores aumentados de IPW.

Cont.

- Suporte para o uso de dados de imputação múltipla (“missing data”.
- Backup de bancos de dados (para objetos criados pelo pacote) com suporte para grandes conjuntos de dados (também com pesos replicados).
- Algum suporte para processamento em paralelo em computadores multicore.

Como o pacote survey funciona

- Para usar o pacote são necessários dois passos:
  - Primeiro é necessário especificar o planejamento (desenho) amostral utilizado, através da função svydesign().
  - Então, o método de análise estatística ou modelo estatístico de interesse é utilizado usando uma função especial que considera o planejamento definido anteriormente. Todas essas funções especiais começam com "svy" e incluem:
    svymean(), svychisq(), svytotal(), svyhist(), svyplot(), svyglm(), svyloglin()

Indicando o planejamento utilizado

- Sintaxe básica:

  svydesign(ids, probs = NULL, strata = NULL, variables = NULL,fpc = NULL, weights = NULL, data = NULL)

  em que

  - ids: fórmula ou dataframe especificando id's (identificadores) de cluster (conglomerados) do maior para o menor nível, ~ 0 ou ~ 1 especifica que não há cluster (conglomerados).
  - probs: fórmula ou dataframe especificando probabilidades de amostragem dos clusters (conglomerados).
  - strata: fórmula ou vetor especificando estratos, use "NULL" para indicar que não há estratos.

Indicando o planejamento utilizado

- variables: fórmula ou frame de dados especificando as variáveis (de interesse). Se "NULL", (todos) os dados fornecidos são utilizados.
- fpc: correção de populações finitas, especificada como o tamanho total da população em cada estrato ou como a fração da população total que foi amostrada (relativa à fração amostral - $(\frac{n}{N})$).
- weights: Fórmula ou vetor especificando pesos de amostragem como uma alternativa ao argumento "prob" (geralmente 1/prob).
- data: dataframe (banco de dados utilizado, deve contar todas as variáveis a serem utilizadas, probs, estratos, variáives, fpc, pesos).

Estimação da média e do total

- Estimação da média: svymean(x, design, na.rm=FALSE,deff=FALSE,...)
- Estimação do total: svytotal(x, design, na.rm=FALSE,deff=FALSE,...)
- x: fórmula, vetor ou matriz.
- design: um objetivo "survey.design" ou "svyrep.design".
- na.rm : se os casos com valores faltantes devem ou não serem desconsiderados.
- deff: retorna o efeito de planejamento (EPA), comparado com a respectiva inferência sob AAS_c ou AAS_s (default AAS_s).