AMOSTRAGEM

Unidade 07

Introdução a Pacotes do R para Amostragem

O Pacote `sampling` do R

Desenvolvido por Alina Mattei e Yves Tillé, da Universidade de Neuchatel, Suíça.

Versão mais atual é a **2.5**.

Disponível no repositório de pacotes do R.

Dispõe de rotinas para selecionar amostras conforme vários métodos usando probabilidades iguais ou desiguais.

Tem funcionalidades para obter amostras estratificadas.

O manual do pacote está disponível em pdf online.

O Pacote `sampling` do R

Funções e planos amostrais disponíveis

- ✓ `srswor`: amostragem aleatória simples **sem** reposição.
- ✓ `srswor1`: amostragem aleatória simples **sem** reposição, método seqüencial.
- ✓ `srswr`: amostragem aleatória simples **com** reposição.
- ✓ Diversos métodos de seleção com probabilidades desiguais.
- ✓ Várias funções auxiliares para amostragem de populações finitas.

Exemplo 1

Considere a população de fazendas de cana de açúcar.

Selecione desta população uma amostra aleatória simples de n=50 fazendas, por AAS.

Use esta amostra para estimar:

a) a proporção de fazendas na região 1;

b) a 'produtividade média' das fazendas, definida como razão entre a produção total e a área total das fazendas;

c) a produção total de fazendas da região 1.

O Pacote `survey` do R

Versão corrente é a 3.26.

Pacote (‘*library*’) elaborado e mantido por Thomas Lumley, da Universidade de Washington (Seattle, EUA).

Livro recém publicado pelo autor apresenta:

- Teoria ‘clássica’ para análise de dados amostrais complexos;
- Facilidades do pacote `survey` para análise de dados;
- Inúmeros exemplos com dados reais.

http://faculty.washington.edu/tlumley/survey/

Princípios Condutores do Desenho do Pacote survey

Facilidade de manutenção e depuração mediante re-utilização de código.

Velocidade e memória não são uma prioridade: só otimiza rotinas quando há um 'caso real de uso' demandando solução.

Rápida liberação de novas versões, de modo que erros e outras infelicidades sejam descobertas e reparadas.

Ênfase em recursos úteis para bioestatísticos (p.ex. calibração, regressão, gráficos exploratórios, análise de sobrevivência).

'Mercado' Pretendido

- Pesquisa em métodos (devido às características de programação do R).
- Ensino (facilita integração com ensino de outros métodos estatísticos, onde R também é usado).
- Análise secundária de dados de pesquisas nacionais (R é familiar a estatísticos não ligados à área de amostragem).
- Planos de duas fases em epidemiologia.

Características e Funcionalidade

- Descrição de planos amostrais: `svydesign()`
- Estatísticas descritivas: médias, totais, quantis, etc.
  ➔ `svymean()`, `svytotal()`, `svyby()`, etc.
- Estimação para domínios.
- Tabelas de contingência: `svychisq()`, `svyloglin()`
- Gráficos: histogramas, diagramas de dispersão, suavizadores.
- Modelos de regressão: `svyglm()`, `svyolr()`
- Calibração e pós-estratificação.

Objetos e Fórmulas

Coleções de informações relacionadas devem ser armazenadas juntas num **objeto**.

Para dados amostrais, isto significa armazenar os metadados relevantes junto dos dados.

A maneira de especificar variáveis num '`data frame`' ou outro objeto do R é através de uma '`formula`':

```
~ a + b + I(c < 5*d)
```

O pacote `survey` sempre usa fórmulas para especificar variáveis num arquivo de dados de pesquisa.

Idéias Básicas de Estimação

Unidades são amostradas com probabilidades conhecidas π_i de uma população de tamanho N, para obter uma amostra de tamanho n.

Definimos um 'indicador de inclusão na amostra' R_i, tomando valor 1 se a unidade i está na amostra e 0 caso contrário.

O problema 'usual' de inferência em amostragem (considerando o plano amostral) é estimar quantidades populacionais definidas caso toda a população fosse observada.

Idéias Básicas de Estimação

A estimação de um total populacional é simples. Um estimador não viciado do total $Y = \sum_{i \in U} y_i$ é dado por:

$$\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i} = \hat{Y}_{HT}$$ ➔ **Estimador de Horvitz-Thompson**

Estimação da precisão (erro padrão) segue diretamente da variância de uma soma de variáveis aleatórias:

$$V_p\left(\hat{Y}_{HT}\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \left(\frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \right)$$

O problema é conhecer as probabilidades de inclusão conjuntas π_{ij}.

Descrevendo um Plano Amostral no `survey`

A função `svydesign()` é a que permite descrever a estrutura de um plano amostral para o pacote `survey`.

Possui recursos para especificar:

- **estratificação**,
- **conglomeração**,
- observações com **pesos desiguais**, para lidar com **probabilidades desiguais** de seleção, e **ajustes** para compensar não resposta e outros ajustes, e
- **métodos** a serem empregados para estimar **erro padrão**.

Depois de aplicada, os metadados sobre o plano amostral são armazenados junto dos dados da pesquisa.

Exemplo 2

Descrevendo para o `survey` a amostra de fazendas citada no exemplo 2.1 (dados de 338 fazendas).

```
fazendas.amostra = svydesign(data=amofaz,
id=~1, strata=NULL, fpc=~Npop)
```

- `amofaz` é o nome do *data.frame* onde estão os dados da amostra de interesse;
- `~1` indica que se trata de uma amostra de unidades elementares (não conglomerada);
- `NULL` indica que não há estratificação;
- `~Npop` indica a coluna com o tamanho da população.

```
> summary(fazendas.amostra)
```

Passos para Usar Pacote `survey`

1. Especificar a estrutura do plano amostral usado para obter os dados que se vai querer analisar ➔ função `svydesign()`.
2. Especificar análise de interesse – no exemplo, função que permite estimar totais populacionais ➔ `svytotal()`.
3. Interpretar e apresentar resultados de interesse obtidos.

Comentários

Especificação da estrutura do plano amostral pode ser feita uma única vez para cada pesquisa ou conjunto de dados.

Análises incorporando plano amostral são tão simples de obter quanto análises ignorando o plano amostral.