R-Ladies Natal

Amostragem utilizando o R

Tayane da Costa Varela

Agosto/2020

O que é amostragem e qual sua importância?

- É o estudo que compreende selecionar uma amostra (parte da população) e analisar os resultados.
- Muitas vezes se torna inviável coletar dados de toda população.
- É de grande importância portar algum conhecimento sobre como interpretar e entender os dados que estão sendo apresentados.
- Dados incorretos e manipulados geram descrédito para estatística, tornando a interpretação de dados estatísticos com desconfiança ou dificuldade para a sociedade, em vez de algo esclarecedor.
- Com isso surge a necessidade de que dados estatísticos sejam divulgados de forma correta, sem manipulação dos dados, especificando como foi realizada a coleta dos dados e de forma objetiva.

Algumas definições básicas

- **Unidade observacional/elemento**: objeto no qual as medidas (dados) são feitas.
- **População alvo**: coleção completa de todas as unidades observacionais (todos elementos) que se deseja estudar.
- **População amostrada**: coleção de todos os possíveis elementos que podem ser escolhidos em uma amostra.
- **Amostra**: qualquer subconjunto da população.
- **Unidade amostral**: elemento ou grupo de elementos que são, na verdade, selecionados para compor a amostra.
- **Cadastro**: lista de unidades amostrais
  Ex.:
  1. Lista dos setores censitários do IBGE
  2. Listas telefônicas (entrevistas por telefone) da cidade ou estado
  3. Lista de endereços dos imóveis da cidade
  4. Lista de matrículas de alunos
- **Notação**
  U = População
  S = Amostra
  N = nº de elementos da população
  n = nº de elementos da amostra

Tendenciosidades

- **Tendenciosidade de seleção**

Quando ocorre?

1. A população amostrada difere muito da amostra selecionada

2. A amostra é feita por conveniência pelo pesquisador ou unidades são escolhidas acidentalmente.

3. Não inclusão de toda a população alvo no cadastro (subcobertura).

4.Falha na obtenção das respostas de todas as unidades escolhidas para a amostra (não-resposta)

- **Tendenciosidade de mensuração:** O instrumento utilizado para fazer as medições das unidades amostrais apresenta falhas ou erros, não representando o valor verdadeiro.

Erros amostrais e não-amostrais

- **Erro amostral:** É um erro proveniente pelo fato de coletar informações apenas de uma amostra em vez de examinar toda a população. Isso ocorre pelo fato de existir variabilidade entre as amostras.

Popularmente conhecido como "margem de erro".

Ex.: Uma margem de erro divulgada ser de 3 pontos percentuais, com nível de confiança de 95%, significa:

"Em 19 de cada 20 amostras possíveis, os resultados diferirão em no máximo 3 pontos percentuais, em qualquer direção, do que teria sido obtido se todos os indivíduos da população fossem entrevistados"

Erros amostrais podem ser controlados a partir de uma escolha apropriada do tamanho da amostra

- **Erro não-amostral:** O erro causado por outras fontes que não seja a variabilidade entre amostras.

1. Subcobertura: pode ser minimizada com revisão do cadastro.

2. Não-resposta: podem ser reduzidos com o uso de métodos probabilísticos.

3. Inacurácia de respostas: Respostas acuradas podem ser obtidas com planejamento cuidadoso e teste do instrumento de obtenção dos dados, e pré-teste da pesquisa (amostra piloto).

Erros amostrais e não-amostrais

- Muitas pesquisas afirmam orgulhosamente que o erro amostral está dentro de uma margem de 3%. Porém, ignoram o grande viés de seleção.

- Em algumas pesquisas o erro amostral pode ser insignificante em comparação com o erro não-amostral.

- Com a abundância de pesquisas mal feitas, não é surpreendente que algumas pessoas sejam céticas em relação a todas as pesquisas. Então, se torna comum ouvir “Minha opinião nunca foi perguntada, como os resultados da pesquisa podem me representar?”. O questionamento público se intensifica depois que uma pesquisa comete um grande erro ao prever resultados de uma eleição, por exemplo.

- Algumas pessoas insistem que apenas um censo completo será satisfatório. Porém, um censo também está sujeito a erros. Em geral, as maiores causas de erro em uma pesquisa são ocultação, falta de resposta e negligência na coleta dos dados.

- Existem três principais justificativas para usar amostragem:

1. Menor custo para realização da pesquisa.

2. Maior velocidade de obtenção da informação.

3. Não destruição de toda a população.

4. Amostragem, em alguns casos, pode ter mais precisão do que um censo.

Tipos de Amostragem

As diversas formas de amostragem podem ser classificadas em

- **amostragem probabilística:** selecionada de maneira que elementos na população tem probabilidades conhecidas (positivas) de inclusão na amostra.

  - Requer disponibilidade de um cadastro

  - Requer uso de mecanismo físico (sorteio, tabela de n.º aleatórios, algoritmo computacional,..)

  Para obter uma amostra probabilística, um ato físico deve ser usado. Este procedimento é conhecido na Estatística como aleatorização.

  Aleatorização da escolha da amostra é feita para criar uma distribuição de probabilidade. Esta distribuição serve de base para o processo inferencial. De brinde, elimina viés e tendências.

- **amostragem não-probabilística:** formada sem conhecimento das probabilidades dos elementos serem incluídos na amostra. Exemplo: Amostragem de conveniência ou acidental, Amostragem de julgamento ou intencional, Amostragem por quotas.

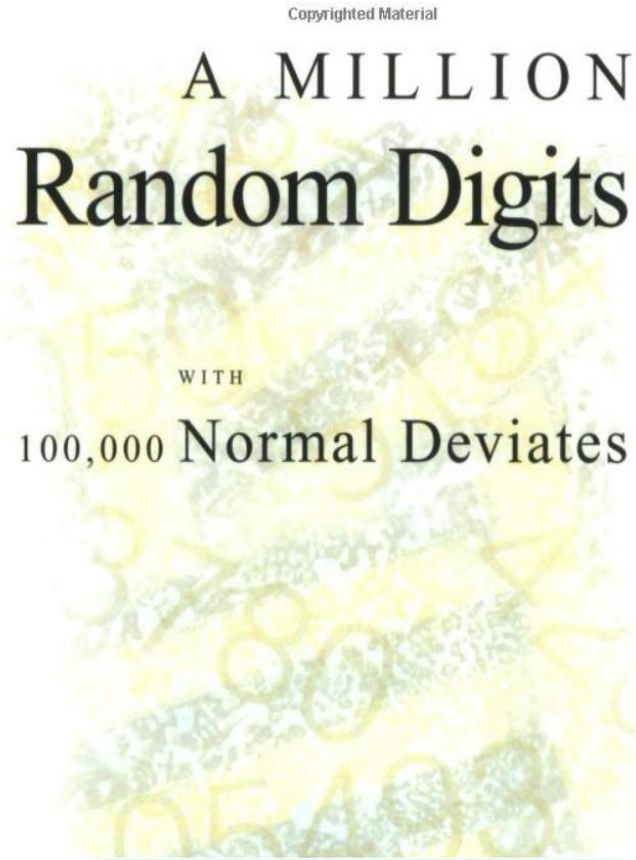
  - São em geral mais simples ou mais baratas de serem selecionadas

  - Propriedades estatísticas dos estimadores são difíceis de serem verificadas

  - São usualmente sujeitas a tendenciosidades de seleção

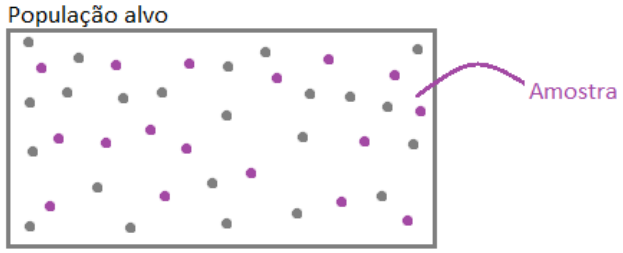
  - Não é recomendado fazer inferência em cima disso

Como eram gerados números aleatórios antigamente

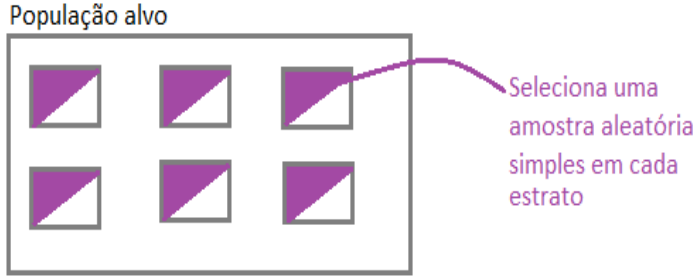


Amostragem probabilística

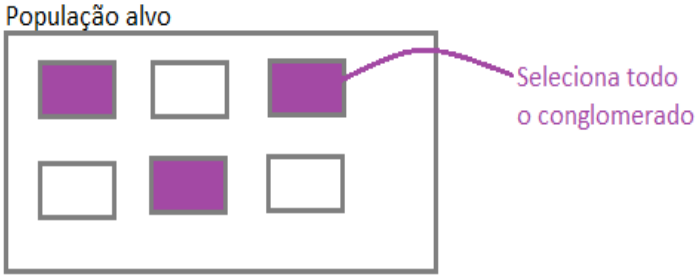
- **Amostragem aleatória simples (AAS):** Obtida de forma que todo subconjunto de tamanho n da população de tamanho N tem a mesma chance de ser selecionado. Pode ser realizada com reposição (AASCR) e sem reposição (AASSR).



- **Amostragem aleatória estratificada (AAE):** Divide-se a população em subpopulações ou estratos.



- **Amostragem aleatória por conglomerados (AAC):** Agrega-se os elementos da população em unidades amostrais conglomerados, uma AAS de tamanho n é feita na seleção dos conglomerados.



Amostragem aleatória simples
Conceitos básicos

- Amostragem aleatória simples é a forma mais básica de amostragem probabilística, ela fornece a base teórica para as formas mais complicadas.

- É selecionada de modo que todo subconjunto possível de n unidades distintas na população tenha a mesma probabilidade de ser selecionado.

- Existe duas formas de se obter uma amostra simples: com reposição, em que a mesma unidade pode ser incluída mais de uma vez na amostra, e sem reposição, em que todas as unidades da amostra são distintas.

- Geralmente, preferimos amostrar sem reposição.

- Para fazer uma AAS, você precisará de uma lista (cadastro) de todas as unidades da população. Cada unidade recebe uma identificação, e uma amostra é selecionada de modo que:

    1. Cada unidade tenha a mesma chance de ser selecionada;

    2. A seleção da unidade não seja influenciada por outras unidades já selecionadas.

Obs.: Na prática, números pseudoaleatórios gerados por computador são geralmente usados para selecionar uma amostra.

Tamanho da amostra para atingir uma margem de erro “e”, com nível de confiança 1 – α (AASSR)

Parâmetro	n_0	n
Média, $\bar{y}_U$	$\frac{z_{\alpha/2}^2 \, s_y^2}{e^2}$	$\frac{n_0}{1 + \frac{n_0}{N}}$
Total, ty	$\frac{N^2 \, z_{\alpha/2}^2 \, s_y^2}{e^2}$	$\frac{n_0}{1 + \frac{n_0}{N}}$
Proporção, p	$\frac{z_{\alpha/2}^2 \, p(1-p)}{e^2}$	$\frac{n_0}{1 + \frac{n_0 - 1}{N}}$

Se não há informação prévia sobre p, adotaremos o cenário de variabilidade máxima nas resposta, isto é, p=1/2. Pode-se usar um valor de proporção anterior, se existir.

Note que estes dois métodos requerem uma estimativa da variância populacional s_y^2. Alguns procedimentos úteis são:

1. Faça um pré-teste (amostra piloto) e calcule a variância amostral.

$$s_s^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2$$

*Obs.: a variância amostral é um estimador consistente para a variância populacional em uma AASSR.

2. Use estudos anteriores ou dados disponíveis na literatura.

3. Se nenhuma informação é disponível, obtenha um palpite para a variância. Algumas vezes uma distribuição hipotética para os dados nos dará informação sobre a variância.

Por exemplo, se a população é normalmente distribuída, você pode não conhecer a variância, mas ainda assim pode se ter uma ideia da amplitude dos dados, você pode então estimar S por:

$$s_y \approx \frac{Amplitude}{4}$$

Etapas

1. Identificação dos objetivos (definição dos objetivos gerais. cadastro, especificação dos parâmetros, população alvo)

2. Planejamento e seleção da amostra (Unidades amostrais, plano amostral, tamanho da amostra, erros)

3. Coleta de informações (Definição do instrumento de mensuração ou coleta de dados)

4. Análise dos resultados

Referências

1. LOHR, S. L. Sampling: design and analysis. Duxbury Press, 2000.
2. Notas de aula da disciplina de amostragem I.