MAT02025 - Amostragem 1

A estimativa do tamanho da amostra

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2023

Introdução

Introdução

- No planejamento de uma pesquisa por amostragem, sempre chega-se a uma etapa em que deve ser tomada uma decisão sobre o **tamanho da amostra**.

Introdução

- **A decisão é importante**.
  - Uma amostra muito grande implica em desperdício de recursos;
  - Uma amostra muito pequena diminui a utilidade dos resultados.
- A decisão nem sempre pode ser feita de forma satisfatória; frequentemente não possuímos **informações** suficientes para ter certeza de que nossa escolha do tamanho da amostra é a melhor.
- A **teoria da amostragem** fornece uma estrutura para pensar de forma inteligente sobre o problema.

Um exemplo

Um exemplo

Um exemplo[1] mostra as etapas envolvidas para se chegar a uma solução.

- Um antropólogo se prepara para estudar os habitantes de alguma ilha.
- Entre outras coisas, ele deseja estimar a **porcentagem de habitantes** pertencentes ao grupo sanguíneo O.
- Foi obtida a cooperação necessária para que seja possível obter uma **amostra aleatória simples**.
- **Qual deve ser o tamanho da amostra?**

[1] Aqui utilizamos um exemplo hipotético, mas que ilustra bem os problemas reais no cálculo do tamanho da amostra.

Um exemplo

Esta pergunta não pode ser respondida sem que antes se responda outra:

- Com que **precisão** o antropólogo deseja saber a porcentagem de pessoas com grupo sanguíneo O?
- Em resposta, ele afirma que ficará satisfeito se a porcentagem estiver correta em $\pm 5\%$, no sentido de que, se a amostra apresentar que 43% dos habitantes estão no grupo sanguíneo O, a porcentagem para toda a ilha **(a verdadeira porcentagem populacional de habitantes)** certamente ficará entre 38% e 48%.

Um exemplo

- Para evitar mal-entendidos, pode ser aconselhável apontar ao antropólogo que **não podemos garantir** absolutamente a precisão dentro de 5%, **exceto se examinarmos todos indivíduos da ilha**.
- Por maior que seja o valor de n, há uma chance de uma amostra **"muito infeliz"** (*azarada*) tenha uma **margem de erro** superior aos 5% desejados.
- O antropólogo responde friamente que está ciente disso, que está disposto a ter uma **chance de** 1 **em** 20 de obter uma "amostra infeliz" e que tudo o que ele pede[2] é o **valor de** n.

[2] Em vez de uma aula sobre estatística **(!)**

Um exemplo

- Agora podemos fazer uma **estimativa aproximada** de n.
- Para simplificar as coisas, a cpf (neste exemplo) é ignorada e a **porcentagem amostral p é considerada normalmente distribuída**.
  - Investigar se essas suposições são razoáveis é uma questão que pode ser verificada quando o n inicial é conhecido.

Um exemplo

- Em termos técnicos, p deve estar dentro dos limites de $(P \pm 5)$, exceto para uma possibilidade em 20.
- Uma vez que p é assumido normalmente distribuído em torno de P, ele estará no intervalo $(P \pm 2\sigma_p)$, **exceto por uma possibilidade em vinte**.
- Além disso, temos que o erro padrão de p é

$$\sigma_p \doteq \sqrt{PQ/n}.$$

- Portanto, podemos escrever que

$$2\sqrt{PQ/n} = 5 \quad \text{ou} \quad \boxed{n = 4PQ/25.}.$$

Um exemplo

Nesse ponto, surge uma **dificuldade comum** a todos os problemas de **estimativa do tamanho da amostra**.

- Uma **fórmula** para n foi obtida, mas n **depende de alguma propriedade da população** que será amostrada.
- Neste caso, a propriedade é a quantidade P, a porcentagem de habitantes no grupo sanguíneo O.
  - Note que esta é justamente a quantidade que gostaríamos de medir.

Um exemplo

Portanto, **perguntamos** ao antropólogo se ele pode nos dar alguma ideia do provável valor de P.

- Ele responde que **a partir de dados anteriores** sobre outros grupos étnicos, e de suas especulações sobre a história étnico-racial desta ilha, ele ficará surpreso se P estiver fora da faixa de 30% a 60%.

Um exemplo

- Esta **informação** é suficiente para fornecer uma resposta útil.
- Para qualquer valor de P entre 30 e 60, o produto PQ está entre 2100 e um máximo de 2500.
  - Este valor máximo de PQ é atingido quando $P = 50$.
- O n correspondente está entre 336 e 400.
  - Por segurança, 400 é considerado como a estimativa inicial de n.

Um exemplo

- As suposições feitas nesta análise podem agora ser reexaminadas.
- Com $n = 400$ e P entre 30 e 60, a distribuição de p deve ser próxima da normal.
- Se a cpf deve ser levada em consideração, então dependemos do número de habitantes da ilha.
  - Se a população exceder 8000 e a fração de amostragem é inferior a 5%, nenhum ajuste para cpf é necessário.

A análise do problema

A análise do problema

As principais etapas envolvidas na escolha do tamanho da amostra são as seguintes.

1. Deve haver alguma declaração sobre o que se espera da amostra.
    - Essa declaração pode ser em termos de **limites de erro desejados**, como no exemplo anterior, ou em termos de alguma decisão ou ação a ser tomada quando os resultados da amostra forem conhecidos.
    - **A responsabilidade** pelo enquadramento da declaração recai principalmente sobre **as pessoas que desejam usar os resultados da pesquisa**, embora frequentemente precisem de orientação para expressar seus desejos em termos numéricos.

A análise do problema

2. Alguma equação que conecte n com a precisão desejada da amostra deve ser encontrada.
   - A equação irá variar de acordo com o **conteúdo da declaração de precisão** e com o **tipo de amostragem que é utilizada**.
   - Uma das vantagens da amostragem probabilística é que ela permite que essa equação seja construída.
3. Esta equação conterá, como parâmetros, certas propriedades desconhecidas da população.
   - **Devem ser "estimados" para dar resultados específicos (concretos)**. Estas estimativas também são chamadas de **valores iniciais**.

A análise do problema

4. Muitas vezes acontece que os dados devem ser publicados para certas subdivisões principais da população **(subpopulações)** e que os limites de erro desejados são estabelecidos para cada subdivisão.
   - **Um cálculo separado é feito para n em cada subdivisão, e o n total é encontrado pela soma dos tamanhos de amostra de cada subgrupo.**
5. Mais de um item ou característica geralmente é medido em uma pesquisa por amostragem: às vezes, o número de itens é grande (proporção de homens/mulheres, proporção de habitantes do grupo sanguíneo O, média da idade, etc.).
   - Se um grau desejado de precisão é prescrito para cada item, os cálculos levam a uma série de valores conflitantes de n, um para cada item.
   - **Algum método deve ser encontrado para reconciliar esses valores.**

A análise do problema

6. O valor escolhido de n deve ser avaliado para ver se é consistente com os recursos disponíveis para colher a amostra.
   - Isso exige uma estimativa do custo, trabalho, tempo e materiais necessários para obter o tamanho proposto da amostra.
   - **Às vezes fica claro que n terá que ser drasticamente reduzido.**
   - Uma difícil decisão deve então ser tomada: se prosseguir com um tamanho de amostra muito menor, reduzindo assim a precisão, ou abandonar os esforços até que mais recursos possam ser encontrados.

A especificação da precisão

A especificação da precisão

- A declaração de precisão desejada pode ser feita fornecendo a quantidade de erro que estamos dispostos a tolerar nas estimativas de amostra.
- Este montante é determinado, da melhor maneira possível, em função dos usos que serão dados os resultados das amostras.
- Às vezes é difícil decidir quanto erro deve ser tolerado, especialmente quando os resultados têm vários usos diferentes.

A especificação da precisão

- Suponha que perguntamos ao antropólogo por que ele desejava que a porcentagem com grupo sanguíneo O fosse correta para 5% em vez de, digamos, 4 ou 6%.
- Ele pode responder que os dados do grupo sanguíneo devem ser usados principalmente para classificação étnico-racial.
- Ele suspeita fortemente que os insulares pertencem a um tipo étnico-racial com P de cerca de 35% ou a um outro tipo com P de cerca de 50%.
- Um limite de erro de 5% na estimativa parecia-lhe pequeno o suficiente para permitir a classificação em um desses tipos.
    - Ele, entretanto, não teria fortes objeções aos limites de erro de 4 ou 6%.

A especificação da precisão

- **Assim, a escolha de um limite de 5% de erro pelo antropólogo foi até certo ponto arbitrária.**
- A esse respeito, o exemplo é típico da maneira pela qual um limite de erro é frequentemente decidido.
- Na verdade, o antropólogo tinha mais certeza do que queria do que muitos outros cientistas e administradores terão.
- **Quando a questão do grau desejado de precisão é levantada pela primeira vez, tais pessoas podem confessar que nunca pensaram a respeito e não têm ideia da resposta.**
- Entretanto, depois de discutido o assunto, eles podem frequentemente indicar pelo menos aproximadamente o tamanho de um limite de erro que lhes parece razoável.

A fórmula para n na amostragem para proporções

n na amostragem para proporções

- **Relembrando:** as unidades são classificadas em duas classes, C e C'.
- Admite-se uma certa margem de erro d na proporção estimada p de unidades na classe C, e há um pequeno risco α de que estejamos dispostos a incorrer de que o erro real seja maior do que d; ou seja, nós queremos

$$\Pr(|p - P| \geq d) = \alpha.$$

- Ou seja, um erro $(p - P)$ maior que d ocorre com probabilidade α.

n na amostragem para proporções

- A **amostragem aleatória simples** é assumida e p é considerado como normalmente distribuído. Da expressão **(2)** da aula 17 **(Teo. 17.2)**, temos:

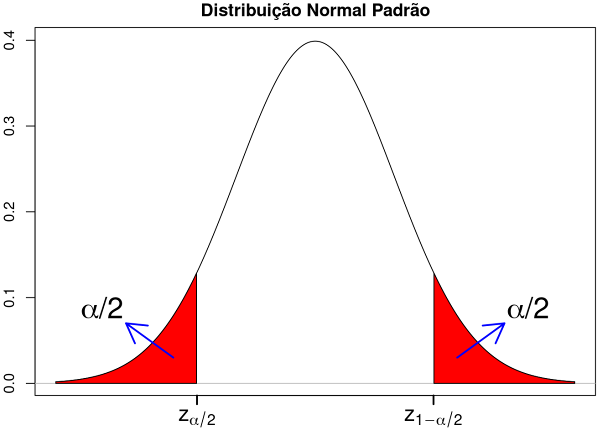
$$\sigma_p = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{PQ}{n}}.$$

- Portanto, a **fórmula que conecta** n com o grau de precisão desejado é

$$d = z_\alpha\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{PQ}{n}},$$

n na amostragem para proporções

em que z_α é a abscissa da curva normal que define uma área α nas caudas.



n na amostragem para proporções

- Resolvendo para n, encontramos

$$n = \frac{\frac{z_\alpha^2 PQ}{d^2}}{1 + \frac{1}{N}\left(\frac{z_\alpha^2 PQ}{d^2} - 1\right)}. \qquad (1)$$

- Para uso prático, uma **estimativa antecipada** p de P é substituída nesta fórmula.
- Se N for grande, uma primeira aproximação é

$$n_0 = \frac{z_\alpha^2 pq}{d^2} = \frac{pq}{V}, \qquad (2)$$

em que $V = (d^2/z_\alpha^2) = pq/n_0 =$ variância desejada da proporção amostral.

n na amostragem para proporções

- Na prática, primeiro calculamos n_0.
  - Se n_0/N for desprezível, n_0 é uma aproximação satisfatória para n de (1).
  - Se não, é aparente na comparação de (1) e (2) que n é obtido como

$$n = \frac{n_0}{1 + (n_0 - 1)/N} \approx \frac{n_0}{1 + (n_0/N)}. \tag{3}$$

Exemplo

- No **exemplo hipotético de grupos sanguíneos**, tivemos

$$d = 0,05, \quad p = 0,5, \quad \alpha = 0,05, \quad z_\alpha \approx 2.$$

Consequentemente,

$$n_0 = \frac{(4)(0,5)(0,5)}{0,0025} = 400.$$

Exemplo

- Suponhamos que haja apenas 3.200 pessoas na ilha. A cpf é necessária e encontramos

$$n = \frac{n_0}{1 + (n_0 - 1)/N} = \frac{400}{1 + 399/3200} = 356.$$

Exemplo

Pacote PracTools

- No R, o pacote PracTools possui funções para a estimativa do tamanho de amostra sob AAS.
  - A função nProp calcula o tamanho de amostra para a proporção conforme as expressões (2) e (3).

```
# install.packages("PracTools")
library(PracTools)

nProp(V0 = (0.05/2)^2, N = 3200, pU = 0.5)
```

```
## [1] 355.6543
```

```
# diferença ao usar z 'arrendondado'
nProp(V0 = (0.05/1.96)^2, N = 3200, pU = 0.5)
```

```
## [1] 343.0804
```

Comentários

- A fórmula para n_0 também é válida se d, p e q forem expressos como **porcentagens** em vez de proporções.
- Como o produto pq aumenta à medida que p se move em direção a $1/2$, ou 50%, uma estimativa conservadora de n é obtida escolhendo para p o valor mais próximo de $1/2$ no intervalo em que se pensa que p provavelmente estará.
- Se p parece estar entre 5 e 9%, por exemplo, assumimos 9% para a estimativa de n.

Comentários

- Às vezes, particularmente ao estimar o número total NP de unidades na classe C, desejamos controlar o **erro relativo** r em vez do erro absoluto em Np.
  - Por exemplo, podemos desejar estimar NP com um erro não superior a 10%. Ou seja, nós queremos

$$\Pr\left(\frac{|Np - NP|}{NP} \geq r\right) = \Pr(|p - P| \geq rP) = \alpha.$$

- Para esta especificação, substituímos rP ou rp para d nas fórmulas (1) e (2). De (2) obtemos

$$n_0 = \frac{z_\alpha^2 pq}{r^2 p^2} = \frac{z_\alpha^2}{r^2}\frac{q}{p}.$$

- A fórmula (3) permanece inalterada.

A fórmula para n com dados contínuos

n com dados contínuos

- Mais comumente, desejamos controlar o **erro relativo** r na estimativa total ou média da população.
- Com uma amostra aleatória simples tendo média $\overline{y}$, queremos

$$\Pr\left(\left|\frac{\overline{y}-\overline{Y}}{\overline{Y}}\right| \geq r\right) = \Pr\left(\left|\frac{N\overline{y}-N\overline{Y}}{N\overline{Y}}\right| \geq r\right) = \Pr(|\overline{y}-\overline{Y}| \geq r\overline{Y}) = \alpha,$$

em que α é uma probabilidade pequena.

- Assumimos que $\overline{y}$ é normalmente distribuído: do **Corolário 10.1** (aula 10), seu erro padrão é

$$\sigma_{\overline{y}} = \sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}.$$

n com dados contínuos

Portanto,

$$r\overline{Y} = z_\alpha \sigma_{\overline{y}} = z_\alpha \sqrt{\frac{N-n}{N}} \frac{S}{\sqrt{n}}$$

- Resolvendo para n, encontramos

$$n = \left(\frac{z_\alpha S}{r\overline{Y}}\right)^2 \Big/ \left[1 + \frac{1}{N}\left(\frac{z_\alpha S}{r\overline{Y}}\right)^2\right].$$

- Observe que a característica da população da qual n depende é seu **coeficiente de variação** $S/\overline{Y}$.
- Isso geralmente é mais estável e fácil de "estimar" antecipadamente que o próprio S.

n com dados contínuos

- Como uma primeira aproximação, tomamos

$$n_0 = \left(\frac{z_\alpha S}{r\overline{Y}}\right)^2 = \frac{1}{C}\left(\frac{S}{\overline{Y}}\right)^2 \tag{4}$$

substituindo uma estimativa antecipada de $(S/\overline{Y})$. A quantidade C é o $(cv)^2$ desejado da estimativa amostral.

n com dados contínuos

- Se n_0/N é não desprezível, calculamos n como em (3)

$$n = \frac{n_0}{1 + (n_0/N)}.$$

- Se em vez do **erro relativo** r quisermos controlar o **erro absoluto** d em $\overline{y}$, tomamos $n_0 = z_{\alpha}^2 S^2/d^2 = S^2/V$, em que V é a variância desejada de $\overline{y}$.

Exemplo

- Em viveiros que produzem árvores jovens para venda, é aconselhável estimar, no final do inverno ou início da primavera, quantas árvores jovens saudáveis podem estar disponíveis, uma vez que isso determina a política de solicitação e aceitação de pedidos.
- Um estudo de métodos de amostragem para a estimativa do número total de mudas foi realizado.
- Os dados a seguir foram obtidos de uma camada de mudas de bordo prateado com 1 pé de largura e 430 pés de comprimento.
- A unidade de amostragem foi de 1 pé do comprimento do leito, de modo que $N = 430$.
- Pela enumeração completa do leito se descobriu que $\overline{Y} = 19$, $S^2 = 85,6$, sendo esses os valores reais da população.

Exemplo

- Com a amostragem aleatória simples, quantas unidades devem ser tomadas para estimar $\overline{Y}$ em 10%, além de uma chance de 1 em 20? De (4) obtemos

$$n_0 = \frac{z_\alpha^2 S^2}{r^2 \overline{Y}^2} = \frac{(4)(85,6)}{(1,9)^2} = 95.$$

- Uma vez que n_0/N não é desprezível, tomamos

$$n = \frac{95}{1 + \frac{95}{430}} = 78.$$

- Quase 20% do leito deve ser contado para se atingir a precisão desejada.

Exemplo

Pacote PracTools

- A função `nCont` calcula o tamanho de amostra com dados contínuos visto anteriormente.

```
# install.packages("PracTools")
library(PracTools)

nCont(S2 = 85.6, ybarU = 19, N = 430, CV0 = (0.10/2))
```

```
## [1] 77.70729
```

```
# diferença ao usar z 'arrendondado'
nCont(S2 = 85.6, ybarU = 19, N = 430, CV0 = (0.10/1.96))
```

```
## [1] 75.168
```

Comentários

- As fórmulas para n fornecidas aqui se aplicam apenas à amostragem aleatória simples em que a média da amostra é usada como a estimativa de $\overline{Y}$.
- As fórmulas apropriadas para outros métodos de amostragem e estimativa são apresentadas com a discussão dessas técnicas **(Amostragem 2)**.

Para casa

- Revisar os tópicos discutidos nesta aula.
- Suponha que você irá realizar um levantamento por amostragem, utilizando amostragem aleatória simples, para estimar a porcentagem da presença de uma certa característica (do seu interesse; do tipo dicotômica) em uma população alvo.
    - Defina a característica de seu interesse e as populações alvo e de estudo.
    - Defina uma margem de erro; justifique a especificação deste valor para os limites de erro.
    - Defina a declaração de confiança das futuras estimativas.
    - Estime o tamanho da amostra para este problema.
    - Elabore conclusões com respeito ao resultado, suposições necessárias e a viabilidade de se executar este levantamento por amostragem.
    - Compartilhe os seus achados no Fórum Geral do Moodle.
- Consultar e estudar a página:

https://sites.google.com/hcpa.edu.br/bioestatistica/softwares-e-aplicativos/pss-health?authuser=0

Próxima aula

- Preparação para as apresentações das ativdades de avaliação 03 e 04.

Por hoje é só!

Bons estudos!