Oficina

Testes de Mercado e Metodologia Quantitativa

Carolina Musso
Universidade de Brasília

2024-12-12

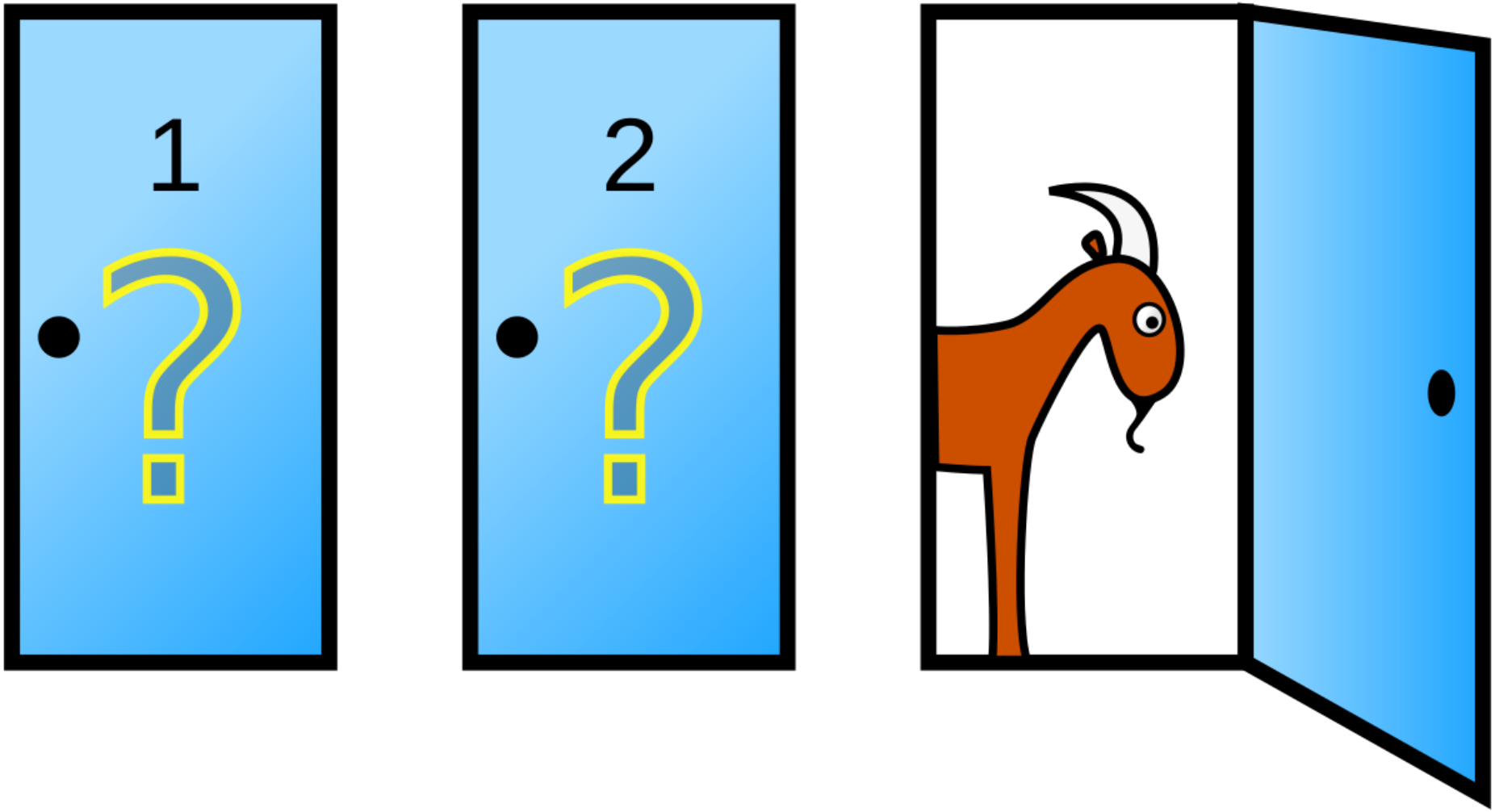
Agenda

- Probabilidade
- Amostragem
- Questionários
- Análise de Dados

Probabilidade ...

- Não é intuitivo.

Problema Monty Hall



Problema dos aniversários

Quantas pessoas são necessárias em um *grupo* para que a probabilidade de *pelo menos duas* delas compartilharem o mesmo *aniversário* seja maior que *90%*?

- 60 pessoas

Porque é esquisito?

- **Viés de confirmação:** focamos no que confirma nossas expectativas.
- **Lógica anecdótica:** eventos raros moldam nossa percepção de frequência.
- **Pensamento de Curto Prazo:** Tendemos a exagerar eventos recentes e ignorar tendências de longo prazo.

Mas!

Compreender probabilidades ajuda a tomar decisões informadas e a diminuir erros.

Introdução à Probabilidade

- Uma função , definida na -álgebra de subconjuntos de ,e com valores entre [0,1], é dita uma probabilidade se ela satisfaz aos axiomas de Komolgorov:
- P() = 1;
- ;
- Se são eventos disjuntos dois a dois () para), então:

$$(A_1, A_2, A_3, \cdots \in \mathcal{A})$$

$$(A_i \cap A_j \neq \emptyset)$$

$$P(\bigcup_{i=1}^{\infty} A_i = \sum_{i=1}^{\infty} P(A_i)$$

- A trinca (, ,) é chamada de espaço de probabilidade.

Amostragem

Amostragem é um processo estatístico de seleção de uma parte **representativa** de uma população para realizar uma **inferência** sobre esta população.

PDAD

Pesquisa Distrital por Amostra de Domicílios

PDAD Ampliada 2023

- 12 municípios da Periferia Metropolitana de Brasília
- 16 outras localidades inseridas no DF e na PMB

Inferencia estatística

- Extrair informações de uma **amostra** para fazer inferências sobre a **população**.
- A teoria e os estudos de probabilidade fornecem a base matemática em amostras *probabilísticas*.
  - Estimar os parâmetros e quantificar a incerteza.
  - Testar hipóteses e avaliar a significância.

O famigerado tamanho da amostra

A fórmula para calcular o tamanho da amostra n ao estimar uma média é dada por:

$$n = \frac{z^2 \cdot \sigma^2}{e^2}$$

- z: ex.: para 95% de confiança; $(z = 1,96)$
- σ^2: variância da população (ou uma estimativa);
- e: margem de erro aceita.

Depende

- **Tamanho da população:** Se a população é pequena, a amostra deve ser ajustada.
- **Distribuição da população:** Se a população não é normal, a amostra deve ser maior.
- **Variabilidade da população:** Se a variabilidade é alta, a amostra deve ser maior.

Número “mágico”: 385-400

Para estimar a uma proporção, podemos usar a variância máxima.

$$n = \frac{z^2 \cdot p(1-p)}{e^2}$$

- variância máxima é $0,25$ (quando $p = 0,5$);

$$n = \frac{1.96^2 \cdot 0.5(1-0.5)}{0.05^2} =$$

Se a população for pequena

- Se a população é pequena (ex: 200).

$$n_{ajustado} = \frac{n_0}{1 + \frac{n_0 - 1}{N}} = \frac{}{1 +}$$

- Lembre-se de que estamos assumindo a variância máxima, então n não pode ser pequeno.

Se sei que a variância é pequena.

- População **homogênea**: Altura dos atletas de uma seleção de basquete.
- Variância pequena: .
- Margem de erro: .

$$\sigma^2 = 4\ cm^2$$

$$e = 2\ cm$$

$$n = \frac{1.96^2 \cdot 4}{2^2} = 3.8416 \approx$$

- Mais que tamanho da amostra
  - O importante é **como** a amostra é **coletada**.

Exemplo

- N = 1000 pessoas de consumidores de um novo streaming. **Quanto** estão dispostos a **pagar** em **média**.
  - DP padrão da população é de 29 reais.
  - Estamos disposos a errar em apenas 2 reais.
  - Com uma confiança de 95%.

Qual o tamanho da amostra necessário nesse caso

- Pela **fórmula inicial**, precisaríamos de 795 pessoas.
- Ajustando para **populações pequenas**, 443.

Amostra com 443 pessoas.

Amostra Aleatória Simples:

- A média estimada é **69.8**.
  - Estimativa **pontual**
- IC 95% **(67.2 , 72.3)**.
  - Estimativa **intervalar**
- A média real é de **68.9**.
  - A média pontual está próxima da real.
  - O IC contém a média real.

Amostra com 443 pessoas.

- **Não** aleatória: Coletada na ordem em que a população foi gerada.
- Média de **41.8**
  - Estimativa pontual
- IC de 95%: **(39.9 e 43.7)**.
  - Estimativa intervalar

Lembre

A média real é de **68.9**.

E se n= 795 ...

- 795 primeiros valores da população.

- A média de **61.4**.
  - Estimativa pontual
- IC de 95%: **(59.5 e 63.3**).
  - Estimativa intervalar

Veja!

Mesmo uma **amostra maior**, a média ainda está **distante** da média real (**round(mean(populacao),1)**) da população por **não ser aleatória**.

Na Prática

Limitações pesquisas amostrais probabilísticas

- **Custo:** Amostragem probabilística é mais cara que a não probabilística.
- **Tempo:** Pode ser mais demorada.
- **Complexidade:** Requer conhecimento técnico.
- **Acesso**: População desconhecida ou não acessível.

E agora?

Por onde começar

É preciso **definir** bem qual é a população de interesse.

- Em geral, valem esforços no sentido de conseguir a lista, com as partes interessadas ou órgãos públicos.
  - Ex: Melhorar bases de dados incompletas ou desatualizadas.

Amostragem estratificada

Pode-se melhorar a representatividade na amostra com o uso de estratificação (por setor, tamanho da empresa, etc.)

Amostragem estratificada

- Dividir a população em subgrupos (estratos) com características semelhantes.
- Outra forma de calcular o **n**, com diferentes possibilidade de *alocação*.

Voltando ao nosso exemplo…

- Na verdade, eram três grupos de consumidores:
  - Sol Nascente (250), Guará (350), Lago Sul (500).
- Agora com **n=150 pessoas**, dividido proporcionalmente e coletado **aleatoriamente**.

- Agora a média é **69**. IC de 95%: **(64.7 e 73.3**).

Veja!

O tamanho n **diminuiu** em relação à AAS, sem perda de precisão.

E se não fosse aleatória.

- A média da amostra estratificada não-aleatória é de **64.6**
  - Média pontual.
- IC de 95%: **(60.3 e 68.9)**

Veja!

A média real da população é de **68.9**. A estimativa já ficou bem melhor do que a coleta não aleatória de ~700 pessoas, mesmo também não sendo aleatória.

Outra possibilidade

- Amostragem sistemática.
  - Sorteia-se o primeiro elemento e cada -ésimo elemento (baseado no tamanho da amostra).
  - Exemplo amostra de 200 pessoas, k = 5.
- Dessa vez, a média da amostra é de **69**.
- IC de 95%: **(65.3 e 72.8)**.

> **Lembre!**
>
> A média real da população é de **68.9**.

Custo e Prazos

- Métodos digitais ou telefônicos são menos custosos que visitas.
- Questionários padrão ajudam na replicação e análise automatizáveis (relatórios estatísticos reprodutíveis);
  - Priorização de perguntas fechadas, facilita a análise.
  - Coleta de respostas por questionários eletrônicos, como **forms**, **surveymonkey**, **redcap** e etc.
- Uso de dados secundários.

Dados secundários: PDAD

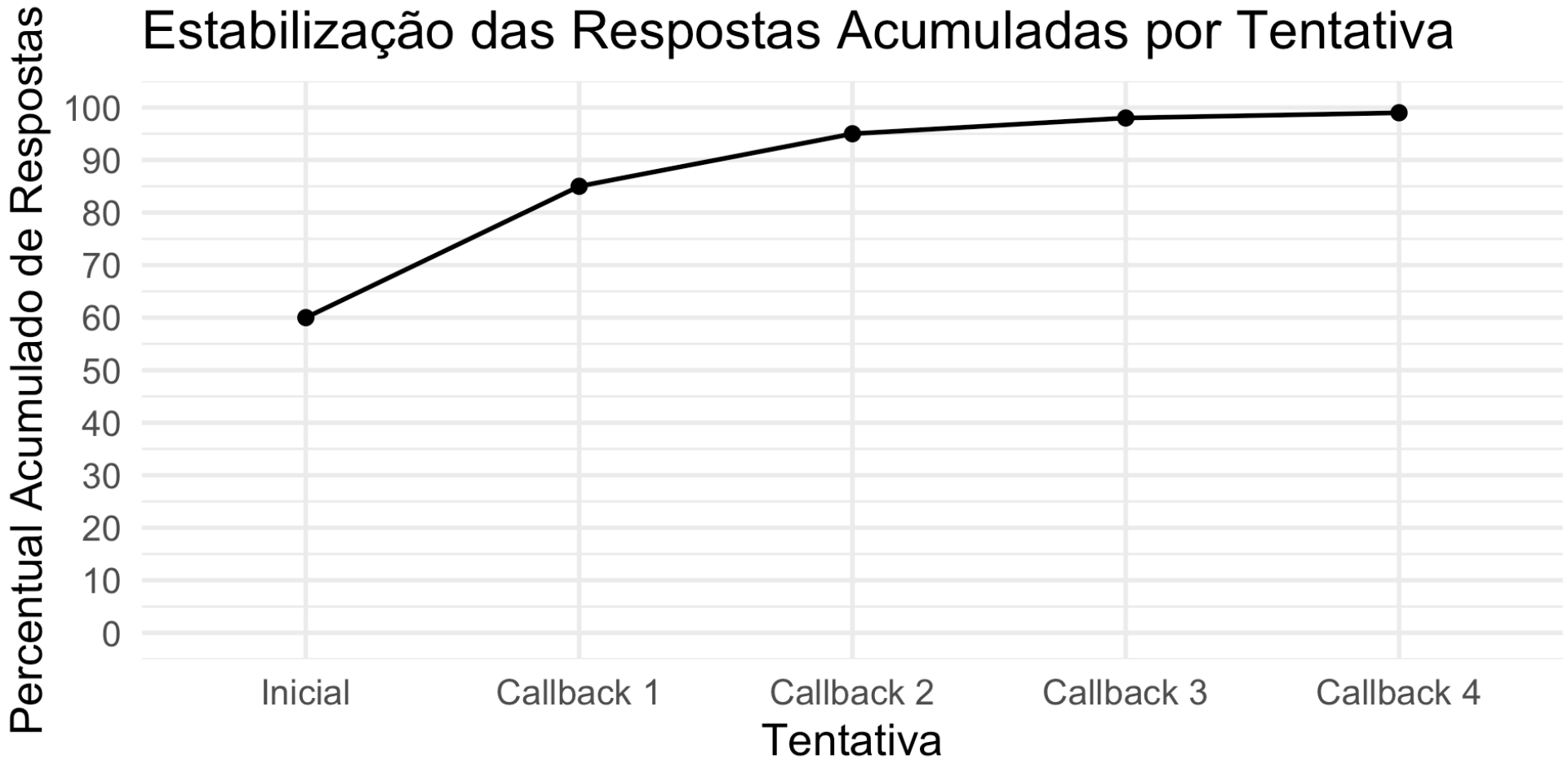
- Pode ajudar a entender o tamanho da população de consumidores, por exemplo.
- Por exemplo, de acordo com a PDAD 2021 :
  - Em 61,6% dos domicílios, havia assinatura serviços on-line, como filmes, músicas, notícias, cursos, esportes etc.;
  - Em 40,4%, havia serviço de TV por assinatura; e
  - 7% assinavam revistas ou jornais impressos

Não respostas

Possibilidades

- **Callbacks:** Contatar novamente os respondentes
  - Aumentar número de respostas.
  - Utilize outros meios de contato.
  - Estabeleça um número máximo de tentativas de contato, considerando o custo-benefício.
  - Priorize subgrupos ou áreas com menores taxas de resposta inicial.

Callbacks


Estabilização das Respostas Acumuladas por Tentativa
Percentual Acumulado de Respostas
100
90
80
70
60
50
40
30
20
10
0
Inicial
Callback 1
Callback 2
Callback 3
Callback 4
Tentativa

Otimização do Questionário

- **Clareza e objetividade**: Ser o mais curto e simples possível.
  - Usar termos claros e acessíveis.
  - Não misturar perguntas.
- Dar preferência a **perguntas fechadas**.
  - Evitar medidas subjetivas.
  - Usar escalas padronizadas

Otimização do Questionário

- Manter **ordem lógica** no questionário.
  - Manter o engajamento do respondente.
  - Evitar fadiga do respondente.
- Evitar perguntas tendenciosas.
- **Piloto** para testar a compreensão e a eficácia.

Exemplos

Exemplo:

“Com que frequência você estuda para as provas das disciplinas?”

( ) Nunca
( ) Raramente
( ) Algumas vezes
( ) Frequentemente
( ) Sempre

“Quantos dias por semana, em média, você estuda para as provas das disciplinas?”

( ) 0 dias
( ) 1-2 dias
( ) 3-4 dias

( ) 5-6 dias
( ) Todos os dias (7 dias)

Exemplo:

“Você está satisfeito com o ensino na sua faculdade?”

( ) Sim

( ) Não

- A pergunta é ampla demais.
- Não capta nuances de opiniões.
- Não há informações sobre quais aspectos precisam de melhorias.

Quão satisfeito você está com os seguintes aspectos do ensino na sua faculdade?
1 (Muito Insatisfeito) - 5 (Muito Satisfeito)

	1	2	3	4	5
Qualidade das aulas	○	○	○	○	○
Disponibilidade de professores para tirar dúvidas	○	○	○	○	○
Infraestrutura das salas de aula.	○	○	○	○	○

Exemplo:

“Há barreiras à livre circulação interestadual do produto X no Brasil? Explique, considerando fatores como custos de transporte, impostos etc.”

- Mistura uma questão fechada com explicações abertas.

Sugestão

- Existem barreiras à livre circulação interestadual do produto X no Brasil? (Sim/Não)
- Quais são os principais fatores que contribuem para essas barreiras? (Opções: Custos de transporte, impostos, regulamentações estaduais, outros).
- Se outros, por favor, especifique.

Exemplo:

“Informe se, caso as Requerentes deste Ato de Concentração viessem a aumentar o preço do produto X após a Operação (por exemplo, em um patamar entre 5% e 10%), sua empresa: (i) continuaria adquirindo os produtos desses fornecedores, ou (ii) substituiria os produtos das Requerentes por produtos de outro(s) fornecedor(es), conseguindo ter sua demanda integralmente atendida. Justifique.”

- Longa, múltiplos parênteses que podem confundir.
- Combina uma pergunta fechada aberta.

Solução Proposta

- Se os preços do produto X fornecido pelas requerentes aumentassem entre 5% e 10% após a operação, como sua empresa responderia? Selecione a alternativa mais provável.
  - Continuaríamos adquirindo os produtos das Requerentes, mesmo com o aumento de preço.
  - Reduziríamos parcialmente as compras desses fornecedores, complementando com outros fornecedores.
  - Substituiríamos completamente os produtos das Requerentes por outros fornecedores.
  - Não é possível atender nossa demanda integralmente com outros fornecedores.

Não respostas

- Maneiras de minimizar impacto de não respostas:
- Se souber de antemão a taxa, aumenta-se **n**.

TABLE 13.3

SMALLEST VALUE OF n FOR GIVEN LIMIT OF ERROR d, WITH RISK $\alpha = 0.05$

% Nonresponse, $100W_2$	d (%) 20	15	10	5
0	24	43	96	384
2	27	50	122	653
4	31	60	166	2000
6	36	75	255	....
8	43	99	521	....
10	53	142	...	....
15	112	...	...	....

Ponderação por frequência de Resposta

- Ajuste os pesos das observações respondidas para refletir a representatividade no total da amostra.

Exemplo:

- Se um grupo (como jovens) tem uma baixa taxa de resposta, as respostas obtidas desse grupo terão maior peso.

Pós-Estratificação

- Divida a amostra em estratos de acordo com características relacionadas à probabilidade de resposta (idade, sexo, região, etc.).
- Ajuste as proporções para refletir a composição populacional conhecida.

Exemplo: Se um estrato deveria representar 30% da população, mas representa apenas 20% na amostra final, multiplique os pesos desse estrato por (1.5).

Estimando médias: Ajustar Viés

- **20% não responderam**, e a maioria deles é homem, que são mais altos (~170cm).
- **80% responderam**, maioria são mulheres, que são mais baixas (~164).

$$Viés = W_{nr} \cdot (Y_r - Y_{nr})$$

Variáveis contínuas

O intervalo é muito amplo, e as **suposições** sobre o comportamento dos não-respondentes são mais complexas.

Para estimativa de proporções

- Ex: proporção de pessoas que concordam com uma decisão.
- Mais fácil, pois está sempre entre [0, 1]
- Podemos fazer intervalos conservadores:
  - Limite inferior (p = 0): Todos responderam “não”.
  - Limite superior (p = 1): Todos esponderam “sim”. . . .

Para estimativa de proporções

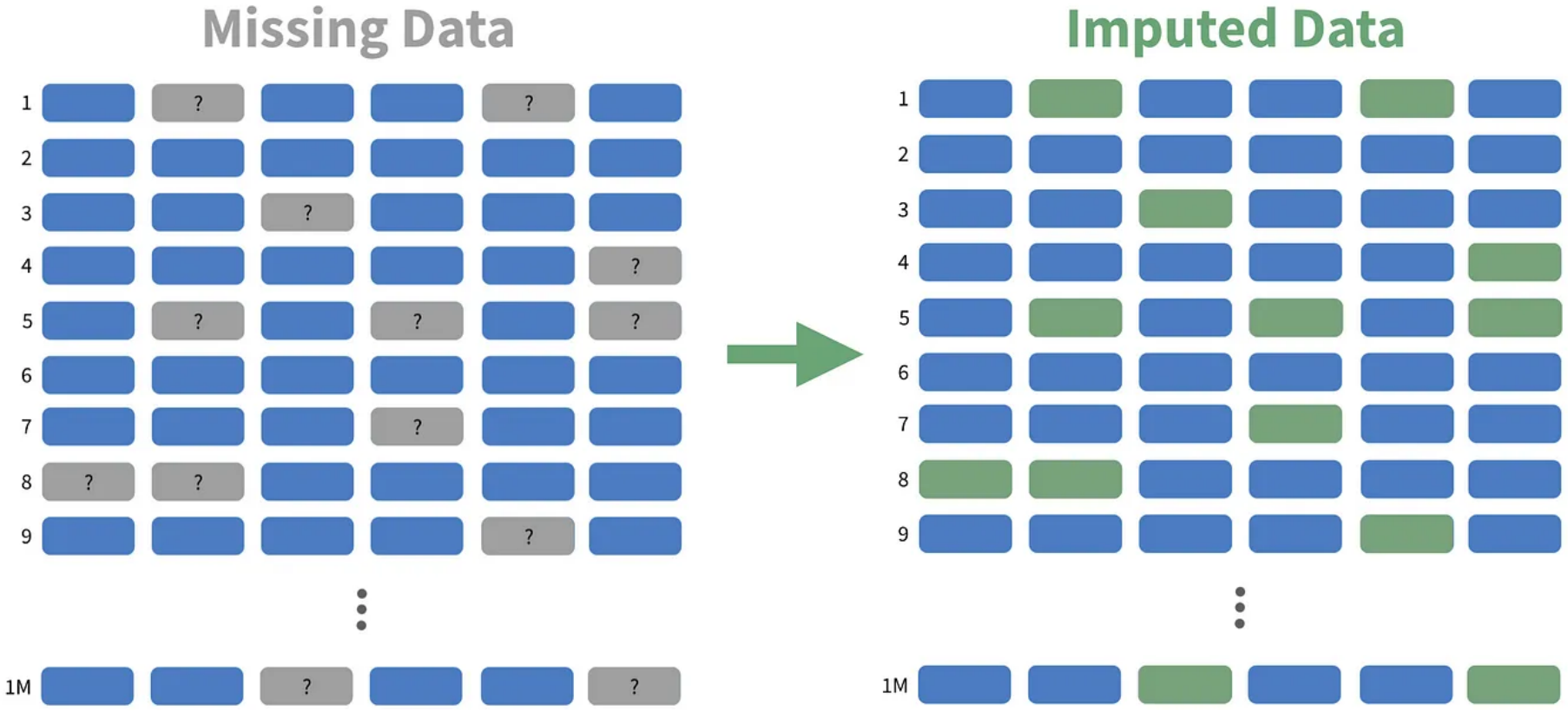
TABLE 13.2

95% CONFIDENCE LIMITS FOR P (%) WHEN $n = 1000$

% Nonresponse, $100W_2$	Sample Percentage, $100p_1$ 5	10	20	50
0	(3.6, 6.4)	(8.1, 11.9)	(17.5, 22.5)	(46.7, 53.2)
5	(3.4, 11.1)	(7.6, 16.3)	(16.5, 26.5)	(44.4, 55.6)
10	(3.2, 15.8)	(7.2, 20.8)	(15.6, 30.4)	(42.0, 58.0)
15	(3.0, 20.5)	(6.8, 25.2)	(14.7, 34.3)	(39.6, 60.4)
20	(2.8, 25.2)	(6.3, 29.7)	(13.7, 38.3)	(37.2, 62.8)

Imputação de Dados Faltantes

Uma possibilidade é **Regressão**: Com base nos respondentes, preveja para os não-respondentes.



Análises

- Não basta fazer a análise descritiva da amostra. É preciso “expandir” a amostra para a população.
- Estimativas **intervalares** são mais robustas que estimativas pontuais.
- SAS
- R: pacote survey

Análise de dados não probabilísticos

Rao, J.N.K. On Making Valid Inferences by Integrating Data from Surveys and Other Sources. Sankhya B 83, 242–272 (2021)

Obrigada!