Visão geral da parte III

Prof. Wagner Hugo Bonat

Departamento de Estatística
Universidade Federal do Paraná

Onde estamos?

1. Estatística descritiva e exploratória (UD1).
2. Probabilidades e variáveis aleatórias (UD2, UD3 e UD4).
3. Inferência estatística (UD5, UD6 e UD7).
4. Métodos estatísticos (UD8).

Figura 1. Foto de Andrea Piacquadio no Pexels.

Resumo

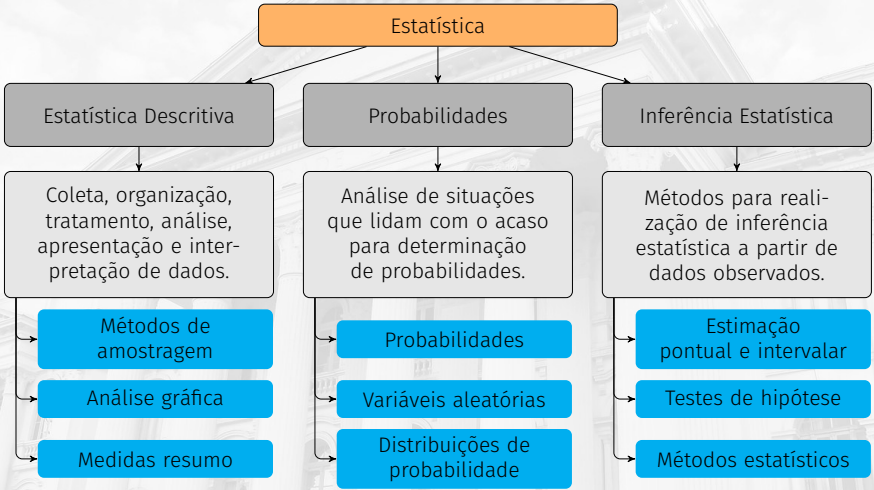

Figura 2. Organização modular da disciplina de estatística básica.

Inferência estatística

- **População** → distribuição de probabilidade.
- Intuição → Como que a v.a. deve se comportar na população.
- Variável → variável aleatória.
- Parâmetros da distribuição de probabilidade → **parâmetros populacionais**.
- Como obter a amostra?
- Como a partir da amostra estimar os parâmetros populacionais?

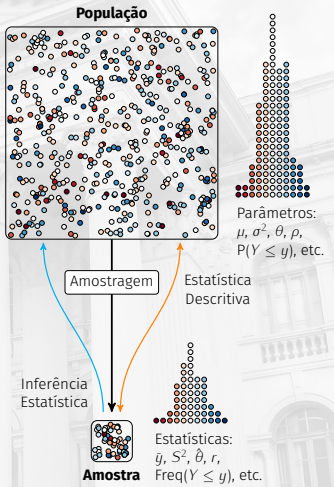

Figura 3. Processo de inferência estatística.

Inferência estatística

- Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?

Inferência estatística

- Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- Formalizando o problema:
  - Qual é a variável aleatória e quais valores ela pode assumir?

Inferência estatística

- Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- Formalizando o problema:
  - Qual é a variável aleatória e quais valores ela pode assumir?
  - Y: desenvolveu anticorpos. Opções SIM ou NÃO.
- Qual a distribuição de probabilidade adequada para esta v.a.?

Inferência estatística

- Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- Formalizando o problema:
  - Qual é a variável aleatória e quais valores ela pode assumir?
  - Y: desenvolveu anticorpos. Opções SIM ou NÃO.
- Qual a distribuição de probabilidade adequada para esta v.a.?
  - Bernoulli com função de probabilidade

$$\mathrm{P}(Y = y) = p^y(1-p)^{1-y}.$$

- Qual o parâmetro de interesse e o que ele significa?

Inferência estatística

- Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- Formalizando o problema:
  - Qual é a variável aleatória e quais valores ela pode assumir?
  - Y: desenvolveu anticorpos. Opções SIM ou NÃO.
- Qual a distribuição de probabilidade adequada para esta v.a.?
  - Bernoulli com função de probabilidade

$$\mathrm{P}(Y = y) = p^y(1-p)^{1-y}.$$

- Qual o parâmetro de interesse e o que ele significa?
  - p: proporção de pessoas que desenvolveram anticorpos.

Pensamento Estatístico

- Como determinar o valor de p?
  - Examinar todos os membros da população e verificar a proporção que desenvolveu anticorpos.
  - Examinar apenas alguns membros da população (amostra) e calcular a proporção que desenvolveu anticorpos.
- Problema: A proporção obtida na amostra não é a mesma obtida na população.
  - Incerteza associada ao valor da proporção devido a termos apenas uma amostra.
  - Como quantificar essa incerteza?
  - Como tomar uma decisão baseada apenas na amostra?
- Descrição probabilística da estatística de interesse $\rightarrow$ **Distribuição amostral.**

Especificação do problema de Inferência

- Y: desenvolveu anticorpos (v.a.).
- Especificação do modelo $Y \sim \text{Ber}(p)$.
- Parâmetro p.
- Informação sobre p através de uma amostra da população.
- Denotamos as amostras por $y_1, \ldots, y_n$.
- Objetivos da inferência estatística:
  - Estimar p baseado apenas na amostra (valor pontual)! Quanto é p na população?
  - Informar o quanto preciso ou creditável é o valor estimado (intervalo de confiança).
  - Decidir sobre possíveis valores de p baseado apenas na amostra.
  - A proporção da população com anticorpos atingiu um patamar desejável?

Especificação do problema de Inferência

- Suponha que coletamos uma amostra (aleatória) de tamanho $n = 10$ e que $y = 7$ pessoas apresentaram anticorpos.
- Qual valor você acha que o parâmetro p assume na população?
- Assumindo observações independentes, sabemos que a soma de v.a. Bernoulli é binomial com $n = 10$ e um parâmetro p desconhecido.
- Podemos calcular a probabilidade de observar $y = 7$ para um valor de p, por exemplo, $p = 0.8$

$$\mathrm{P}(Y = 7 | n = 10, p = 0{,}80) = \binom{10}{7} 0{,}80^7 (1 - 0{,}80)^{10-7} = 0{,}2013.$$

Especificação do problema de Inferência

- Para qualquer outro valor de p

$$\mathrm{P}(Y = 7|n = 10, p) = \binom{10}{7} p^7 (1-p)^{10-7},$$

variando p temos a **função de verossimilhança**

$$L(p) \equiv \mathrm{P}(Y = 7|n = 10, p) = \binom{10}{7} p^7 (1-p)^{10-7}.$$

- **Ideia**: Se p for um determinado valor, **qual a probabilidade** de observar o que eu realmente observei na amostra.

Pensamento frequentista

- Se o experimento for repetido um número grande de vezes e a cada realização $\hat{p}$ for obtido, o que aconteceria?
- $\hat{p}$ é uma variável aleatória.
- Se é variável aleatória, então tem distribuição de probabilidade que descreve o seu comportamento.
  - Qual é a sua distribuição?
  - Qual o seu valor esperado?
  - Qual a sua variância?

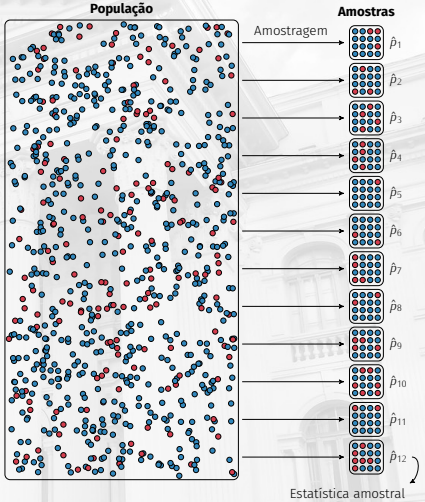

Figura 4. Ilustração da distribuição amostral.

Ilustração computacional

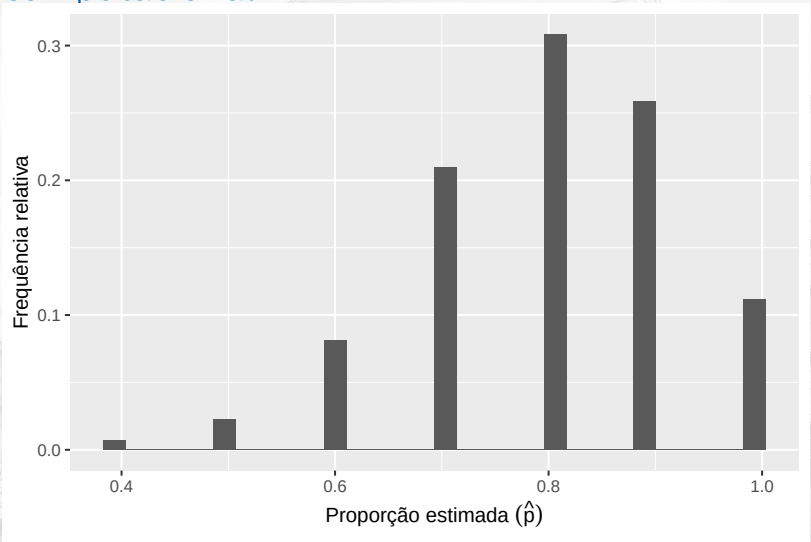

Figura 5. Distribuição amostral da proporção.

Distribuição amostral

- Veja que mesmo se o valor verdadeiro for $p = 0{,}8$ existe uma probabilidade não desprezível de observarmos 7 pessoas com anticorpos entre as 10 avaliadas.
- A incerteza associada ao valor de p no caso de apenas 10 observações é grande.
- Como podemos diminuir esta incerteza?

Distribuição amostral

- Veja que mesmo se o valor verdadeiro for $p = 0{,}8$ existe uma probabilidade não desprezível de observarmos 7 pessoas com anticorpos entre as 10 avaliadas.
- A incerteza associada ao valor de p no caso de apenas 10 observações é grande.
- Como podemos diminuir esta incerteza?
- Solução: Aumentar o número de observações.

Ilustração computacional

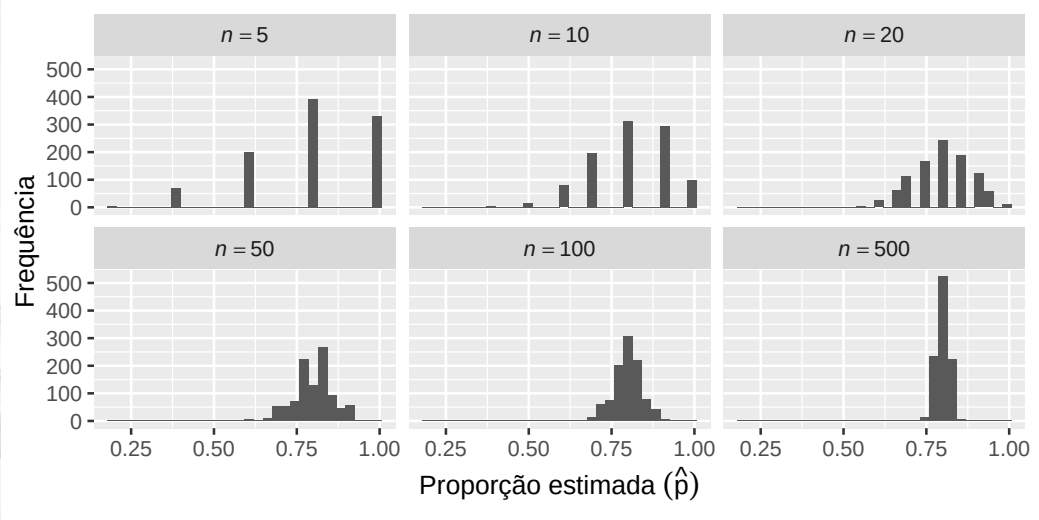

Figura 6. Efeito de aumentar o tamanho da amostra na distribuição amostral da proporção estimada.

Estatística frequentista

- Temos o procedimento, mas e como faremos as replicações do experimento em termos práticos?

Estatística frequentista

- Temos o procedimento, mas e como faremos as replicações do experimento em termos práticos?
- Não faremos!!
- Estimador é função da variável aleatória.
- Portanto, tem distribuição de probabilidade.
- A **distribuição amostral** do estimador pode ser usada para estudar o que aconteceria caso o estudo fosse replicado um número muito grande de vezes.
- Distribuição exata de um estimador é dificil de se obter.
- O Teorema Central do Limite oferece uma aproximação para amostras grandes (assintótica).

Reforçando os conceitos

- Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- Precisamos saber como a altura dos alunos se distribui.

Reforçando os conceitos

- Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- Precisamos saber como a altura dos alunos se distribui.
- Formalizando o problema.
  - Qual é a variável aleatória e quais valores ela pode assumir?

Reforçando os conceitos

- Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- Precisamos saber como a altura dos alunos se distribui.
- Formalizando o problema.
  - Qual é a variável aleatória e quais valores ela pode assumir?
  - $Y \in \Re_+$ - Altura dos alunos da UFPR.
- Qual a distribuição de probabilidade adequada para esta v.a.?

Reforçando os conceitos

- Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- Precisamos saber como a altura dos alunos se distribui.
- Formalizando o problema.
  - Qual é a variável aleatória e quais valores ela pode assumir?
  - $Y \in \Re_+$ - Altura dos alunos da UFPR.
- Qual a distribuição de probabilidade adequada para esta v.a.?
  - Normal, Gama, Lognormal, Normal Inversa, Weibul, etc.
- Qual o parâmetro de interesse e o que ele significa?

Reforçando os conceitos

- Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- Precisamos saber como a altura dos alunos se distribui.
- Formalizando o problema.
  - Qual é a variável aleatória e quais valores ela pode assumir?
  - $Y \in \Re_+$ - Altura dos alunos da UFPR.
- Qual a distribuição de probabilidade adequada para esta v.a.?
  - Normal, Gama, Lognormal, Normal Inversa, Weibul, etc.
- Qual o parâmetro de interesse e o que ele significa?
  - Altura média e variabilidade da altura dos alunos da UFPR.

Juntando dados e probabilidades

- Suponha que uma amostra (AAS) de tamanho $n = 293$ foi obtida.

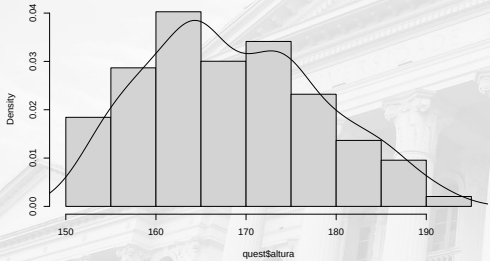

Figura 7. Histograma da altura dos alunos UFPR.

- Qual modelo é o mais provável de ter gerado essa amostra?

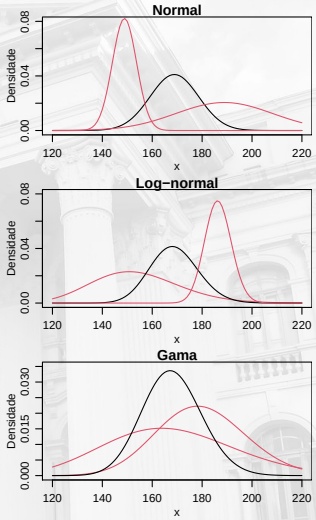

Figura 8. Distribuições de probabilidades candidatas.

Escolhendo o modelo

- Modelo Normal
  - Notação $Y \sim \mathrm{N}(\mu, \sigma^2)$;
  - $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$;
  - $\mathrm{E}(Y) = \mu$ e $\mathrm{V}(Y) = \sigma^2$.
- Quais os valores de μ e σ^2 devo usar?
- Podemos usar os equivalentes amostrais?
  - $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ e $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2$.
- Como medir a incerteza em $\hat{\mu}$ e $\hat{\sigma}^2$, sendo que temos apenas uma amostra? → Distribuição amostral.
- E para os outros modelos? → Métodos para estimação de parâmetros.

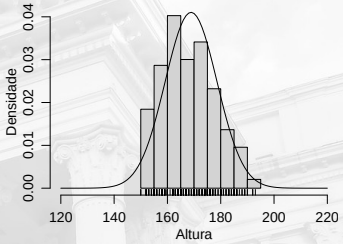

Figura 9. Ajuste da distribuição Normal para a variável altura.

Distribuição amostral

- Objeto de inferência (*frequentista*).
- A estimativa pontual é um resumo desta distribuição.
- Intervalos entre quantis representam a incerteza sobre o valor estimado.
- Comparam-se estimadores concorrentes pelas características de suas distribuições amostrais.
- E para tudo isto:
  é preciso saber como estimar.

Figura 10. Distribuição amostral de diferentes estimadores de um parâmetro.

Resumo

- Modelo → comportamento da natureza.
- Parâmetros do modelo → parâmetros populacionais de interesse.
- Qual modelo melhor descreve os dados?
- Assumimos um modelo → parâmetros são desconhecidos.
- Baseado na amostra → encontrar os parâmetros compatíveis com a amostra.
- Descrever a incerteza → **distribuição amostral**.

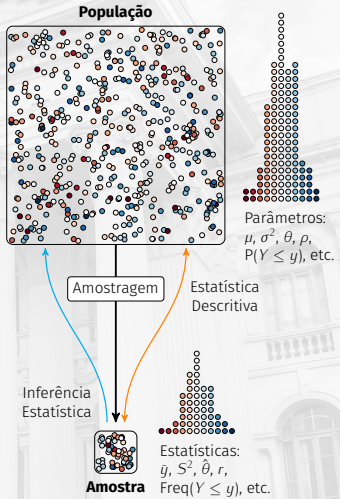

Figura 11. Processo de inferência estatística.