Inferência Estatística

Visão geral e ilustração computacional

Prof. Me. Lineu Alberto Cavazani de Freitas

Departamento de Estatística
Laboratório de Estatística e Geoinformação

Introdução

Introdução

Já discutimos os conceitos de **população**, **amostra** e **inferência**.

- **População**: conjunto de todos os elementos que compartilham alguma característica comum que temos interesse em estudar.
- **Amostra**: subconjunto da população.
- **Inferência**: ramo da Estatística que tem como objetivo estudar a **população** por meio de evidências fornecidas por uma **amostra**.

Introdução

- Muitas vezes estamos interessados em **quantidades populacionais**, contudo trabalhar com a população pode ser custoso ou até mesmo impossível.
- A solução é trabalhar com um subconjunto da população, isto é, uma **amostra**.
- O objetivo das técnicas de **amostragem** é gerar um subconjunto que seja **representativo** em relação a população para estimar quantidades de interesse (uma média, uma variância, uma proporção, etc).

Introdução

- Contudo é intuitivo notar que, caso se repita o processo de amostragem, uma amostra diferente da inicial será obtida.
- Consequentemente, as medidas de interesse calculadas (média, variância, etc.) em diferentes amostras não serão iguais.
- Isto quer dizer que mesmo o procedimento de amostragem estando correto sempre haverá aleatoriedade envolvida e os valores calculados com base na amostra são **candidatos** à quantidade na população.

Introdução

- Devido à **natureza aleatória**, todas as quantidades associadas à amostra devem receber **tratamento probabilístico**.

- Levando isso em conta, são objetivos da inferência estatística:
  1. Estimar quantidades com base apenas na amostra (**estimativa pontual**).
  2. Avaliar o quão preciso ou creditável é o valor estimado (**intervalo de confiança**).
  3. Decidir sobre possíveis valores da quantidade baseado apenas na amostra (**teste de hipótese**).

Alguns conceitos importantes

Alguns conceitos importantes

- **Parâmetro**: uma medida numérica que descreve alguma característica da população.
  - Na maior parte dos casos são desconhecidos.
  - Usualmente representados por letras gregas (θ, μ, σ, etc).
- **Espaço paramétrico**: conjunto de valores que um parâmetro pode assumir.
- **Estimador**: função da variável aleatória.
  - Cálculo efetuado com os elementos da amostra com a finalidade de representar (estimar) um parâmetro na população.
  - Usualmente representados por letras gregas com acento circulflexo ($\hat{\theta}$, $\hat{\mu}$, $\hat{\sigma}$, etc).

Alguns conceitos importantes

- **Estimativa/estatística**:
  - Valores numéricos assumidos pelos estimadores.
  - Realização de um estimador.
  - Uma função dos valores observados.
  - Função que define o estimador avaliada em dados observados.
  - Um número.
- **Estimativa pontual**: um único valor numérico como candidato para o parâmetro de interesse.
- **Estimativa intervalar**: intervalo conjunto de valores "plausíveis" para o parâmetro de interesse.

Um exemplo prático

- Suponha que temos interesse em estimar a idade média de alunos de um curso de graduação.
- Calcular a média populacional é muito custoso, por isso, tomou-se uma amostra da população.
- Com essa amostra foi usado um estimador para chegar a uma estimativa da média populacional baseada na amostra.
- Complementar à estimativa pontual foi construído um intervalo de confiança para esta estimativa.
- A coordenação do curso tem interesse em avaliar se existe evidência suficiente nos dados que permite afirmar que a idade média é menor que 22 anos. Tarefa que pode ser cumprida por meio de um teste de hipóteses.

Diferentes paradigmas

- Existem diferentes paradigmas de Inferência Estatística, baseados em formas diferentes de sistematizar os problemas.
- Outros paradigmas são o Bayesiano e Verossimilhancista.
- Os problemas de Inferência também podem ser abordados por meio de técnicas computacionalmente intensivas, como Bootstrap.
- Nesta disciplina nosso foco será o paradigma frequentista, que tem a distribuição amostral como objeto central.

Distribuição amostral

Distribuição amostral

- Suponha que estamos interessados em uma variável aleatória na população, denotada por Y (por exemplo, o peso dos indivíduos).
- Desta variável aleatória tomamos uma amostra de tamanho n, que denotaremos por $Y_1, Y_2, \cdots, Y_n$ (por exemplo, uma amostra de pesos dos indivíduos da população).
- Em geral, consideramos que esta amostra é aleatória simples com reposição para garantir que os elementos da amostra sejam independentes e identicamente distribuídos (iid).

Distribuição amostral

- Suponha que temos interesse em uma quantidade populacional θ (por exemplo, o peso médio dos indivíduos).
- Para obter o real valor de θ precisaríamos avaliar toda a população, na maior parte das vezes isso é inviável.
- A solução é estimar θ por meio de um estimador $\hat{\theta}$ que é uma função das variáveis aleatórias constituintes da amostra, isto é, $\hat{\theta} = f(Y_1, Y_2, ..., Y_n)$.

Distribuição amostral

- Estimativas são **variáveis aleatórias** (sabemos o que pode acontecer, mas não o que vai acontecer).
- Variáveis aleatórias têm distribuição de probabilidade.
- A **distribuição de probabilidades de estimativas** é chamada de **distribuição amostral**.
- Para estudar um **parâmetro**, usamos a distribuição amostral.
- No **paradigma frequentista** pensamos no que aconteceria se diversas amostras fossem tomadas e em cada amostra a quantidade de interesse fosse obtida.

Distribuição amostral

- Imagine que:
  - Coletamos diversas amostras.
  - Em cada amostra calculamos o estimador de interesse (uma média, por exemplo).
  - Se obtivermos a distribuição empírica deste estimador, temos tudo que precisamos para fazer inferência.
- A distribuição amostral pode ser usada para avaliar o que aconteceria se o estudo fosse replicado um grande número de vezes.
- A distribuição amostral é o objeto de inferência (frequentista).

Distribuição amostral

- A **estimativa pontual** é um resumo da distribuição amostral.
- Intervalos entre **quantis** representam a incerteza sobre o valor estimado.
- Contudo, na prática temos apenas **uma** amostra.
- Mas diversas estatísticas de interesse tem distribuições amostrais conhecidas, como média, variância e proporção.
- Usamos estas distribuições baseadas no valor obtido da amostra.

Principais estimadores e estimativas

Principais estimadores e estimativas

- Denote os parâmetros média, variância e proporção de certa característica na população por μ, σ^2 e p, respectivamente.

- Os estimadores "naturais" para estas quantidades são as correspondentes média, variância e proporção calculadas na amostra.

$$\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$\hat{p} = \frac{\text{número de sucessos}}{n}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Teorema Central do Limite

Teorema Central do Limite

- A média é uma das quantidades de maior interesse em contextos práticos.
- A distribuição amostral da média é conhecida graças ao Teorema Central do Limite (TCL).
- Segundo o teorema, quanto maior o tamanho da amostra, a distribuição da média amostral se comporta segundo um modelo Normal.

Teorema Central do Limite

- Suponha uma amostra aleatória de tamanho n retirada de uma população com média μ e variância σ^2 finita.
- Note que o modelo da variável aleatória não é especificado.
- A amostra $(Y_1, Y_2, \cdots, Y_n)$ consiste de n variáveis aleatórias independentes e identicamente distribuidas.

Teorema Central do Limite

Segundo o teorema:

$$\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \right) \xrightarrow{D} Z \sim N(0,1), \text{ para } n \to \infty$$

De forma alternativa:

$$\bar{Y} \sim N(\mu; \sigma^2/n)$$

O resultado pode ser generalizado para proporções:

$$\hat{p} \sim N\left(p; \frac{p(1-p)}{n}\right)$$

Distribuições amostrais

Média e variância

- Se σ^2 é conhecido
  - $\hat{\mu} \sim N(\mu; \sigma^2/n)$.
  - $(\hat{\mu} - \mu)/(\sigma/\sqrt{n}) \sim N(0; 1)$.
- Se σ^2 é desconhecido
  - $(\hat{\mu} - \mu)/(S/\sqrt{n}) \sim t_{n-1}$.
  - $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Proporção

- $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$.

Ilustração computacional

Ilustração computacional

- Um questionário foi aplicado a alunos de uma turma de Estatística com diversas questões sobre características dos alunos.
- Uma das questões perguntava o peso dos alunos. Consideraremos que a turma é uma população e temos interesse em fazer inferência sobre o peso médio desta população.
- Para isso:
  1. Tomamos diversas amostras.
  2. Para cada amostra calculamos o peso médio.
  3. Considerando o vetor de médias, construimos a distribuição amostral.
  4. Com base na distribuição amostral empírica, fazemos inferência (estimativa pontual e intervalar).
- Neste caso, sabemos o verdadeiro peso médio. Logo, podemos verificar se nossa inferência foi bem sucedida.

Ilustração computacional

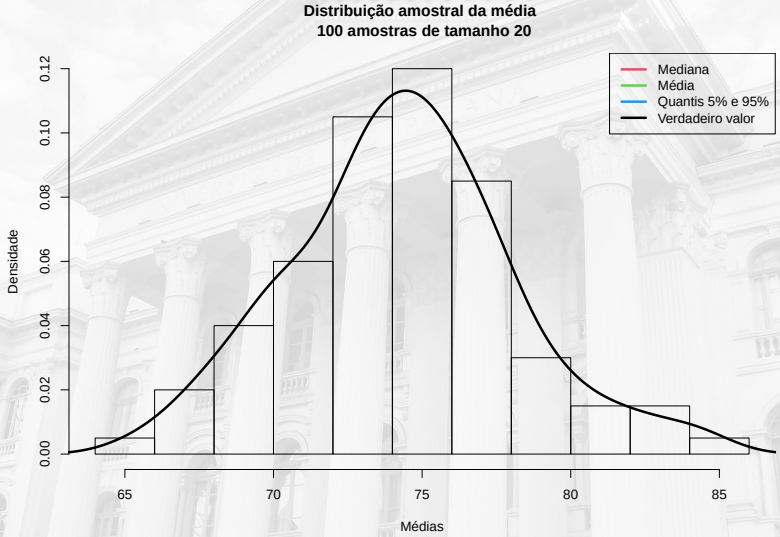


Ilustração computacional

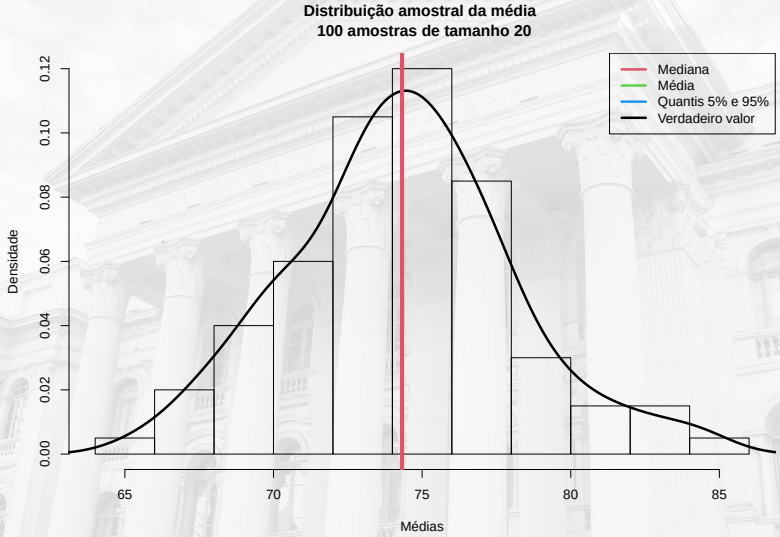


Ilustração computacional

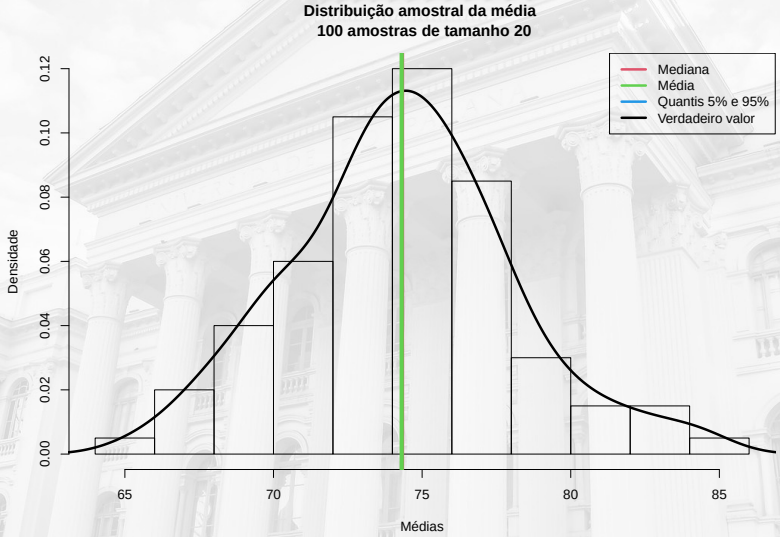


Ilustração computacional

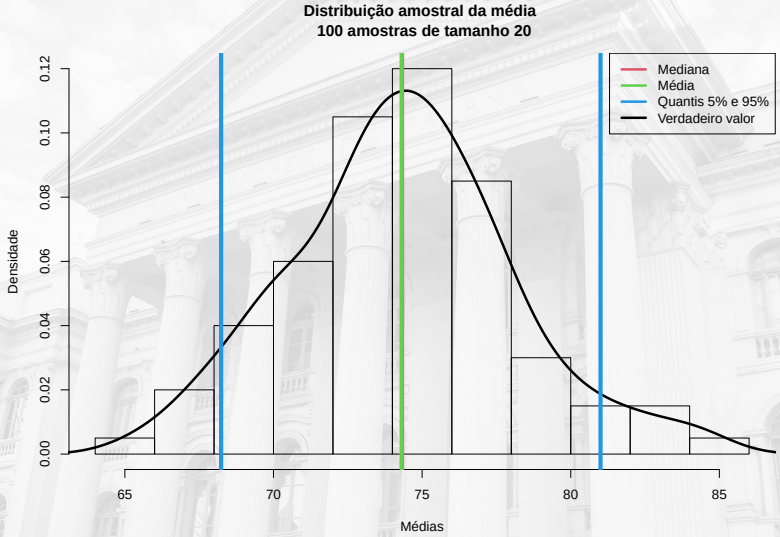


Ilustração computacional

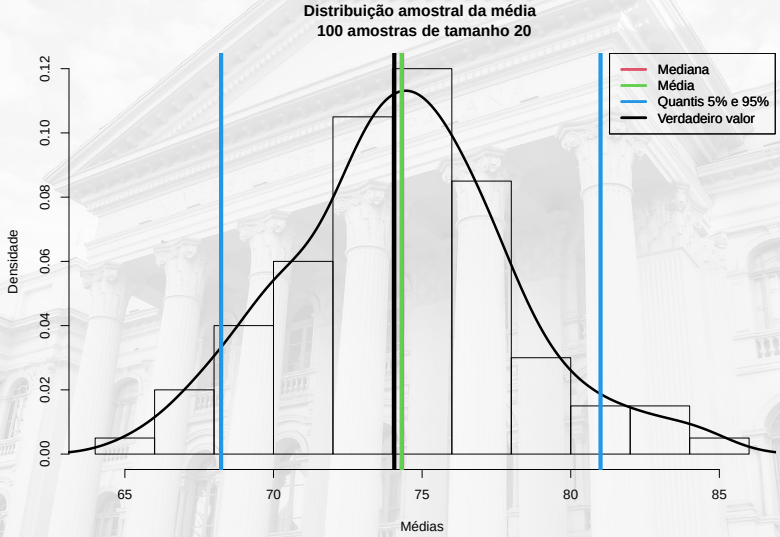


Ilustração computacional

- Os resultados mostram que não precisamos olhar a população para ter uma estimativa satisfatóriamente próxima do verdadeiro valor do parâmetro de interesse.
- Contudo esta estratégia é inviável na prática, pois necessita de várias amostras.
- Verificamos que a distribuição amostral é simétrica.
- O teorema central do limite garante que esta distribuição é Normal.

Ilustração computacional

- Na prática (para média) podemos usar uma distribuição normal centrada na estimativa de uma única amostra.
- Com base nesta distribuição amostral estimada, fazemos inferência.
- Os quantis desta distribuição garantem a confiança. Neste caso tomaremos os quantis 5% e 95% da distribuição estimada.
- Se replicarmos o procedimento 100 vezes, esperamos que em 10 vezes o intervalo dado pelos quantis não contenham o valor do parâmetro.

Ilustração computacional

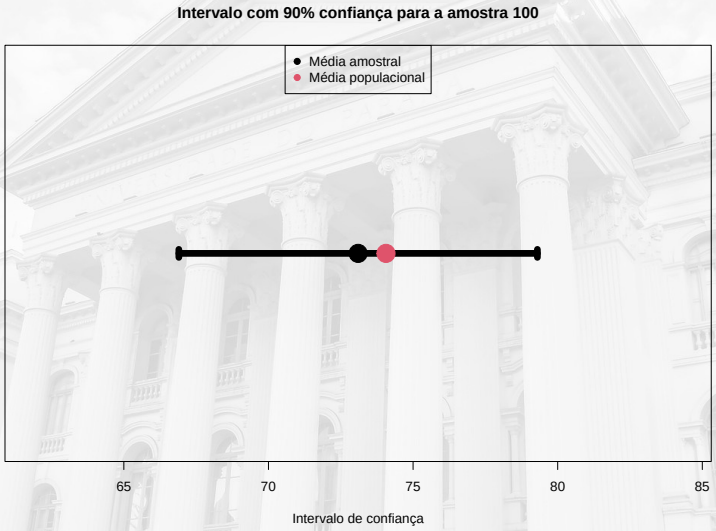


Ilustração computacional

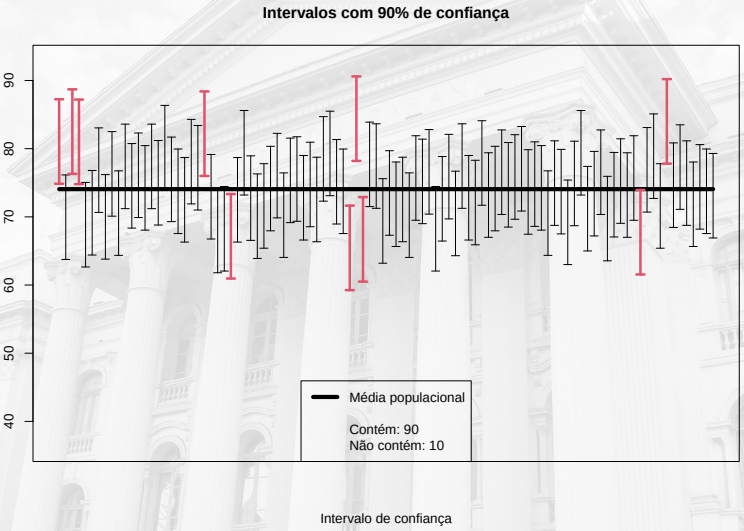
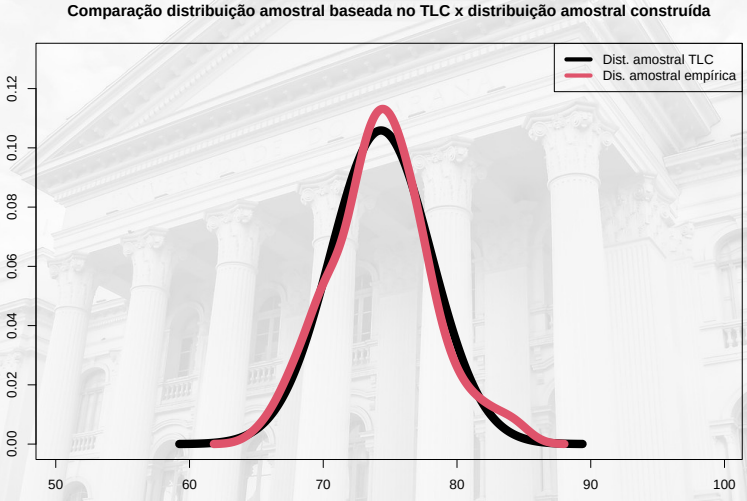


Ilustração computacional



O que foi visto:

- Introdução à Inferência estatística.
- Conceitos importantes.
- Distribuição amostral.
- Distribuição amostral da média.
- Ilustração computacional.

Próximos assuntos:

- Estimação pontual e intervalar.
  - Média com variância conhecida.
  - Média com variância desconhecida.
  - Proporção.