BIOESTATÍSTICA

M.I. Eng. Biomédica

2015-2016

Aula Teórica 3

Representação de dados

- Indicadores numéricos
  - Localização
  - Dispersão
  - Distribuição

Representação de dados

- Medidas de localização
  - Média aritmética

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Representação de dados

- Medidas de localização
  - Média aritmética: a soma dos desvios à média é nula.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Representação de dados

- Medidas de dispersão
  - Desvio padrão

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Representação de dados

- Medidas de dispersão
  - Desvio absoluto médio

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Representação de dados

- Medidas de dispersão
  - Coeficiente de variação

$$CV = \frac{s}{\bar{x}}$$

Representação de dados

- Variância
  - média dos quadrados dos desvios em relação à média

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Variância corrigida

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Representação de dados

- Variância
  - média dos quadrados dos desvios em relação à média

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Variância corrigida

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Representação de dados

- Estatísticas de ordem
  - Aplicável a variáveis quantitativas ou ordinais

$$x_1 \leq x_2 \leq x_3 \cdots \leq x_n$$

Representação de dados

- Estatísticas de ordem
  - Extremos

Representação de dados

- Estatísticas de ordem
  - Mediana
    - Valor da colecção que tem 50% das observações

$$M = \begin{cases} x_{k+1} & se \quad n = 2k+1 \\ (x_k + x_{k+1})/2 & se \quad n = 2k \end{cases}$$

Representação de dados

- Estatísticas de ordem
  - Quantis
    - O quantil de ordem α ($0 < \alpha < 1$) é o valor da colecção que tem $\alpha\, n$ observações

Representação de dados

- Estatísticas de ordem
  - Quantis
    - A mediana é o quantil de ordem $\alpha = 0{,}5$

Representação de dados

- Estatísticas de ordem
  - Quantis
    - Percentis: $\alpha = 0{,}01; 0{,}02; \cdots; 0{,}99$
    - Decis: $\alpha = 0{,}1; 0{,}2; \cdots; 0{,}9$
    - Quartis: $\alpha = 0{,}25; 0{,}50; 0{,}75$

Representação de dados

- Estatísticas de ordem
  - Amplitude de variação

$$AIV = x_n - x_1$$

Representação de dados

- Estatísticas de ordem
  - Amplitude interquartil

$$AIQ = Q_3 - Q_1$$

Inferência

- Na teoria da probabilidade
  - Modelo probabilístico → probabilidade de resultados

> A probabilidade de obter um falso negativo com um teste de gravidez é de 0.01. Qual a probabilidade de em 100 mulheres grávidas testadas haver no máximo 2 testes falsos negativos?

- Na inferência estatística
  - Dados/observações → modelo

> A empresa SAIBAJÁ lançou um novo teste de gravidez. Para afeirir da percentagem de falsos negativos decidiu testar o produto em 100 mulheres grávidas.

Inferência

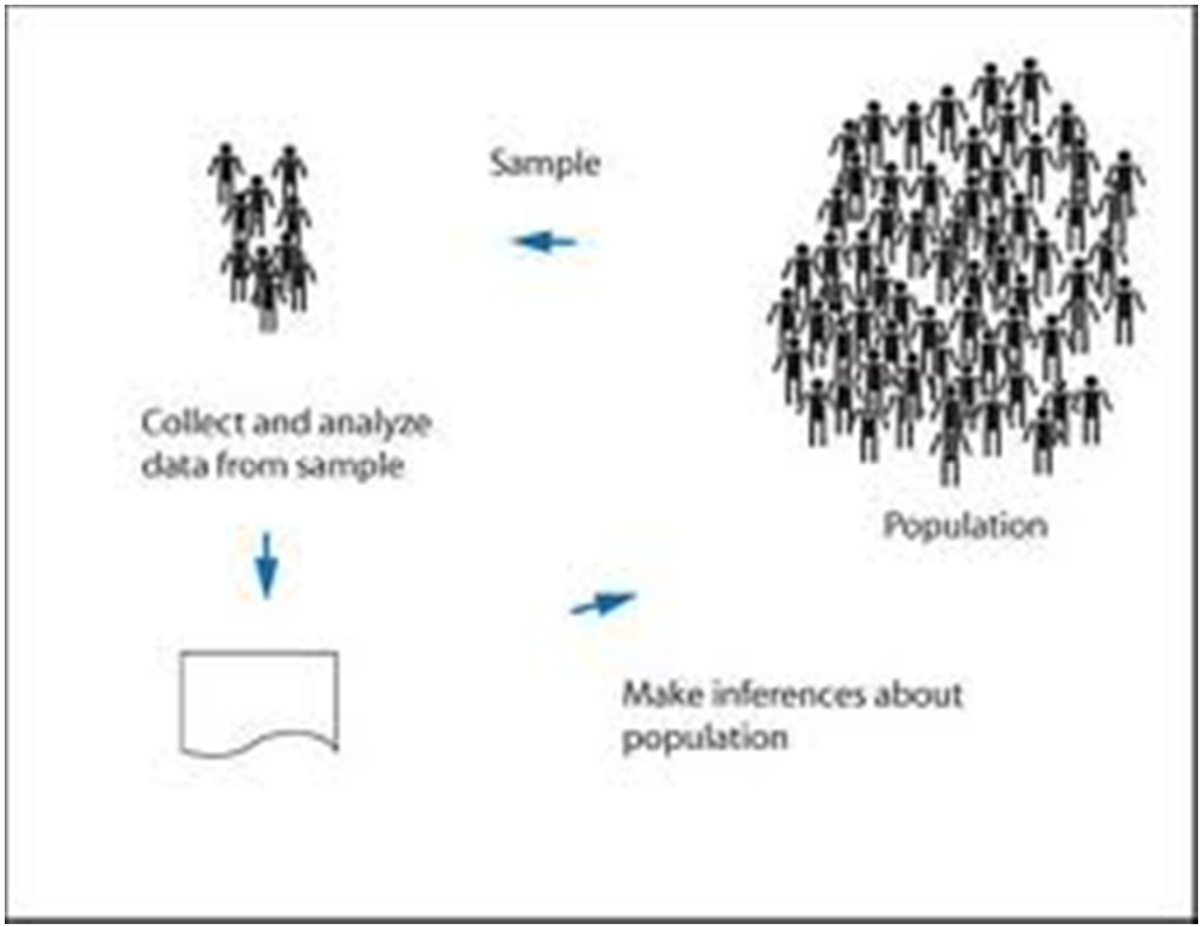
- O observador pode ter interesse em:
  - estimar a partir da observação da amostra a proporção de testes falsos negativos;
  - construir, com base na amostra, um intervalo que, com razoável confiança, contenha o valor desconhecido dessa proporção;
  - proposta a hipótese de que a proporção de testes falsos negativos é inferior a 1%, averiguar em que termos os dados suportam essa proposição;

Inferência

- A inferência é um problema central na estatística:
  - partindo de um conjunto de dados pretende-se 'determinar' as propriedades da distribuição que está na sua base.

- A inferência estatística é usualmente dividida em
  - Estimação
    - parâmetros específicos da população
  - Decisão (testes de hipóteses)
    - decisão sobre o valor de um determinado parâmetro da população;

Inferência


Sample
Collect and analyze data from sample
Population
Make inferences about population

Amostragem

- Censo
  - informação relativa a todos os elementos da população;

- Amostragem:
  - analisa-se um subconjunto da população

Amostragem

- Vantagens da amostragem
  - impossível a recolha de todos os elementos da população em:
    - populações infinitas ou com elevado nº de elementos;
    - quando o estudo das características de cada elemento conduz à sua destruição;
  - O estudo cuidadoso de uma amostra conduz a resultados mais fidedignos do que o estudo sumário de toda a população;
  - Menor custo e obtenção de resultados em tempo oportuno;
  - Problemas de ordem ética devem ser tidos em consideração:
    - estudo de novos medicamentos ou de novas técnicas cirúrgicas;
    - técnicas invasivas

Amostragem

- Amostras de conveniência
  - são, muitas vezes, as únicas possíveis de obter, principalmente quando se trata de populações raras, mal conhecidas, geograficamente mal determinadas;
  - perigo de tendenciosidade, logo inadequadas para produzir inferência;
- Amostragem aleatória, casual ou probabilística
  - é a que garante melhor representatividade;
  - é necessário possuir uma listagem de todos os elementos da população de modo a que a probabilidade de qualquer elemento da população ser seleccionado seja conhecida à priori (≠0.)
  - extremamente difícil obter tal amostragem, mas possível obter uma aproximação

Amostragem

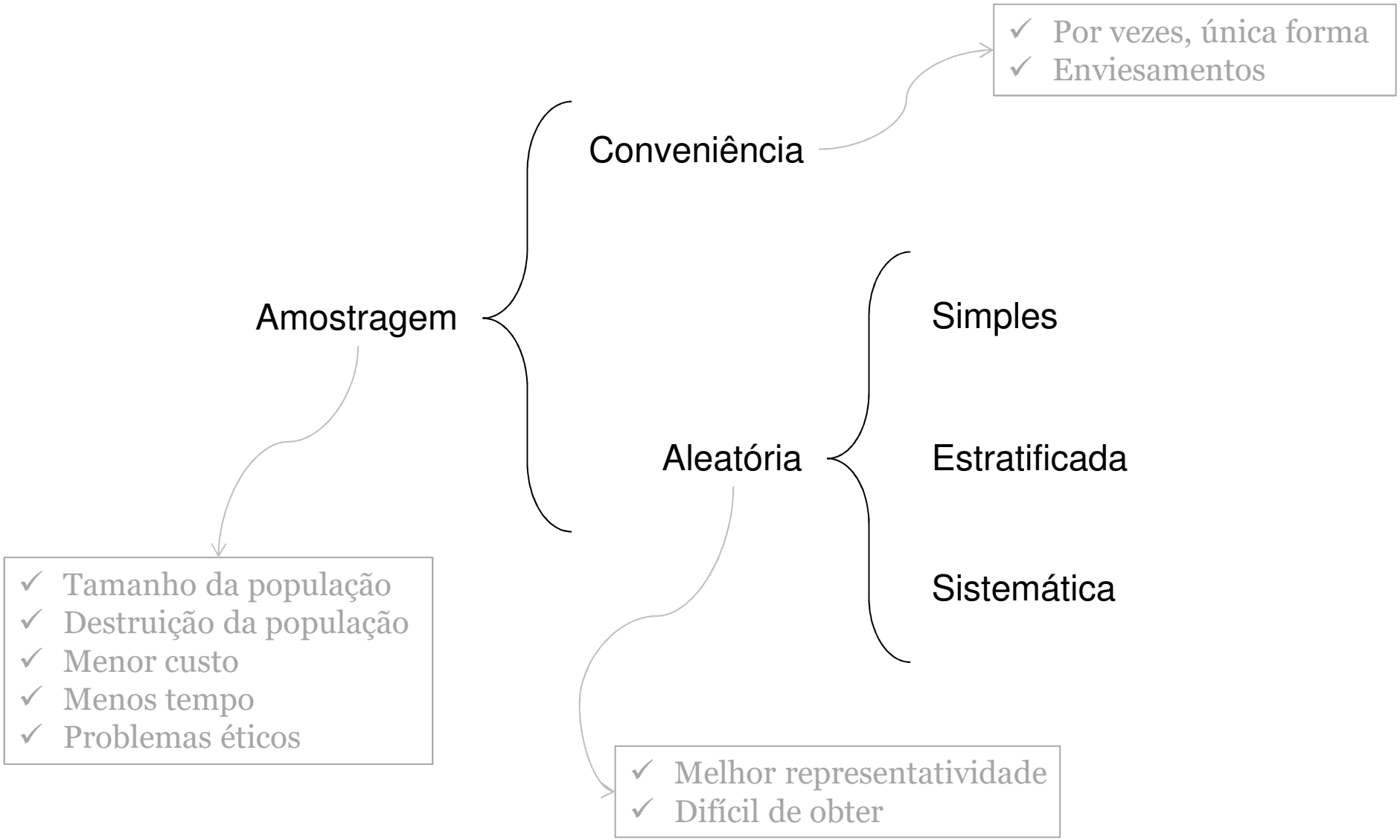
- Amostragem aleatória
  - Simples:
    - todos os elementos têm igual probabilidade de serem seleccionados (1/N) por sorteio;
    - este método não é muito usado dado que é difícil obter populações réplica;

Amostragem

- Amostragem aleatória
  - Estratificada
    - quando se conhece a estrutura da população,
    - conduz a amostras representativas de menor dimensão;
    - a população é dividida em estratos, grupos homogéneos relativamente a uma característica (ex: sexo), e dentro de cada estrato seleccionam-se os elementos duma forma aleatória simples, de acordo com a proporção de cada grupo na população;

Amostragem

- Amostragem aleatória
  - Sistemática ou quase aleatória
    - apenas o 1º elemento da amostra é escolhido aleatoriamente, e os restantes são determinados de modo sistemático pela razão N/n (N – dimensão da população; n – dimensão da amostra);
    - o 1º elemento pode ser obtido por uma tabela de nºs aleatórios no intervalo [1, N/n], e os restantes por adição de N/n (valores arredondados ao menor inteiro);


Amostragem
Conveniência
✓ Por vezes, única forma
✓ Enviesamentos
Aleatória
Simples
Estratificada
Sistemática
✓ Tamanho da população
✓ Destruição da população
✓ Menor custo
✓ Menos tempo
✓ Problemas éticos
✓ Melhor representatividade
✓ Difícil de obter

Estimação

- O objectivo da estimação é estimar parâmetros de uma população teórica a partir de estatísticas obtidas numa amostra representativa dessa população.

- Se se extraírem *n* amostras de uma população cuja função de probabilidade (densidade) depende de um parâmetro (e.g. a média: μ) do qual se desconhece o verdadeiro valor, é necessário estimá-lo, com um determinado grau de
  - precisão (estimação por pontos)
  - confiança (estimação por intervalos)

Estimação

- Um estimador natural para estimar a média , μ, de uma população é a média amostral:

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

- O melhor estimador do desvio padrão:

$$s^* = \sqrt{\frac{n-1}{n}}\, s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Estimação

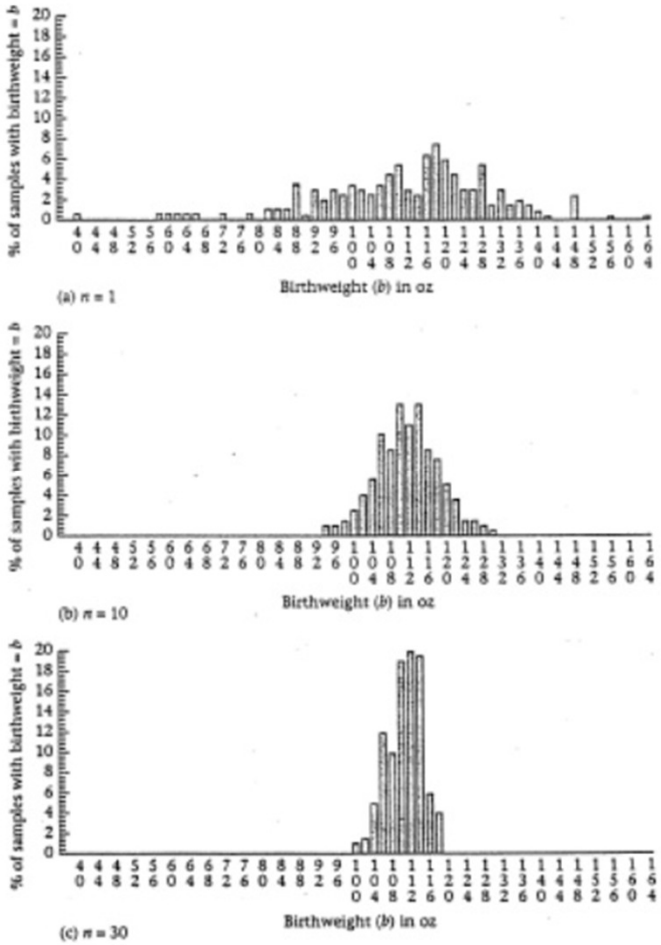
- Considere-se todas as possíveis amostras com tamanho *n* que se poderiam retirar da população;
  - As médias obtidas para cada uma dessa amostras seriam, previsivelmente, diferentes;
  - Assim, a amostra 'colhida' deve ser tomada como representativa de todas as amostras (de tamanho *n*) possíveis;

- Seja $X_1, X_2, \cdots, X_n$ uma amostra casual de população para a qual existe média μ. Então para a média amostral $\bar{X}$, $E(\bar{X}) = \mu$.

Estimação

- A média amostral é um estimador da média da população qualquer que seja o tamanho da amostra.
- Quanto maior o tamanho da amostra melhor a estimativa?

Estimação



N = 1

range ≈100

N = 10

range = 32

N = 30

range = 20

Estimação

$$Var(\bar{X}) = Var\left\{\sum_{i=1}^{n} \frac{X_i}{n}\right\} = \frac{1}{n^2}\, Var \sum_{i=1}^{n} X_i = \frac{1}{n^2}\,(n\,\sigma^2)$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Estimação

- O erro padrão da média ou simplesmente o erro padrão é dado por $\sigma/\sqrt{n}$.
- O erro padrão representa o desvio padrão estimado para um conjunto de médias amostrais de amostras de (tamanho igual a n) de uma população com variância σ^2.

Teorema do limite central

- Dada a sucessão de variáveis aleatórias independentes e identicamente distribuidas (iid), X_1, X_2, ..., X_n, com média μ e variância σ^2, então quando n $\rightarrow\infty$, a função da distribuição da variável aleatória,

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\,\mu}{\sqrt{n}\,\sigma}$$

tende para uma função de distribuição $N(0,1)$, ou seja a distribuição assimptótica ou aproximada de Z_n é $N(0,1)$.

Teorema do limite central

- Se se tomar um número muito elevado de amostras, cada uma com n observações, a distribuição da médias das amostras tende para uma distribuição normal, com média μ e variância σ^2/n.

SIMULAÇÃO