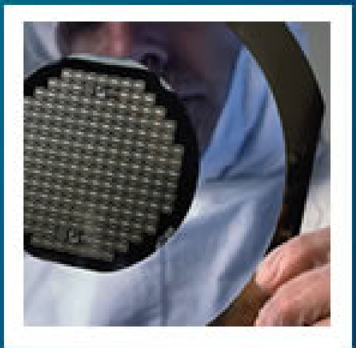
Bioestatística

Aula 5

Teoria da Amostragem e Teoria da Estimação

Prof. Tiago A. E. Ferreira



Teoria da Amostragem

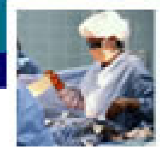
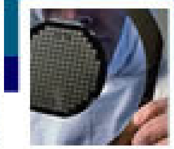
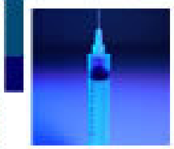
- Com já visto, desejamos uma amostra que seja representativa de uma dada população.

Amostra		**População**
Parte	Inferir	Todo
Particular		Geral

Leis gerais aplicáveis a toda população

  – Observe que pode haver diferenças entre os valores reais (população) e os valores estimados (amostra)

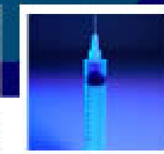
Teoria da Amostragem

- Na prática, quase sempre ocorrerá alguma discrepância entre os resultados estimados (amostra) e os resultados reais (população)
  - Se esta discrepância, ou erro, for no entanto tratável estatisticamente, ainda há possibilidade de inferirmos conclusões verídicas da população a partir da amostra.
    - O estudo das relações entre amostras e população é o objetivo da **Teoria da Amostragem.**

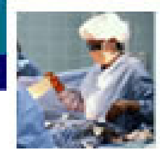
Técnicas de Extração de Amostras

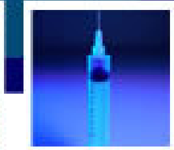




- Basicamente, as técnicas de amostragem podem ser:
  - Amostragem com Reposição (ACR)
    - Todo elemento selecionado para a amostra é devolvido ao repositório antes da seleção do próximo elemento da amostra
      - Dada uma população de tamanho **N** e amostras de tamanho **n, é** possível construir **N^n amostras.**
  - Amostragem sem Reposição (ASR)
    - Um elemento selecionado para a amostra não volta para o repositório antes do término da construção da amostra.
      - Dada uma população de tamanho **N** e amostras de tamanho **n, é** possível construir **π_n(N) amostras,** onde π_n(N) é a função fatorial incompleta, dada por: $$\pi_n(N) = N(N-1)(N-2)\cdots(N-n)$$

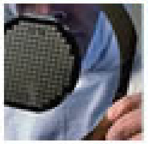
Amostragem





- Dada uma população
  - Desta população são extraídas **n** amostras
    - Os elementos destas **n** amostras podem variar, não sendo necessariamente os mesmo de uma amostra para a outra.
    - Assim, os **parâmetros amostrais** devem ser diferentes entras as **n** amostras e aos parâmetros populacionais.
    - A grande questão é:
      - "Estas diferenças seguem ou não um comportamento regular e previsível?"

Distribuição Amostral das Médias

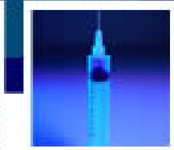

- A partir da DAM (Distribuição Amostral das Médias) é possível calcular parâmetros característicos para a distribuição
  - Como a média e a variância da DAM
    - Para uma amostragem com reposição (ou população infinita):

$$\mu_{DAM} = \mu \qquad \sigma^2_{DAM} = \frac{\sigma^2}{n}$$

  - Ou seja, a média da DAM é igual a média populacional e a variância da DAM é a variância populacional dividida pelo tamanho da amostra

Exemplo





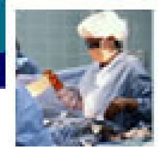
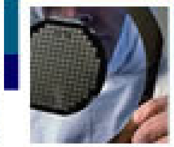
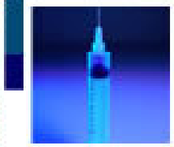
- Dado um população hipotética formada por 3 indivíduos
  - Cujas taxas de glicose sejam 80, 100, 120 (mg/100ml)
  - Assim, a média das taxas de glicose para a população é:

$$\bar{x} = \frac{80+100+120}{3} = 100(mg/100ml)$$

  - E a variância populacional é:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = \frac{(80-100)^2 + (100-100)^2 + (120-100)^2}{3} = 266{,}67(mg/100ml)$$

Exemplo

- Calculando todas as possíveis amostras de 2 elementos com reposição

Amostra	Indivíduos	Médias Amostrais
1	(80,80)	80
2	(80,100)	90
3	(80,120)	100
4	(100,80)	90
5	(100,100)	100
6	(100, 120)	110
7	(120,80)	100
8	(120,100)	110
9	(120,120)	120

Calculando obtemos,

$$\mu_{DAM} = \frac{80+90+\cdots+120}{9} = 100(mg/100ml)$$

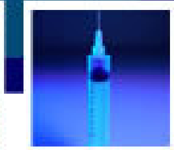
$$\sigma^2_{DAM} = \frac{(80-100)^2+\cdots+(120-100)^2}{9} = 133{,}33(mg/100ml)$$

$$\frac{\sigma^2}{2} = \frac{266{,}67}{2} \cong 133{,}33(mg/100ml)$$

$$N[\mu,\sigma] \therefore DAM \rightarrow N\left[\mu, \frac{\sigma}{\sqrt{n}}\right]$$

Distribuição Amostral das Médias

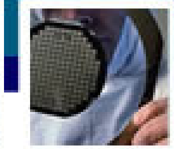
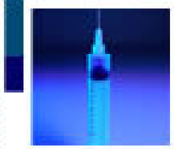




- Assim,
  - Se a população tem distribuição normal, caracterizada por uma média μ e uma desvio padrão σ
  - Então, as médias amostrais terão também uma distribuição normal com média μ e uma desvio padrão $\sigma/\sqrt{n}$
    - Gerando uma v.a. padronizada da média das médias amostrais, gera-se uma distribuição t-student:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

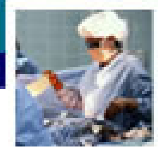
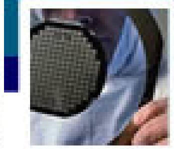
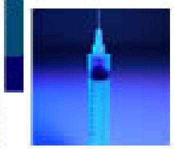
Distribuição Amostral das Médias

- Para o caso de amostragem sem reposição:

$$\mu_{DAM} = \mu \qquad \sigma^2_{DAM} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

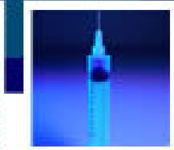
Distribuição Amostral das Proporções

- Também é possível trabalhar com variáveis não numéricas
  - Como, por exemplo, "percentual de fumantes" e "não fumantes", ou, sexo "Feminino" e "Maculino", ou ...
    - A distribuição estatística das proporções de uma v.a. extraídas de todas as amostras de tamanho **n** de uma população de tamanho **N** é denominada de **Distribuição Amostral das Proporções (DAP).**

Distribuição Amostral das Proporções





- Para uma população infinita (com reposição) com distribuição binomial,

$$\underset{N \to \infty}{P} \to B\left[p, q = 1 - p\right]$$

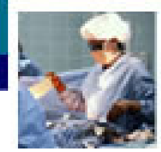
  – Os parâmetros da DAP são:

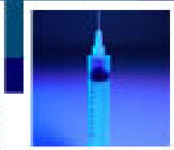
$$\mu_{DAP} = p \qquad \sigma^2_{DAP} = \frac{pq}{n}$$

  – Para população finita (sem reposição)

$$\mu_{DAP} = p \qquad \sigma^2_{DAP} = \frac{pq}{n}\left(\frac{N - n}{N - 1}\right)$$

Distribuição Amostral das Diferenças ou das Somas

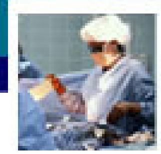


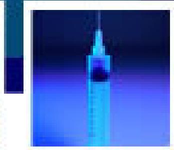


- Dadas duas populações A e B
  - Suponha que sejam calculadas as diferenças entre as médias de todas as combinações possíveis de amostras de tamanho $\mathbf{n_A}$ e $\mathbf{n_B}$ das por. A e B
    - Então tem-se uma **Distribuição Amostral das Diferenças (DAD)** entre as médias, ou analogamente, uma **Distribuição Amostral das Somas (DAS)** entre as médias, com médias:

$$\mu_{DAD} = \mu_A - \mu_B \qquad \mu_{DAS} = \mu_A + \mu_B$$

Distribuição Amostral das Diferenças ou das Somas





- A variância tanto para a DAD quanto para a DAS é:

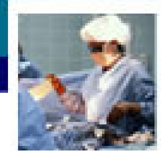
$$\sigma^2_{DADS} = \frac{\sigma^2_A}{n_A} + \frac{\sigma^2_B}{n_B}$$

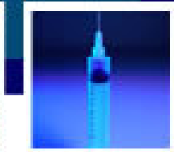
- A ideia da Distribuição Amostral das Diferenças ou das Somas pode ser aplicada a qualquer parâmetro ou estatística

Teoria da Estimação

- Estimativa
  - De modo geral, uma estimativa é uma tentativa de avaliar algum dado desconhecido
  - Desta forma, uma estimativa sempre terá duas características:
    - A estimativa é uma aproximação do dado verdadeiro
    - A estimativa sempre contém um componente de incerteza.

Estimativa Pontual





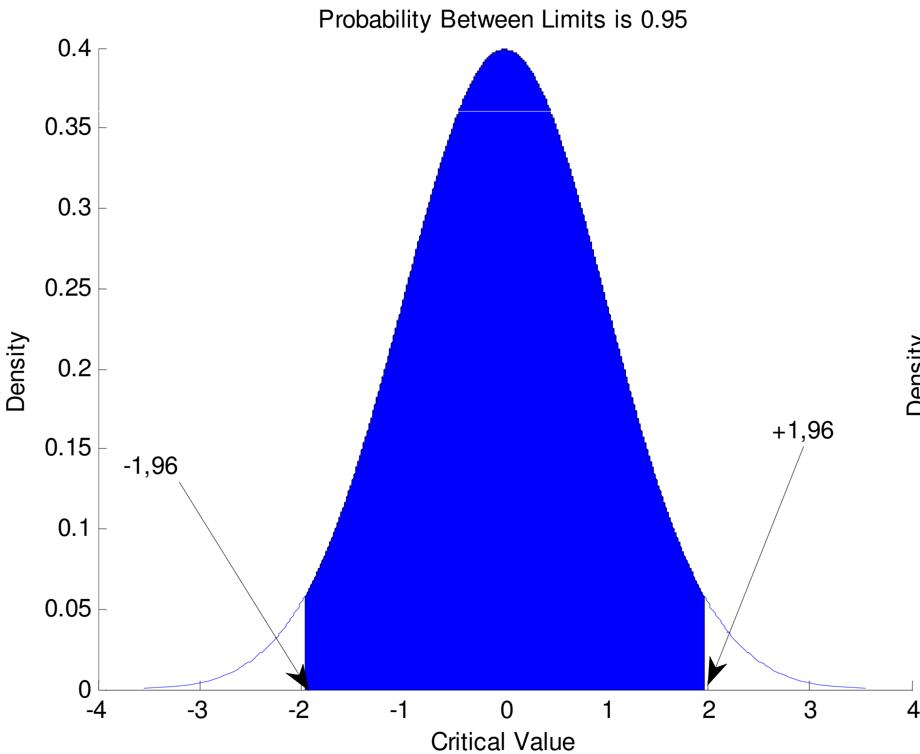
- Um estimativa é dita ser pontual se esta se baseia em apenas um único ponto.
  - Exemplo:
    - "A média da taxa de glicose de diabéticos foi estimada em 200 mg/100ml"
    - Este é uma estimativa pontual pois considera possível apenas um valor único para a média de todos os diabéticos.
      - Embora esta estimativa seja muito precisa, suas chances de serem verdadeiras são quase nulas!

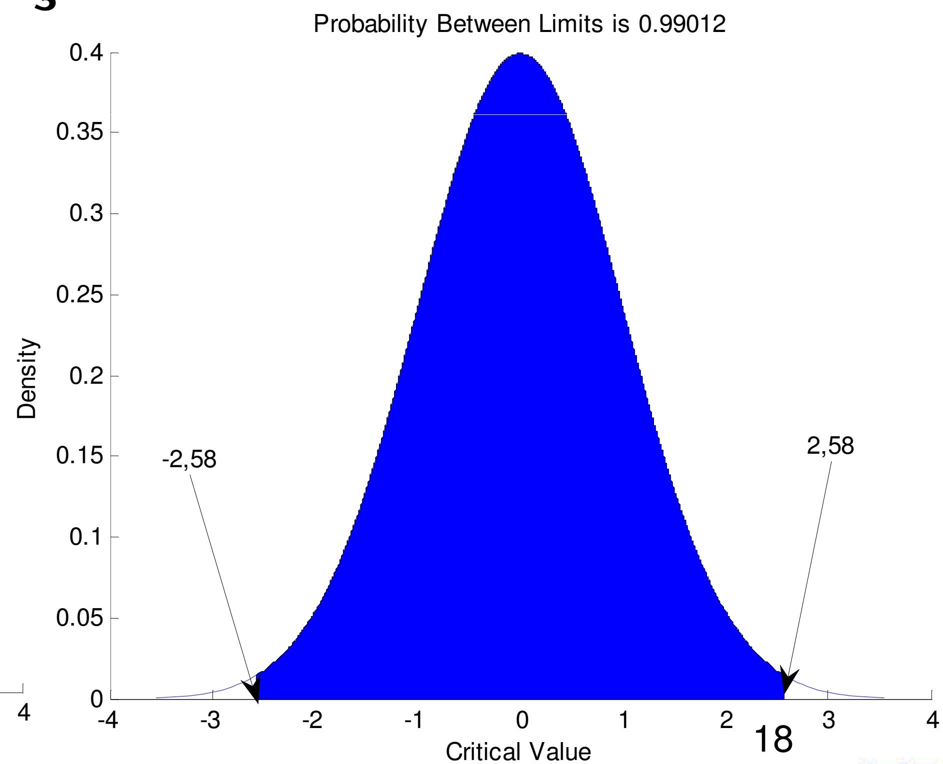
Estimativa por Intervalo

- Uma estimativa mais factível de ocorrer pode ser descrita por um intervalo.
  - Este procedimento acarreta em um **Intervalo de confiança.**
  - Por exemplo:
    - A taxa média de glicose em diabéticos está entre 180 e 220 mg/100ml, em um nível de confiança de 90%.

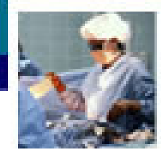
Intervalo de Confiança para Média Populacional

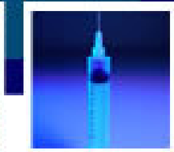
- A ideia do intervalo de confiança é que ele englobe uma certa quantidade de observações com uma dada confiança ou certeza.





Intervalo de Confiança para Média Populacional



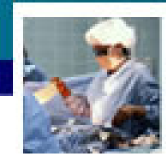


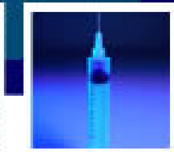
- O Intervalo de Confiança associado a um grau de certeza (GC) para a média populacional, quando o desvio padrão populacional σ é conhecido, é dado por:

$$IC_{GC}(\mu) \rightarrow \bar{x} \mp z_{GC} \cdot \frac{\sigma}{\sqrt{n}}$$

  – Esta expressão depende:
    - Do grau de Confiança $\mathbf{z_{GC}}$
    - Do Desvio Padrão Populacional $\boldsymbol{\sigma}$
    - Do tamanho da Amostra **n**

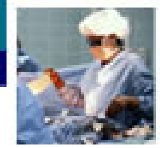
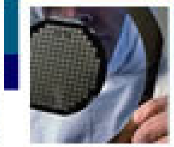
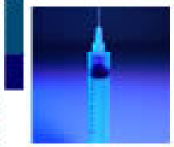
Intervalo de Confiança para Média Populacional





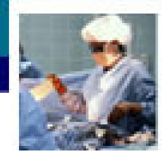
- Observe que:
  - O termo $\sigma/\sqrt{n}$ é **erro padrão** da média
  - Grau de confiança muito elevado implica em uma grande amplitude do intervalo de confiança
    - Deseja-se que a estimativa precisa, ou seja, alto grau de confiança (confiabilidade) e pequeno intervalo de confiança
      - Entretanto estas características são concorrentes

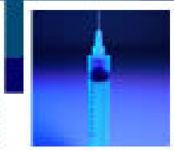
Tamanho da Amostra

- É possível determinar o tamanho de uma amostra para que esta satisfaça um certo intervalo de confiança.
  - Foi visto que a média populacional pode ser estimada a partir das médias amostrais com um erro de $z_{GC}\sigma/\sqrt{n}$
  - Assim é possível fixar uma margem máxima de erro, ou uma proporção ε

Tamanho da Amostra





- Assim,

$$\frac{z_{GC} \cdot \sigma}{\sqrt{n}} \leq \varepsilon \cdot \bar{x}$$

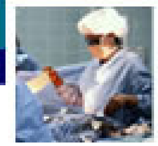
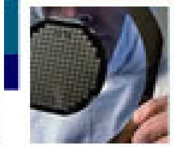
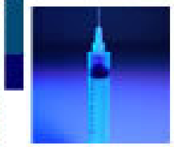
$$ou$$

$$n \geq \left(\frac{z_{GC} \cdot \sigma}{\varepsilon \cdot \bar{x}} \right)^2$$

Exemplo

- Deseja-se estimar o diâmetro pupilar de coelhos adultos normais, a partir de uma amostra de 12 animais, cuja média foi 5,2 mm e considerando que o desvio padrão é 1,2 mm. Empregando o grau de confiança de 95% (z_{GC} = 1,96)
  – IC95% → 5,2 (mm) ± 12(mm)/$\sqrt{12}$
  – IC95% → 5,2 (mm) ± 0,68 (mm)
    - Ou seja, pode-se ter uma confiança de 95% de que a média pupilar esteja entre 4,52 mm e 5,88 mm

Desvio Padrão Populacional Estimado

- Na realidade, na grande maioria dos casos, também não conhecemos o desvio padrão populacional
  - Mais podemos estimá-lo através do desvio padrão amostral

$$s = \sqrt{\frac{\sum (x - \bar{x})}{n-1}}$$

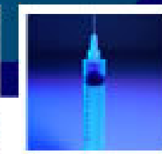
  - Assim

$$IC_{GC}(\mu) \rightarrow \bar{x} \mp t_{GC,gl} \cdot \frac{s}{\sqrt{n}}$$

    - Onde agora é utilizada a distribuição t-student com *gl graus de liberdade*

Exemplo



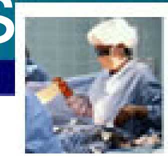


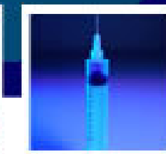
- Para o exemplo anterior da medida do diâmetro pupilar do coelhos,
  - x=5,2 mm
  - s=0,69 mm
  - Assim, $t_{95\%,11}$=2,20, logo:

$$IC_{95\%}(\mu) \to 5{,}2 \mp 2{,}10 \cdot \frac{0{,}69}{\sqrt{12}} = 5{,}2 \mp 0{,}44$$

$$IC_{95\%}(\mu) \to 4{,}76(mm) \le \mu \le 5{,}64(mm)$$

Intervalo de Confiança para a Diferença de Médias Populacionais



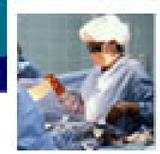


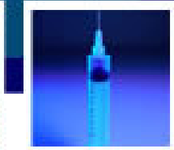
- Dada duas populações A e B
  - O Intervalo de Confiança para a diferença das médias, para um certo GC, é dado por

$$IC_{GC}(\mu_A - \mu_B) \rightarrow (\bar{x}_A - \bar{x}_B) \mp t_{GC,gl} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

$$sendo \quad gl = n_A + n_B - 2$$

Exemplo





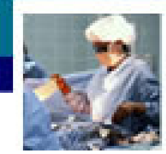
- Suponha que medimos o diâmetro pupilar de dois grupos de coelhos, um grupo normal e o outro após estímulo doloroso.
  - Grupo 1: Normal

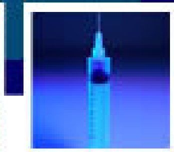
Coelho	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DP(mm)	5,0	5,5	5,0	4,5	4,5	6,0	6,5	5,5	5,5	5,0	5,5	4,0

  - Grupo 2: tratamento doloroso

Coelho	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DP(mm)	9,0	10,0	8,0	8,0	8,0	9,0	8,5	10,0	8,0	10,5

Exemplo





- Para o grupo 1:
  - Média = 5,2 mm
  - S = 0,69 mm
- Para o grupo 2:
  - Média = 8,9 mm
  - S = 0,97 mm
- Pergunta-se:
  - Qual a alteração pupilar de um coelho ao receber um estímulo doloroso?

Exemplo

- Com confiabilidade de 95%:
  - Onde *gl=10+12-2=20* e $t_{95\%,20} = 2{,}09$

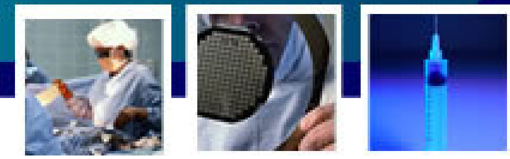
$$IC_{95\%}(\mu_A - \mu_B) \rightarrow (8{,}9 - 5{,}2) \mp 2{,}09 \cdot \sqrt{\frac{0{,}97^2}{10} + \frac{0{,}69^2}{12}}$$

$$IC_{95\%}(\mu_A - \mu_B) \rightarrow 3{,}7(mm) \mp 0{,}76(mm)$$

$$IC_{95\%}(\mu_A - \mu_B) \rightarrow 2{,}94(mm) \le (\mu_A - \mu_B) \le 4{,}46(mm)$$

- *Assim, é possível afirmar, com um nível de confiança de 95%, que o estímulo doloroso aumenta o diâmetro pupilar de 2,94 a 4,46 mm*

Amostras Pareadas



- Amostras pareadas ou pares de amostras são dados referentes a um mesmo conjunto de indivíduos, tomados em situações distintas
  - Geralmente, estas duas situações são denominadas *antes* e *depois*.
    - De forma geral, deseja-se sabe se estas duas amostras são iguais ou diferentes!

Amostras Pareadas

- Dada uma situação onde foram medidas as variáveis:

ANTES	DEPOIS
$X_{A,1}$	$X_{B,1}$
...	...
$X_{A,n}$	$X_{B,n}$

  – Assim, o intervalo de confiança é:

$$IC_{GC}\left(\Delta_{A-B}\right) \rightarrow \bar{d}_{A-B} \mp t_{GC,gl} \frac{s_d}{\sqrt{n}}$$

$$onde,$$

$$d_i = x_{A,i} - x_{B,i} \qquad \bar{d}_{A-B} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)}$$

Exemplo

- Considere 10 pacientes febris que foram tratados com um antitérmico.

Paciente	1	2	3	4	5	6	7	8	9	10
Temp. (ºC) Antes	38,0	39,5	38,0	38,7	37,5	39,0	38,0	37,8	36,7	39,0
Temp. (ºC) Depois	36,5	37,8	37,0	37,5	36,0	38,5	37,0	36,5	36,0	37,5

– Quais as diferenças?

- $d_1 = x_{A,1} - x_{B,1} = 38{,}0 - 36{,}5 = 1{,}5$ ºC
- E assim por diante...
  - – 1,7ºC; 1,0 ºC; 1,2ºC; 1,5ºC; 0,5ºC; 1,0ºC; 1,3ºC; 0,7ºC; 1,5ºC

Exemplo

- As médias das diferenças "antes-depois":

$$\bar{d}_{A-B} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1{,}5 + 1{,}7 + \cdots + 1{,}5}{10} = 1{,}19^{o}\,C$$

- O Desvio Padrão para as diferenças

$$s_i = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)} = \sqrt{\frac{(1{,}5-1{,}19)^2 + \cdots + (1{,}5-1{,}19)^2}{10-1}} = 0{,}39^{o}\,C$$

- O IC com GC=95%

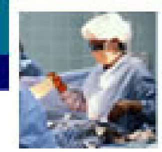
$$IC_{95\%}\left(\Delta_{A-B}\right) \to 1{,}19 \mp 2{,}26 \cdot \frac{0{,}39}{\sqrt{10}} = 1{,}19^{o}\,C \mp 0{,}28^{o}\,C$$

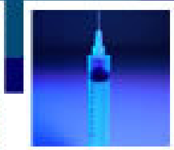
Intervalo de Confiança para Proporções Populacionais

- Quanto temos alguma proporção binária em uma população, caímos em uma distribuição binomial, com probabilidade *p* e *q*.
  - Assim, dada um proporção π, o intervalo de confiança será:

$$IC_{GC}(\pi) \rightarrow p \mp t_{GC.gl} \cdot \sqrt{\frac{p \cdot q}{n}}$$

Exemplo





- Dado um certo levantamento sobre alcoolismo mostrou que, de 30 entrevistados, 18 afirmaram ingerir bebidas alcoólicas com frequência. Qual a estimativa, com GC=95%, da proporção de indivíduos que bebem habitualmente?
  - *p=18/30=0,6; q=1-p=0,4; gl=30-1=29;* $t_{95\%,29} = 2,05$

$$IC_{95\%}(\pi) \rightarrow 0,6 \mp 2,05 \cdot \sqrt{\frac{0,6 \cdot 0,4}{30}} = 0,6 \mp 0,1833$$

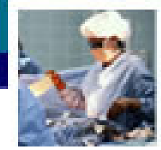
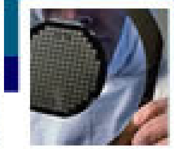
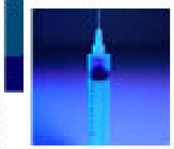
$$IC_{95\%}(\pi) \rightarrow 41,65\% \leq \pi \leq 78,33\%$$

Intervalo de Confiança para Diferenças de Proporções Populacionais

- Da mesma forma que temos diferenças para médias, é possível termos diferenças para proporções populacionais para as populações A e B (π_A - π_B).
  - Assim,

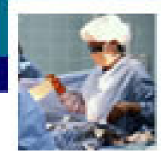
$$IC_{GC}(\pi_A - \pi_B) \rightarrow (p_A - p_B) \mp t_{GC,gl} \cdot \sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}$$

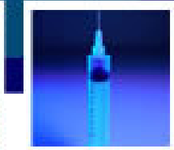
Comparando duas Grandezas

- Ao montarmos os IC para Diferenças, **com um GC**, é possível responder se duas grandezas são estatisticamente iguais ou diferente!
  - Se o IC contiver o valor "Zero", então não é possível distinguir diferença entre as grandezas
  - Caso contrário, as duas grandezas são distintas!

Referências





- Livro Texto: Bioestatística. Teoriaia e Computacional (Héctor G. Arango). Guanaba
  - Capítulo 7
- Leituras Complementares:
  - Livro: Bioestatística. Princípios e Aplicações (Sidia M. Callegari-Jaqcques)
    - Capítulo 5, 8 e 9.
  - Livro: Estatística Básica (Wilton de O. Bussab e Pedro A. Morettin). Saraiva
    - Capítulo: 11
  - Livro: Probabilidade. Aplicações à Estatística (Paul L. Meyer). LTC
    - Capítulos 14