
FIAP
MBA

FIAP MBA+

MBA em DATA SCIENCE & ARTIFICIAL INTELLIGENCE

STATISTICS WITH R

profregina.bernal@fiap.com.br
reginabernal@terra.com.br

Dra. Regina Tomie Ivata Bernal

Cientista de Dados na área da Saúde

Formação Acadêmica:

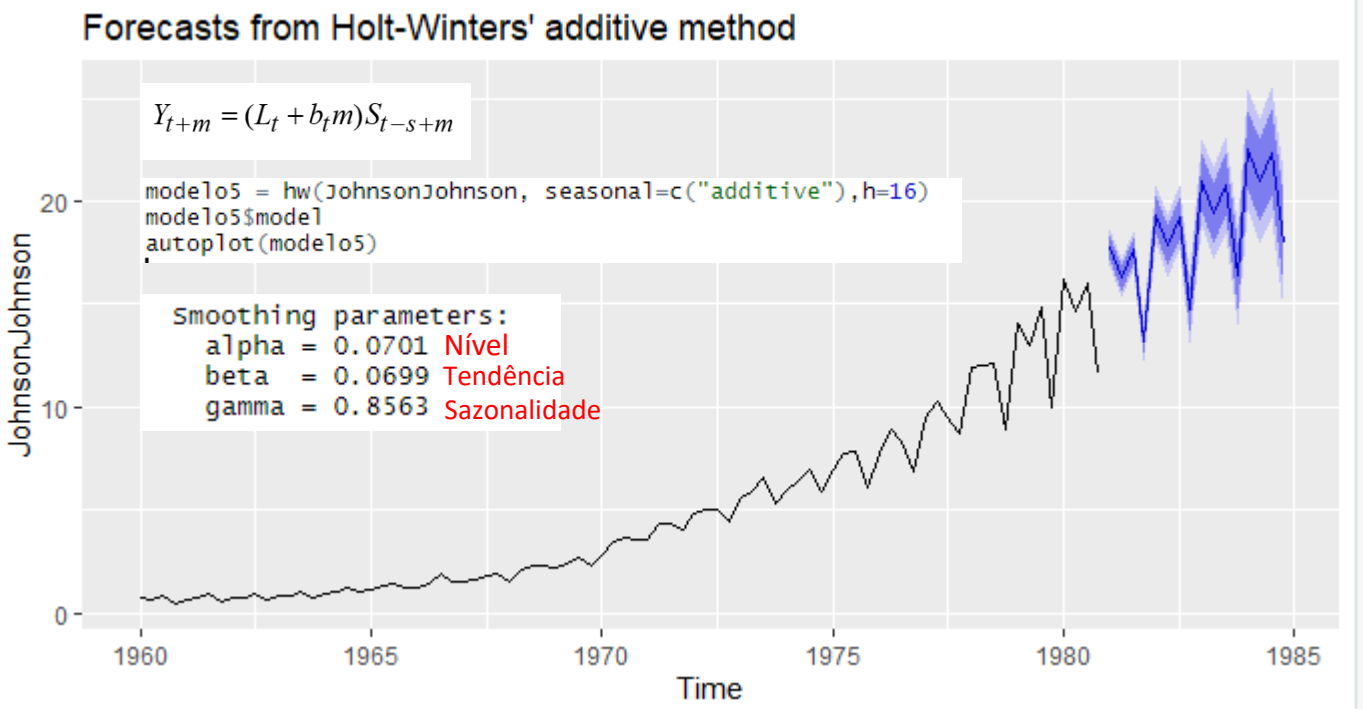
- Estatístico - UFSCar
- Mestre em Saúde Pública – FSP/USP
- Doutor em Ciências – Epidemiologia - FSP/USP

Atividades Profissionais:

- Professora de pós-graduação na FIAP
- Consultora externa da SVS/MS
- Cientista de Dados em Saúde

EXEMPLO

Exemplo: Quarterly Earnings per Johnson & Johnson Share

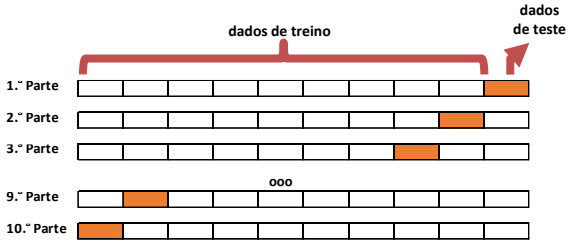


EXEMPLO

APLICAÇÃO DA **AMOSTRAGEM**

Validação Cruzada

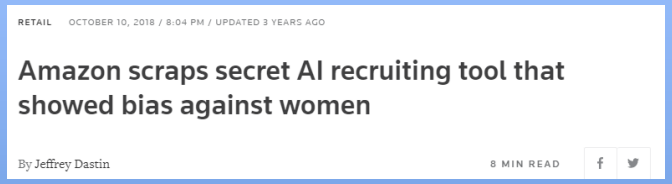
- Dividir os dados em partes iguais e utilizar:
  - uma fração delas para treinar o algoritmo com um hiperparâmetro;
  - outra parte testar a sua predição



Seleção do hiperparâmetro com melhor performance ➔ definição do algoritmo com esse hiperparâmetro nos dados de treino.
Fazer o mesmo para todos os algoritmos.
A única forma de saber qual o algoritmo de melhor performance é testando todos.

EXEMPLO

APLICAÇÃO DE IA



RETAIL OCTOBER 10, 2018 / 8:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

...

That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

Fonte: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

EXEMPLO

APLICAÇÃO DE MACHINE LEARNING

Ethical Implications Of Bias In Machine Learning

Adrienne Yapo
Bentley University
175 Forest St.
Waltham, MA 02452
adrienne.yapo@gmail.com

Joseph Weiss
Bentley University
175 Forest St.
Waltham, MA 02452
jweiss@bentley.edu

1.2. Machine learning algorithm bias

Although machine learning algorithms can produce numerous benefits to individuals, consumers, businesses, investors, the government, and society at large, recent research has uncovered many instances of bias in machine learning algorithms that have troubling implications and deleterious consequences.

1.3. Machine learning in the criminal justice system

Yet perhaps the most troubling incidents of bias in machine learning to date are unfolding in the criminal justice system. Consider the following statement from then U.S. Attorney General Eric Holder on the Sentencing Reform and Corrections Act of 2015:

Table 1: Disproportionate incarceration rates

Source: ***Propublica*** **analysis from Broward County, Florida**

Prediction Fails Differently for Black Defendants

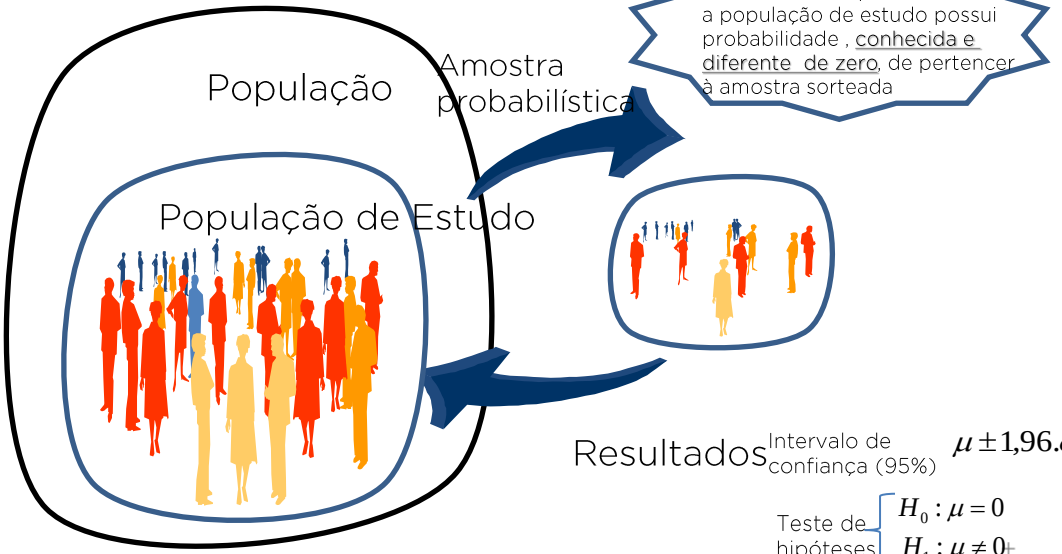
	White	African American
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

"Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes."

Fonte: https://scholarspace.manoa.hawaii.edu/bitstream/10125/50557/paper0670.pdf

INFERÊNCIA ESTATÍSTICA

INFERÊNCIA **ESTATÍSTICA**



Resultados

Intervalo de confiança (95%) $\mu \pm 1{,}96.\sigma_{\bar{x}}$

Teste de hipóteses $\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$

INFERÊNCIA **ESTATÍSTICA**

- Amostragem

O que é necessário garantir?

- Que a amostra seja representativa da população

A amostra deve possuir as mesmas características básicas da população, no que diz respeito às variáveis que desejamos pesquisar.

INFERÊNCIA **ESTATÍSTICA**

- Amostragem: Conceitos

População

É um conjunto de indivíduos, objetos, ou elementos que possuem pelo menos uma característica em comum.

Característica em comum é o atributo usado como critério de reconhecimento ou de inclusão.

Parâmetros: μ (média) , σ (desvio padrão), P (proporção)

Amostra

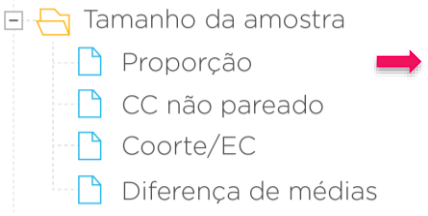
É um subconjunto da população.

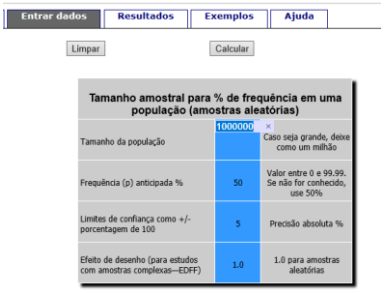
Estimativas: $\bar{x}$ (média) , s (desvio padrão),
p (proporção)

Unidade amostral é um objeto (pessoa ou coisas) sobre o qual coletaremos as medidas.

Cálculo do Tamanho da Amostra

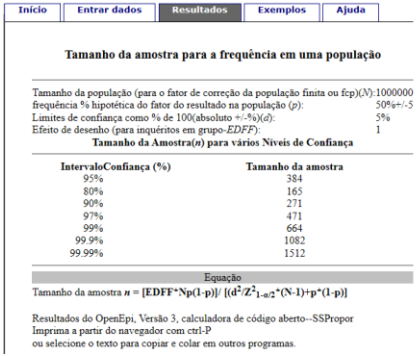
Calculadora estatística: http://www.openepi.com/SampleSize/SSPropor.htm





Cálculo do Tamanho da Amostra

Calculadora estatística: http://www.openepi.com/SampleSize/SSPropor.htm



Início | Entrar dados | Resultados | Exemplos | Ajuda

Tamanho da amostra para a frequência em uma população

Tamanho da população (para o fator de correção da população finita ou fcp)(*N*): 1000000
frequência % hipotética do fator do resultado na população (*p*): 50%+/-5
Limites de confiança como % de 100(absoluto +/-%)(*d*): 5%
Efeito de desenho (para inquéritos em grupo-*EDFF*): 1

Tamanho da Amostra(*n*) para vários Níveis de Confiança

IntervaloConfiança (%)	Tamanho da amostra
95%	384
80%	165
90%	271
97%	471
99%	664
99.9%	1082
99.99%	1512

Equação

Tamanho da amostra $n = [EDFF*Np(1-p)]/ [(d^2/Z^2_{1-\alpha/2}*(N-1)+p*(1-p)]$

Resultados do OpenEpi, Versão 3, calculadora de código aberto--SSPropor
Imprima a partir do navegador com ctrl-P
ou selecione o texto para copiar e colar em outros programas.

INFERÊNCIA **ESTATÍSTICA**

Tipos de amostragem

- ✓ PROBABILÍSTICA
  - ✓ ALEATÓRIA SIMPLES
  - ✓ SISTEMÁTICA
  - ✓ ESTRATIFICADA
  - ✓ CONGLOMERADO
- ✓ NÃO PROBABILÍSTICA (INTENCIONAL)
  - ✓ COTAS
  - ✓ PROCURA
  - ✓ INTERNET PANELS (RIVER SAMPLING, RESPONDENT-DRIVEN SAMPLING)

Procedimento de Sorteio de Amostra

✓ Aleatória Simples

Cadastro de clientes da Empresa XPTO

ID	tel_mov	tel_fixo	sexo	idade	anos_estudo
1	1	1	2	63	1
2	1	1	2	28	12
3	1	1	2	75	12
4	1	1	2	34	12
5	1	1	4	26	9
6	1	1	2	25	11
7	0	1	2	30	16
8	0	0	4	75	5
9	0	0	2	40	12
10	1	1	2	58	16
11	1	1	4	39	1
12	1	1	2	46	12
13	1	1	4	46	16
14	1	1	2	22	14
15	1	1	2	20	13
16	1	1	4	18	12
17	0	0	4	76	5
18	0	0	2	34	1
19	1	1	2	50	16
20	1	1	4	44	16
21	1	1	2	19	13
22	1	1	4	18	12
23	0	1	2	89	12
24	0	1	4	87	12
25	0	1	2	76	12
26	0	1	4	74	6

Sorteio de uma amostra aleatória simples composta de cinco clientes

- n = 5
- gerar cinco números aleatórios entre 1 a N

Usando o Excel:

Função geradora de número aleatório:

= aleatórioentre(1;N)

Resultado

Contador	Sorteio
1	15
2	25
3	4
4	9
5	8

Procedimento de Sorteio de Amostra

✓ Sistemática

Cadastro de clientes da Empresa XPTO

ID	tel_mov	tel_fixo	sexo	idade	anos_estudo	Contador
1	1	1	2	63	1	1
2	1	1	2	28	12	2
3	1	1	2	75	12	3
4	1	1	2	34	12	4
6	1	1	2	25	11	5
7	0	1	2	30	16	6
9	0	0	2	40	12	7
10	1	1	2	58	16	8
12	1	1	2	46	12	9
14	1	1	2	22	14	10
15	1	1	2	20	13	11
18	0	0	2	34	1	12
19	1	1	2	50	16	13
21	1	1	2	19	13	14
23	0	1	2	89	12	15
25	0	1	2	76	12	16
5	1	1	4	26	9	17
8	0	0	4	75	5	18
11	1	1	4	39	1	19
13	1	1	4	46	16	20
16	1	1	4	18	12	21
17	0	0	4	76	5	22
20	1	1	4	44	16	23
22	1	1	4	18	12	24
24	0	1	4	87	12	25
26	0	1	4	74	6	26

Sorteio de uma amostra sistemática composta de cinco clientes (n=5)

Usando o Excel:

1) Arquivo ordenado por Sexo
2) Intervalo=N/n = 26/5 = 5,2
3) Sorteio do inicio casual : gerar um número aleatório entre 1 e 5,2 (intervalo)
   =aleatório()*(5.2 – 1) + 1

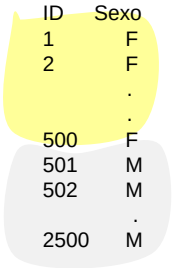
obs: copie e cole especial (valor)

Contador	Fórmula	Sorteio
1	A1 = IC	3.315336
2	A2 = A1+INT	8.515336
3	A3 = A2+INT	13.71534
4	A4=A3+INT	18.91534
5	A5=A4+INT	24.11534

Procedimento de Sorteio de Amostra

✓ Exemplo de Amostra Estratificada

Uma empresa emprega 2000 engenheiros do sexo masculino e 500 do sexo feminino. O departamento de RH deseja pesquisar opiniões desses profissionais sobre o sistema de avaliação de desempenho da empresa. Para dar uma atenção adequada à opinião feminina, a amostra será composta por 200 engenheiros e 200 engenheiras.

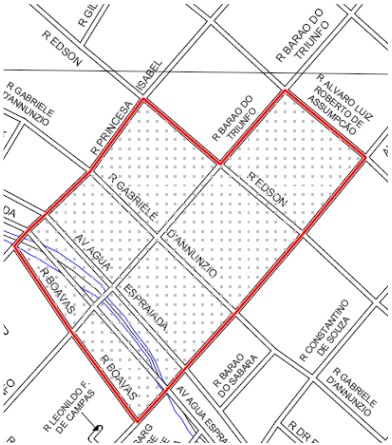


Estrato 1 — Amostra aleatória simples de tamanho 200

Estrato 2 — Amostra aleatória simples de tamanho 200

Procedimento de Sorteio de Amostra

✓ Exemplo de Amostra por Conglomerado



Setor censitário

↓

± 300 domicílios

↓

Indivíduos

Procedimento de Sorteio de Amostra

- ✓ Usos

Procedimento de Sorteio da Amostra	Pesos amostrais(*)
Aleatória Simples	Não
Sistemática	Não
Estratificada	Sim
Conglomerado	Sim

(*) análise estatística

- ✓ Amostra por Cotas
  - pesos amostrais

OUTROS TIPOS DE **AMOSTRAGEM**

EXEMPLO

ConVid Pesquisa de Comportamentos

Objetivos

Geral

- Descrever as mudanças nos estilos de vida, nas atividades de rotina, na situação de trabalho, e nos cuidados à saúde, e avaliar o estado de ânimo dos brasileiros no período de isolamento social/quarentena consequente à pandemia de coronavírus.

Fonte: https://convid.fiocruz.br/

OUTROS TIPOS DE **AMOSTRAGEM**

EXEMPLO

Amostra

O convite aos participantes é feito por um procedimento de amostragem em cadeia. Na primeira etapa, os pesquisadores do estudo escolheram um total de 200 outros pesquisadores de diferentes estados do Brasil. Adicionalmente, cada pesquisador do estudo escolheu 20 pessoas da sua rede social, totalizando 400 pessoas selecionadas para participar. As pessoas escolhidas na primeira etapa são chamadas de influenciadores ou sementes.

Com vistas a conseguir diversidade na rede virtual, os influenciadores enviam o link da pesquisa para pelo menos 12 pessoas das suas redes sociais, obedecendo a uma estratificação por sexo, faixa de idade (18-39; 40-59; 60+) e grau de escolaridade (ensino médio incompleto ou menos; ensino médio completo ou mais), isto é, são convidadas pelo menos 3 pessoas em cada uma das categorias formada pela combinação das três categorias:

Fonte: https://convid.fiocruz.br/

OUTROS TIPOS DE **AMOSTRAGEM**

EXEMPLO

A essas pessoas, é solicitado que elas convidem outras três pessoas de suas redes sociais e assim por diante, compondo a rede ConVid.

Além disso, as informações sobre o estudo serão divulgadas por meio de comunicados à imprensa, comunicação social das instituições de pesquisas participantes, secretarias estaduais de saúde, e mídias sociais.

Adicionalmente, o link da pesquisa ficará disponível nas instituições de pesquisa dos influenciadores e os funcionários das instituições poderão responder ao questionário.

Fonte: https://convid.fiocruz.br/

OUTROS TIPOS DE **AMOSTRAGEM**

EXEMPLO

Sexo	Faixa de idade	Grau de escolaridade	No. de convites
Masculino / Feminino	18 - 39	Médio incompleto ou menos	3 ou mais
		Médio completo ou menos	
	40 - 59	Médio incompleto ou menos	
		Médio completo ou menos	
	60 ou mais	Médio incompleto ou menos	
		Médio completo ou menos	

Fonte: https://convid.fiocruz.br/

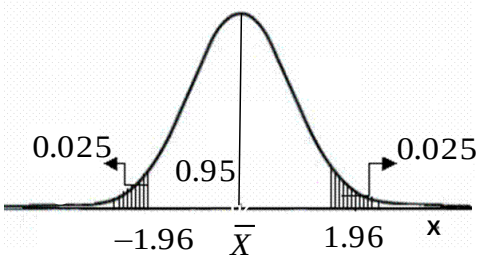
INFERÊNCIA **ESTATÍSTICA**

Usos de amostragem

- ✓ Pesquisa eleitoral
- ✓ Pesquisa com clientes
- ✓ Desenvolvimento de modelos estatísticos
  - ✓ Amostra de desenvolvimento
  - ✓ Amostra de validação

Intervalo de Confiança

- Intervalos de confiança $P[(\bar{x} - 1.96.dp(\bar{x}) \leq \bar{X} \leq \bar{x} + 1.96.dp(\bar{x})] = 0.95$



Distribuição da média amostral segundo o modelo normal com parâmetros ($\bar{x}; dp(\bar{x})$)

O uso da distribuição normal como modelo para a distribuição da média amostral possibilita esperar que 95% das estimativas sejam diferentes do valor populacional por no máximo 1.96 desvios padrão.

Intervalo de Confiança

A figura 8.5 mostra 100 intervalos de confiança (95%) para a média de glicose no sangue de uma população. Os intervalos com estimativas menores que 34(g/l) ou maiores que 36(g/l) também não contém o valor 35.

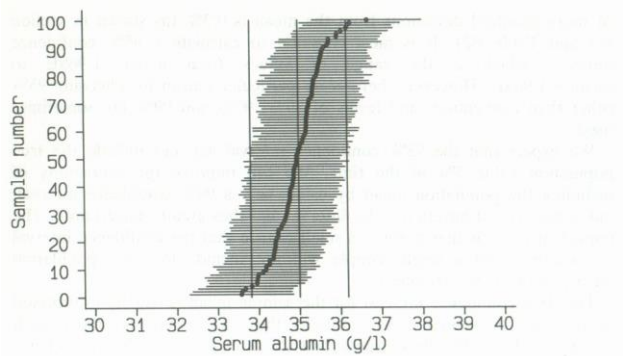
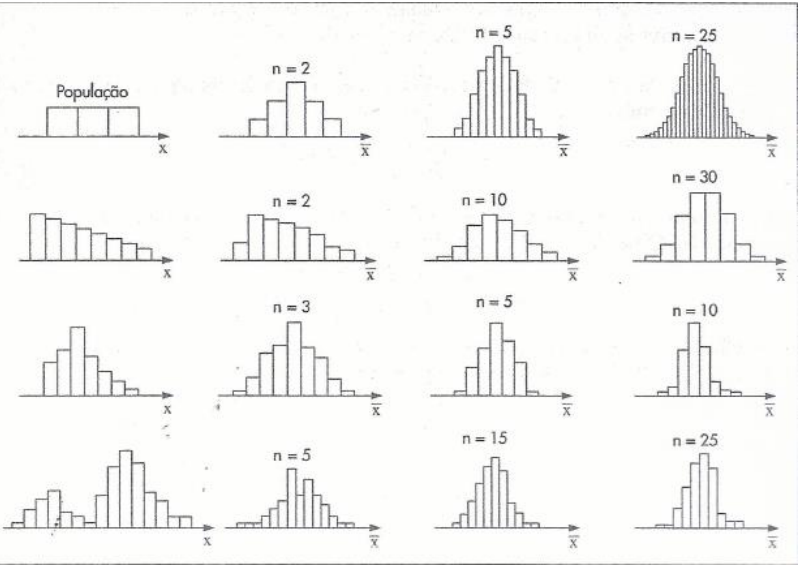


Figure 8.5 Confidence intervals from Figure 8.4 ordered by the magnitude of the mean of the random sample.

Fonte:Altman D.G. Practical Statistics for Medical Research.pag.164

Teorema do Limite Central



Para amostras aleatórias simples (X_1, ..., X_n) retiradas de uma população com média μ e variância σ^2 finita, a distribuição da média amostral $\overline{X}$ aproxima-se, para n grande, de distribuição normal com média μ e variância $\frac{\sigma^2}{n}$.

Fonte: BUSSAB, W.O.; MORETTIN, P. A., Estatística Básica, 5a. ed., São Paulo: Saraiva, 2006. Página 273.

Medidas de Dispersão

Desvio Padrão X Erro Padrão (Std Error)

- Desvio Padrão: variabilidade das observações em relação à média de uma amostra.
- Erro Padrão (Std Error) : variabilidade entre as amostras

$$Erropadrao = \frac{desviopadrão}{\sqrt{n}}$$

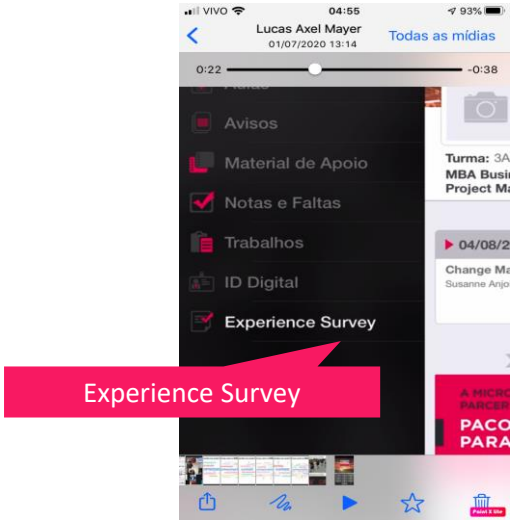
A grande finalidade do conhecimento não é conhecer, mas agir.

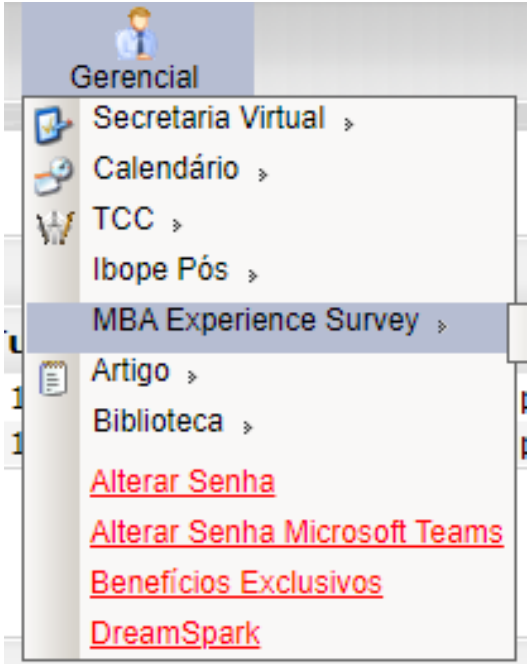
T. Huxley

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADO

 / Regina T. I. Bernal

FIAP


FIAP