FIAP

MBA

MBA EM DATA SCIENCE & AI

STATISTICS WITH R

Avaliação da disciplina

Avaliação	Peso
Listas de Exercícios (indiv.)	0.5
Projeto Integrado (2 a 4)	0.5
Aprenda R no R (indiv.)	+1 pto

Aprenda R no R

- Conjunto de lições de R em que deve-se seguir um tutorial da linguagem.

  - Lições feitas em casa;
  - No próprio ambiente R Studio
  - Banco de dados NoSQL na nuvem
  - Micro serviço com autenticação

Aprenda R no R

FIAP MBA+

Comandos de referência

```
# Instala pacote swirl
install.packages("swirl")
library(swirl)
select_language(language = 'portuguese')
```

```
# Instala curso
library(swirl)
uninstall_course('Aprenda_R_no_R')
install_course_github('elthonf','Aprenda_R_no_R')
```

```
# Inicia os cursos interativos
swirl()
```

```
# Outros comandos
library(swirl)
bye()
info()
Sys.setlocale("LC_ALL", 'en_US.UTF-8')
```

Aprenda R no R

```
| Toda a prática está rendendo frutos!

  |=========================================| 100%
| Gostaria de informar ao professor sobre a
| conclusão desta lição

1: Sim
2: Não
```

```
| Qual o cÃ³digo da sua turma?
```

(Usar FIAP-5DTS)

```
| Qual seu cÃ³digo de aluno?
```

Consultar a nota utilizando o Link abaixo

Visualizando notas

user	Comandos básicos do R	Logicos	Sequencias Numéricas	Vetores	Valores ausentes	Data e Hora	Filtrando vetores	Matrizes e tabelas	Amostras e Simulações	Funções	Manipulação com dplyr	Gráficos Básicos	K-Médias	Gráficos Analíticos	Total	Nota
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0

AULA 2

Estatística descritiva

Introdução a probabilidade

Distribuição de probabilidades

Importando Dados

O pacote {readr} do tidyverse é utilizado para importar arquivos de texto, como .txt ou .csv, para o R. Para carregá-lo, rode o código:

```
library(readr)
```

O {readr} transforma arquivos de textos em tibbles usando as funções:

read_csv(): para arquivos separados por vírgula.
read_rds(): para arquivos do tipo rds

Importando Dados

```
imdb <- read_rds("imdb.rds")
```

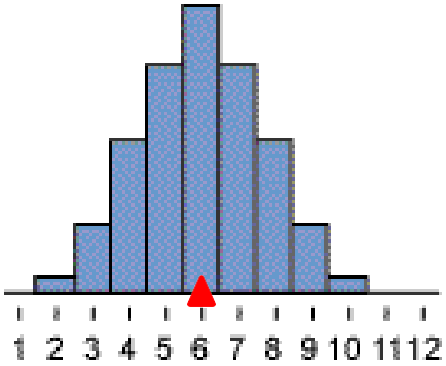
	titulo	ano	diretor	duracao	cor	generos	pais	classificacao
1	Avatar	2009	James Cameron	178	Color	Action|Adventure|Fantasy|Sci-Fi	USA	A partir de 13 anos
2	Pirates of the Caribbean: At World's End	2007	Gore Verbinski	169	Color	Action|Adventure|Fantasy	USA	A partir de 13 anos
3	The Dark Knight Rises	2012	Christopher Nolan	164	Color	Action|Thriller	USA	A partir de 13 anos
4	John Carter	2012	Andrew Stanton	132	Color	Action|Adventure|Sci-Fi	USA	A partir de 13 anos
5	Spider-Man 3	2007	Sam Raimi	156	Color	Action|Adventure|Romance	USA	A partir de 13 anos
6	Tangled	2010	Nathan Greno	100	Color	Adventure|Animation|Comedy|Family|Fantasy|Musical|Roma...	USA	Livre
7	Avengers: Age of Ultron	2015	Joss Whedon	141	Color	Action|Adventure|Sci-Fi	USA	A partir de 13 anos
8	Batman v Superman: Dawn of Justice	2016	Zack Snyder	183	Color	Action|Adventure|Sci-Fi	USA	A partir de 13 anos
9	Superman Returns	2006	Bryan Singer	169	Color	Action|Adventure|Sci-Fi	USA	A partir de 13 anos
10	Pirates of the Caribbean: Dead Man's Chest	2006	Gore Verbinski	151	Color	Action|Adventure|Fantasy	USA	A partir de 13 anos
11	The Lone Ranger	2013	Gore Verbinski	150	Color	Action|Adventure|Western	USA	A partir de 13 anos
12	Man of Steel	2013	Zack Snyder	143	Color	Action|Adventure|Fantasy|Sci-Fi	USA	A partir de 13 anos
13	The Chronicles of Narnia: Prince Caspian	2008	Andrew Adamson	150	Color	Action|Adventure|Family|Fantasy	USA	Livre
14	The Avengers	2012	Joss Whedon	173	Color	Action|Adventure|Sci-Fi	USA	A partir de 13 anos
15	Pirates of the Caribbean: On Stranger Tides	2011	Rob Marshall	136	Color	Action|Adventure|Fantasy	USA	A partir de 13 anos
16	Men in Black 3	2012	Barry Sonnenfeld	106	Color	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi	USA	A partir de 13 anos
17	The Amazing Spider-Man	2012	Marc Webb	153	Color	Action|Adventure|Fantasy	USA	A partir de 13 anos
18	Robin Hood	2010	Ridley Scott	156	Color	Action|Adventure|Drama|History	USA	A partir de 13 anos
19	The Hobbit: The Desolation of Smaug	2013	Peter Jackson	186	Color	Adventure|Fantasy	USA	A partir de 13 anos

Medidas Resumo

São estatísticas que resumem, em um único valor, a tendência central (média, mediana, moda), a variabilidade (variância, desvio padrão) e a forma da distribuição (simétrica ou assimétrica) da variável.

Medidas Resumo

Distribuição simétrica



Distribuição do tempo de uso de internet (horas)

Medidas de tendência central:

- Média
- Mediana
- Moda

Indicam o centro da distribuição de frequências ou a região de maior concentração de frequência na distribuição.

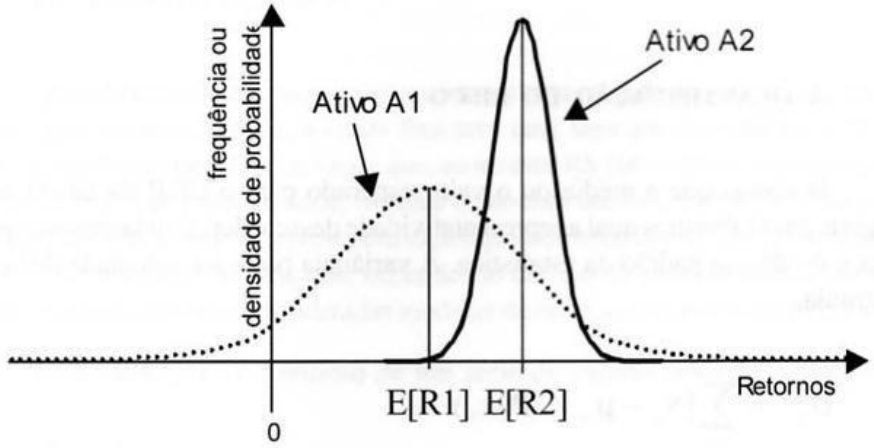
Medidas de dispersão:

- Variância
- Desvio padrão

Indicam o grau de homogeneidade dos valores, até que ponto eles se encontram concentrados ou dispersos da média.

Medidas Resumo

Decisão pela média



Qual ativo você escolheria para investir? Justifique sua escolha.

Medidas Resumo

Exemplo 2

Durante uma verificação de qualidade no conteúdo de seis recipientes de café instantâneo, foram obtidas as seguintes notas:

6,03 5,59 6,40 6,00 5,99 6,02

Qual a média e a mediana encontrada?

Média aritmética: $\bar{x} = \sum^{n}$ ➡ $x = \frac{6{,}03+5{,}59+6{,}40+6{,}00+5{,}99+6{,}02}{6}$ ➡ $\bar{x} = 6{,}00$

Mediana: 5,59 5,99 6,00 6,02 6,03 6,40

$$mediana = \frac{6{,}00 + 6{,}02}{2} = 6{,}01$$

Medidas Resumo

Exemplo 1

Durante uma verificação de qualidade no conteúdo de seis recipientes de café instantâneo, foram obtidas as seguintes notas:

6,03 5,59 6,40 6,00 5,99 6,02

Qual a média e a mediana encontrada? $\bar{x} = 6{,}00 \quad mediana = 6{,}01$

Suponha que o terceiro valor tenha sido incorretamente medido e que na verdade seja de 6,04. Determine novamente a nota média e mediana.

Média aritmética: $x = \frac{6{,}03 + 5{,}59 + 6{,}04 + 6{,}00 + 5{,}99 + 6}{}$

Mediana: 5,59 5,99 6,00 6,02 6,03 6,04

$$mediana = \frac{6{,}00 + 6{,}02}{2} = 6{,}01$$

Medidas Resumo

Comparação entre Média, Mediana e Moda

	VANTAGENS	LIMITAÇÕES	TIPO DE VARIÁVEIS
MÉDIA	Reflete todos os valores da amostra	É influenciada por valores extremos	Contínua e discreta
MEDIANA	Menos sensível a valores extremos que a média	Mais difícil de ser determinada para grande quantidade de dados	Contínua e discreta
MODA	Representa um valor típico	Não tem função em certos conjuntos de dados	Contínua, discreta, nominal e ordinal

Medidas Resumo

MEDIDAS DE **POSIÇÃO - MÉDIA**

- Média Aritmética Simples:

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

- Média Aritmética Ponderada:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot F_i}{n}$$

- Média Geométrica (evolução):

$$Mg = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

- Média Quadrática:

$$\bar{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

Medidas Resumo

Moda

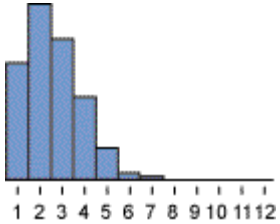
```
library(dplyr)
moda <- function(vetor){
posModa <- vetor %>% table %>% data.frame() %>%
arrange(desc(Freq))
modaVec <- posModa %>%  filter(Freq >= posModa[1, 2])
moda <- modaVec[,1] %>% as.character %>% as.numeric()  moda}

vetor <- c(1,2.1, 2.1,3,3,5,6)
moda(amostra)
amostra <- sample.int(10, 15, replace = TRUE)
moda(amostra)
```

Medidas Resumo

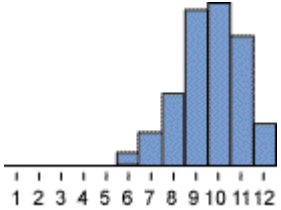
Decisão pela média ?????

Assimétrico à direita



Média > Mediana

Assimétrico à esquerda



Média < Mediana

Medidas Resumo

- Outras Medidas de Posição

Decis: dividem um conjunto de dados em dez partes iguais.

Percentis (P): dividem a série em cem partes, de modo que p% ficam abaixo dele (P).

Quartis: dividem a série em quatro partes iguais.

Medidas Resumo

- Medidas de Dispersão

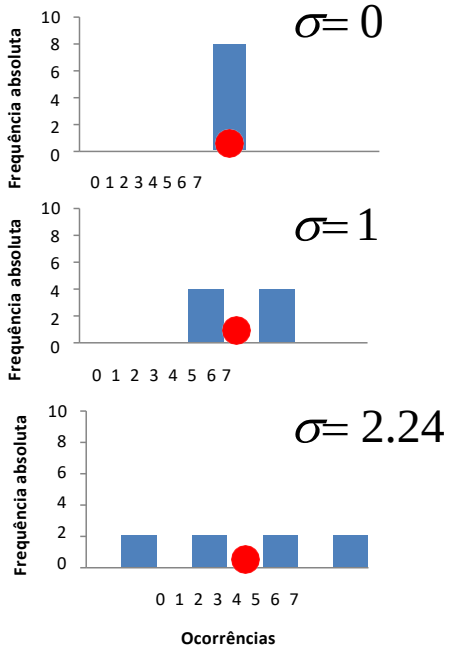
Exemplo 8:

A: 4, 4, 4, 4, 4, 4, 4 ,4

B: 3, 3, 3, 3, 5, 5, 5, 5

C: 1, 1, 3, 3, 5, 5, 7, 7

Qual o desvio padrão?



Medidas de Dispersão

Medidas de Dispersão: variância e desvio padrão Exemplo C

X	Média	(X-Média)	$(X\text{-Média})^2$
1	4	-3	9
1	4	-3	9
3	4	-1	1
3	4	-1	1
5	4	1	1
5	4	1	1
7	4	3	9
7	4	3	9
Soma	-	0	40

Variância:

$$\sigma^2 = \frac{40}{8} = 5$$

Desvio padrão:

$$\sigma = \sqrt{\sigma^2} = \sqrt{5} = 2.24$$

Medidas de Dispersão

O quanto os pontos (dados) estão distantes da média (ponto central)

- variância da população $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n}$
- variância da amostra $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Medidas de Dispersão

No R...

```
x <- c(1, 1, 3, 3, 5, 5, 7, 7)

sd(x) (desvio-padrão)

sd(x)² (variância)
```

Medidas de Dispersão

No R...

```
x <- c(1, 1, 3, 3, 5, 5, 7, 7)

sd(x) (desvio-padrão)

sd(x)² ou var(x) (variância)
```

Como calcular a variância populacional?

Exercícios

1) Leia a base imdb.rds e selecione as variáveis **ano, duração e orçamento**.
2) Calcule média, dp, 1° Quartil, Mediana, 3° Quartil, min, máx.
3) Dado a variável y<-c(445, 530, 540, 510, 570, 530, 545, 545, 505, 535, 450, 500, 520, 460, 430, 520, 520, 430, 535, 535, 475, 545, 420, 495, 485, 570, 480, 495, 470, 490); Calcule as medidas de resumo do item 2.

Medidas Resumo

Medidas de Assimetria

As medidas de assimetria referem-se à forma da curva que representa a distribuição de frequência. A assimetria é o afastamento da simetria de uma frequência.

- Curvas de frequência simétrica ou em forma de sino: caracterizam-se pelo fato das observações equidistantes do ponto central terem a mesma frequência (curva normal)

- Curvas de frequência moderadamente assimétricas ou desviadas: a cauda de um lado da ordenada máxima é mais longa do que do outro. Se o ramo mais alongado fica à direita, a curva é dita de assimetria positiva, enquanto que, se ocorre o inverso, diz-se que a curva é de assimetria negativa.

Medidas Resumo

Coeficientes de Assimetria (Skewness)

➔ $As = \frac{m^3}{\sigma^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{[\frac{1}{n}(x_i - \bar{x})^2\}^{\frac{3}{2}}}$

As=0 ➔simétrica
As>0 ➔assimétrica positiva
As<0 ➔ assimétrica negativa

Índice de Assimetria (Pearson)

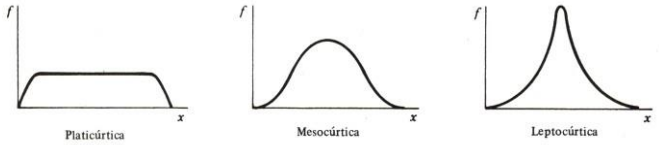
➔ $A = \frac{\text{média} - \text{moda}}{\text{desviopadrão}}$

|A|<0,15➔simétrica
0,15<|A|<1➔ assimetria moderada
|A|>1 ➔ assimetria forte

Medidas Resumo

Medidas de Assimetria

- Curtose: grau de achatamento em relação a uma curva Normal
  - Leptocúrtica (afilado) ➔ K>3
  - Mesocúrtica ➔K=3
  - Platicúrtica (achatado) ➔K<3



Medidas Resumo

Outras Medidas de Dispersão

- Coeficiente de Variação
- Amplitude
- Amplitude Inter-Quartílica

Medidas Resumo

Outras Medidas de Dispersão

Coeficiente de variação (CV)

É o quociente entre o desvio padrão e a média.

$$CV = \frac{\sigma}{\overline{x}}$$

Vantagem: caracterizar a dispersão dos dados em termos relativos a seu valor médio.

Medidas Resumo

Qual o coeficiente de variação?

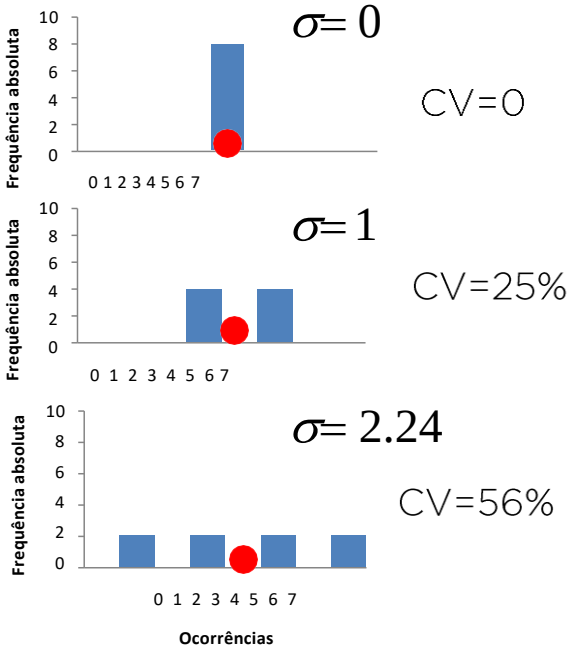
- Medidas de Dispersão

Exemplo 8:

A: 4, 4, 4, 4, 4, 4, 4 ,4

B: 3, 3, 3, 3, 5, 5, 5, 5

C: 1, 1, 3, 3, 5, 5, 7, 7



Medidas Resumo

Outras Medidas de Dispersão

Amplitude

É definida como a diferença entre o maior e o menor valor de um conjunto de dados.

Fortemente relacionado com a dispersão dos dados.

A amplitude pode levar a erros de avaliação, pois não representa o conjunto dos dados.

Muitas vezes reflete muito mal a dispersão dos mesmos.

Medidas Resumo

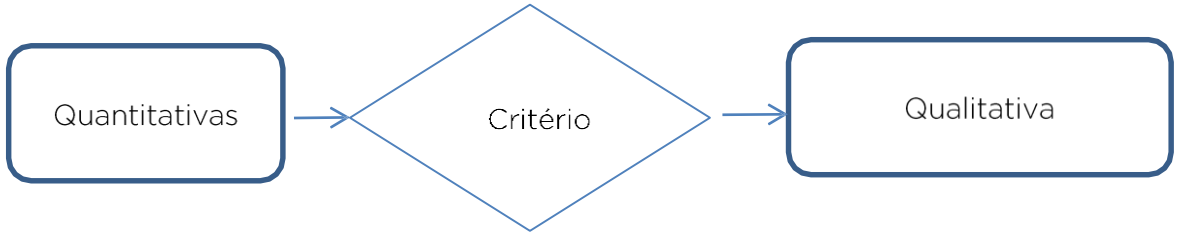
- Outras Medidas de Dispersão

Amplitude Inter-quartílica

É a diferença entre o terceiro e o primeiro quartil (Q3-Q1).

Usada em análise exploratória de dados – gráficos Box Plot.

Transformando variáveis quantitativas em qualitativas



Exemplo:

Anos de estudo

Critério	Grau instrução
0	Analfabeto
[1 - 9]	Fundamental
[10 – 12]	Médio
>= 13	Superior

Distribuição de Frequência

O número de vezes que ocorreram valores em cada classe ou valores chama-se frequência absoluta. O conjunto das ocorrências, com correspondentes frequências absolutas (FA) e relativas (FR), define a distribuição de frequências da variável. Conhecer o comportamento da variável.

Distribuição etária dos trabalhadores da Empresa XXX, 01/05/2019

Faixa etária	Frequency	Percent	Cumulative Frequency	Cumulative Percent
00 - 17	19052	33,8	19052	33,8
18 - 29	16143	28,6	35195	62,4
30 - 39	13710	24,3	48905	86,7
40 - 49	5773	10,2	54678	96,9
50 - 59	1559	2,8	56237	99,7
60 - 69	174	0,3	56411	100,0
Acima 69	13	0,0	56424	100,0
Total	56424	100,0		

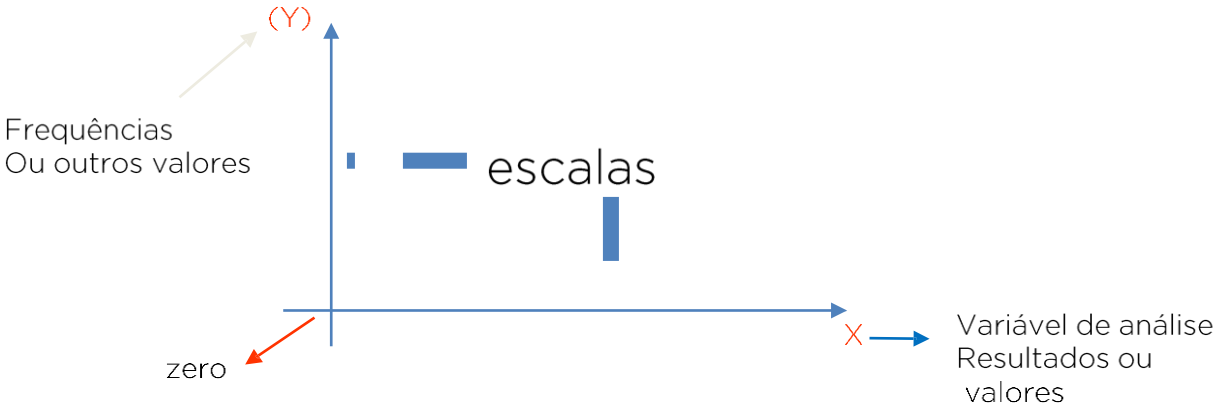
Apresentação Gráfica dos Dados

As regras básicas de elaboração de um gráfico são:

- simplicidade
- clareza
- veracidade

Apresentação Gráfica dos Dados

➢ EIXOS CARTESIANOS



Como medir incerteza? (Probabilidade?)

Experimento Aleatório: procedimento que, ao ser repetido sob as mesmas condições, pode fornecer resultados diferentes.

Exemplos:

1. Resultado do lançamento de um dado equilibrado;
2. Sorteado um estudante da escola e perguntar se ele é fumante ou não.
3. Tipo sanguíneo de um habitante escolhido ao acaso;
4. Um lote de ações é comprado por R$ 100,00. Você deseja observar o preço que esse lote de ações pode ser vendido daqui a um ano;
5. Dois motoristas em uma rodovia do estado de São Paulo são selecionados aleatoriamente e verifica-se se estão usando o cinto de segurança.

Espaço Amostral (Ω): Conjunto de todos os resultados possíveis de um experimento aleatório.

1. Lançamento de um dado equilibrado.
   $\Omega = \{1, 2, 3, 4, 5, 6\}$
2. Sorteado um estudante da escola e perguntar se ele é fumante ou não.
   $\Omega = \{\text{Fumante}, \text{Não Fumante}\}$
3. Tipo sanguíneo de um habitante de Osasco escolhido ao acaso.
   $\Omega = \{A, B, AB, O\}$
4. Um lote de ações é comprado por R$ 100,00. A que preço esse lote de ações pode ser vendido em um ano.
   $\Omega = \{x \in \mathbb{R} \mid x \geq 0\}$
5. Se um motorista estiver usando cinto de segurança usaremos a letra C, caso contrário S. $\Omega = \{CC, CS, SC, SS\}$

Probabilidade: é uma medida da incerteza associada aos resultados do experimento aleatório.

Deve fornecer a informação de quão verossímil é a ocorrência de um particular evento.

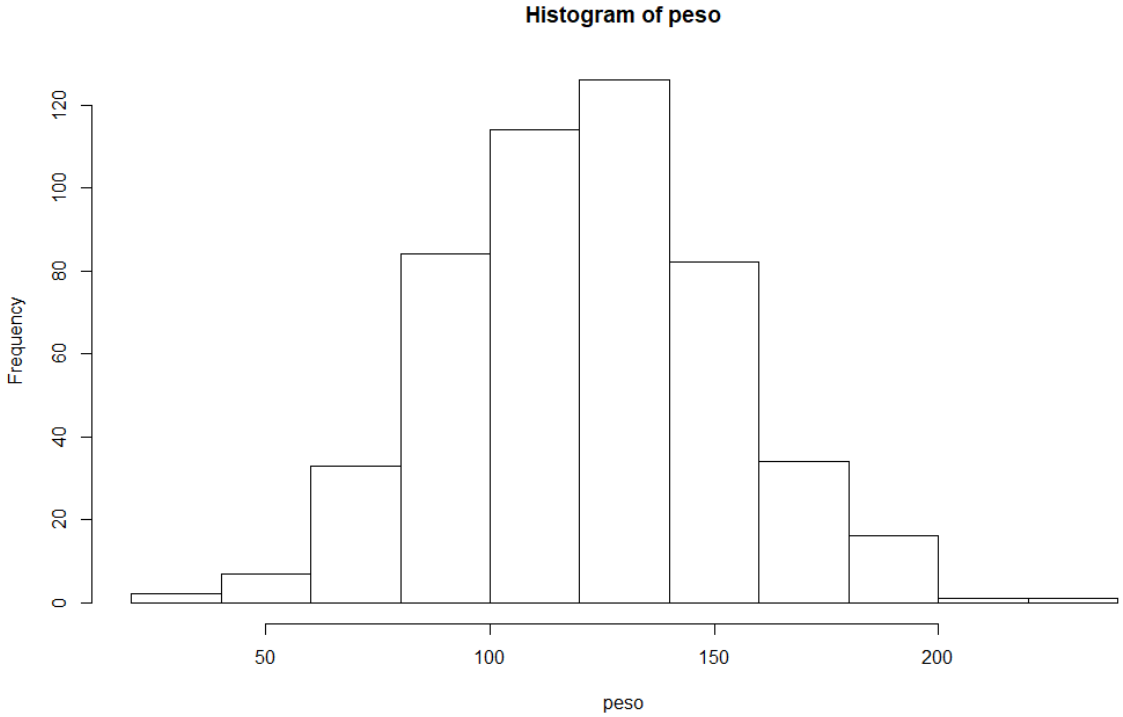
Como atribuir probabilidade aos eventos do espaço amostral?

Temos duas abordagens possíveis:

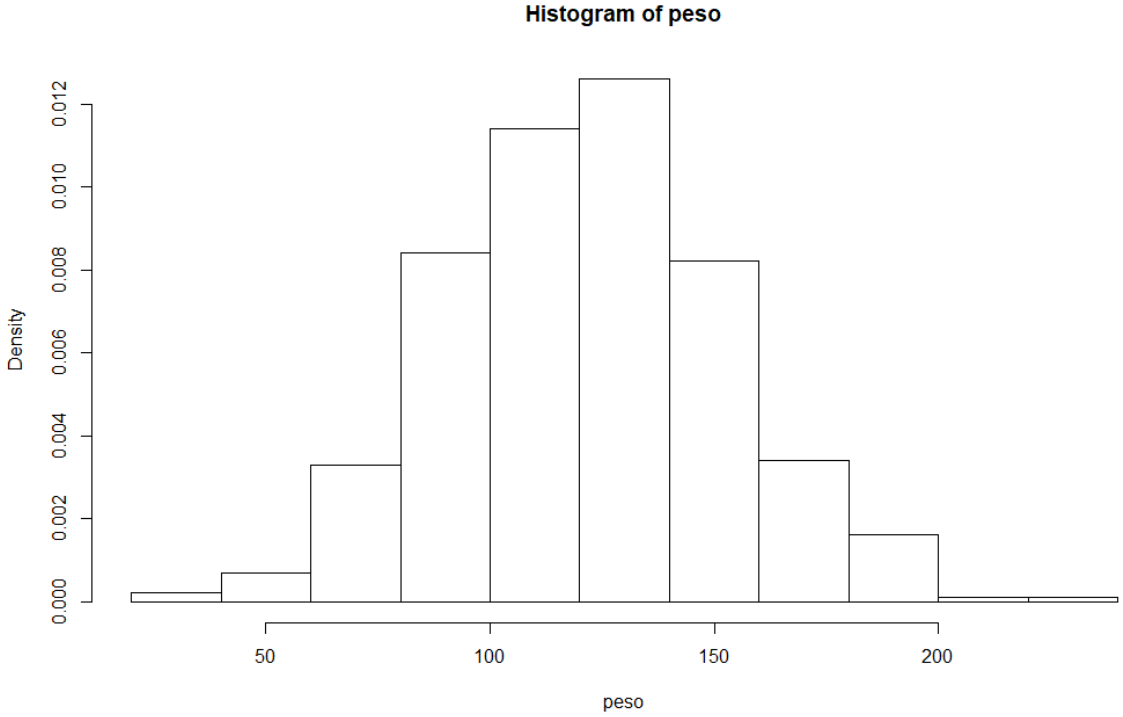
Freqüências de ocorrências de um evento: é o número de vezes que esse evento ocorre dividido pelo total de vezes que o experimento é realizado.

Suposições teóricas: nessa abordagem a atribuição de probabilidade a um evento e feita baseando-se em características teóricas do experimento.

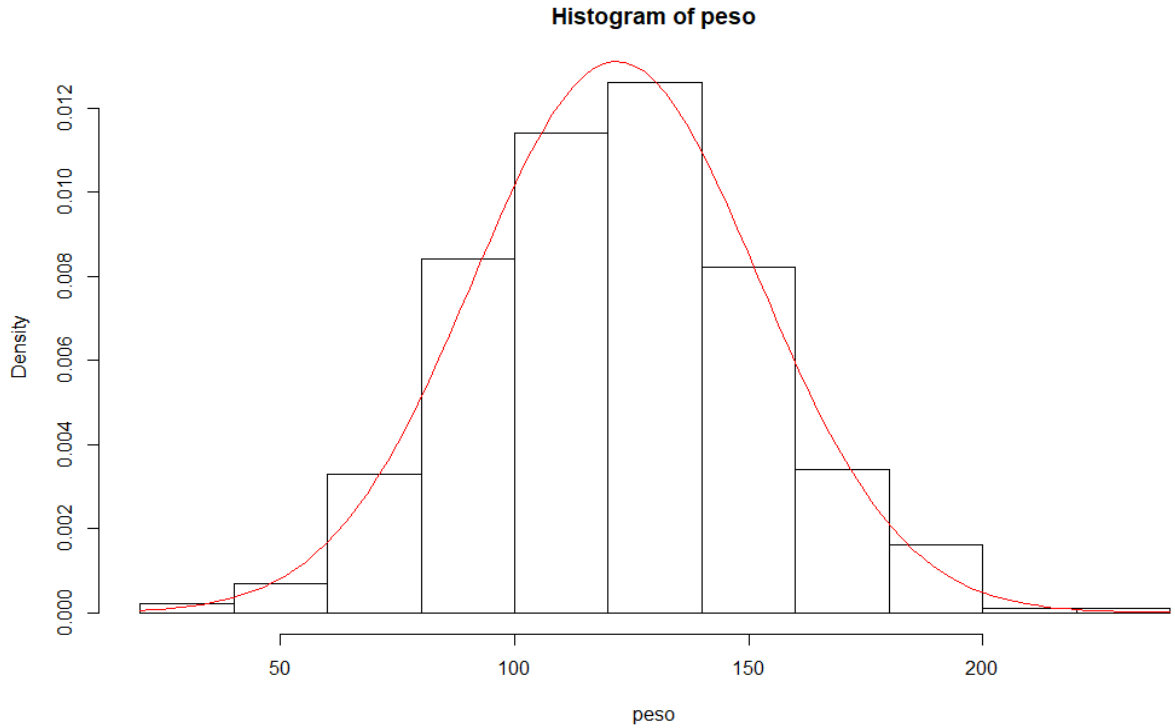
Variável peso: Frequência



Variável peso: Probabilidade



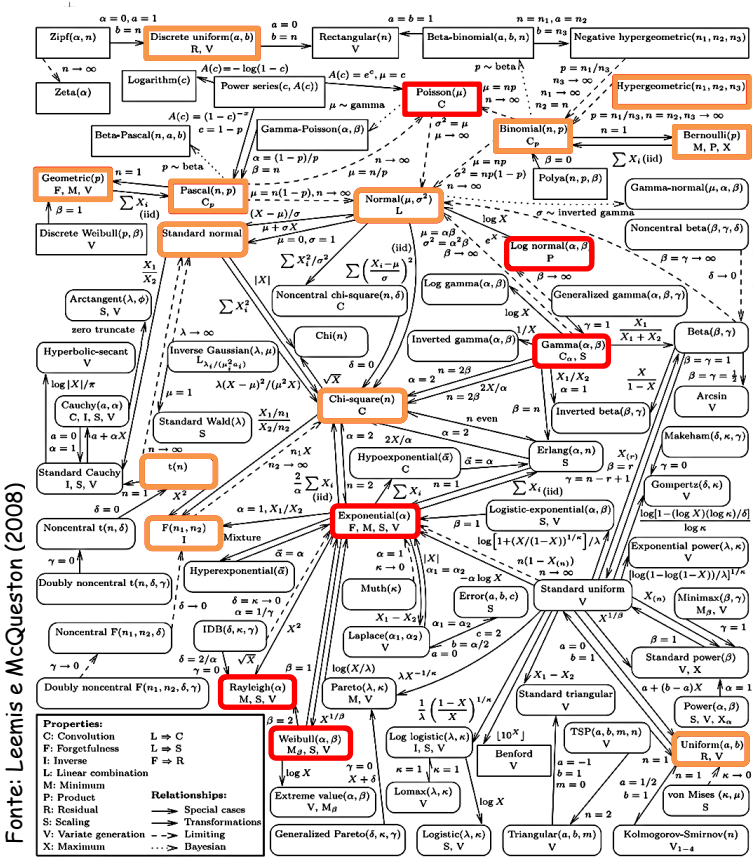
Variável peso: Fitada por uma N(120,30)



Variável Aleatória

- Uma quantidade **X**, associada a cada possível resultado do espaço amostral **Ω**, é denominada **variável aleatória**, se assume valores em um conjunto, com certa probabilidade **P.**
- Dizemos que a ocorrência de eventos segue uma **distribuição de probabilidade.**
- Assume-se que as observações de uma amostra são oriundas de uma variável aleatória cuja distribuição é **conhecida ou não**.

Distribuições de Probabilidade



Fonte: Leemis e McQueston (2008)

Quantas funções que descrevem variáveis aleatórias existem?

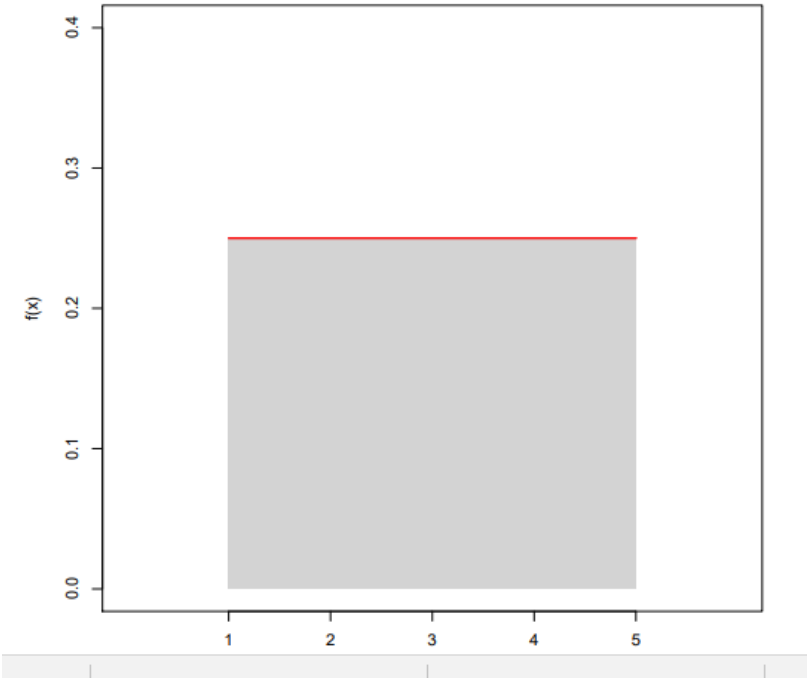
V.A. Discreta	V.A. Contínua
• Uniforme Discreta	• Uniforme
• Bernoulli	• Normal
• Binomial	• Exponencial
• Geométrica	• Log-Normal
• Binomial Negativa ou Pascal	• Triangular
• Hipergeométrica	• Beta
	• Gamma

O que é importante saber:

- Tipo de v.a. (discreta ou contínua)
- Escopo da v.a. (mínimo e máximo)
- Função de Distribuição de Probabilidade e seus parâmetros
- A média (medida de tendência central)
- A variância (medida de dispersão)

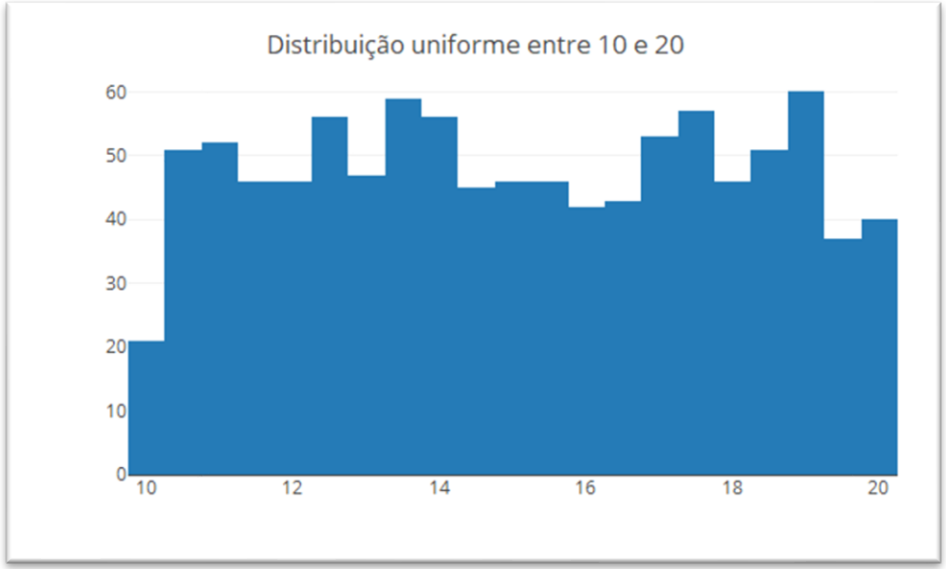
Distribuição de Probabilidade

- Familia Uniforme



Distribuição de Probabilidade

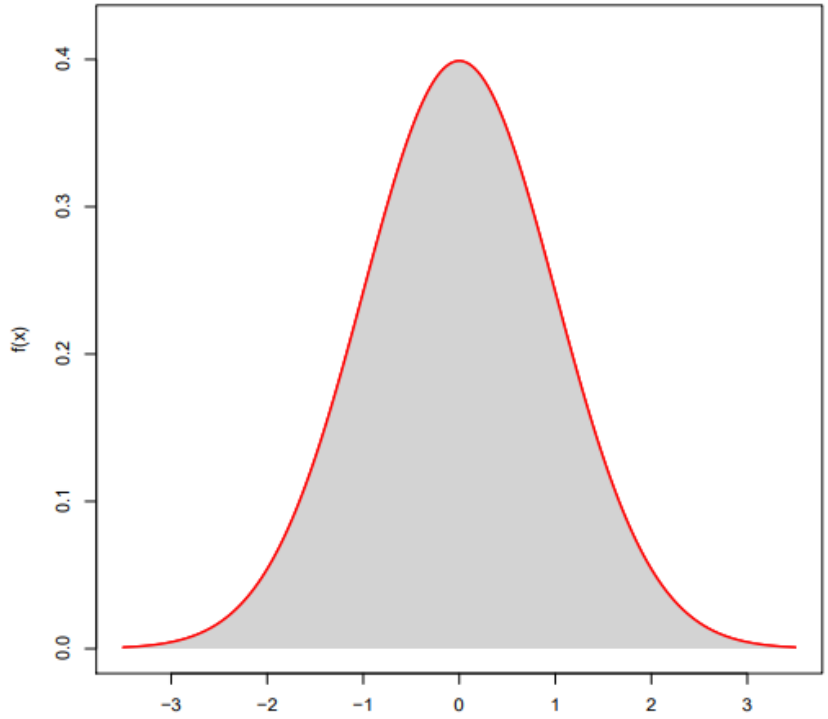
- Familia Uniforme



Gerando números aleatórios

- Família Uniforme
  - `runif`: Para obter número aleatórios seguindo uma distribuição uniforme a partir de um mínimo min e um máximo max
  - `dunif` - Avalia a probabilidade uniforme de um valor (dado um mínimo min e dado um máximo max)
  - `punif` - Avalia a probabilidade ACUMULADA uniforme de um valor (dado min e max)
    - Esta função deve formar uma reta

Distribuição Normal



Distribuição Normal

Distribuição Normal

Se X é uma variável aleatória com distribuição normal de média μ e variância σ^2, a função densidade de probabilidade de X é definida por

Distribuição Normal

FIAP MBA+

Distribuição Normal

Se X é uma variável aleatória com distribuição normal de média μ e variância σ^2, a função densidade de probabilidade de X é definida por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{\sigma^2}(x-\mu)^2},$$

para $-\infty < x, \mu < +\infty$ e $\sigma > 0$. Notação: $X \sim N(\mu, \sigma^2)$.

Distribuição Normal

FIAP MBA+

Padronização

Se $X \sim N(\mu, \sigma^2)$ e $Z \sim N(0, 1)$ (normal padrão), então

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right),$$

ou seja, todos os cálculos podem ser feitos pela normal padrão.

Distribuição Normal

Cálculo de probabilidades

Por exemplo, a probabilidade $A = P(0 \leq X \leq 1)$ pode ser calculada pela diferença

$$P(X \leq 1) - P(X \leq 0) = 0,841 - 0,5 = 0,341.$$

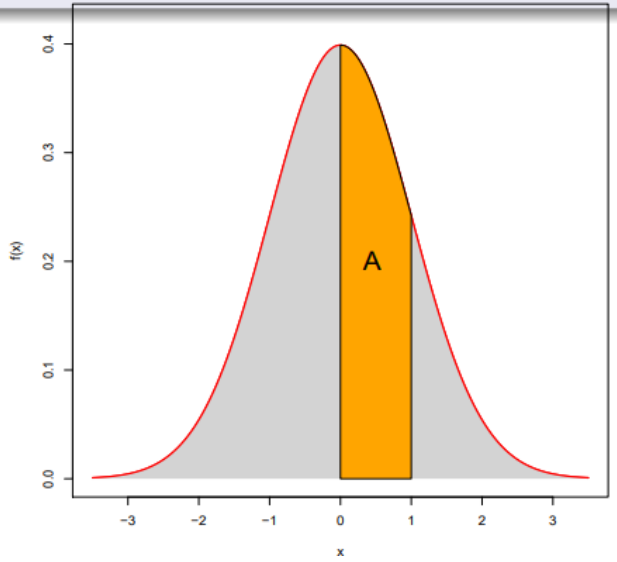
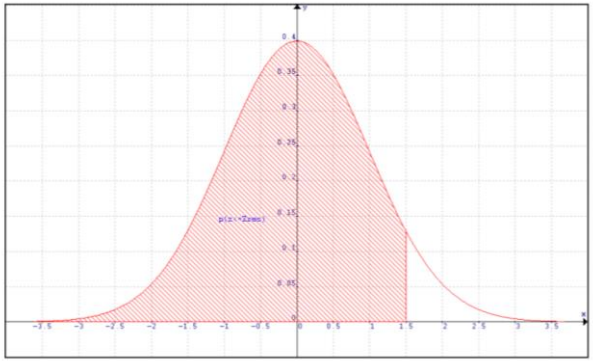


Tabela da distribuição Normal

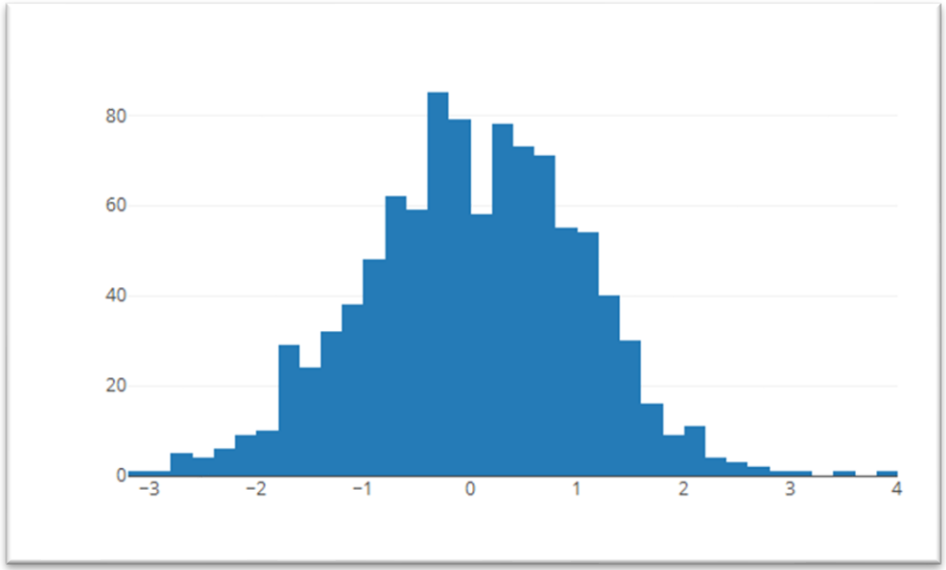


P(Z<1.38)=?

Zres até a 1ª casa decimal	Segunda casa decimal de Zres									
	0	1	2	3	4	5	6	7	8	9
0,0	**0,5000**	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
...	...	...	...	...	...	...	...	...	...	...
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	**0,9162**	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
...	...	...	...	...	...	...	...	...	...	...

Gerando números aleatórios

- Família Normal



Trabalhando no R

- Família normal
  - `rnorm`: Para obter número aleatórios seguindo uma distribuição normal
  - `dnorm` - Avalia a probabilidade da normal de um valor (dada a média μ e o desvio padrão σ)
  - `pnorm` - Avalia a probabilidade ACUMULADA da normal de um valor (dada μ e σ)
    - Esta função deve formar uma curva sigmóide!

Exemplo

```
rnorm(100, 0, 10)

dnorm(5, 0, 10)

pnorm(5, 0, 10)
```

Exercício

01. Uma empresa produz televisores de dois tipos, tipo *A* (comum) e tipo *B* (luxo), e garante a restituição da quantia paga se qualquer televisor apresentar defeito grave no prazo de seis meses. O tempo para ocorrência de algum defeito grave nos televisores tem distribuição normal sendo que, no tipo *A*, com média de 10 meses e desvio padrão de 2 meses e no tipo *B*, com média de 11 meses e desvio padrão de 3 meses. Os televisores de tipo *A* e *B* são produzidos com lucro de 1200 u.m. e 2100 u.m. respectivamente e, caso haja restituição, com prejuízo de 2500 u.m. e 7000 u.m., respectivamente.

(a) Calcule as probabilidades de haver restituição nos televisores do tipo *A* e do tipo *B*.

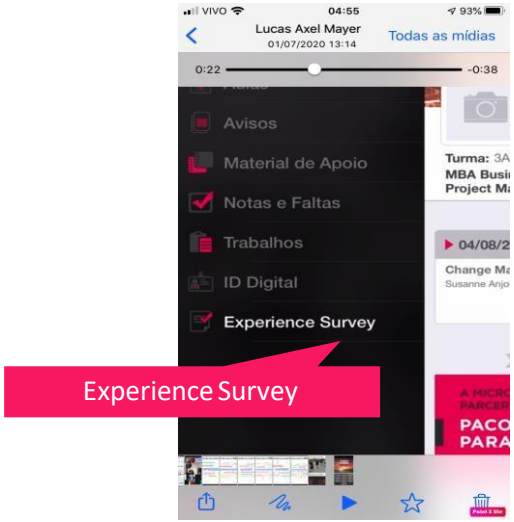
(b) Calcule o lucro médio para os televisores do tipo *A* e para os televisores do tipo *B*.

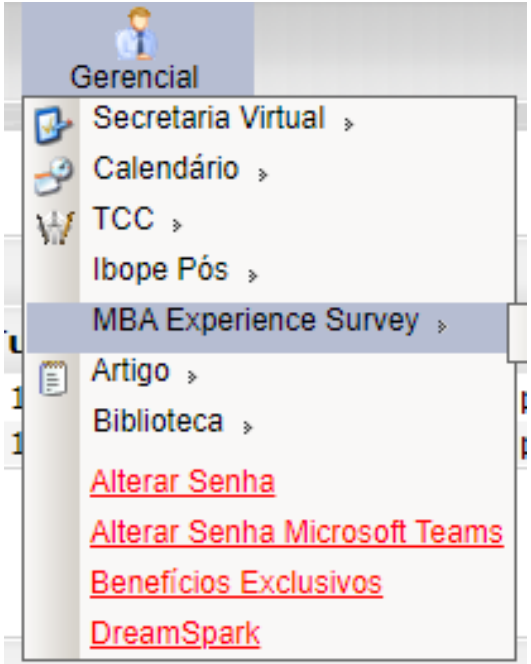
(c) Baseando-se nos lucros médios, a empresa deveria incentivar as vendas dos aparelhos do tipo *A* ou do tipo *B*?

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

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





OBRIGADO

profleandro.ferreira@fiap.com.br

FIAP MBA+


FIAP