

Mestrado Profissional em Avaliação e
Monitoramento de Políticas Públicas

Métodos Quantitativos I

Aula 2: Descrevendo Variáveis

Professores: Daniel Grimaldi e Arthur Bragança

3º Trimestre - 2025

Conceitos básicos

Variável

- ❖ No contexto de pesquisa empírica, **uma variável é um conjunto de realizações (ou observações) de um mesmo fenômeno.**
 - ❖ renda mensal dos brasileiros;
 - ❖ idade dos moradores do plano piloto;
 - ❖ emprego dos alunos da ENAP
- ❖ As variáveis podem representar fenômenos quantitativos ou qualitativos...
- ❖ Elas sempre possuem uma distribuição
 - ❖ Uma distribuição é uma descrição da frequência com que determinada variável assume um valor (ou conjunto de valores) específicos

Definição formal

Denominamos de variável aleatória (V.A.) qualquer função $X : \Omega \rightarrow \mathbb{R}$. Ou seja, é **uma função do espaço amostral Ω nos reais para a qual é possível calcular a probabilidade de ocorrência**.

Sendo X uma V.A., sua **função de distribuição** é definida por:

$$F_X(x) = P(X \in (-\infty, x]) = P(X \leq x))$$

Sempre vale que: $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$ e $F(x) \leq F(y)$ sempre que $x \leq y$, $\forall x, y \in \mathbb{R}$

Tipos de variáveis

- Formalmente, a **diferenciação relevante é entre variáveis aleatórias discretas e contínuas**
 - **discreta:** assume somente um número enumerável de valores (finito ou infinito)
 - **contínua:** podem assumir um número não enumerável de valores
- Mas existem outras categorizações comuns: ordinais vs. categóricas; quantitativa vs. qualitativa

Variáveis Discretas

Uma V.A. X é **discreta** se há uma associação entre probabilidades p_1, p_2, \dots, p_k e um conjunto de valores possíveis x_1, x_2, \dots, x_k **mediada por um função de probabilidade** $p(x)$, que satisfaz as seguintes propriedades:

- i. $0 \leq p(x_i) \leq 1, \forall i = 1, 2, 3, \dots, k;$
- ii. $\sum_{i=1}^k p(x_i) = 1$

Exemplos: resultados de um dado, número de carros na garagem da ENAP, número de eleitores que aprova o presidente da república etc.

Notas: ¹ Quando uma V.A. discreta assume muitos valores possíveis, ela começa a se parecer muito com uma V.A. contínua - mas isso não significa que ela se torna contínua.

Variáveis Contínuas

Uma V.A. X é **contínua** se há uma associação entre probabilidades p_1, p_2, \dots, p_k e um conjunto de valores possíveis x_1, x_2, \dots, x_k **mediada por uma função de densidade** $f(x)$, que satisfaz as seguintes propriedades:

- i. $f(x) \geq 0 \quad \forall x \in \mathbb{R};$
- ii. $\int_{-\infty}^{\infty} f(x) dx = 1$

Exemplos: salário, renda *per capita*, preço do dólar etc.

Notas: ¹ Quando a variável é contínua, não é possível associar uma probabilidade positiva a um valor particular, mas pode-se atribuir uma probabilidade para um intervalo arbitrariamente pequeno de valores. ² Na prática, uma variável contínua pode assumir valores em toda a reta real, enquanto uma variável discreta somente pode assumir um conjunto enumerável de valores

Principais distribuições teóricas

Uniforme Discreta

Uma V.A. segue um modelo **Uniforme discreto** com valores x_1, x_2, \dots, x_k se tem uma **função de probabilidade** $p(x)$ dada por:

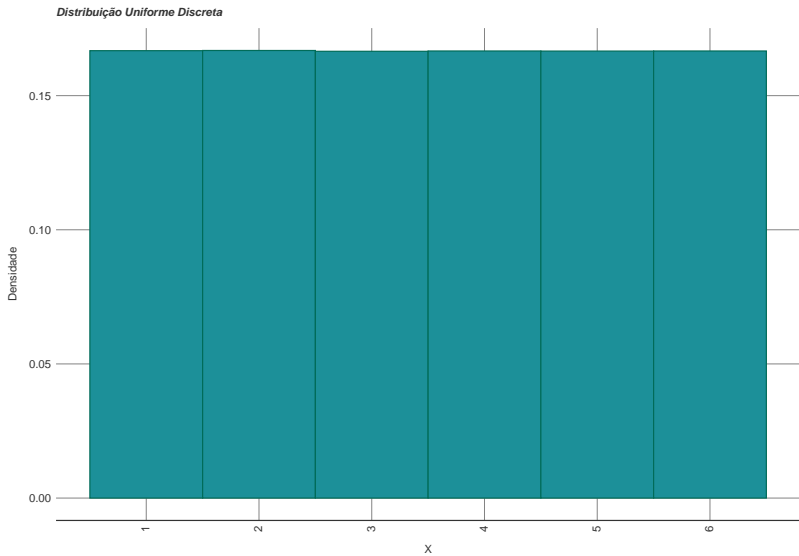
$$p(x_i) = \frac{1}{k} \quad \forall i = 1, 2, 3, \dots, k$$

Exemplo: o valor observado para 1 lançamento de um dado não viesado.

Notas: ¹ Na prática, o modelo uniforme atribui uma mesma probabilidade de ocorrência para cada valor possível. ² Nesse caso dizemos que $X \sim U_d[E]$ sendo $E = \{1, \dots, 6\}$

Distribuição Uniforme Discreta

```
data <- data.frame(X = rdunif(10000000, 1, 6))
fig <- ggplot(data) +
  geom_histogram(aes(X, y=..density..),
                 color=cores$verde_escuro,
                 fill=cores$verde_claro, bins=6) +
  scale_x_continuous(breaks=seq(1, 6)) +
  labs(subtitle="Distribuição Uniforme Discreta",
       y="Densidade") +
  tema_base_fundobranco()
```



Bernoulli

Uma V.A. segue um modelo **Bernoulli** se assume apenas os valores 0 e 1. Sua **função de probabilidade** $p(x)$ é dada por:

$$p(x_i) = \begin{cases} p & \text{se } x_i = 1 \\ 1 - p & \text{se } x_i = 0 \end{cases}$$

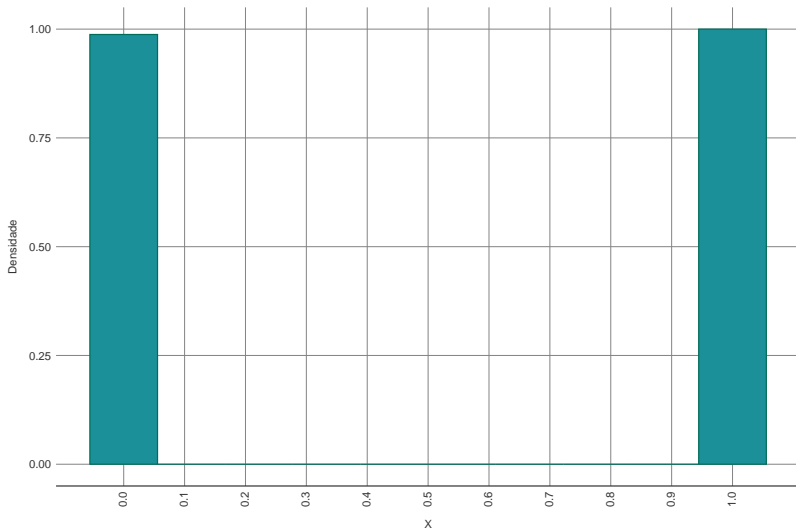
Exemplo: o resultado do lançamento de uma moeda.

Notas: ¹ Nesse caso, dizemos que $X \sim \text{Bernoulli}(p)$.

Distribuição Bernoulli

```
data <- data.frame(X = rbern(10000, 0.5))
fig <- ggplot(data) +
  geom_histogram(aes(X, y=..ndensity..),
                 color=cores$verde_escuro,
                 fill=cores$verde_claro,
                 bins=10) +
  scale_x_continuous(breaks=seq(0, 1, by=0.1)) +
  labs(title="Distribuição Bernoulli",
       y="Densidade") +
  tema_base_fundobranco()
```

Distribuição Bernoulli



Binomial

Uma V.A. segue um modelo **Binomial** se ela representa a quantidade total de sucessos obtidos por meio da realização de n ensaios de Bernoulli. Sua **função de probabilidade** $p(x)$ é dada por:

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

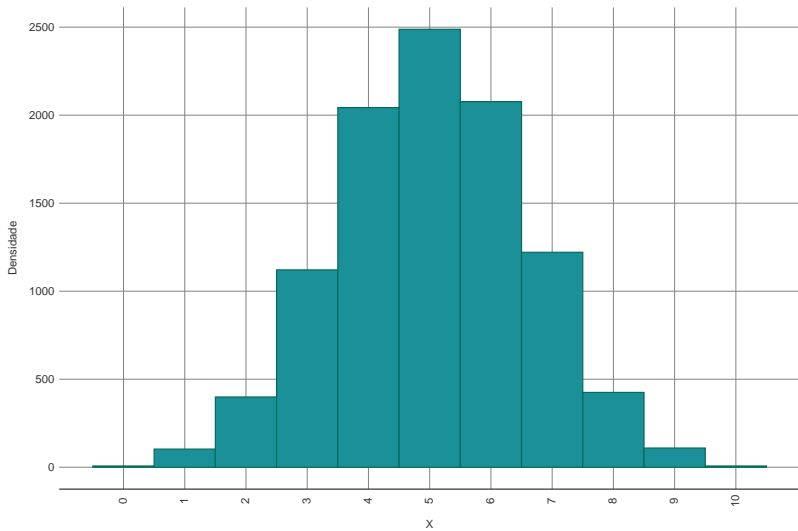
Exemplo: A quantidade de “caras” após 10 lançamentos de uma moeda.

Notas: ¹ Nesse caso, dizemos que $X \sim B(n, p)$.

Distribuição Binomial

```
data <- data.frame(X = rbinom(10000, 10, 0.5))
fig <- ggplot(data) +
  geom_histogram(aes(X),
                 color=cores$verde_escuro,
                 fill=cores$verde_claro,
                 bins=11) +
  scale_x_continuous(breaks=seq(0, 10, by=1)) +
  labs(title="Distribuição Binomial",
       y="Densidade") +
  tema_base_fundobranco()
```


Distribuição Binomial



Exponencial

Uma V.A. X segue um modelo **Exponencial** se sua **função densidade** $f(x)$ é dada por:

$$f(x) = \lambda e^{-\lambda x} I_{(0, \infty)}$$

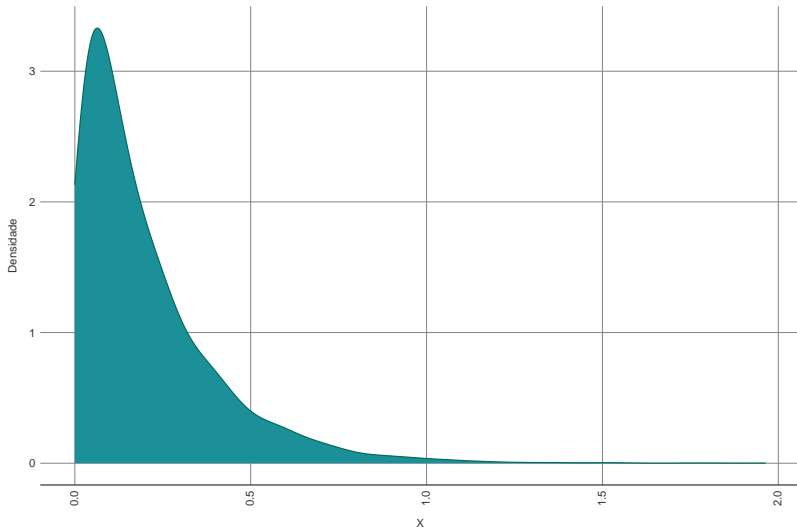
Exemplo: A quantidade de ligações que um serviço de atendimento ao consumidor recebe em 1 hora.

Notas: ¹ Nesse caso, dizemos que $X \sim \text{Exp}(\lambda)$.

Distribuição Exponencial

```
data <- data.frame(X = rexp(10000, 5))
fig <- ggplot(data) +
  geom_density(aes(X),
               color=cores$verde_escuro,
               fill=cores$verde_claro,
               adjust=2) +
  labs(title="Distribuição Exponencial",
       y="Densidade") +
  tema_base_fundobranco()
```

Distribuição Exponencial



Normal

Uma V.A. X segue um modelo **Normal** se sua **função densidade** $f(x)$ é dada por:

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

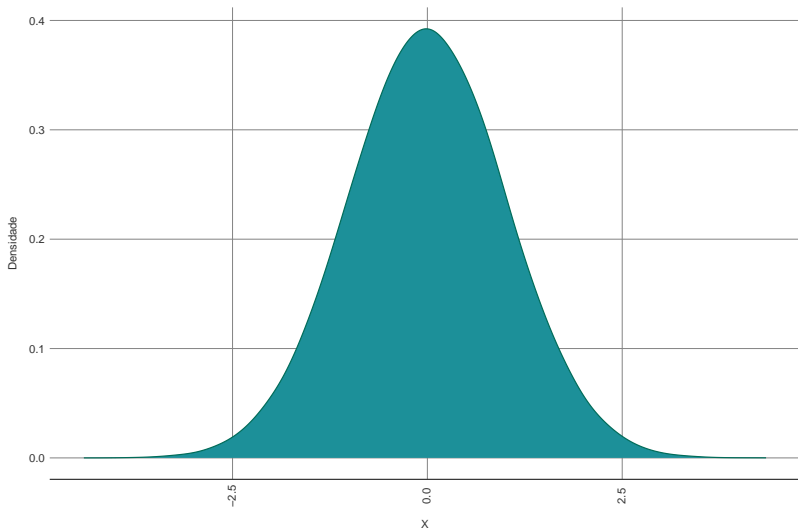
Exemplo: Número de beneficiários do PBF em um município.

Notas: ¹ Nesse caso, dizemos que $X \sim N(\mu, \sigma^2)$. ² Sempre que $X \sim N(\mu, \sigma^2)$, $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. ³ Quando $\log(X) \sim N(\mu, \sigma^2)$, dizemos que X tem distribuição log-normal. ⁴ A distribuição Normal é imensa em estatística. Ela serve como modelo para quantidades de interesse em Inferência Estatística e também é usada em aproximações.

Distribuição Normal

```
data <- data.frame(X = rnorm(100000, 0, 1))
fig <- ggplot(data) +
  geom_density(aes(X),
               color=cores$verde_escuro,
               fill=cores$verde_claro,
               adjust=2) +
  labs(title="Distribuição Normal",
        y="Densidade") +
  tema_base_fundobranco()
```

Distribuição Normal



Valor Esperado e Variância

Definição de Parâmetro

Formalmente, um **parâmetro é uma constante que caracteriza uma família de distribuições** e, a partir disso, caracteriza uma população que tenha um Processo Gerador de Dados (PGD) orientado por essa distribuição.

- ✚ Uma Normal, por exemplo, é caracterizada por ter valor esperado (ou esperança) μ e variância σ^2 .
- ✚ Portanto, qualquer realização de um PGD que siga uma $N(\mu, \sigma^2)$ terá a influência desses parâmetros.
- ✚ Diversos parâmetros podem ser relevantes para caracterizarmos uma distribuição, mas a **esperança e a variância são os mais comumente utilizados**.

Valor Esperado

Intuitivamente, o valor esperado (ou esperança) de uma variável X ($E(X)$) **equivale à soma de todos os valores possíveis de X , ponderados pelas suas respectivas probabilidades.**

Formalmente:

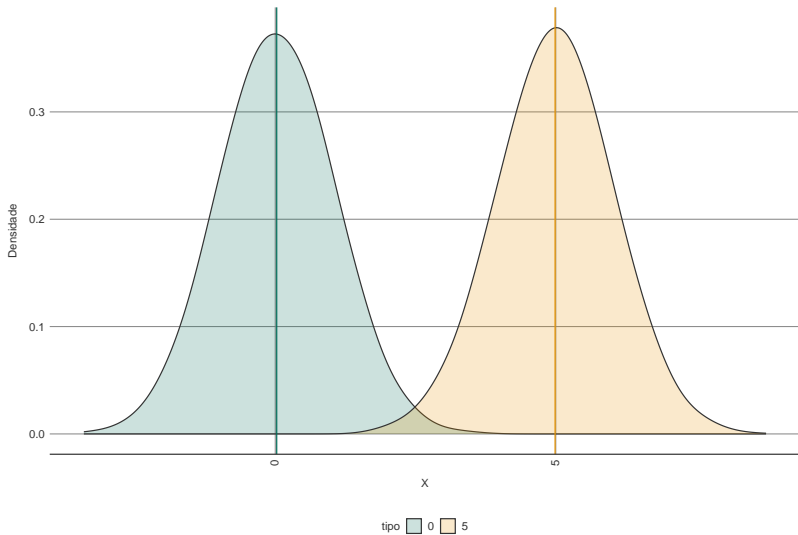
$$E(X) \equiv \mu_x = \begin{cases} \sum_i x_i p_X(x_i), & \text{se } X \text{ tem distribuição discreta} \\ \int_{-\infty}^{\infty} x f(x) d_x = 1, & \text{se } X \text{ tem distribuição contínua} \end{cases}$$

O valor esperado equivale ao momento de ordem 1 de uma V.A. e ele pode ser interpretado como uma medida de centralidade da distribuição.

Centralidade

```
N <- 5000
data_0 <- data.frame(tipo="0",
                     X=rnorm(N))
data_5 <- data.frame(tipo="5",
                     X=rnorm(N, 5))
data <- rbind(data_0, data_5)
fig <- ggplot(data) +
  geom_density(aes(X, fill=tipo),
               color=cores$cinza_escuro,
               alpha=0.2, adjust=2) +
  scale_fill_manual(values=c(cores$verde_escuro,
                             cores$amarelo_escuro)) +
  geom_vline(xintercept = mean(data_5$X),
             color=cores$amarelo_escuro) +
  geom_vline(xintercept = mean(data_0$X),
             color=cores$verde_escuro) +
  labs(title="Diferentes centralidades",
       y="Densidade") +
  tema_base_fundobranco()
```

Diferentes centralidades



Propriedades do Valor Esperado

Sendo a e b constantes e X e Y duas V.As. quaisquer, vale que:

- i. $E(a) = a$;
- ii. $E(aX) = aE(X)$;
- iii. $E(X + b) = E(X) + b$;
- iv. $E(X + Y) = E(X) + E(Y)$; e
- v. !Se X e Y são independentes
 $\Rightarrow E(XY) = E(X) + E(Y)$

Variância

Intuitivamente, a variância é uma medida de quão distante, na média, uma realização específica de uma variável tende a estar do centro da distribuição. Formalmente:

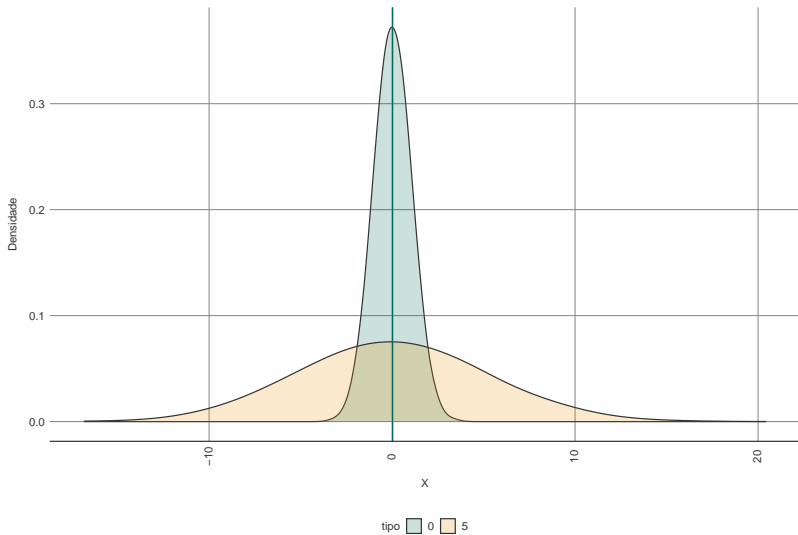
$$Var(X) \equiv \sigma_x^2 = E[(X - \mu_x)^2]$$

A variância equivale ao momento de ordem 2 de uma V.A. e ela pode ser interpretada como uma medida de dispersão da distribuição. O desvio-padrão (σ) é a raiz quadrada da variância e tem a mesma unidade de medida da variância.

Dispersão

```
N <- 5000
data_1 <- data.frame(tipo="0",
                     X=rnorm(N, sd=1))
data_5 <- data.frame(tipo="5",
                     X=rnorm(N, sd=5))
data <- rbind(data_0, data_5)
fig <- ggplot(data) +
  geom_density(aes(X, fill=tipo),
              color=cores$cinza_escuro,
              alpha=0.2, adjust=2) +
  scale_fill_manual(values=c(cores$verde_escuro,
                             cores$amarelo_escuro)) +
  geom_vline(xintercept = mean(data_5$X),
            color=cores$amarelo_escuro) +
  geom_vline(xintercept = mean(data_0$X),
            color=cores$verde_escuro) +
  labs(title="Diferentes dispersoes",
       y="Densidade") +
  tema_base_fundobranco()
```

Diferentes dispersões



Propriedades da Variância

Sendo a e b constantes e X e Y duas V.As. quaisquer, vale que:

- i. $Var(X) \geq 0$;
- ii. $Var(a) = 0$;
- iii. $Var(aX) = a^2 Var(X)$;
- iv. $Var(X + b) = Var(x)$; e
- v. Se X e Y são independentes
 $\Rightarrow Var(X + Y) = Var(X) + Var(Y)$

Parâmetro e estimador

População e Amostra

- ❖ **População:** grupo completo (exaustivo) de indivíduos que se deseja estudar
- ❖ **Amostra:** um subconjunto da população
 - ❖ usamos a amostra porque raramente podemos estudar toda a população de interesse;
 - ❖ existem diferentes processos de amostragem;
 - ❖ a amostra pode ou não ser representativa da população

Estimadores

Um **estimador** é uma função que associa a uma amostra um **número (estimativa)**, com o objetivo de determinar o valor de um **parâmetro populacional**.

- ❖ Como a amostra é um conjunto de realizações de V.A., **todo estimador é também uma V.A.**. Logo, ele tem medidas de centralidade e dispersão
- ❖ Usamos $\hat{\cdot}$ como notação para um estimador. Por exemplo, $\hat{\mu}_x$ é um estimador de μ_x
- ❖ Existem diferentes famílias de estimadores (máxima verossimilhança, método de momentos etc.). Não vamos nos aprofundar nisso, mas saibam que: **todo parâmetro de interesse pode ter múltiplos estimadores**

Lei dos Grandes Números

Sejam Y_1, Y_2, \dots, Y_N V.As. com esperança finita e y_1, y_2, \dots, y_N um conjunto de realizações dessas V.As.. Então, pela Lei dos Grandes Números (LGN), vale que:

$$\frac{1}{N} \sum_{i=1}^N y_i - \mu_y \xrightarrow{p} 0$$

Intuitivamente: a média amostral $(\frac{1}{N} \sum_{i=1}^N y_i)$ converge para a média populacional (μ_y) conforme aumenta o tamanho da amostra.

Teorema do Limite Central

Sejam Y_1, Y_2, \dots, Y_N V.As. iid. com esperança μ e variância σ^2 , e y_1, y_2, \dots, y_N um conjunto de realizações dessas V.As.. Então, pelo Teorema do Limite Central (TLC), vale que:

$$\frac{\frac{1}{N} \sum_{i=1}^N y_i - \mu_y}{\sigma \sqrt{n}} \xrightarrow{d} N(0, 1)$$

Intuitivamente: a distribuição da média amostral converge para uma distribuição normal conforme aumenta o tamanho da amostra.

Hands on

Caso Desenrola

- ❖ Suponha que o governo queira desenvolver um programa para reduzir o endividamento da população.
- ❖ Para tanto, ele pretende criar um fundo público, que será usado para pagar 50% do valor das dívidas de pessoas físicas, desde que o endividado aceite pagar pelos 50% restante.
- ❖ O governo precisa definir o montante total a ser aportado nesse fundo. Contudo, o governo não sabe exatamente o número de endividados nem o valor médio da dívida deles.
- ❖ Como podemos ter uma ideia do tamanho amostral necessário para uma boa estimativa?

Simulando um PGD

- ❖ Primeiro, vamos definir nossa tolerância para uma *boa estimativa*
 - ❖ Suponhamos que um erro de $\pm 2\%$ no valor de aporte do fundo
- ❖ Agora vamos definir alguns parâmetros para o nosso PGD
 - ❖ Suponhamos uma população de 200 milhões
 - ❖ Suponhamos que a proporção de endividados na população seja de 15%
 - ❖ Suponhamos que, dentre os endividados, o valor esperado da dívida - $E(D)$ - é igual BRL 130 + ϵ e $\epsilon \sim EXP(\lambda)$ e que $\lambda = 5 * 10^{-5}$.

Amostra: 100 mil

Simulação

```
set.seed(seed)
sample_n=100000
data <- data.frame("id"=paste("cpf",
                              1:sample_n,
                              sep="_")) %>%
  mutate(endividado = rbinom(sample_n, 1, 0.15),
         divida = endividado * (130+rexp(sample_n, 5*10^(-5)))
head(data, 3)
```

##	id	endividado	divida
## 1	cpf_1	0	0.00
## 2	cpf_2	0	0.00
## 3	cpf_3	1	10939.57

Endividamento médio

```
# Endividamento populacional
```

```
mu <- 130 + 1/(5*10^(-5))
```

```
mu
```

```
## [1] 20130
```

```
# Endividamento médio amostral
```

```
mu_hat <- sum(data$divida)/sum(data$endividado)
```

```
mu_hat
```

```
## [1] 20169.45
```

```
# Erro percentual no endividamento médio
```

```
(mu_hat - mu)*100 / mu
```

```
## [1] 0.1959679
```

Proporção de endividados

```
endiv_d <- 0.15
endiv_d_hat <- sum(data$endividado)/nrow(data)
# Proporção de endividados
endiv_d_hat

## [1] 0.14926

# Erro percentual na proporção de endividados
(endiv_d_hat - endiv_d)*100 / endiv_d

## [1] -0.4933333
```

Aporte no fundo

```
aporte <- 200000000 * endiv_d * mu
# Valor necessário do aporte (em milhões)
aporte/1000000
```

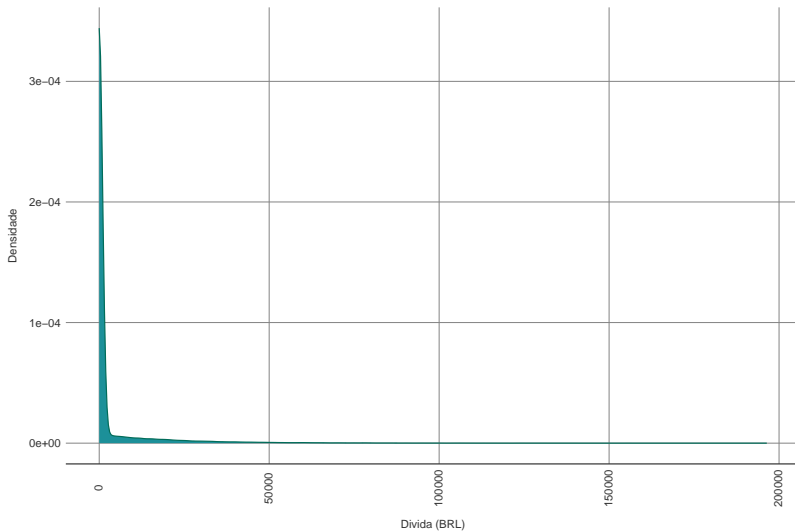
```
## [1] 603900
```

```
aporte_hat <- 200000000 * endiv_d_hat * mu_hat
# Valor estimado do aporte (em milhões)
aporte_hat/1000000
```

```
## [1] 602098.4
```

```
# Erro percentual no aporte
(aporte_hat - aporte)*100 / aporte
```

```
## [1] -0.2983322
```



Amostra: 20

Simulação

```
set.seed(seed)
sample_n=20
data <- data.frame("id"=paste("cpf",
                              1:sample_n,
                              sep="_")) %>%
  mutate(endividado = rbinom(sample_n, 1, 0.15),
         divida = endividado * (130+rexp(sample_n, 5*10^(-5)))
head(data, 3)
```

##	id	endividado	divida
## 1	cpf_1	0	0.000
## 2	cpf_2	0	0.000
## 3	cpf_3	1	4837.632

Endividamento médio

```
# Endividamento populacional
```

```
mu <- 130 + 1/(5*10^(-5))
```

```
mu
```

```
## [1] 20130
```

```
# Endividamento médio amostral
```

```
mu_hat <- sum(data$divida)/sum(data$endividado)
```

```
mu_hat
```

```
## [1] 15428.75
```

```
# Erro percentual no endividamento médio
```

```
(mu_hat - mu)*100 / mu
```

```
## [1] -23.35444
```

Proporção de endividados

```
endiv_d <- 0.15
endiv_d_hat <- sum(data$endividado)/nrow(data)
# Proporção de endividados
endiv_d_hat

## [1] 0.35

# Erro percentual na proporção de endividados
(endiv_d_hat - endiv_d)*100 / endiv_d

## [1] 133.3333
```

Aporte no fundo

```
aporte <- 200000000 * 0.5 * endiv_d * mu
# Valor necessário do aporte (em milhões)
aporte/1000000
```

```
## [1] 301950
```

```
aporte_hat <- 200000000 * 0.5 * endiv_d_hat * mu_hat
# Valor estimado do aporte (em milhões)
aporte_hat/1000000
```

```
## [1] 540006.3
```

```
# Erro percentual no aporte
(aporte_hat - aporte)*100 / aporte
```

```
## [1] 78.83964
```

Convergência pela LGN

Criação de amostra, para um dado N

```
gen_data <- function(sample_size){  
  data <- data.frame("id"=paste("cpf",  
                                1:sample_size,  
                                sep="_")) %>%  
  mutate(N=sample_size,  
         endividado = rbinom(sample_size, 1, 0.15),  
         divida = endividado * (130+rexp(sample_size, 5*10^4)))  
  data  
}
```

Cálculo de estimativa

```
get_estimativa <- function(data){  
  
  N = first(data$N)  
  mu_hat = sum(data$divida)/sum(data$endividado)  
  endiv_d_hat <- sum(data$endividado)/nrow(data)  
  aporte_hat <- 200 * 0.5 * endiv_d_hat * mu_hat  
  
  estimativa <- data.frame(  
    "N"=N,  
    "mu_hat"=mu_hat,  
    "endiv_d_hat"=endiv_d_hat,  
    "aporte_hat"=aporte_hat  
  )  
  estimativa  
}
```

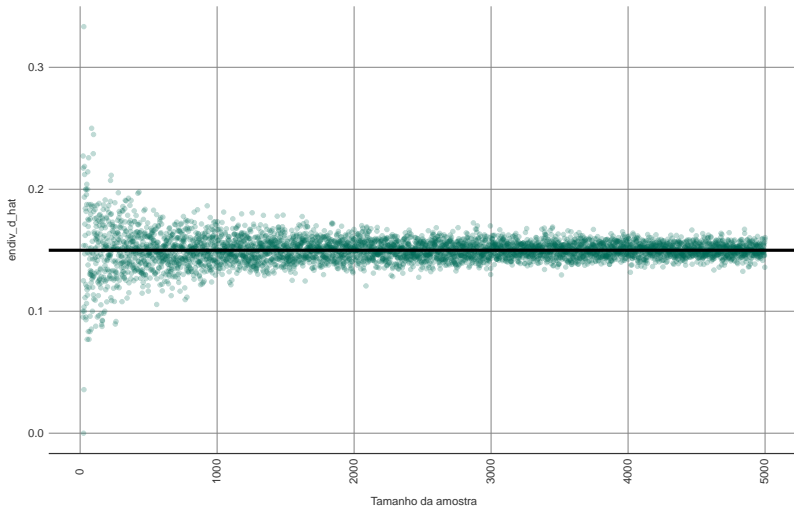
Conjunto de estimativas

```
get_estimativas <- function(sample_size){  
  estimativa <- get_estimativa(gen_data(sample_size))  
}  
estimativas <- lapply(20:5000, get_estimativas) %>%  
  rbindlist()  
head(estimativas, 4)
```

##	N	mu_hat	endiv_d_hat	aporte_hat
##	<int>	<num>	<num>	<num>
## 1:	20	8475.952	0.1000000	84759.52
## 2:	21	6801.356	0.0952381	64774.82
## 3:	22	8677.519	0.2272727	197216.35
## 4:	23	21304.127	0.2173913	463133.19

Convergência da média amostral para o parâmetro populacional

Proporção de endividados



Convergência pelo TLC

Ajustando a função do PGD

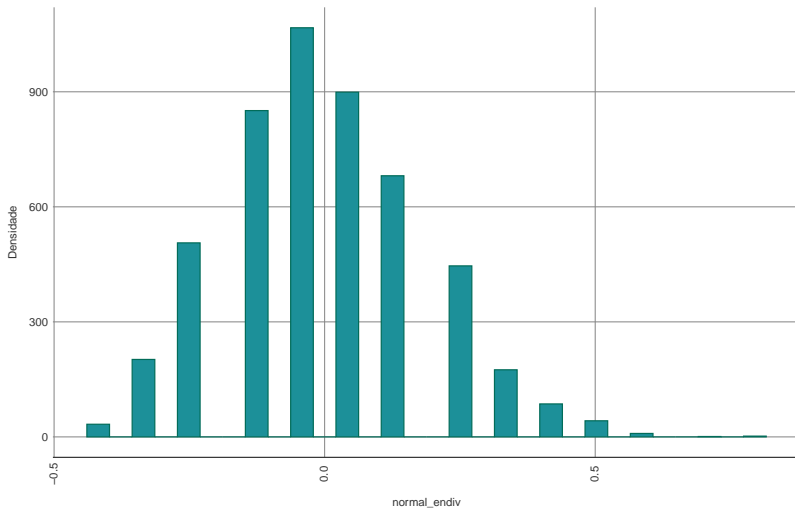
```
gen_data <- function(i, sample_size){  
  data <- data.frame("id"=paste("cpf",  
                                1:sample_size,  
                                sep="_")) %>%  
  mutate(i=i,  
         N=sample_size,  
         endividado = rbinom(sample_size, 1, 0.15),  
         divida = endividado * (130+rexp(sample_size, 5*104))  
  data  
}
```

Função para análise de frequência

```
gen_normal_endiv <- function(sample_size, n_samples=5000){  
  # parametros teóricos  
  mu <- 0.15  
  sigma_endiv <- sqrt(sample_size * 0.15 * (1-0.15))  
  
  data <- lapply(1:n_samples, gen_data, sample_size=sample_size) %>%  
    rbindlist() %>%  
    mutate(normal_endiv =  
             sqrt(sample_size)*((endividado - 0.15)/(sigma_endiv))) %>%  
    group_by(i) %>%  
    summarise(normal_endiv = mean(normal_endiv))  
  
  fig <- ggplot(data) +  
    geom_histogram(aes(normal_endiv),  
                   color=cores$verde_escuro,  
                   fill=cores$verde_claro) +  
    labs(title="Distribuição Dívida Normalizada",  
         subtitle=paste0("N=", sample_size),  
         y="Densidade") +  
    tema_base_fundobranco()  
  
  plot(fig)  
}
```

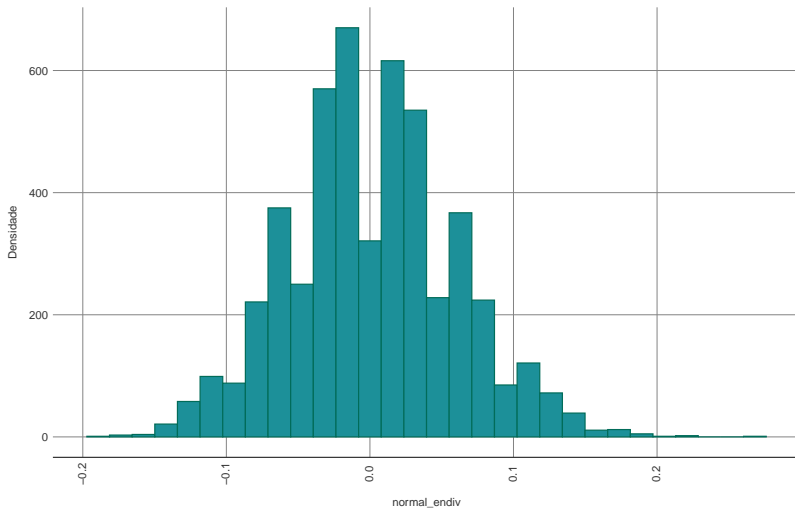
Distribuição Dívida Normalizada

N=30



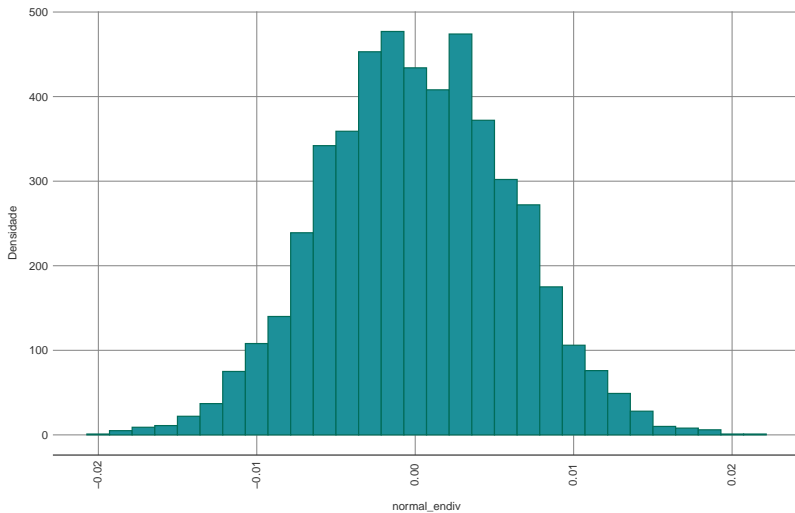
Distribuição Dívida Normalizada

N=300



Distribuição Dívida Normalizada

N=30000



Lista 1

Lista I: Orientações gerais

- ❖ Objetivo: replicar a Figura do Slide 57, mas para valor estimado para o aporte
- ❖ Vocês vão partir do arquivo modelo da *Lista I*, que é um arquivo tipo Rmd
 - ❖ Lembrem de renomear o arquivo para um nome com padrão “Lista_1_nome” antes de submeter!
 - ❖ submetam o relatório gerado (html) e o código (Rmd) antes da próxima aula, enviando-os para o email daniel.sgrimaldi@outlok.com.br

Lista I: Especificações

- ❖ Você(s) vão escrever a função `exec_lista_1`, que terá **obrigatoriamente 2 inputs**
 - ❖ `nome` (nome do aluno);
 - ❖ `max_sample_size` (tamanho da maior amostra a ser considerado)
- ❖ A função `exec_lista_1` terá **obrigatoriamente 3 outputs**
 - ❖ `nome`: um objeto tipo *character*, com o nome do aluno;
 - ❖ `data`: um objeto tipo *data frame*, com os resultados das simulações
 - ❖ `fig`: uma figura tal qual aquela do slide 57, mas para os valores estimados de aporte

Aluno

```
lista_1$nome
```

```
## [1] "Daniel Grimaldi"
```

Base de dados gerada

```
head(lista_1$data, 20)
```

```
##           N  aporte_hat  aporte
##      <int>      <num>    <num>
##  1:    20         NaN 3.0195e+11
##  2:   21 340851410481 3.0195e+11
##  3:   22  85083253103 3.0195e+11
##  4:   23 302938232030 3.0195e+11
##  5:   24 126155529346 3.0195e+11
##  6:   25 374263389477 3.0195e+11
##  7:   26 508778084271 3.0195e+11
##  8:   27 358687736780 3.0195e+11
```