Estatística Aplicada a Ciências Ambientais

# Distribuições de Probabilidade (pt. 2)

Daniel Detzel
detzel@ufpr.br

# Agenda

Distribuições contínuas

- distribuição Normal
- demais distribuições

Aplicação: análise de frequência de cheias

Comentários sobre a estimação de parâmetros

Traçado da FDA empírica

# DISTRIBUIÇÕES DE PROBABILIDADE

## distribuições contínuas

# Distribuições de probabilidade | distribuições contínuas

Distribuições contínuas mais utilizadas:

- Normal
- Log-Normal
- Uniforme
- Exponencial
- Gama
- Gumbel
- Generalizada de Valores Extremos (GEV)
- Weibull
- Fréchet
- Beta, etc.

# Distribuições de probabilidade | distribuição Normal

Distribuição Normal:

Também conhecida como distribuição de Gauss (ou distribuição gaussiana)

desenvolvida para o tratamento de erros aleatórios de medidas experimentais

Empregada para descrever uma variável aleatória que flutua simetricamente em torno de um valor central

A mais importante distribuição contínua de probabilidades

base para testes de hipótese, intervalos de confiança, regressão e correlação

Possui propriedades relevantes aplicáveis a diversos dados (mesmo não normais)

# Distribuições de probabilidade | distribuição Normal

A distribuição Normal é definida por:

$$f_X(x) = \frac{1}{\sqrt{2\pi\theta_1^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right], \text{para } -\infty \leq x \leq \infty$$

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\theta_1^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right] dx$$

onde

$\theta_1$ e $\theta_2$ parâmetros da distribuição

# Distribuições de probabilidade | distribuição Normal

O valor esperado e a variância da distribuição Normal são:

$$E[X] = \mu = \theta_1$$

$$VAR[X] = \sigma^2 = \theta_2^2$$

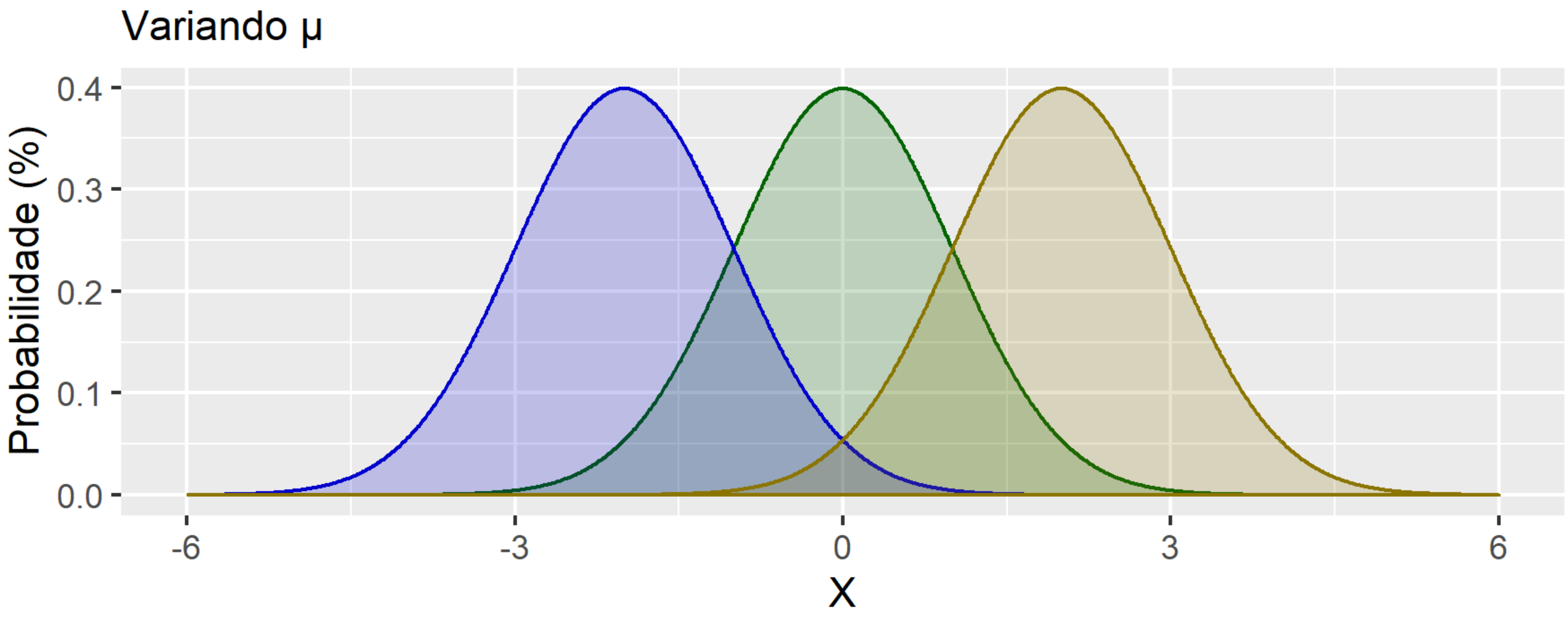
$$\gamma = 0$$

Portanto, pode-se escrever:

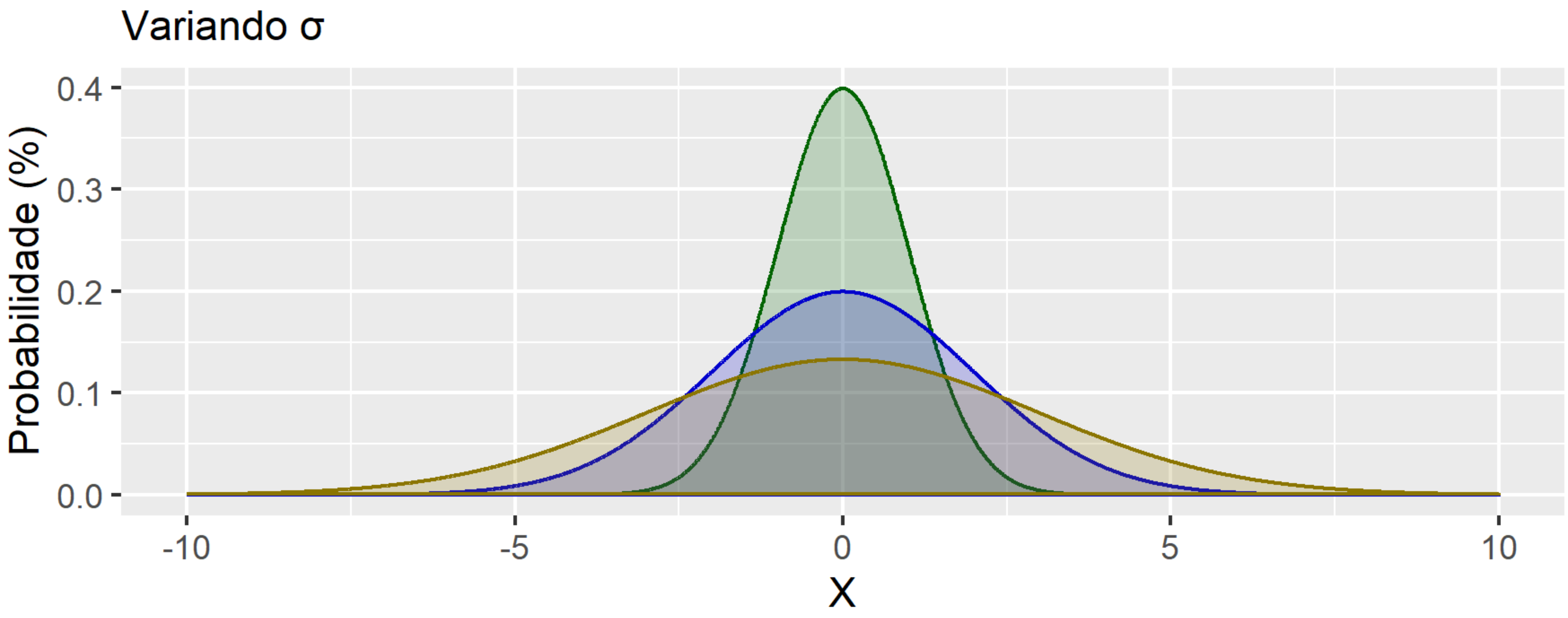
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \text{para } -\infty \leq x \leq \infty$$

# Distribuições de probabilidade | distribuição Normal

Sensibilidade quanto ao parâmetro média:



Sensibilidade quanto ao parâmetro desvio padrão:



# Distribuições de probabilidade | distribuição Normal

A FDP (e a FDA) não tem solução analítica, portanto requer integração numérica

o problema é que essa integração deve ser feita para cada par de parâmetros $\theta_1$ e $\theta_2$, o que é inconveniente

Neste ponto, a propriedade reprodutiva da distribuição se torna conveniente

qualquer combinação linear de $n$ variáveis aleatórias independentes e normalmente distribuídas também segue uma distribuição normal

ex.: se $X$ é normalmente distribuída e $Y = aX + b$, $Y$ também seguirá uma distribuição normal

Portanto, pode-se trabalhar com a variável reduzida $Z$:

$$Z = \frac{X - \mu}{\sigma}$$

Nesse caso, $Z \sim N(0,1)$

lê-se: $Z$ é normalmente distribuído com média 0 e desvio padrão unitário

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right], \text{para } -\infty \leq x \leq \infty$$

# Distribuições de probabilidade | distribuição Normal

A distribuição normal em função da variável reduzida recebe o nome de distribuição normal padrão

sua integração numérica é possível e fornecida em tabelas (ou softwares)

| *z* | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **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 | 0,5606 | 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 |
| **0,3** | 0,6179 | 0,6217 | 0,6255 | 0,6293 | 0,6331 | 0,6368 | 0,6406 | 0,6443 | 0,6480 | 0,6517 |
| **0,4** | 0,6554 | 0,6591 | 0,6628 | 0,6664 | 0,6700 | 0,6736 | 0,6772 | 0,6808 | 0,6844 | 0,6879 |
| **0,5** | 0,6915 | 0,6950 | 0,6985 | 0,7019 | 0,7054 | 0,7088 | 0,7123 | 0,7157 | 0,7190 | 0,7224 |
| **0,6** | 0,7257 | 0,7291 | 0,7324 | 0,7357 | 0,7389 | 0,7422 | 0,7454 | 0,7486 | 0,7517 | 0,7549 |
| **0,7** | 0,7580 | 0,7611 | 0,7642 | 0,7673 | 0,7704 | 0,7734 | 0,7764 | 0,7794 | 0,7823 | 0,7852 |
| **0,8** | 0,7881 | 0,7910 | 0,7939 | 0,7967 | 0,7995 | 0,8023 | 0,8051 | 0,8078 | 0,8106 | 0,8133 |
| **0,9** | 0,8159 | 0,8186 | 0,8212 | 0,8238 | 0,8264 | 0,8289 | 0,8315 | 0,8340 | 0,8365 | 0,8389 |
| **1,0** | 0,8413 | 0,8438 | 0,8461 | 0,8585 | 0,8508 | 0,8531 | 0,8554 | 0,8577 | 0,8599 | 0,8621 |
| **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 |

Imagem: adaptada de Naghettini e Pinto (2007, p. 135)

## No R

```
pnorm(x,mean = mu,sd = sigma)
pnorm(0.1,mean = 1,sd = 0)
[1] 0.5398278
```

# Distribuições de probabilidade | distribuição Normal

## Desvios-padrão em relação à média

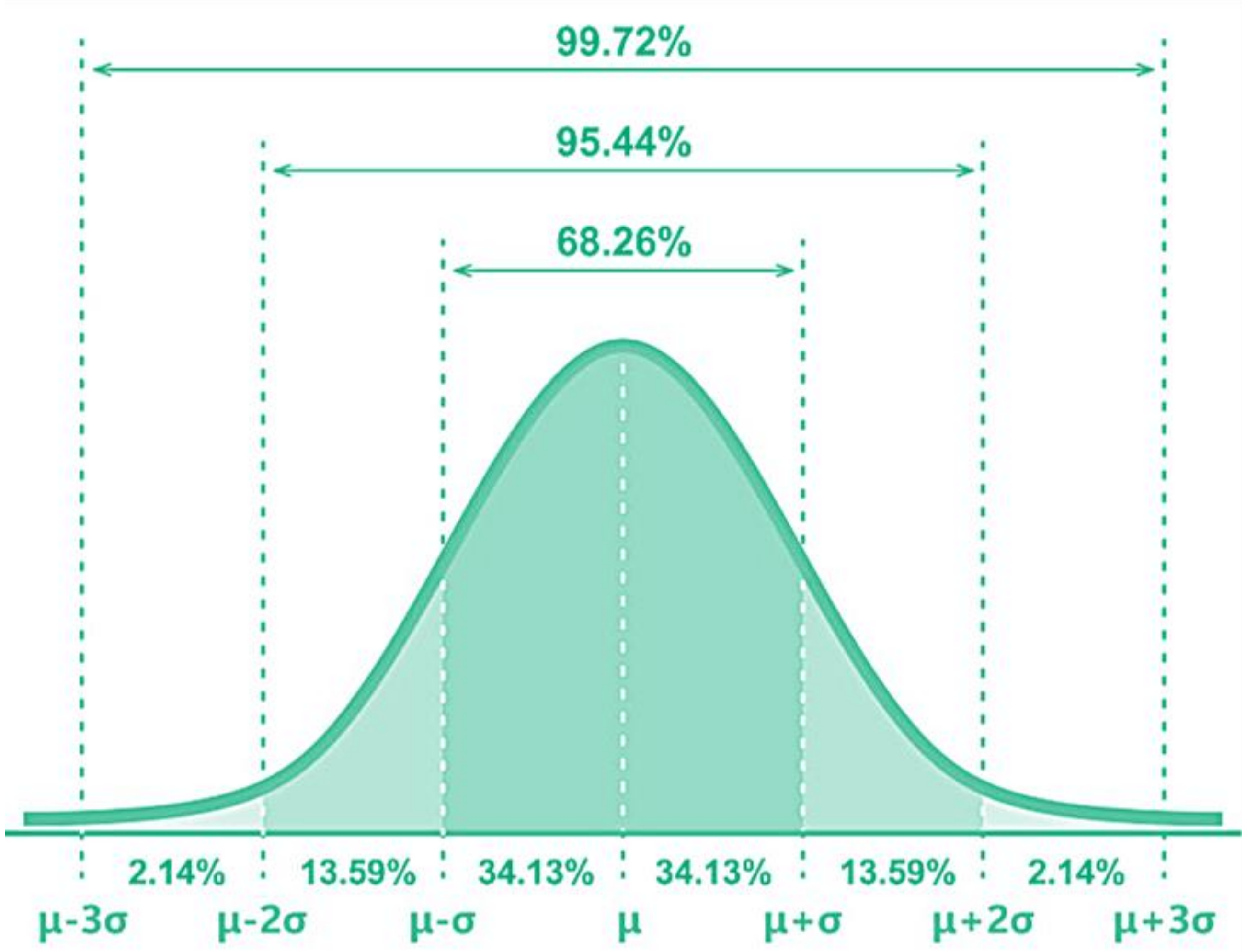



Imagem: adaptada de https://www.simplypsychology.org/z-score.html

Apenas 0,28% da área sob a curva está fora do limite de $3\sigma$

A probabilidade de um valor ser inferior a $\mu - 3\sigma$ é de 0,14%

Na prática, se $X$ tem média maior do que $3\sigma$ a probabilidade de $X$ assumir valores negativos é mínima

# Distribuições de probabilidade | distribuição Normal

Exemplo: Considere uma amostra com uma distribuição normal tal que N(12,25). Qual é a probabilidade de que $X$, retirada dessa amostra, esteja entre 15,6 e 20,4?

Solução:

O que se pede é $\text{prob}(15{,}6 \leq x \leq 20{,}4)$.

A amostra tem parâmetros: $\bar{X} = 12$ e $s = 25$. Assim, as variáveis reduzidas ficam:

$$z_1 = \frac{15{,}6 - 12}{25} = 0{,}12;\, z_2 = \frac{20{,}4 - 12}{25} = 1{,}08$$

# Distribuições de probabilidade | distribuição Normal

Portanto, $\text{prob}(15{,}6 \leq x \leq 20{,}4) \equiv \text{prob}(0{,}12 \leq z \leq 1{,}08)$

$$\text{prob}(0{,}12 \leq z \leq 1{,}08) = \text{prob}(1{,}08) - \text{prob}(0{,}12)$$

| $z$ | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 0,5606 | 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 |
| 0,3 | 0,6179 | 0,6217 | 0,6255 | 0,6293 | 0,6331 | 0,6368 | 0,6406 | 0,6443 | 0,6480 | 0,6517 |
| 0,4 | 0,6554 | 0,6591 | 0,6628 | 0,6664 | 0,6700 | 0,6736 | 0,6772 | 0,6808 | 0,6844 | 0,6879 |
| 0,5 | 0,6915 | 0,6950 | 0,6985 | 0,7019 | 0,7054 | 0,7088 | 0,7123 | 0,7157 | 0,7190 | 0,7224 |
| 0,6 | 0,7257 | 0,7291 | 0,7324 | 0,7357 | 0,7389 | 0,7422 | 0,7454 | 0,7486 | 0,7517 | 0,7549 |
| 0,7 | 0,7580 | 0,7611 | 0,7642 | 0,7673 | 0,7704 | 0,7734 | 0,7764 | 0,7794 | 0,7823 | 0,7852 |
| 0,8 | 0,7881 | 0,7910 | 0,7939 | 0,7967 | 0,7995 | 0,8023 | 0,8051 | 0,8078 | 0,8106 | 0,8133 |
| 0,9 | 0,8159 | 0,8186 | 0,8212 | 0,8238 | 0,8264 | 0,8289 | 0,8315 | 0,8340 | 0,8365 | 0,8389 |
| 1,0 | 0,8413 | 0,8438 | 0,8461 | 0,8585 | 0,8508 | 0,8531 | 0,8554 | 0,8577 | 0,8599 | 0,8621 |
| 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 |

No R

```
pnorm(1.08,mean = 0,sd = 1)
[1] 0.8599289

pnorm(0.12,mean = 0,sd = 1)
[1] 0.5477584
```

# Distribuições de probabilidade | distribuição Normal

$$\therefore \text{prob}(1{,}08) - \text{prob}(0{,}12) = 0{,}8599 - 0{,}5478 = 0{,}3121$$

Assim, a probabilidade de que $X$ esteja entre 15,6 e 20,4 é de 31,2%.

# Distribuições de probabilidade | distribuição Normal

Teorema do Limite Central: a soma, ou a média da soma de VA independentes e igualmente distribuídas tende a uma distribuição normal conforme o tamanho da amostra aumenta

Em geral, $n$ não precisa ser tão grande para a convergência acontecer

Ex.: experimento com um dado:

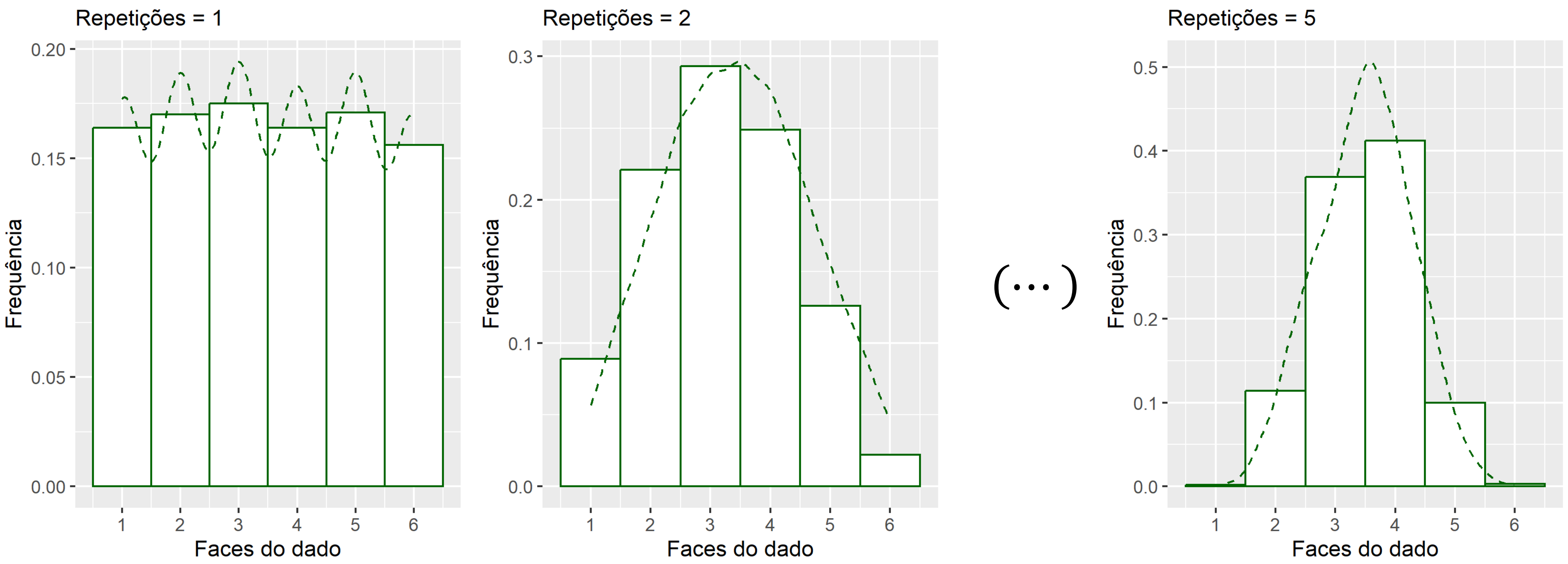
1. Lançar 1000 vezes e anotar os resultados
2. Lançar outras 1000 vezes e fazer a média dos resultados entre (1) e (2)

(...)

5. Lançar outras 1000 vezes e fazer a média dos resultados entre (1), (2), ..., (5)

# Distribuições de probabilidade | distribuição Normal

## Resultados



Em outros casos, $n$ precisa ser suficientemente grande

como regra geral, adota-se $n = 30$

Na prática, a aplicabilidade do teorema está na possibilidade de convergência para uma normal da soma, ou média, de um número suficientemente grande de componentes aleatórios sem ter o conhecimento das distribuições individuais (ou distribuições marginais)

Ex. de aplicação: distribuição dos totais anuais precipitados em um local

a distribuição normal pode ser candidata por força do teorema do limite central

Distribuição Log-Normal:

Utilizada para variáveis aleatórias resultantes de ações multiplicativas de componentes aleatórios independentes, ou seja:

$$X = X_1 \cdot X_2 \cdot \cdots X_n$$

Uma variável $Y = \ln X = \ln X_1 + \ln X_2 + \cdots + \ln X_n$, por força do teorema do limite central, irá tender a uma variável normalmente distribuída (com $n$ suficientemente grande)

Assim, $X \sim LN\left(\mu_{\ln(X)}, \sigma_{\ln(X)}\right)$

# Distribuições de probabilidade | distribuição Log-Normal

A distribuição log-normal é definida por:

$$f_X(x) = \frac{1}{x\sigma_{\ln(X)}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu_{\ln(X)}}{\sigma_{\ln(X)}}\right)^2\right], \text{para } x > 0$$

$$F_X(x) = \int_0^x \frac{1}{x\sigma_{\ln(X)}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu_{\ln(X)}}{\sigma_{\ln(X)}}\right)^2\right] dx$$

onde

$\mu_{\ln(X)}$ e $\sigma_{\ln(X)}$ parâmetros da distribuição

# Distribuições de probabilidade | distribuição Log-Normal

O valor esperado e a variância da distribuição log-normal são:

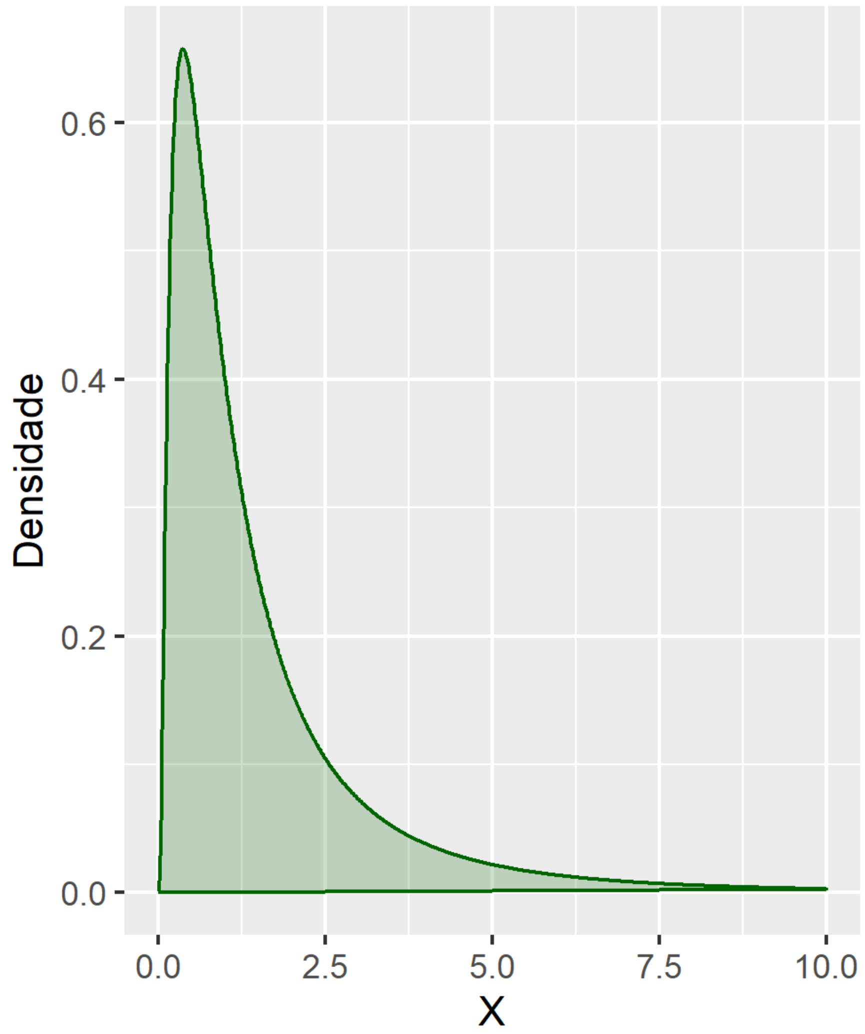
$$E[X] = \mu_X = \exp\left[\mu_{\ln(X)} + \frac{\sigma^2_{\ln(X)}}{2}\right]$$

$$VAR[X] = \sigma_X^2 = \mu_X^2\left[\exp\left(\sigma^2_{\ln(X)}\right) - 1\right]$$

$$\gamma = 3CV_X + (CV_X)^3$$

onde

$$CV_X = \sqrt{\exp\left(\sigma^2_{\ln(X)}\right) - 1}$$



# Distribuições de probabilidade | distribuição Log-Normal

Há ainda a distribuição log-normal a três parâmetros, definida por:

$$f_X(x) = \frac{1}{(x-\Delta)\sigma_{\ln(X)}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x-\Delta)-\mu_{\ln(X)}}{\sigma_{\ln(X)}}\right)^2\right], \text{para } x > 0$$

onde

$$E[X] = \mu_X = \Delta + \exp\left[\mu_{\ln(X)} + \frac{\sigma^2_{\ln(X)}}{2}\right]$$

$$\Delta = \frac{x_{\text{mín}} \cdot x_{\text{máx}} - \tilde{x}^2}{x_{\text{mín}} + x_{\text{máx}} - 2\tilde{x}}$$

$$VAR[X] = \sigma_X^2 = \left[\exp\left(\sigma^2_{\ln(X)}\right) - 1\right] \cdot \exp\left(2\mu_{\ln(X)} + \sigma^2_{\ln(X)}\right)$$

# Distribuições de probabilidade | distribuição Log-Normal

O parâmetro de deslocamento Δ é usado quando se precisa de um limite inferior para variável aleatória

A aplicação da equação proposta para estimação de Δ pode render resultados fisicamente impossíveis se a assimetria da série for negativa

nesse caso, seu valor pode ser atribuído (ex.: $\Delta = x_{\min}$)

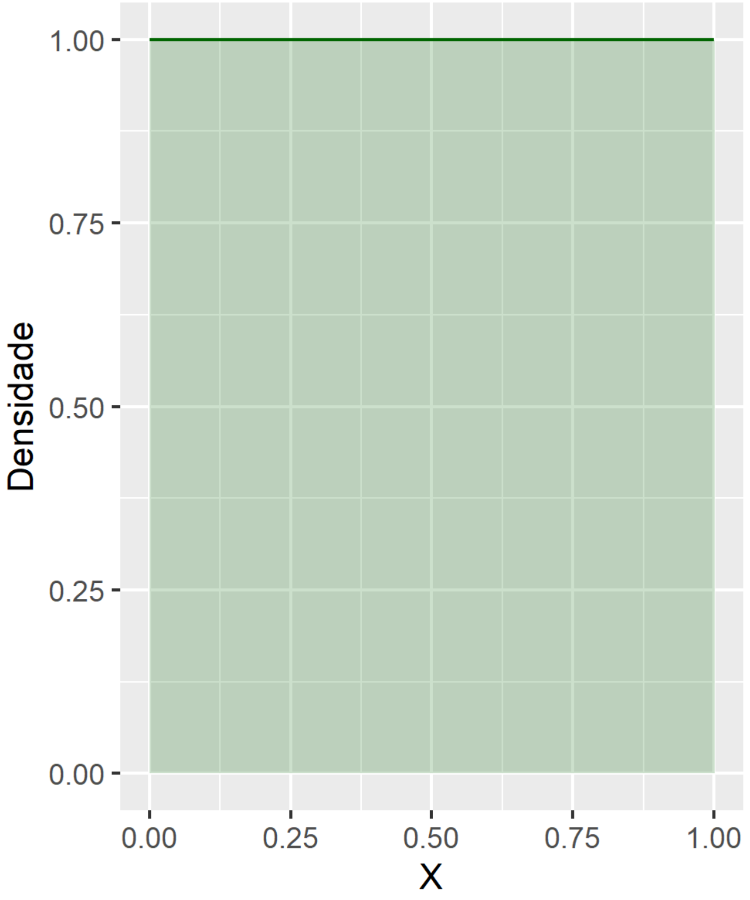
Na prática, o uso da distribuição log-normal é feito aplicando-se uma transformação logarítmica na variável original

essa forma dispensa o uso das equações anteriores

Distribuição Uniforme:

Muito usada em processos de simulação, para a geração de números pseudoaleatórios

$$f_X(x) = \frac{1}{(\beta - \alpha)}, \alpha \leq x \leq \beta$$




onde

$\alpha$ assume $x_{\text{min}}$

$\beta$ assume $x_{\text{máx}}$

## Distribuição Exponencial:

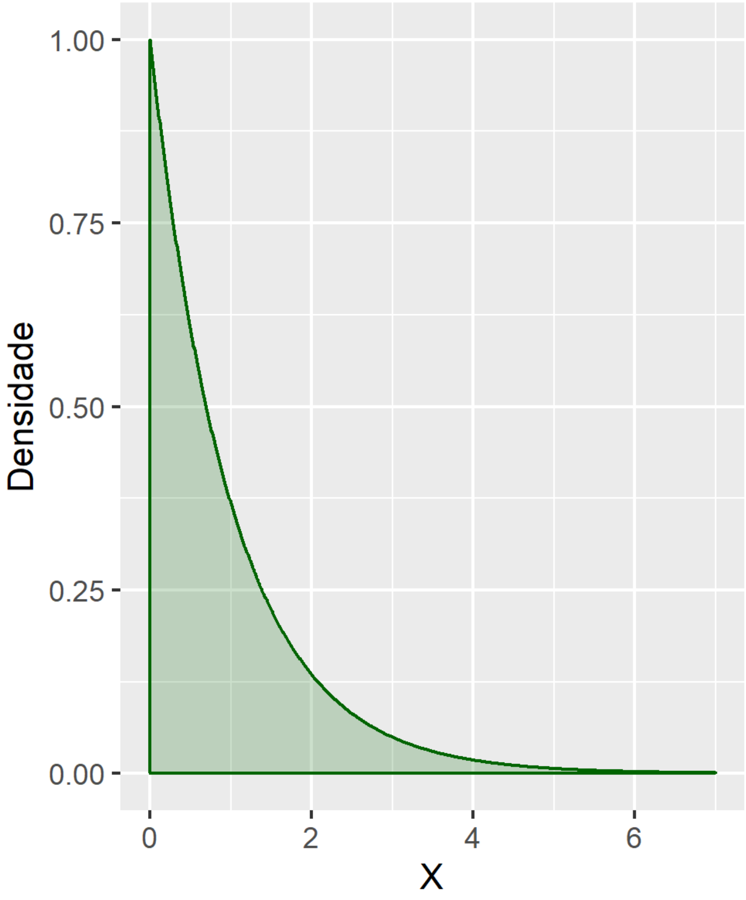
Uma das distribuições mais simples com aplicações em ciências naturais

$$f_X(x) = \lambda \exp(-\lambda x), x \geq 0$$

$$F_X(x) = 1 - e^{-\lambda x}, x \geq 0$$

$$E[X] = \frac{1}{\lambda}$$

$$VAR[X] = \frac{1}{\lambda^2} \qquad \gamma = 2$$



# Distribuições de probabilidade | distribuição Gumbel

Distribuição Gumbel: (para máximas)

Muito utilizada em análise de frequência de cheias

$$f_X(x) = \frac{1}{\alpha}\exp\left[-\frac{x-\beta}{\alpha} - \exp\left(-\frac{x-\beta}{\alpha}\right)\right]$$

$$F_X(x) = \exp\left[-\exp\left(-\frac{x-\beta}{\alpha}\right)\right],$$

$$\text{para } -\infty \leq x \leq \infty, -\infty \leq \beta \leq \infty, \alpha > 0$$

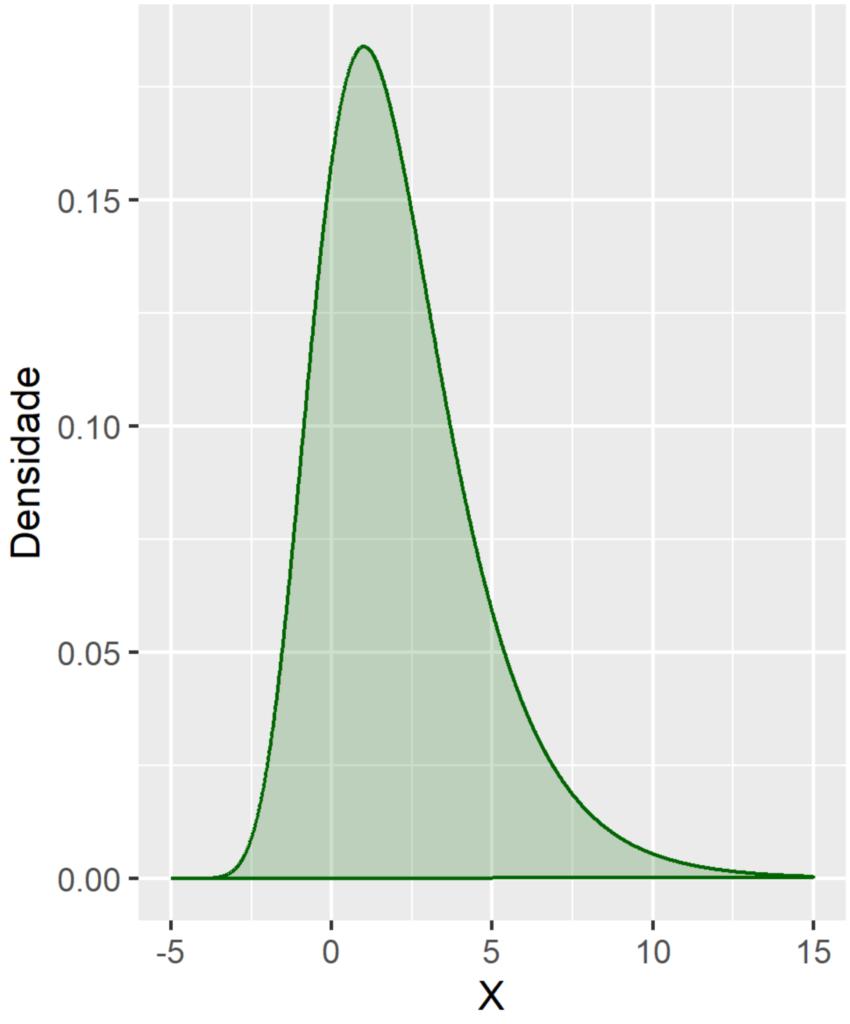
# Distribuições de probabilidade | distribuição Gumbel

O valor esperado, variância e assimetria da distribuição Gumbel, são:

$$E[X] = \beta + 0{,}5772\alpha$$

$$VAR[X] = \frac{\pi^2\alpha^2}{6}$$

$$\gamma = 1{,}1396$$



# Distribuições de probabilidade | distribuição GEV

Distribuição GEV: (generalizada de valores extremos)

Incorpora três formas assintóticas de valores extremos máximos

$$f_X(x) = \frac{1}{\alpha}\left[1 - \kappa\left(\frac{x-\beta}{\alpha}\right)\right]^{1/(\kappa-1)} \exp\left\{-\left[1 - \kappa\left(\frac{x-\beta}{\alpha}\right)\right]^{1/\kappa}\right\}$$

$$F_X(x) = \exp\left\{-\left[1 - \kappa\left(\frac{x-\beta}{\alpha}\right)\right]^{1/\kappa}\right\}$$

onde $\kappa$, $\alpha$ e $\beta$ são os parâmetros de forma, escala e posição, respectivamente

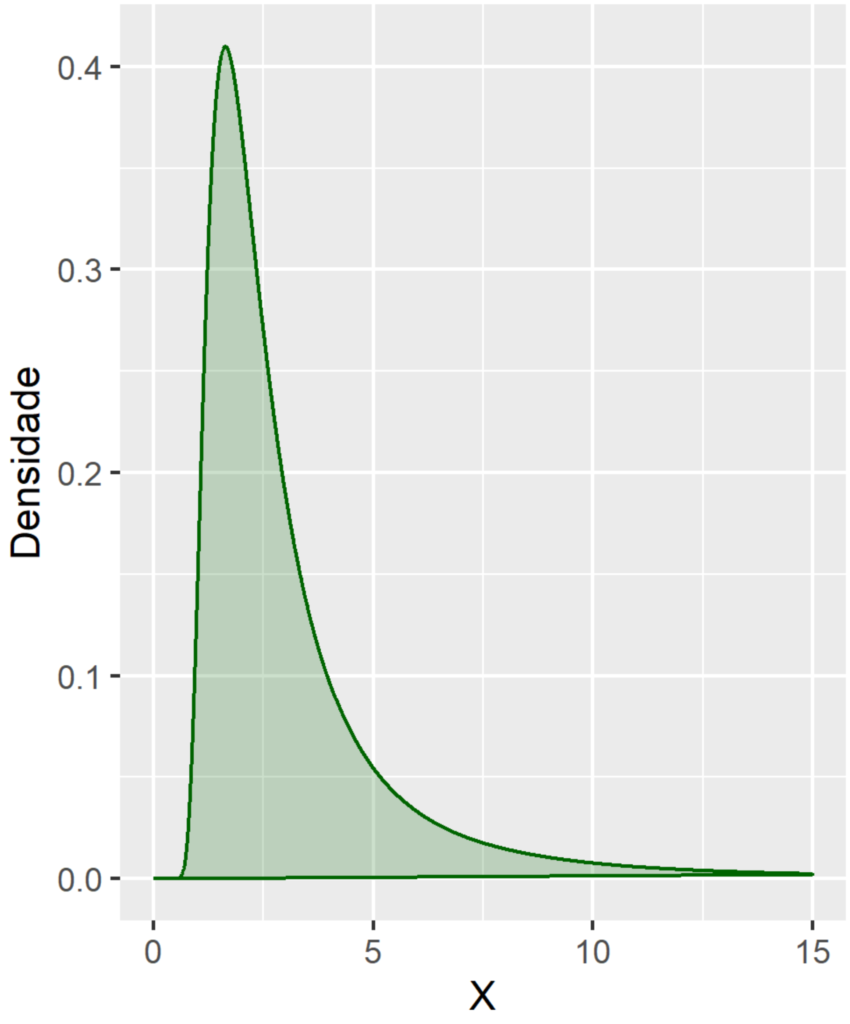
# Distribuições de probabilidade | distribuição GEV

O valor e o sinal de $\kappa$ define a forma assintótica dos valores extremos

se $\kappa < 0$: GEV Tipo II

se $\kappa > 0$: GEV Tipo III

se $\kappa = 0$: distribuição de Gumbel



# Distribuições de probabilidade | distribuição GEV

O valor esperado, variância e assimetria da distribuição GEV, são:

$$E[X] = \beta + \frac{\alpha}{\kappa}[1 - \Gamma(1+\kappa)]$$

$$VAR(X) = \left(\frac{\alpha}{\kappa}\right)^2 [\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]$$

$$\gamma = \langle \text{sinal de } \kappa \rangle \frac{-\Gamma(1+3\kappa) + 3\Gamma(1+\kappa)\Gamma(1+2\kappa) - 2\Gamma^3(1+\kappa)}{[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]^{3/2}}$$

# Distribuições de probabilidade

aplicação: análise de frequência de cheias

# Distribuições de probabilidade | análise de frequência de cheias

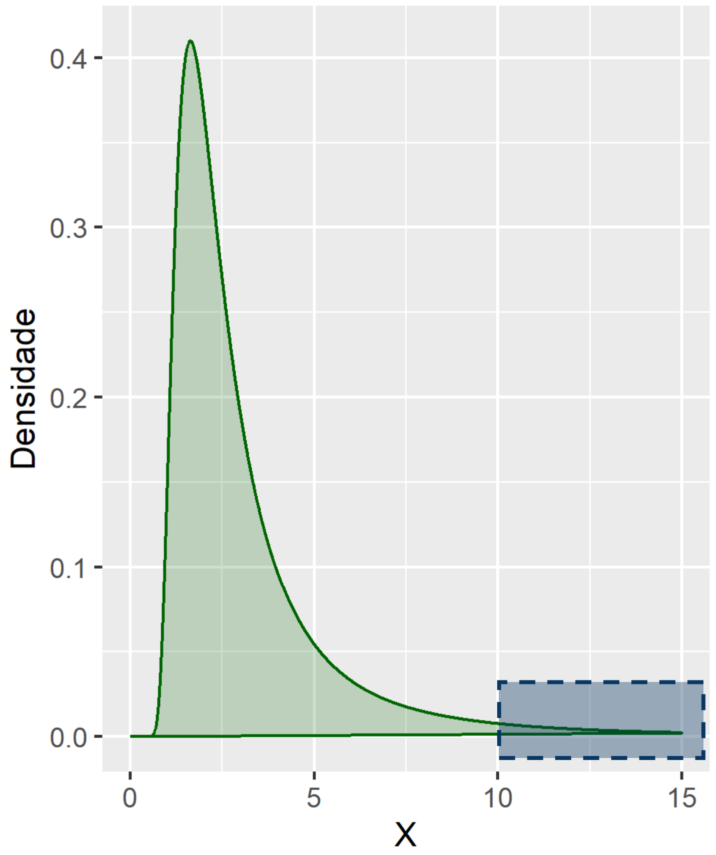
A análise de frequência de cheias tem por objetivo estimar valores de vazão para tempos de retorno elevados, em geral maiores do que a amostra

(essa análise pode ser feita para outras variáveis também)

A atenção se volta para a cauda direita das distribuições

menores probabilidades de ocorrência

maiores valores de vazão



# Distribuições de probabilidade | análise de frequência de cheias

As estimativas são obtidas por meio dos quantis das distribuições

quantil: ponto na distribuição cuja área à esquerda tem probabilidade $p$

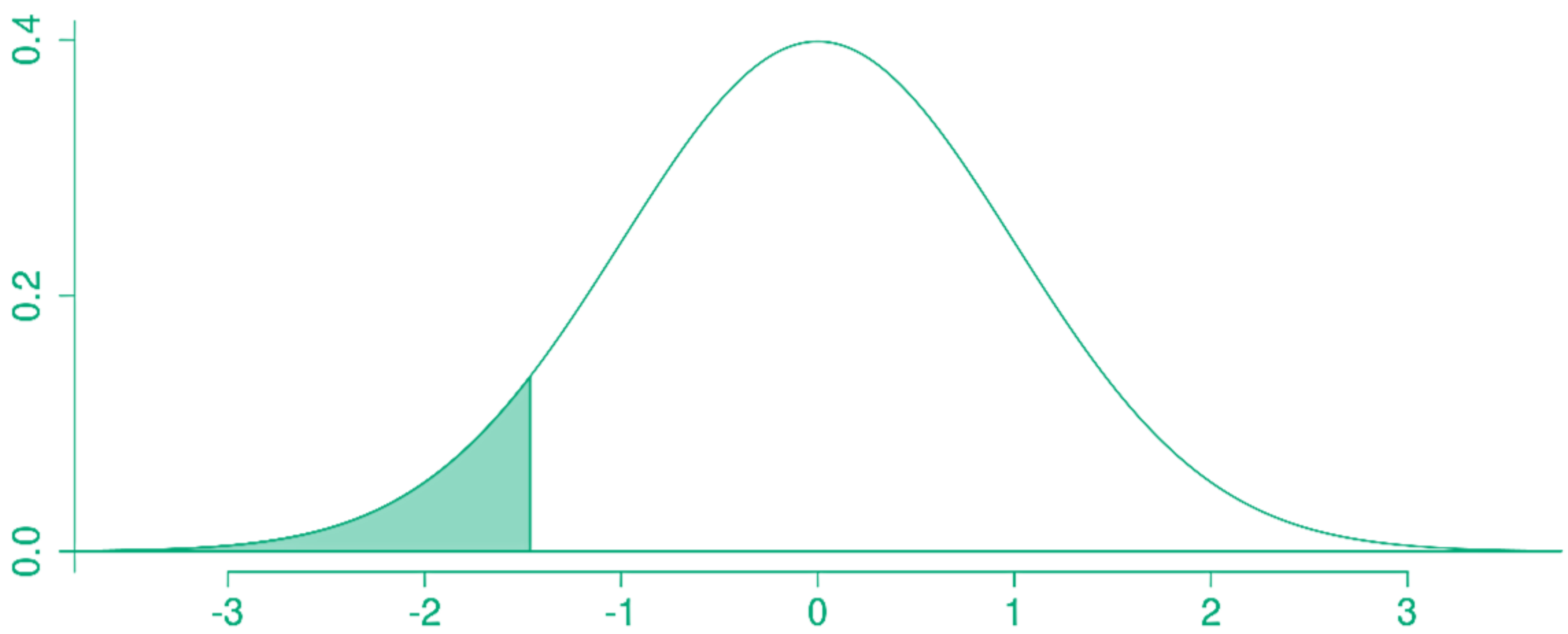



Imagem: adaptada de https://https://passeioaleatorio.shinyapps.io/dist_calc/

# Distribuições de probabilidade | análise de frequência de cheias

Matematicamente, é preciso inverter a FDA da distribuição para explicitar $x$

Ex.: distribuição exponencial

FDA: $F_X(x) = 1 - e^{-\lambda x}$

Quantil: $F_X^{-1}(x) = x = -\frac{\ln F_X(x)}{\lambda}$

Na sequência, atribui-se valores a $F_X(x)$

lembrando que $F_X(x) = 1/T$

# Distribuições de probabilidade | análise de frequência de cheias

Contudo, na análise de frequência de cheias o interesse está na probabilidade de excedência de um valor de vazão, ou seja:

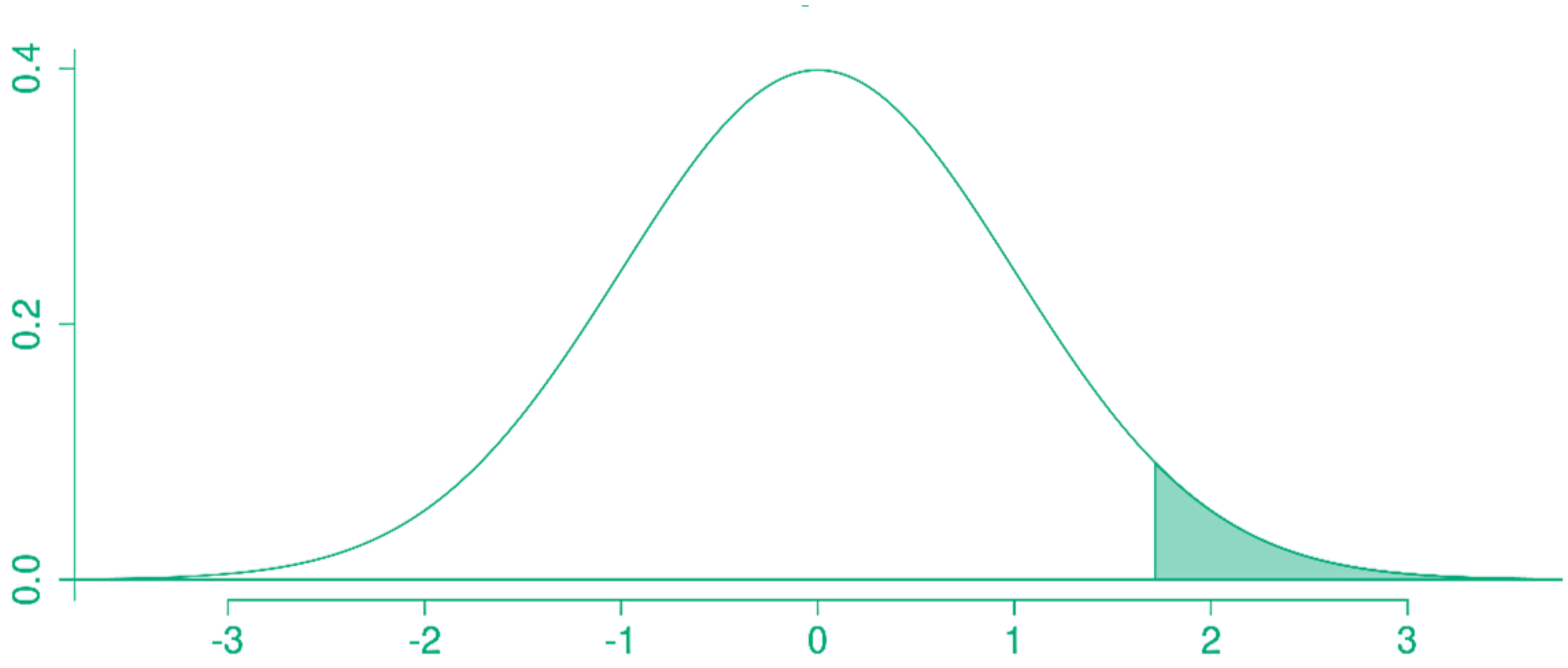



Imagem: adaptada de https://https://passeioaleatorio.shinyapps.io/dist_calc/

Assim, o que se quer calcula é $F'_X(x) = 1 - F_X(x)$

# Distribuições de probabilidade | análise de frequência de cheias

Exercício (proposto): Na análise de uma amostra de vazões de um determinado rio, determinou-se que sua média é de 60 m³/s e seu desvio padrão 12 m³/s. Determine o valor da cheia com 1000 anos de tempo de retorno para as distribuições exponencial e Gumbel.

Solução:

Para resolver o problema é preciso deduzir as equações dos quantis para o caso de probabilidade de excedência ($F'_X(x)$).

Após isso, substitui-se $F'_X(x) = 1/T$.

# Distribuições de probabilidade | análise de frequência de cheias

Para a distribuição exponencial:

$$F_X'(x) = 1 - \left(1 - e^{-\lambda x}\right)$$

Resolvendo para $x$ e substituindo $F_X'(x) = 1/T$:

$$x = -\frac{\ln(1/T)}{\lambda}$$

# Distribuições de probabilidade | análise de frequência de cheias

Para a distribuição de Gumbel:

$$F'_X(x) = 1 - \left(\exp\left[-\exp\left(-\frac{x-\beta}{\alpha}\right)\right]\right)$$

Resolvendo para $x$ e substituindo $F'_X(x) = 1/T$:

$$x = -\alpha \ln\left[-\ln\left(1 - {}^{1}/_{T}\right)\right] + \beta$$

# Distribuições de probabilidade | análise de frequência de cheias

O próximo passo é estimar os parâmetros das distribuições

Para a distribuição exponencial:

$$E[X] = \frac{1}{\lambda} \rightarrow \lambda = \frac{1}{E[X]}$$

Fazendo $E[X] = \bar{X}$, chega-se a:

$$\lambda = \frac{1}{60}$$

# Distribuições de probabilidade | análise de frequência de cheias

Para a distribuição de Gumbel:

$$E[X] = \beta + 0{,}5772\alpha$$
$$VAR[X] = \frac{\pi^2\alpha^2}{6}$$

Fazendo $E[X] = \bar{X}$ e $VAR[X] = s^2$, chega-se a:

$$\alpha = 9{,}356$$
$$\beta = 54{,}599$$

# Distribuições de probabilidade | análise de frequência de cheias

Por fim, basta substituir os valores nas equações dos quantis:

Para a distribuição exponencial:

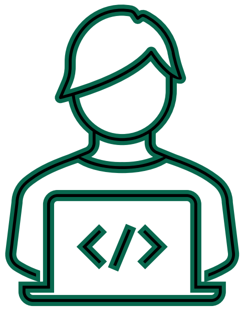
$$x = -\frac{\ln\left(1/1000\right)}{1/60} \cong 415\ \mathrm{m^3/s}$$

Para a distribuição de Gumbel:

$$x = -9{,}356 \ln\left[-\ln\left(1 - {}^{1}\!/_{1000}\right)\right] + 54{,}599 \cong 119\ \mathrm{m^3/s}$$

# Distribuições de probabilidade | análise de frequência de cheias

No R:




## Exponencial:

```
qexp(p, rate = lambda, lower.tail = FALSE)
qexp(1/1000, rate = 1/60, lower.tail = FALSE)
[1] 414.4653
```

## Gumbel: (é preciso o pacote “`ordinal`”)

```
qgumbel(p,location = mu, scale = sigma,lower.tail = FALSE)
qgumbel(1/1000,location = 60, scale = 12,lower.tail = FALSE)
[1] 142.8871
```

Obs.: a diferença entre os resultados para Gumbel se deve ao método utilizado para a estimação dos parâmetros.

# DISTRIBUIÇÕES DE PROBABILIDADE

considerações sobre estimação de parâmetros

# Distribuições de probabilidade | estimação de parâmetros

Os parâmetros condicionam as distribuições de probabilidade.
Matematicamente:

$$F_X(x) \equiv F_X(x|\theta_1, \theta_2, \ldots, \theta_m)$$

onde

$\theta_i$ parâmetros (populacionais) da distribuição ($i = 1,2, \ldots, m$)

Existem diferentes métodos para estimar parâmetros e eles podem produzir valores diferentes para um mesmo parâmetro de uma mesma distribuição

# Distribuições de probabilidade | estimação de parâmetros

A razão é que o estimador $\widehat{\theta}_i$ é função da variável aleatória em questão, o que o torna também uma variável aleatória (portanto, possui dist. de probabilidade)

a notação $\hat{\theta}_i$ denota que o valor é o estimador do parâmetro populacional $\theta_i$

$$F_X(x) \equiv F_X(x|\theta_1, \theta_2, \ldots, \theta_m)$$

Quanto maior o tamanho da amostra, mais $\widehat{\theta}_i$ se aproxima de $\theta_i$

Alternativamente, se várias amostras forem usadas para obter $\widehat{\theta}_i$, o valor médio de todas as estimativas se aproxima de $\theta_i$

# Distribuições de probabilidade | estimação de parâmetros

Os estimadores possuem propriedades importantes:

Não tendenciosidade:
Um parâmetro não tendencioso é aquele que $E[\hat{\theta}] = \theta$

Consistência:
Um estimador consistente é aquele cuja probabilidade de $\hat{\theta}$ ser diferente de $\theta$ por um fator qualquer $\varepsilon$, se aproxima de zero conforme o tamanho da amostra aumenta

Eficiência:

Um estimador eficiente é aquele cuja variância é menor ou igual à variância de outro estimador qualquer

Suficiência:

Um estimador suficiente é aquele que usa toda a informação relevante da amostra, tal que nenhuma outra informação pode ser adicionada por um estimador alternativo

As propriedades auxiliam na caracterização dos métodos de estimação, que podem variar a depender da amostra disponível e da própria distribuição

Os três principais métodos de estimação são:

- método dos momentos
- método da máxima verossimilhança
- método dos momentos-L

A ideia aqui não é explorar a matemática por trás dos métodos (pois existem softwares disponíveis para isso), mas trazer dicas da aplicabilidade de cada um

Método dos momentos:

Tido como o método mais simples para estimação de parâmetros

São de qualidade inferior aos demais métodos (menos eficientes)

Perdem para o método da máxima verossimilhança principalmente em distribuições de dois ou três parâmetros

Por outro lado, para amostras pequenas eles podem ser tão bons (ou até melhores) do que outros estimadores

# Distribuições de probabilidade | estimação de parâmetros

Método da máxima verossimilhança:
É o método mais eficiente (produz estimadores de menor variância)

São estimadores assintoticamente consistentes, suficientes e não tendenciosos

Entretanto, tem aplicabilidade comparável (ou inferior) a outros métodos no caso de amostras pequenas

Exige maior esforço computacional por requerer a solução de sistemas de equações diferenciais parciais

# Distribuições de probabilidade | estimação de parâmetros

Método dos momentos-L:

Alternativa mais eficiente do que o método dos momentos e comparáveis ao método da máxima verossimilhança

São menos sujeitos a variações amostrais (portanto, são robustos)

Exigem menor esforço computacional do que o método da máxima verossimilhança

Boa opção para amostras pequenas

# Distribuições de probabilidade

## distribuições empíricas

# Distribuições de probabilidade | distribuições empíricas

As distribuições empíricas se referem às distribuições dos dados sob análise (portanto, são distribuições amostrais)

São úteis para avaliar a adequação de distribuições teóricas

- ferramentas gráficas para análise de adequação de ajustes

Aqui são abordadas as FDAs empíricas

- a avaliação das FDPs empíricas foi mostrada anteriormente e consiste em histogramas e gráficos de densidade

# Distribuições de probabilidade | distribuições empíricas

Para o traçado da FDA empírica, são precisos dois passos

1. Ordenar a série de forma crescente
2. Atribuir posições de plotagem $p_i$, as quais são definidas por:

$$p_i = \frac{i - a}{n + 1 - 2a}$$

onde

$a$ constante a ser adotada de acordo com o método escolhido

## Valores de $a$ de acordo com a motivação

| **Denominação** | **Valor de $a$** | **Motivação** |
|---|---|---|
| Weibull | 0 | Probabilidades de excedência não tendenciosas para todas as distribuições |
| Blom | 0,375 | Distribuição normal |
| Cunnane | 0,40 | Probabilidades de excedência aproximadamente não tendenciosas para todas as distribuições |
| Gringoten | 0,44 | Distribuição de Gumbel |

Portanto, utilizando as posições de Weibull, tem-se:

$$p_i = \frac{i}{n+1}$$

# Distribuições de probabilidade | distribuições empíricas

Exemplo: FDA empírica da série de vazões médias anuais em Foz do Areia, rio Iguaçu (1931 a 2021)

Original

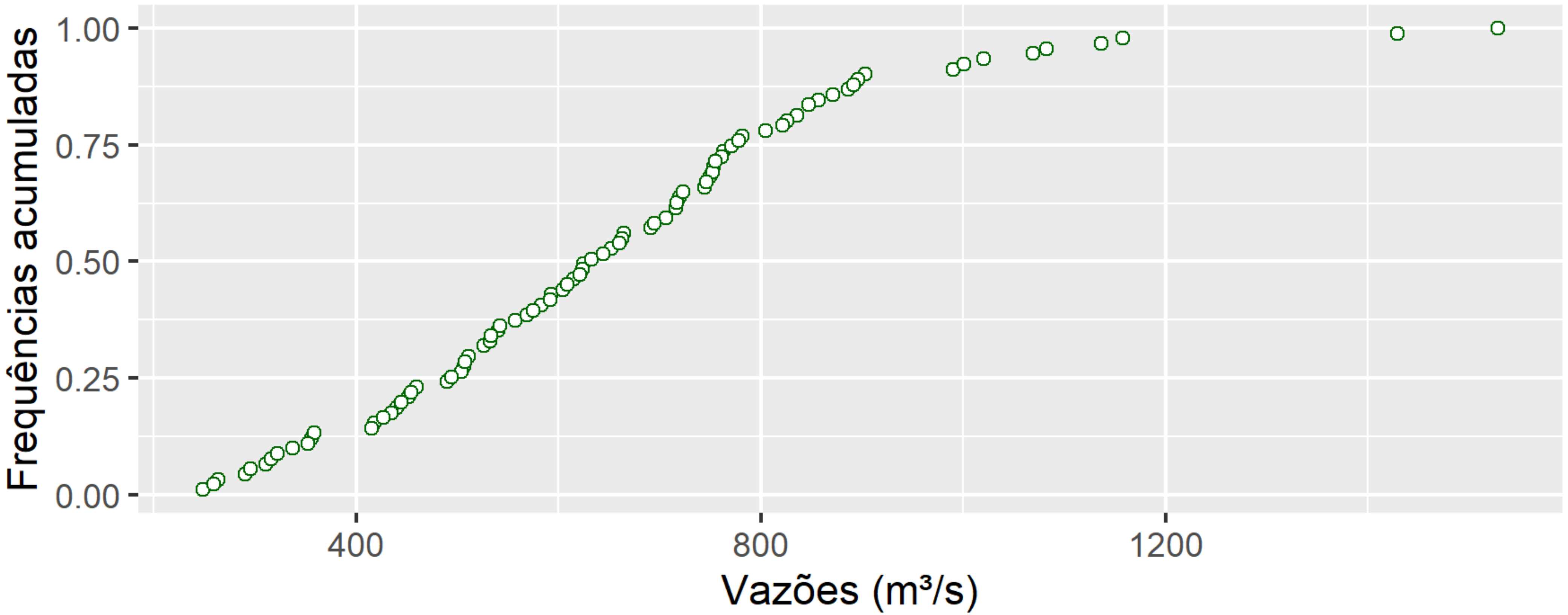
| Anos | $\bar{Q}$ (m³/s) |
|---|---|
| 1931 | 720 |
| 1932 | 753 |
| 1933 | 249 |
| ... | ... |
| 2019 | 533 |
| 2020 | 296 |
| 2021 | 427 |

Ordenada

| Anos | $\bar{Q}$ (m³/s) | $q_i$ |
|---|---|---|
| 1933 | 249 | 0,011 |
| 2006 | 259 | 0,022 |
| 1968 | 264 | 0,033 |
| ... | ... | ... |
| 1990 | 1158 | 0,967 |
| 1998 | 1429 | 0,978 |
| 1983 | 1528 | 0,989 |

# Distribuições de probabilidade | distribuições empíricas

## FDA empírica resultante



# Distribuições de probabilidade | distribuições empíricas

O ajuste a distribuições teóricas pode ser verificado graficamente

Para traçar a FDA teórica:

1. Estimar os parâmetros da distribuição desejada
2. Inverter a FDA da distribuição, atribuindo as posições de plotagem no lugar das probabilidades $F_X(x)$

Para a distribuição normal: (necessário auxílio de tabelas ou softwares)

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

# Distribuições de probabilidade | distribuições empíricas

Exemplo: FDA teórica da série de vazões médias anuais em Foz do Areia, rio Iguaçu (1931 a 2021)

Parâmetros:

$\bar{X} = 656$ m³/s

$s = 244$ m³/s

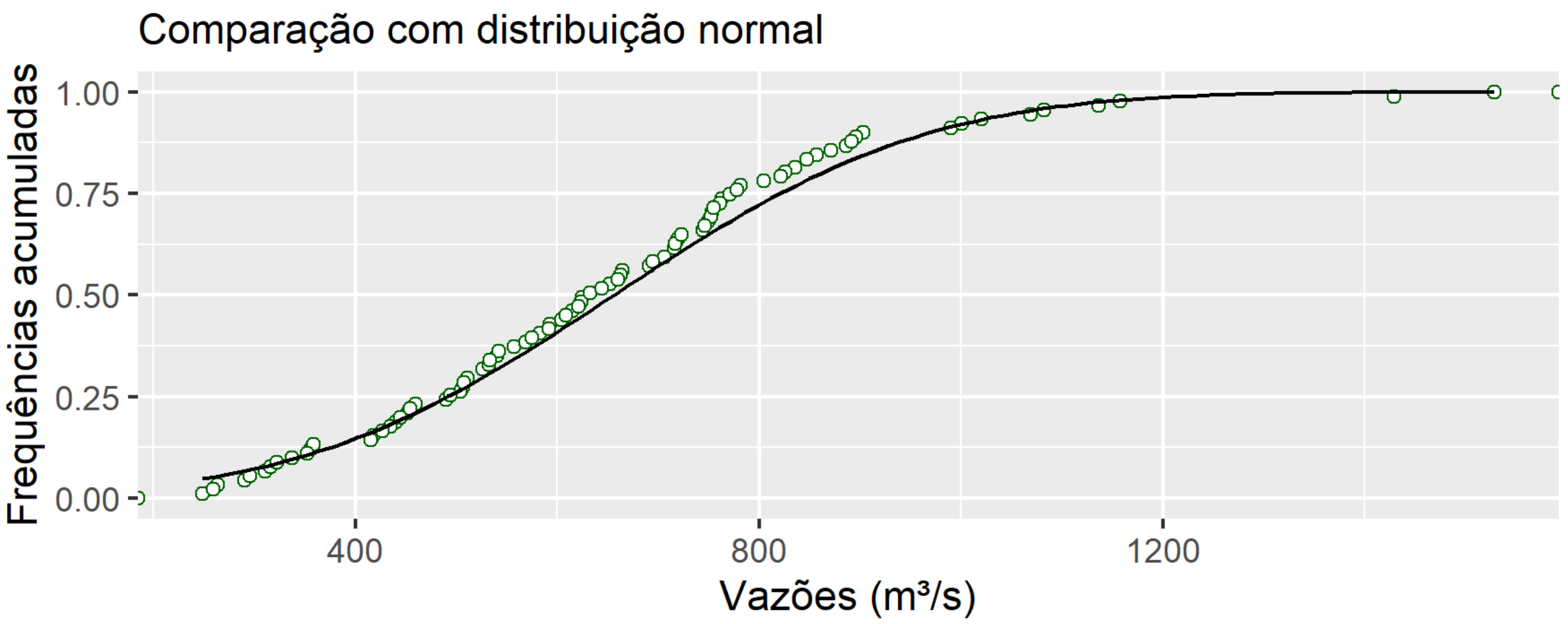
| $q_i$ | $\bar{Q}$ (m³/s) |
|---|---|
| 0,011 | 97 |
| 0,022 | 164 |
| 0,033 | 207 |
| ... | ... |
| 0,967 | 1105 |
| 0,978 | 1148 |
| 0,989 | 1215 |

No R:

```
qnorm(qi, mean = 656, sd = 244)
```

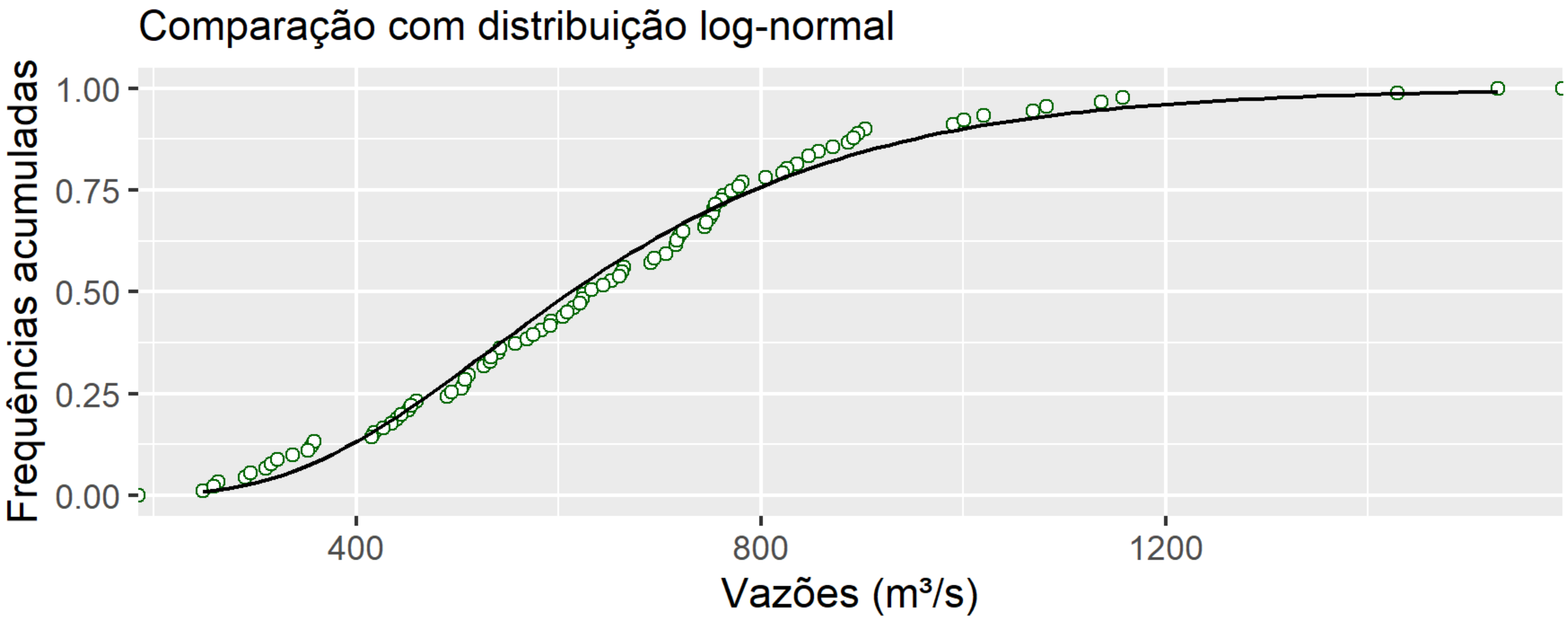
# Distribuições de probabilidade | distribuições empíricas

## FDA empírica vs. teórica resultante



# Distribuições de probabilidade | distribuições empíricas

O mesmo procedimento pode ser aplicado a outras distribuições



# Revisão

A distribuição normal é importante por suas propriedades, mesmo que não tenha aplicabilidade direta em séries de ciências naturais

- reprodutibilidade
- nº de desvios-padrão em relação à média
- Teorema do Limite Central

Distribuições contínuas diversas

- aqui focadas em formas assimétricas

Análise de frequência de cheias

- uma das diversas aplicações das distribuições contínuas

Distribuições empíricas

- úteis para a seleção de modelos teóricos candidatos

# Estatística Aplicada a Ciências Ambientais

Daniel Detzel
detzel@ufpr.br