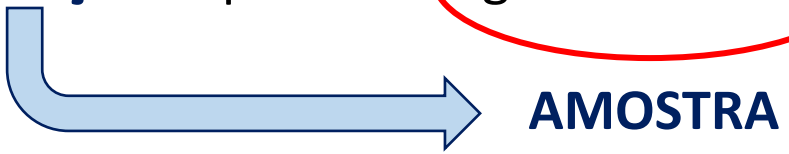
# Estatística Básica e Introdução ao R

Profª. Dra. Natalia Giordani

# 2. Distribuições e amostras

- Estatística
  - Fazer **inferências** sobre a distribuição de alguma variável em uma determinada

**POPULAÇÃO** a partir de alguns elementos dela



**AMOSTRA**

**Modelos probabilísticos**

# 2. Distribuições e amostras

- Estatística
  - Se pudermos supor que distribuição de probabilidades de certa variável possa ser descrita por um modelo probabilístico específico nosso problema se reduz a **estimar os parâmetros** dessa distribuição
  - Há vários modelos probabilísticos
    - Variáveis discretas: função de probabilidade
    - Variáveis contínuas: função densidade de probabilidade

# 2. Distribuições e amostras

- Modelos probabilísticos para variáveis discretas

| Modelo | Parâmetros | Exemplo de uso |
|---|---|---|
| Binomial | n, p | Decisão: comprar / não comprar; clicar / não clicar |
| Poisson | λ | Eventos por unidade de tempo: nº de chamadas telefônicas de uma central em 1h |

# 2. Distribuições e amostras

- Modelos probabilísticos para variáveis contínuas

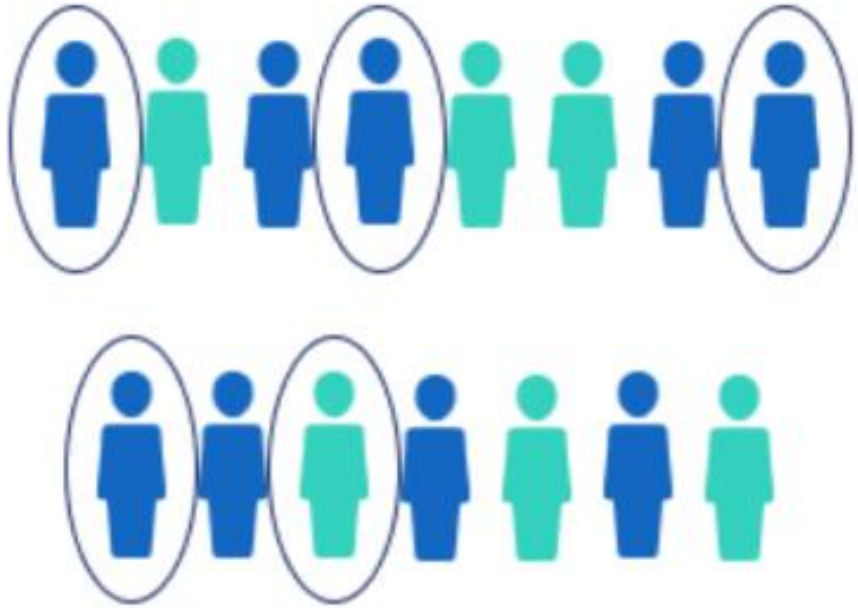
| Modelo | Parâmetros | Exemplo de uso |
|---|---|---|
| Normal | $\mu$, $\sigma$ | Peso de recém-nascidos |
| Exponencial | $\alpha$ | Distância entre um evento e o próximo: tempo entre visitas a um site |
| t-Student | n | Base de referência para distribuição de médias amostrais, diferenças entre duas médias,... |

# 2. Distribuições e amostras

- Dados Amostrais
  - Subconjunto de um conjunto maior (população)
  - Para inferência deve **satisfazer algumas condições**
    - Amostragem probabilística - seleção aleatória
    - Exemplos: Amostragem Aleatória Simples (AAS) e Amostragem Estratificada

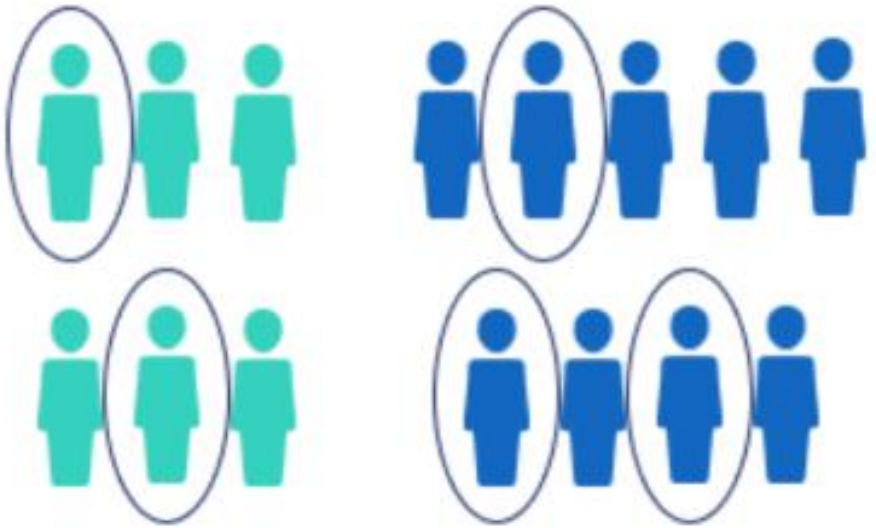
# 2. Distribuições e amostras

- Amostragem Aleatória Simples
  - Todos os membros da população tem a mesma probabilidade de ser incluídos na amostra



# 2. Distribuições e amostras

- Amostragem Estratificada
  - População é dividida em estratos (homogêneos dentro e heterogêneos entre) e é realizada AAS em cada um

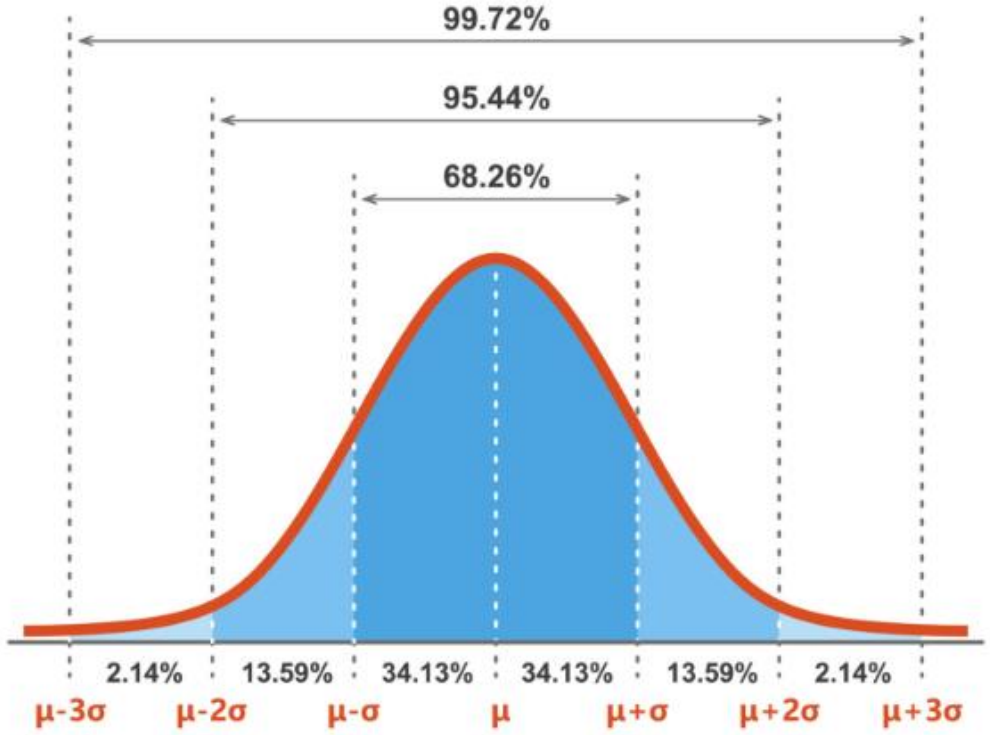


# 2. Distribuições e amostras

- Especificação modelo para inferência
  - Escolha de um modelo probabilístico para representar a variável de interesse (x) na população
- Como?
  - Possibilidade: histograma dos dados da amostra *vs* histogramas teóricos de modelos probabilísticos candidatos
  - Alternativa mais utilizada: **gráficos QQ** (QQ plots)
    - Pontos representam os quantis obtidos das distribuições amostral e teórica
    - Se os dados amostrais forem compatíveis com o modelo probabilístico proposto os pontos devem estar dispostos em torno de uma reta

# 2. Distribuições e amostras

- **Gráficos QQ**
  - Distribuição Normal é requisito para muitos métodos
  - Características:
    - Formato de sino
    - Simétrica em relação a média
    - A média e mediana tem o mesmo valor
    - A área sob a curva representa 1 ou 100%
    - Aproximadamente 68% dos valores de X estão entre os pontos (μ – σ) e (μ + σ)
    - Aproximadamente 95% dos valores de x estão entre os pontos (μ – 2σ) e (μ + 2σ)
    - Aproximadamente 99,7% dos valores de x estão entre pontos (μ – 3σ) e (μ + 3σ)

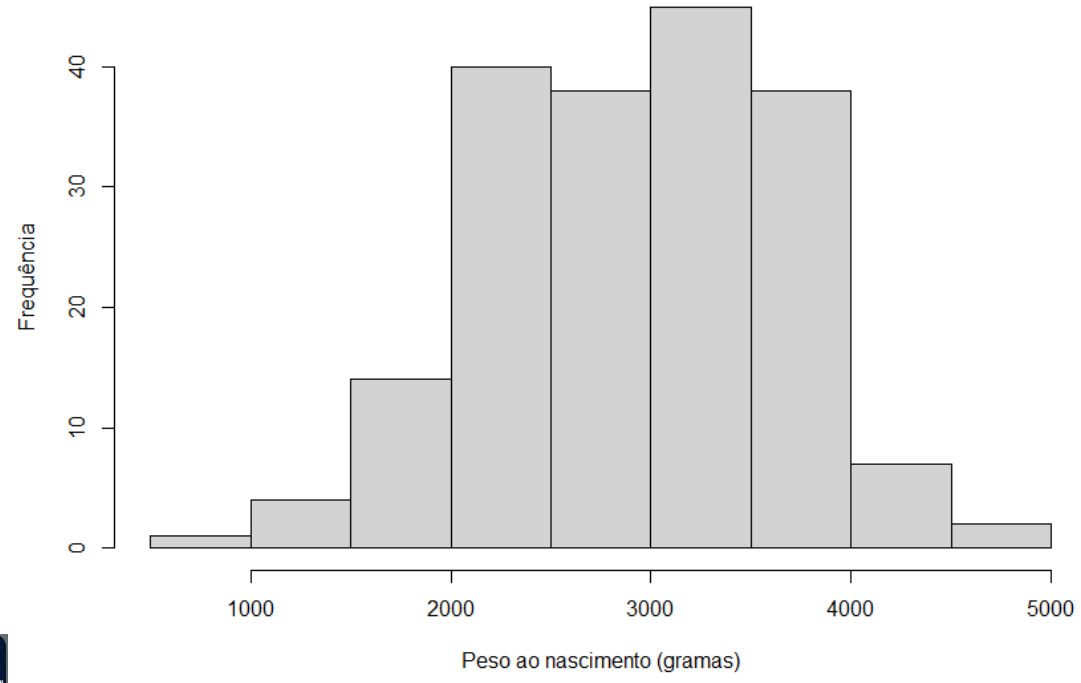


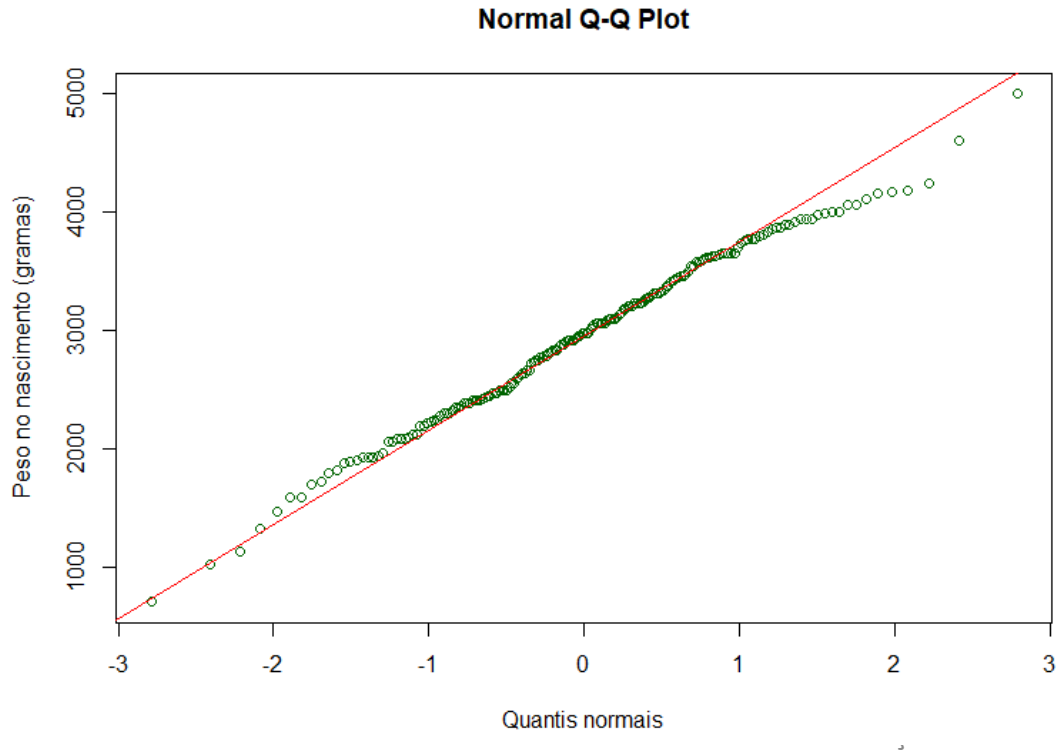
# 2. Distribuições e amostras

- **Gráficos QQ**
  - Distribuição Normal é requisito para muitos métodos
  - Características:
    - Resultado importante dado pelo **Teorema do limite central**: para qualquer que seja a distribuição da variável de interesse, a distribuição das médias amostrais tenderá a uma distribuição Normal à medida que o tamanho de amostra cresce.

# 2. Distribuições e amostras

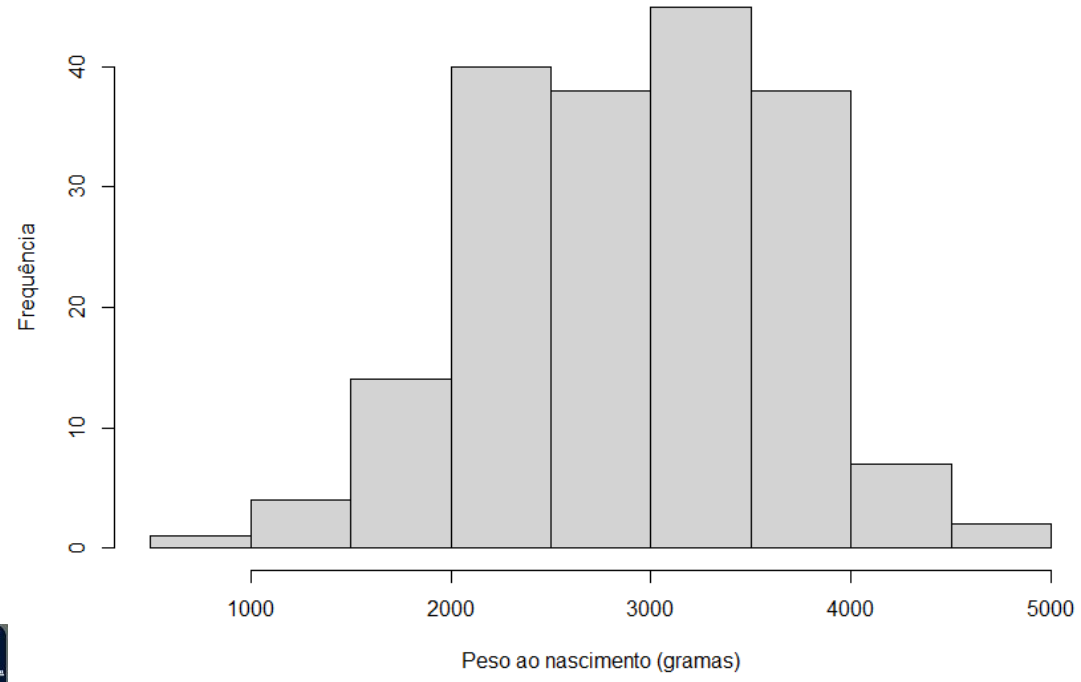
- **Gráficos QQ**
  - Exemplo: peso de recém-nascidos

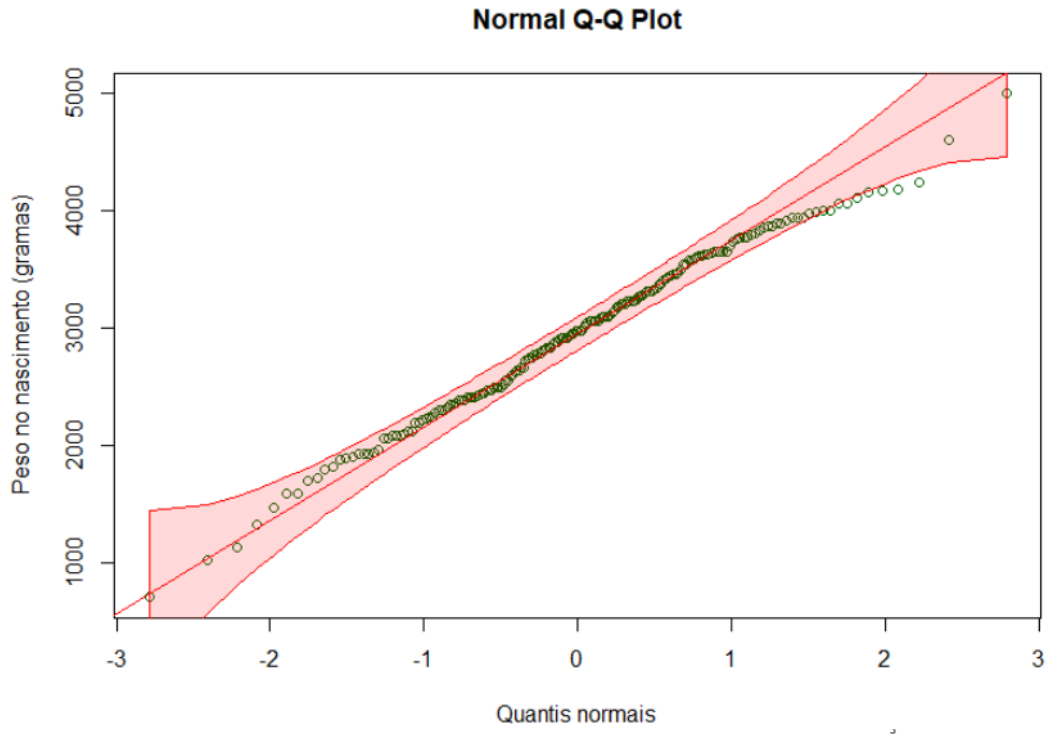


# 2. Distribuições e amostras

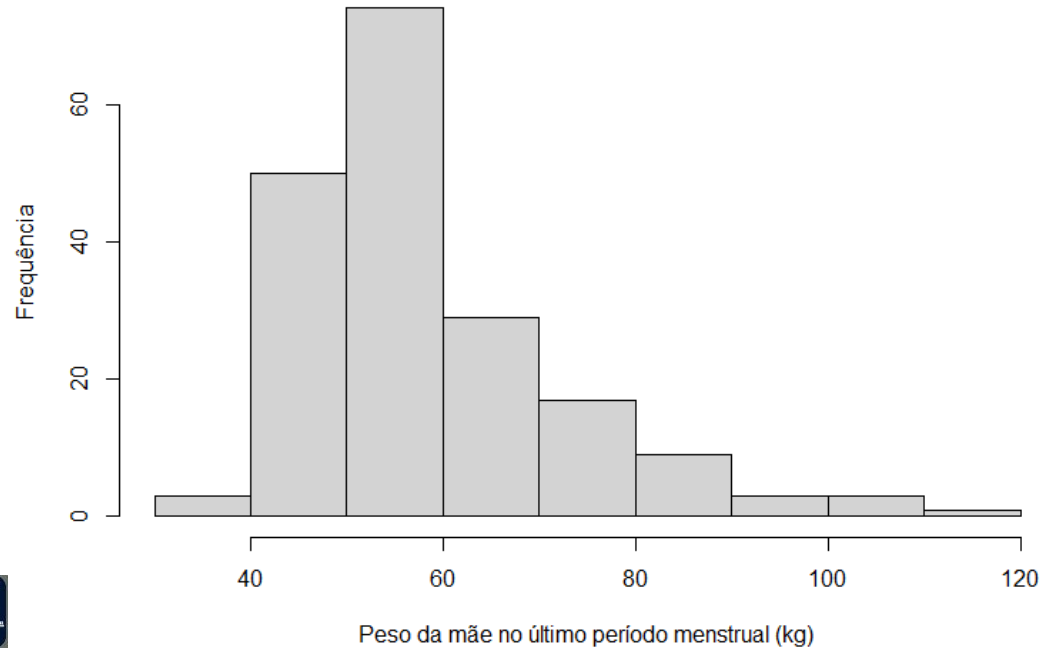
- **Gráficos QQ**
  - Exemplo: peso de recém-nascidos

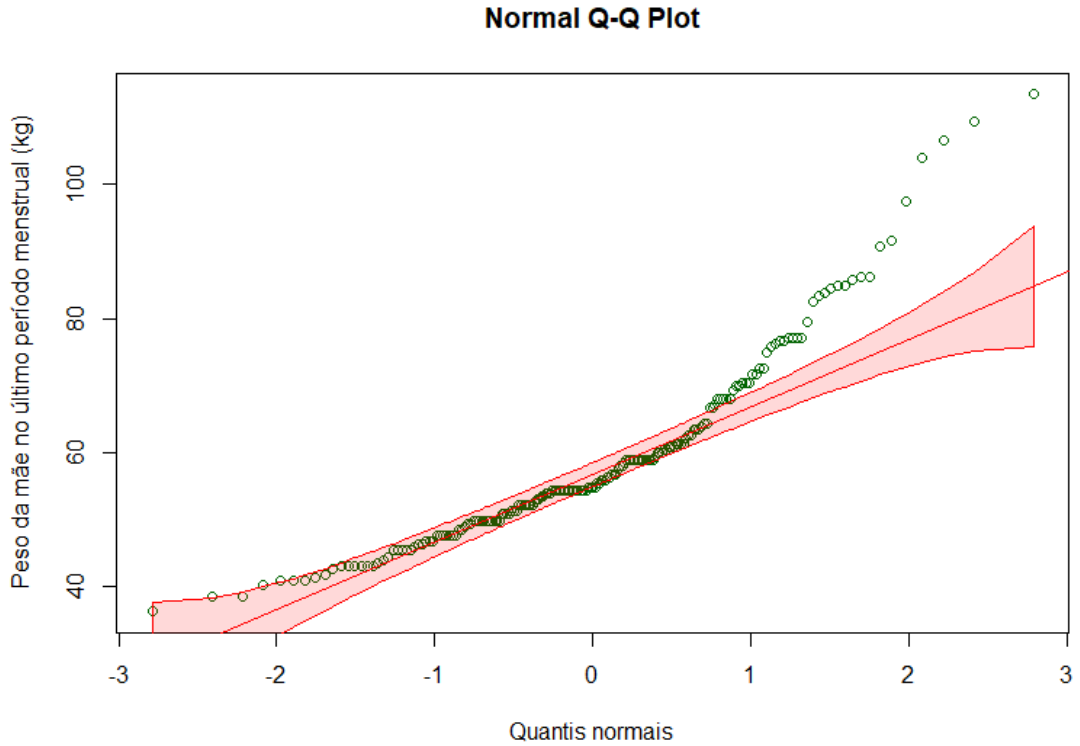


# 2. Distribuições e amostras

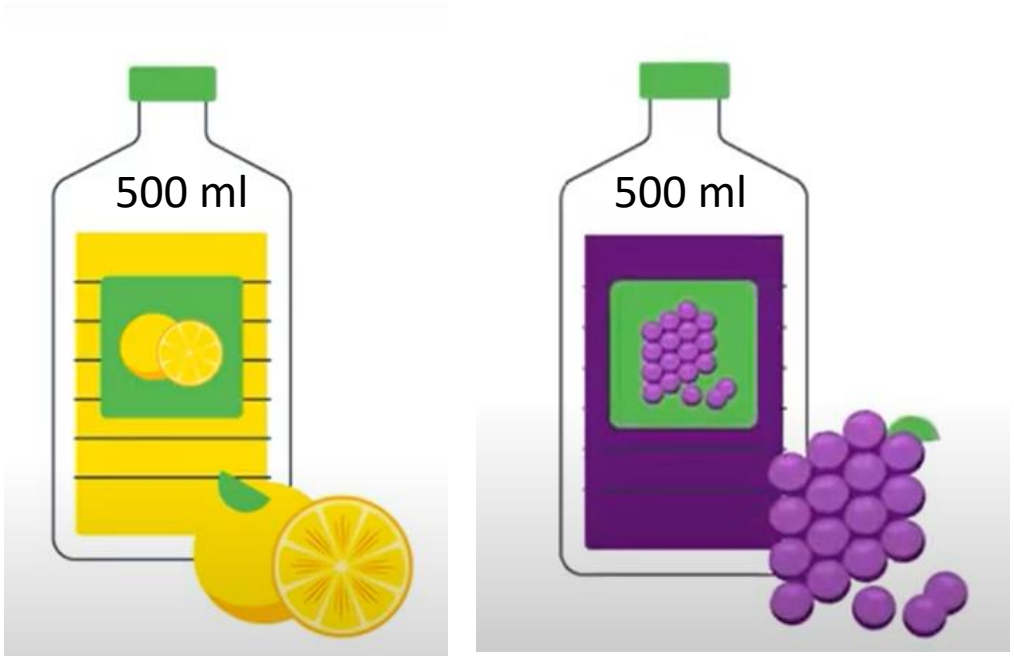
- **Gráficos QQ**
  - Exemplo: peso da mãe no último período menstrual (kg)



# 2. Distribuições e amostras

- **Desvio padrão e erro padrão**



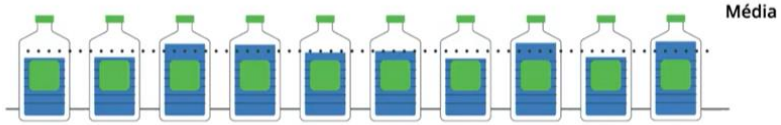

# 2. Distribuições e amostras

- **Desvio padrão e erro padrão**

- Selecionar 10 garrafas de cada suco (aleatoriamente) e verificar seu volume com um medidor

| Garrafa | Suco | Qtde (ml) |
|---|---|---|
| 1 | Uva | 446 |
| 2 | Uva | 450 |
| 3 | Uva | 554 |
| ⋮ | ⋮ | ⋮ |
| 18 | Laranja | 506 |
| 19 | Laranja | 502 |
| 20 | Laranja | 495 |

| Suco | Média |
|---|---|
| Uva | 500 |
| Laranja | 500 |



# 2. Distribuições e amostras

- **Desvio padrão e erro padrão**

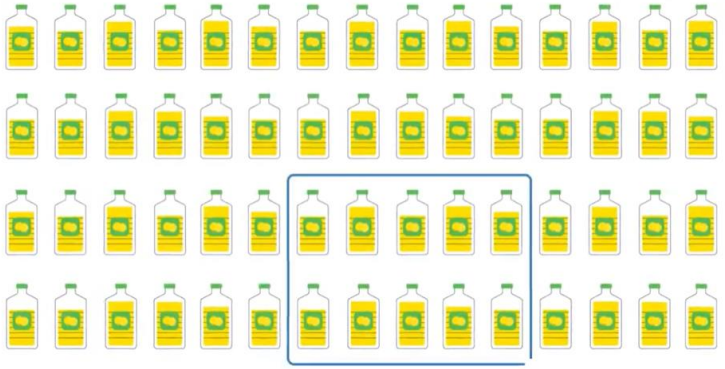
- Selecionar 10 garrafas de cada suco (aleatoriamente) e verificar seu volume com um medidor

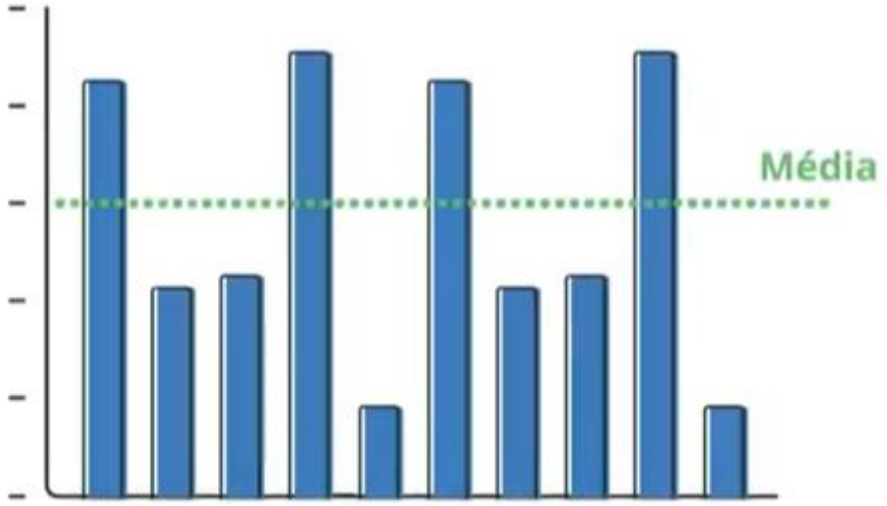
| Suco | Média | Desvio Padrão |
|---|---|---|
| Uva | 500 | 52,7 |
| Laranja | 500 | 6 |

  - Quanto menor o desvio padrão: mais concentrados próximos a média estão as observações (mais homogênea é a amostra)

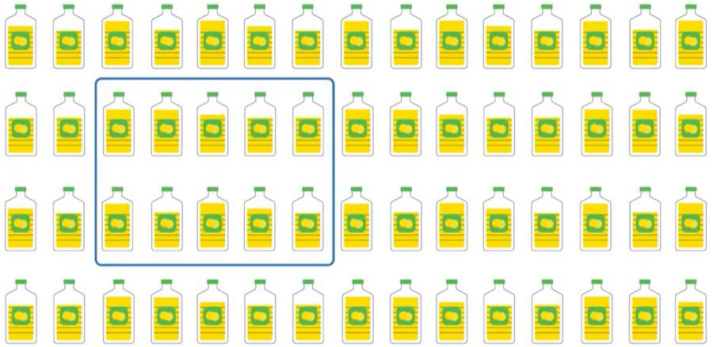
# 2. Distribuições e amostras

- **Desvio padrão e erro padrão**
  - Amostra considerada






  - E se mudasse a amostra, o que aconteceria com a média?



$$\text{Erro padrão} = \frac{desvio\ padrão}{\sqrt{n}}$$

# 2. Distribuições e amostras

- **Desvio padrão e erro padrão**

| Suco | Média | Desvio Padrão | Erro padrão |
|---|---|---|---|
| Uva | 500 | 52,7 | $\frac{52,7}{\sqrt{10}} = 16.7$ |
| Laranja | 500 | 6 | $\frac{6}{\sqrt{10}} = 1.9$ |

# 2. Distribuições e amostras

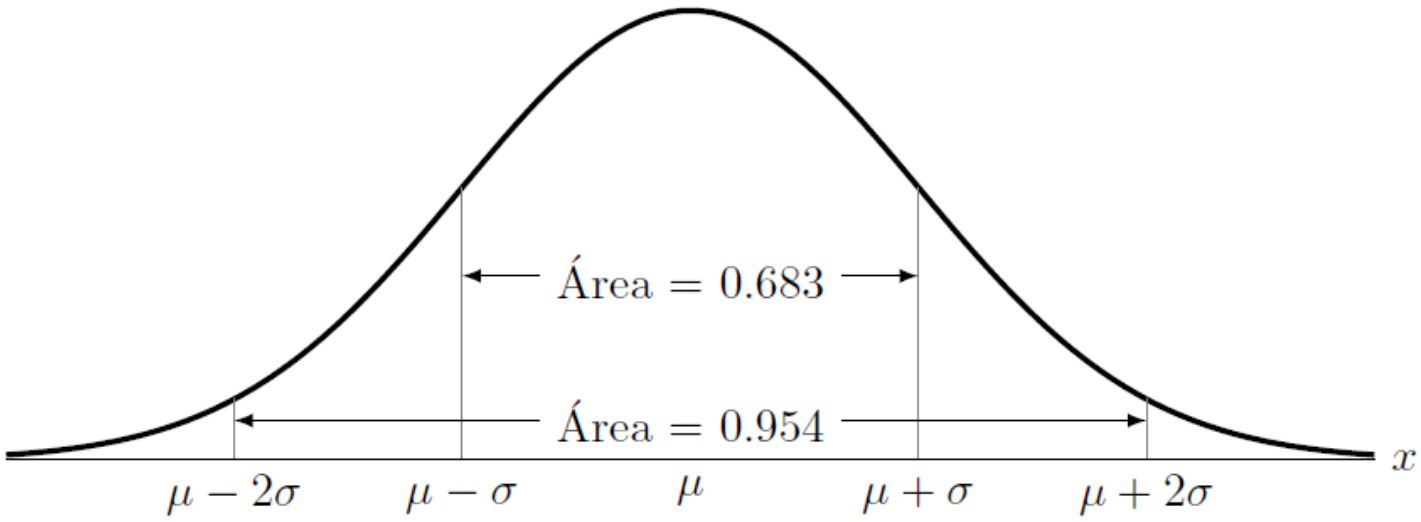
- **Intervalo de confiança e tamanho da amostra**
  - Dados para análise estatística são, geralmente, provenientes de **variáveis observadas em unidades de investigação** obtidas de uma população de interesse **através de** um processo de **amostragem**
  - Interesse?
    - Descrever e resumir dados da amostra
    - Fazer inferência
      - Não temos dúvidas sobre os resultados da amostra
      - Mas, o resultado pode ser extrapolado para a população?

# 2. Distribuições e amostras

- **Intervalo de confiança e tamanho da amostra**
  - Margem de erro
    - Medida da incerteza na extrapolação amostra – população
    - Dependente de:
      - Processo amostral
      - Desvio padrão (S)
      - Tamanho de amostra (n)
      - $\mathrm{me} = k\mathrm{S}/\sqrt{n}$
        - AAS, distribuição normal, 95,4% de confiança: $k = 2$

# 2. Distribuições e amostras

- **Intervalo de confiança e tamanho da amostra**
  - Margem de erro




- Confiança de 95%
  - $\text{me} = (1{,}96 * \text{S})/\sqrt{n}$
- Quando o tamanho da **amostra é suficientemente grande**, podemos utilizar esses valores mesmo que a distribuição de onde foi obtida a amostra não é normal

# 2. Distribuições e amostras

- **Intervalo de confiança e tamanho da amostra**
  - Com base na margem de erro
    - **Intervalo de 95% de confiança para média (IC 95%) =** $\bar{X} \pm 1{,}96S/\sqrt{n}$
  - Exemplo: peso de 5000 recém nascidos (kg)

| ID_bebe | peso |
|---|---|
| 1 | 3,2 |
| 2 | 2,5 |
| ... | ... |
| 5000 | 3,9 |

- Média amostral = $\bar{X}$ = 2,80 kg
- Desvio padrão amostral = S = 0, 5 kg
- IC 95% = [2,80 – (1,96*0,5)/$\sqrt{5000}$ ; 2,80 + (1,96*0,5)/$\sqrt{5000}$]

  = [2,80 – 0,014; 2,80 + 0,014; ]

  = [2,79; 2,81]

# 2. Distribuições e amostras

- **Intervalo de confiança e tamanho da amostra**
  - Exemplo: pesquisa eleitoral para candidato A - 150 entrevistados declararam apoio

| ID_entrevistado | apoia_candidato |
|---|---|
| 1 | 0 |
| 2 | 1 |
| … | … |
| 500 | 0 |

- Média amostral de X = proporção de eleitores favoráveis ao candidato na amostra ($\hat{p}$) = $\frac{150}{500}$ = 0,30
- **IC 95% para proporção: $\hat{p} \pm 1,96\sqrt{\hat{p}(1-\hat{p})/n}$**
- IC 95% = [0,30 – 1,96*0,02; 0,30 + 1,96*0,02]
  = [0,30 – 0,04; 0,30 + 0,04]
  = [0,259; 0,342]
  = [25,9%; 34,2%]

# 2. Distribuições e amostras

- **Intervalo de confiança e tamanho da amostra**
  - Qual é o tamanho da amostra necessário para que meus resultados tenham precisão $\varepsilon$??
    - Depende... O que você quer estimar?
      - Média
        - $n = (1{,}96S/\varepsilon)^2$
      - Proporção
        - $n = (1{,}96\sqrt{\hat{p}(1-\hat{p})}/\varepsilon)^2$

# 2. Distribuições e amostras

- **Intervalo de confiança e tamanho da amostra**
  - Qual é o tamanho da amostra necessário para que meus resultados tenham precisão $\varepsilon$??
    - Depende... O que você quer estimar?
      - Média
        - $n = (1{,}96S/\varepsilon)^2$

        Piloto, estudos parecidos..
        Regra de bolso? [max(X) – min(X)]/4
      - Proporção
        - $n = (1{,}96\sqrt{\hat{p}(1-\hat{p})}/\varepsilon)^2$

        Pior cenário: $\hat{p}$ = 0,5 -> $\hat{p}$(1- $\hat{p}$) = 0,25 -> $\sqrt{0{,}25} = 0{,}5$
        Considerando 1,96 ~ 2
        n = $(1/\varepsilon)^2$
        Exemplo: pesquisas de intenção de voto, onde $\varepsilon$ = 3 pontos percentuais
        n = $(1/0{,}03)^2$ = 1.111

# 2. Distribuições e amostras

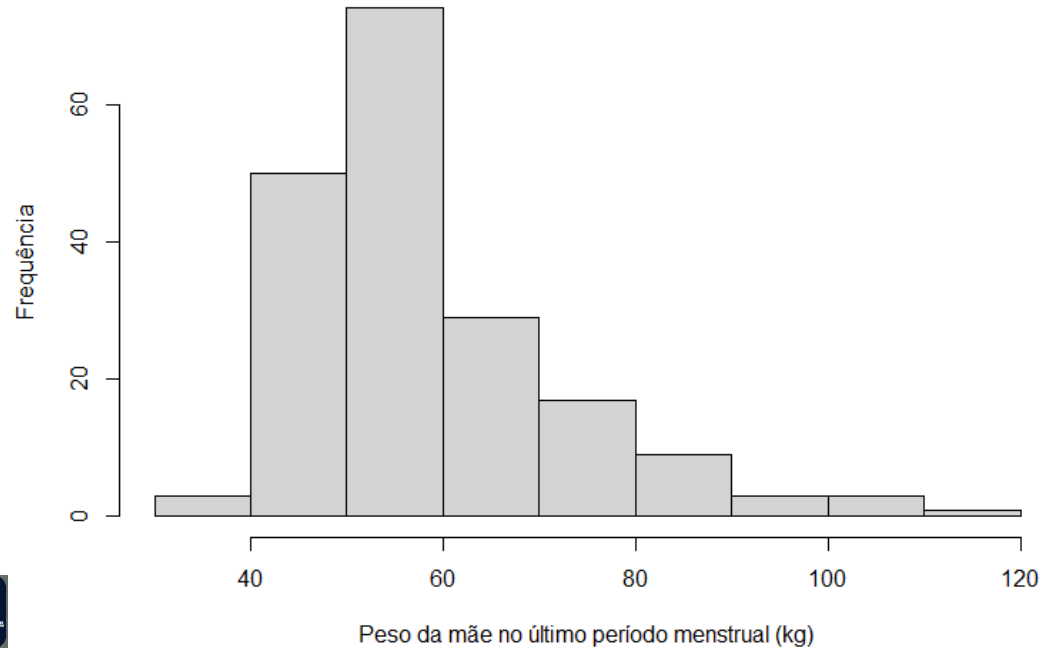
- **Intervalo de confiança e tamanho da amostra**
  - Qual é o tamanho da amostra necessário para que meus resultados tenham precisão $\varepsilon$??
    - Calculadoras de tamanho de amostra
      - WinPepi

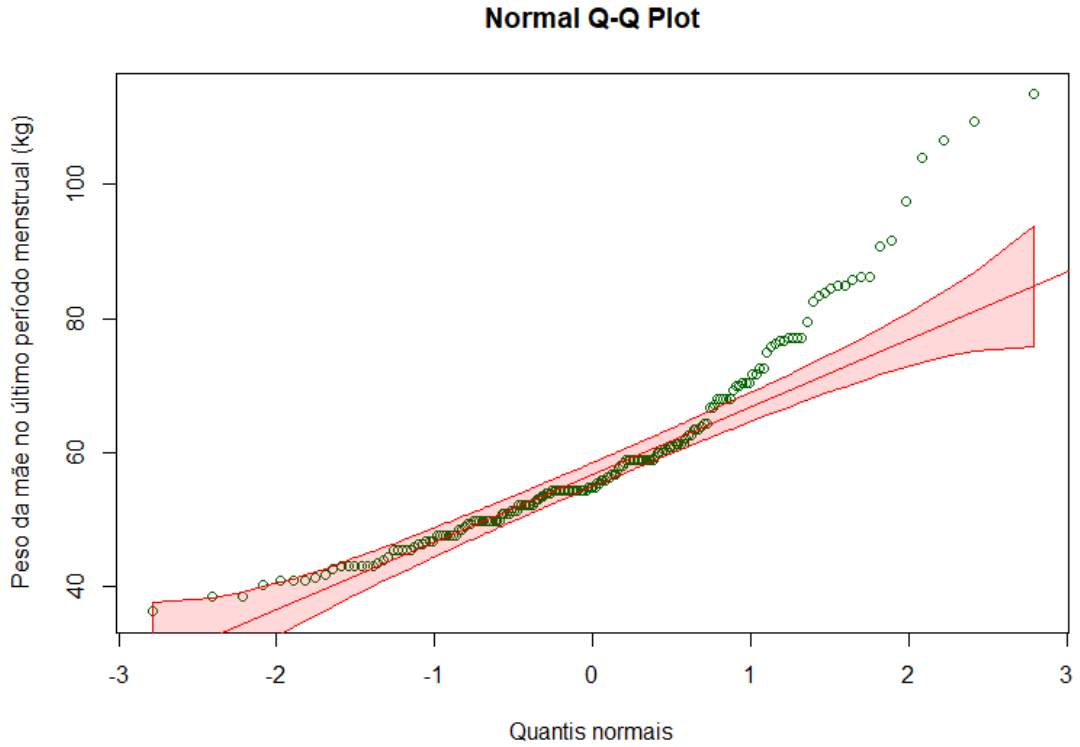
# 2. Distribuições e amostras

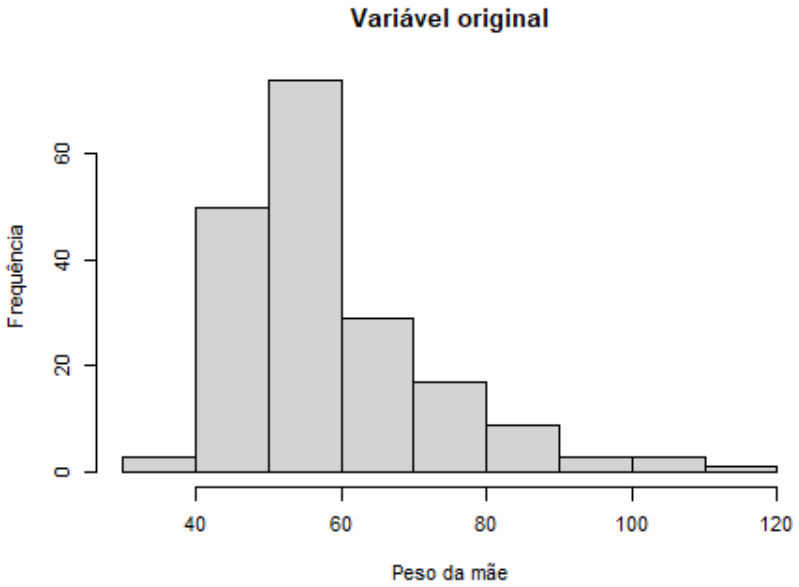
- **Transformações de variáveis**
  - Distribuição Normal é suposição de diversos métodos estatísticos
  - Na prática é comum a distribuição dos dados na amostra ser assimétrica e conter valores atípicos
  - O que fazer?
    - Transformação de dados a fim de obter uma distribuição mais simétrica
      - log(x)
      - -1/x
      - $\sqrt{x}$
      - $\sqrt[3]{x}$

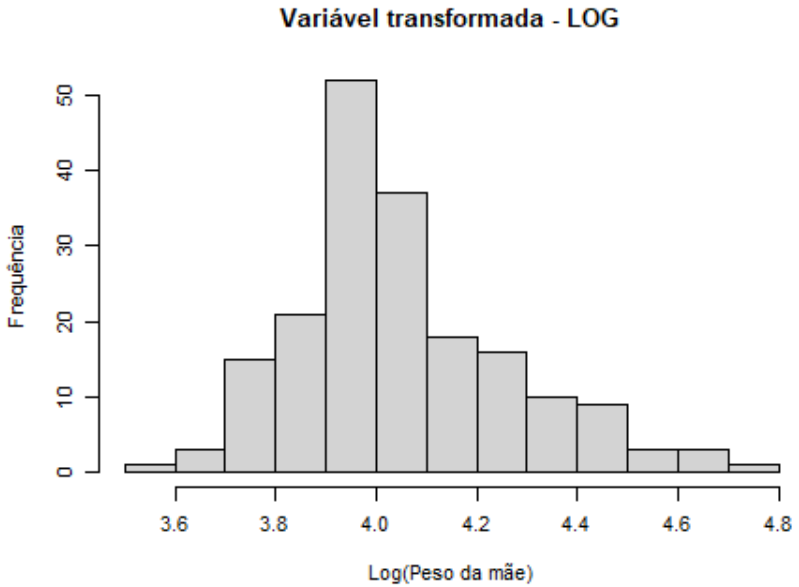
# 2. Distribuições e amostras

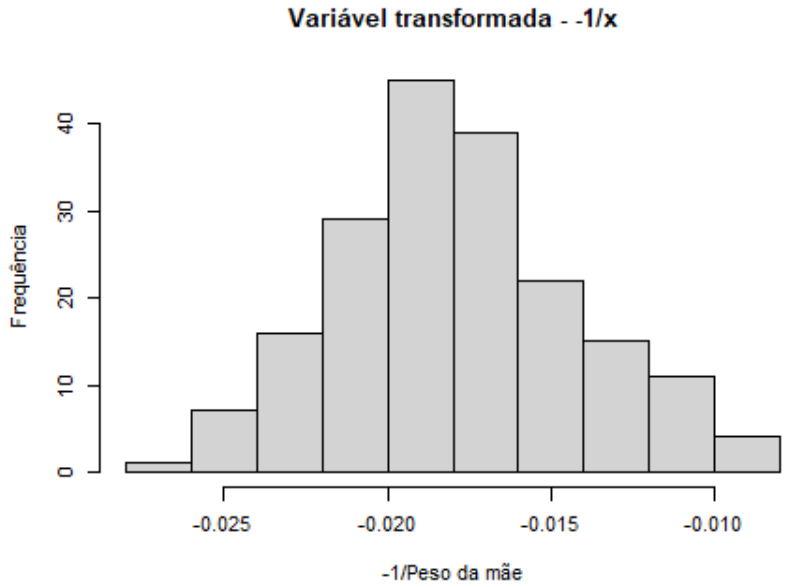
- **Transformações de variáveis**
  - Exemplo: peso da mãe no último período menstrual (kg)

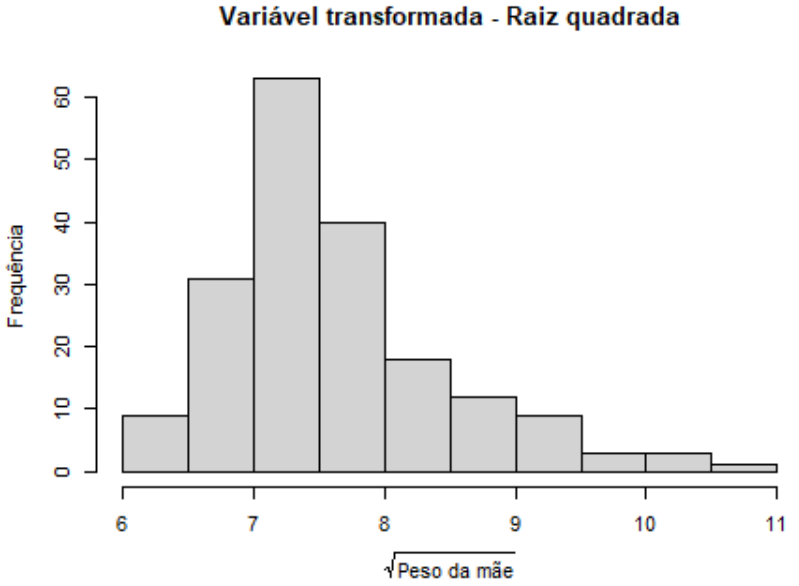


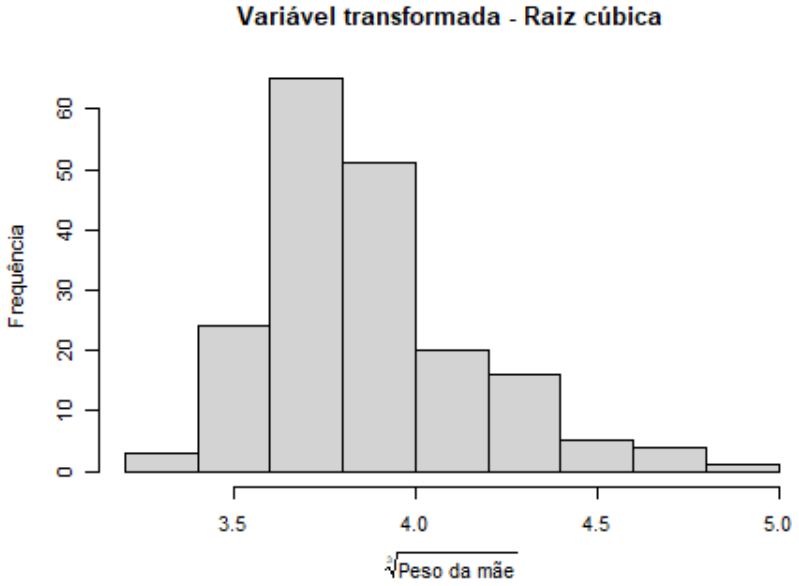



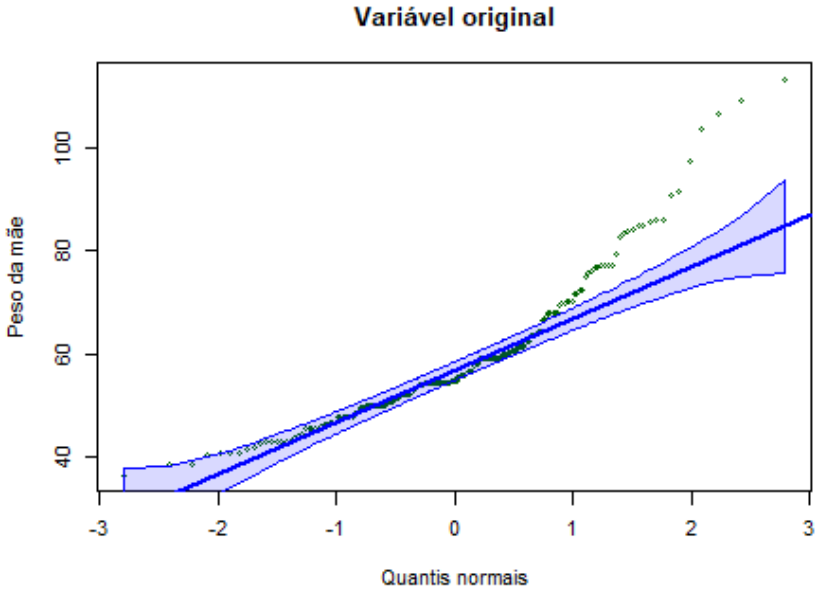

Variável original
Frequência
0
20
40
60
40
60
80
100
120
Peso da mãe

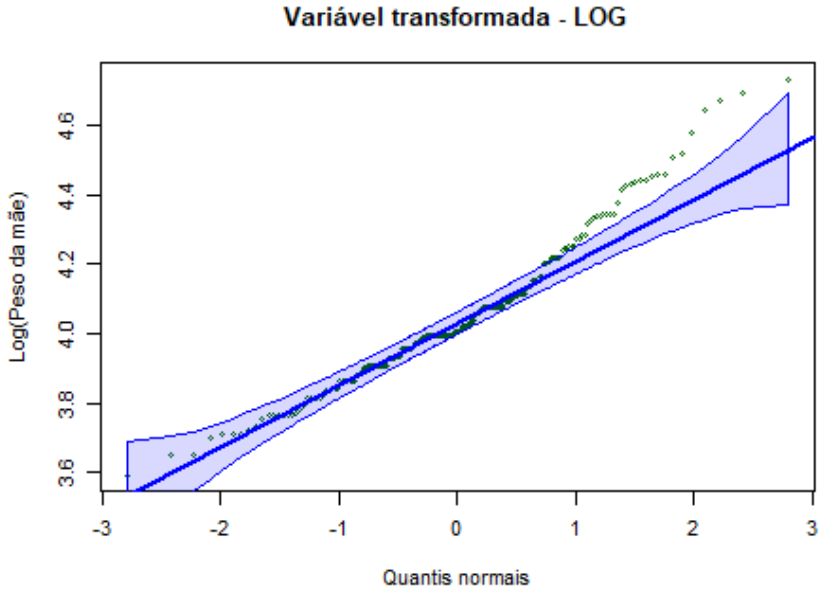

Variável transformada - LOG
Frequência
0
10
20
30
40
50
3.6
3.8
4.0
4.2
4.4
4.6
4.8
Log(Peso da mãe)

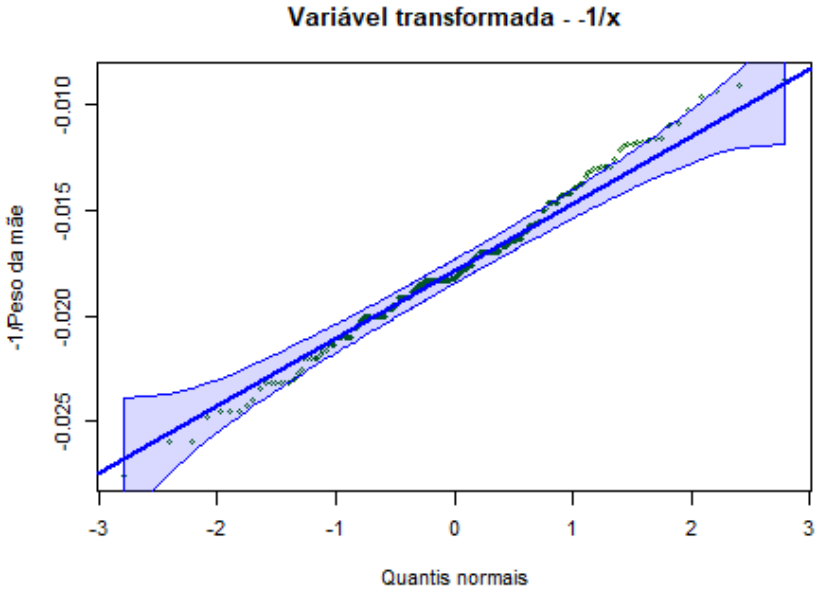

Variável transformada - -1/x
Frequência
0
10
20
30
40
-0.025
-0.020
-0.015
-0.010
-1/Peso da mãe

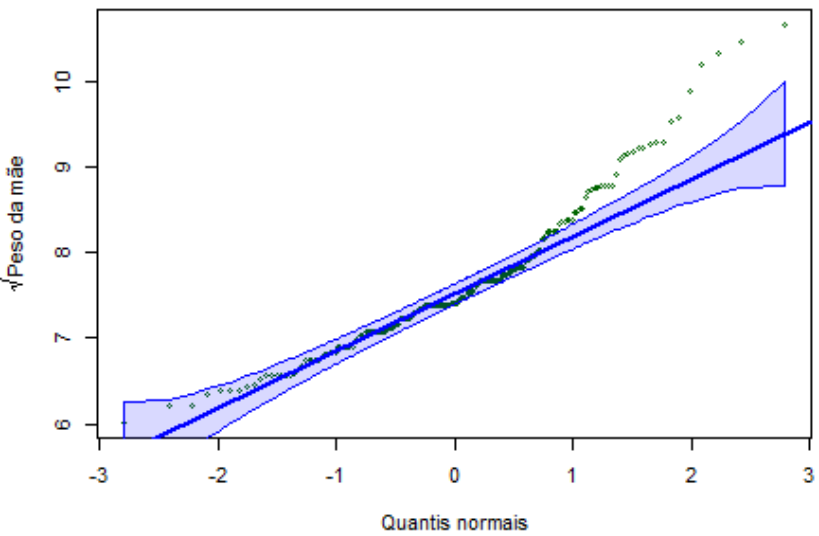

Variável transformada - Raiz quadrada
Frequência
0
10
20
30
40
50
60
6
7
8
9
10
11
√Peso da mãe

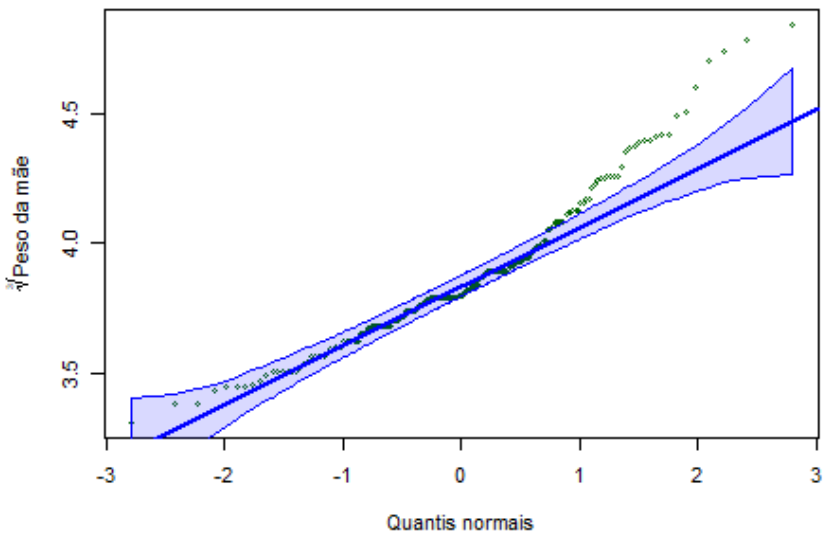

Variável transformada - Raiz cúbica
Frequência
0
10
20
30
40
50
60
3.5
4.0
4.5
5.0
∛Peso da mãe


Variável original
Peso da mãe
Quantis normais


Variável transformada - LOG
Log(Peso da mãe)
Quantis normais


Variável transformada - -1/x
-1/Peso da mãe
Quantis normais


Variável transformada - Raiz quadrada
√Peso da mãe
Quantis normais


Variável transformada - Raiz cúbica
∛Peso da mãe
Quantis normais

# 2. Distribuições e amostras

- **Transformações de variáveis**
  - Padronização (z-escore)
    - Média 0, desvio padrão 1
    - $\mathrm{z} = \frac{x - \bar{x}}{s}$
  - Normalização (min-max)
    - Intervalo 0 e 1
    - $\mathrm{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$

# 2. Distribuições e amostras

- **Transformações de variáveis**
  - Alterar tipo de variável: criar categorias de valores a partir de uma variável numérica
    - Ex.: faixa etária, faixa de renda