Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

06 setembro 2022

Atividade 4.2 - Análise de dados - comparação entre duas populações

Profª. Ana Maria Nogales

Métodos Estatísticos 2

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

1. A partir de sua amostra dos resultados do SAEB 9o. ano, considere amostras de tamanho 20 e 200. Para cada amostra, explore a associação entre as seguintes variáveis:

- NOTA_MT e LOCALIZAÇÃO (Urbana e Rural)
- NOTA_LP e Ano de nascimento (2001 ou antes ; 2002 ou depois)

Para avaliar essas relações construa os gráficos adequados e medidas de posição e variabilidade segundo categorias das variáveis qualitativas. Você diria que a proficiência em matemática é maior em escolas urbanas? Existe diferença entre as proficiências em língua portuguesa segundo grupo de idade do estudante? Considere os testes: t de Student, Wilcoxon-Mann-Whitney e Kolmogorov-Smirnov. (Não se esqueça de testar normalidade e homocedasticidade - fazer referência, se vc já analisou esses aspectos anteriormente)

2. Para a amostra de tamanho 20, verifique se há diferença entre as notas de língua portuguesa e matemática. Considere os testes: t de Student, Wilcoxon e Sinais. Comente os resultados.

Testes: Proeficiência em matemática pela localização da escola:

Amostra n=20

```
##
##  Welch Two Sample t-test
##
## data:  rural20$NOTA_MT and urbana20$NOTA_MT
## t = 1.0757, df = 5.847, p-value = 0.3244
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.97426  71.36819
## sample estimates:
## mean of x mean of y
##  268.0597  246.3628

##
##  Wilcoxon rank sum exact test
##
## data:  rural20$NOTA_MT and urbana20$NOTA_MT
## W = 42, p-value = 0.3847
## alternative hypothesis: true location shift is not equal to 0

##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  rural20$NOTA_MT and urbana20$NOTA_MT
## D = 0.4375, p-value = 0.5214
## alternative hypothesis: two-sided
```

Amostra n=200

```
##
##  Welch Two Sample t-test
##
## data:  rural200$NOTA_MT and urbana200$NOTA_MT
## t = -1.2507, df = 29.639, p-value = 0.2208
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -32.180493   7.742997
## sample estimates:
## mean of x mean of y
##  246.6873  258.9060

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  rural200$NOTA_MT and urbana200$NOTA_MT
## W = 1757, p-value = 0.1826
## alternative hypothesis: true location shift is not equal to 0

##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  rural200$NOTA_MT and urbana200$NOTA_MT
## D = 0.21402, p-value = 0.2509
## alternative hypothesis: two-sided
```

Proeficiência em língua portuguesa pelo ano de nascimento:

Amostra n=20

```
##
##  Welch Two Sample t-test
##
## data:  antes20$NOTA_LP and depois20$NOTA_LP
## t = -0.3601, df = 6.2827, p-value = 0.7306
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -74.80237  55.42761
## sample estimates:
## mean of x mean of y
##  253.8469  263.5343

##
##  Wilcoxon rank sum exact test
##
## data:  antes20$NOTA_LP and depois20$NOTA_LP
## W = 33, p-value = 0.7354
## alternative hypothesis: true location shift is not equal to 0

##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  antes20$NOTA_LP and depois20$NOTA_LP
## D = 0.33333, p-value = 0.7701
## alternative hypothesis: two-sided
```

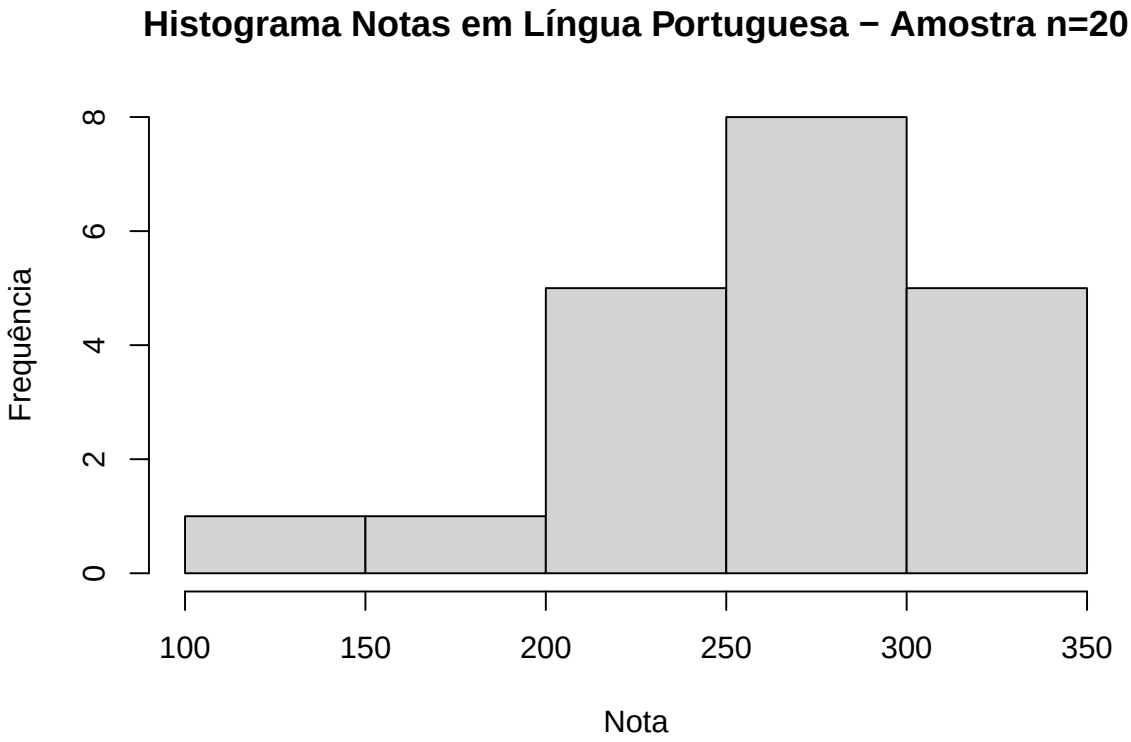
Amostra n=200

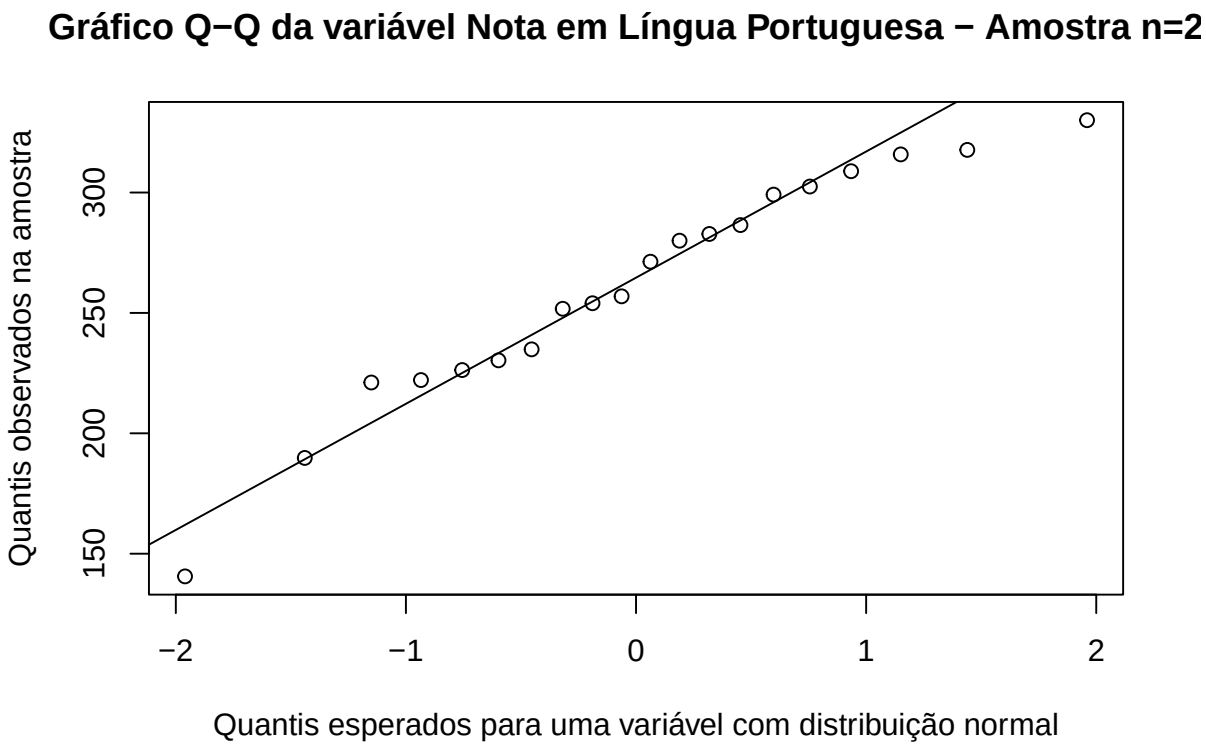
```
##
##  Welch Two Sample t-test
##
## data:  antes200$NOTA_LP and depois200$NOTA_LP
## t = -4.389, df = 53.05, p-value = 5.454e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -57.03949 -21.25853
## sample estimates:
## mean of x mean of y
##  230.2571  269.4061

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  antes200$NOTA_LP and depois200$NOTA_LP
## W = 1727, p-value = 5.069e-05
## alternative hypothesis: true location shift is not equal to 0

##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  antes200$NOTA_LP and depois200$NOTA_LP
## D = 0.36661, p-value = 0.0003884
## alternative hypothesis: two-sided
```

Gráficos

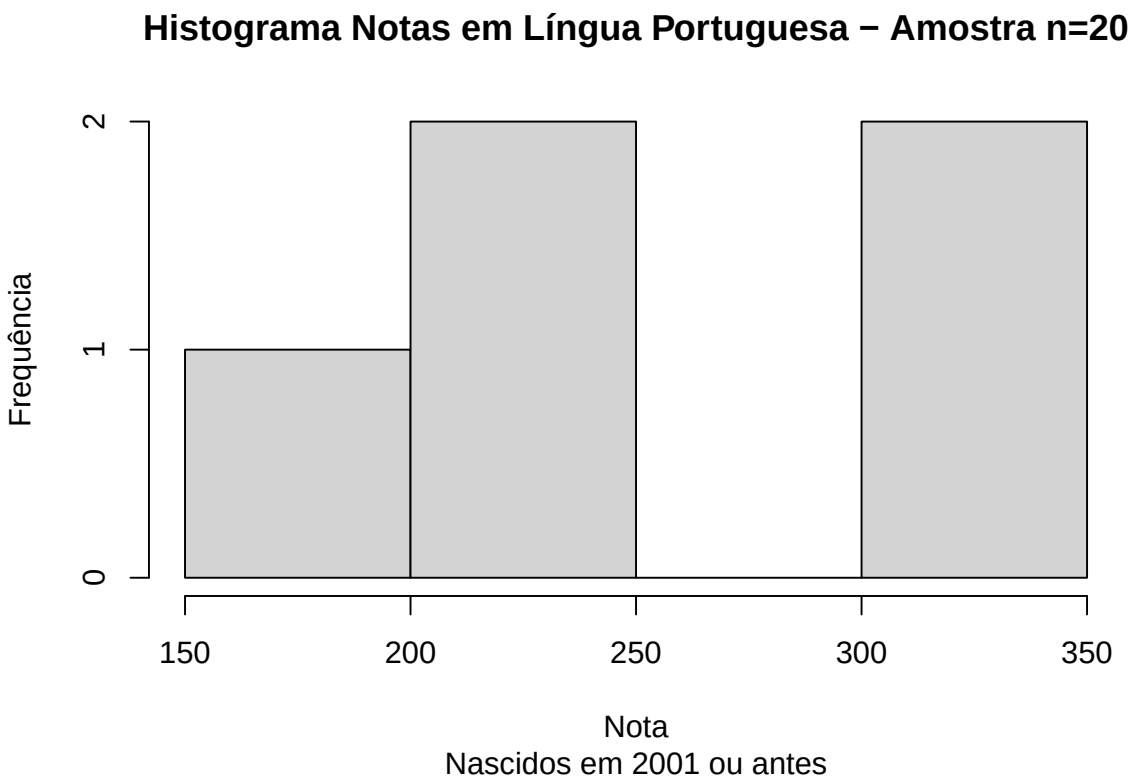
Amostra n=20, variável Nota em língua portuguesa

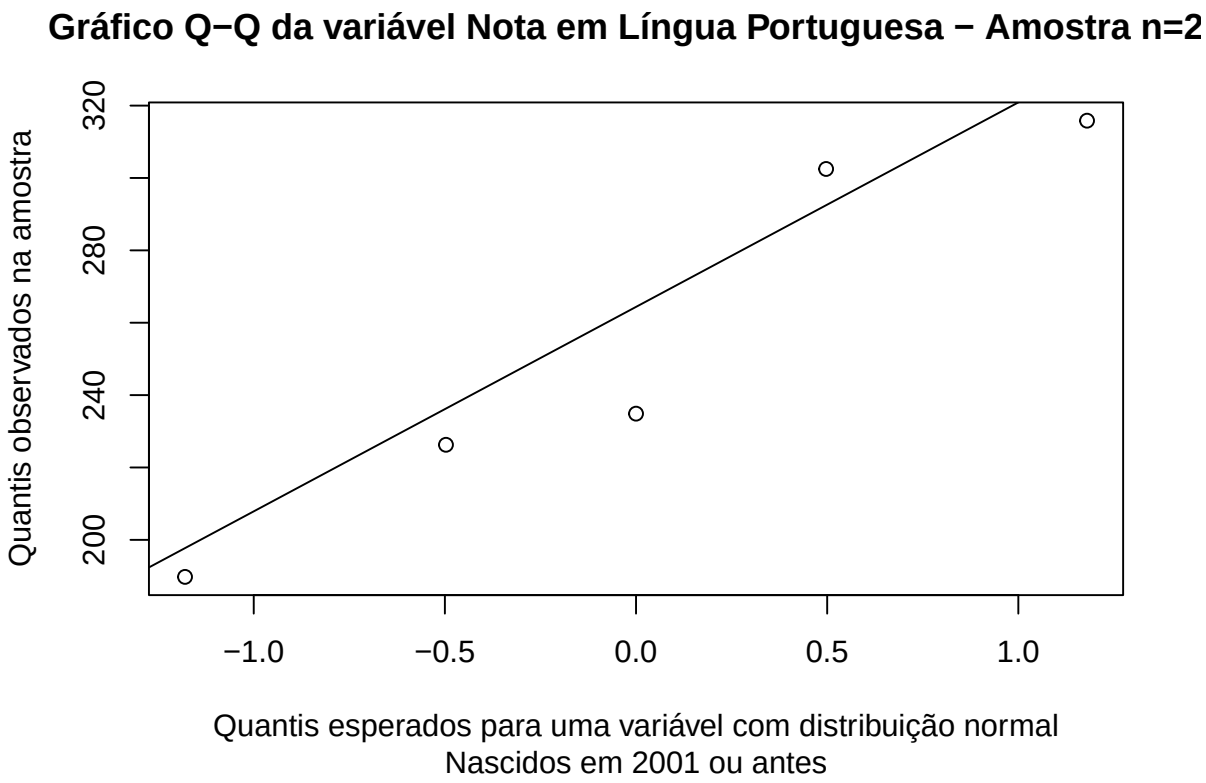
Ano de nascimento indiscriminado



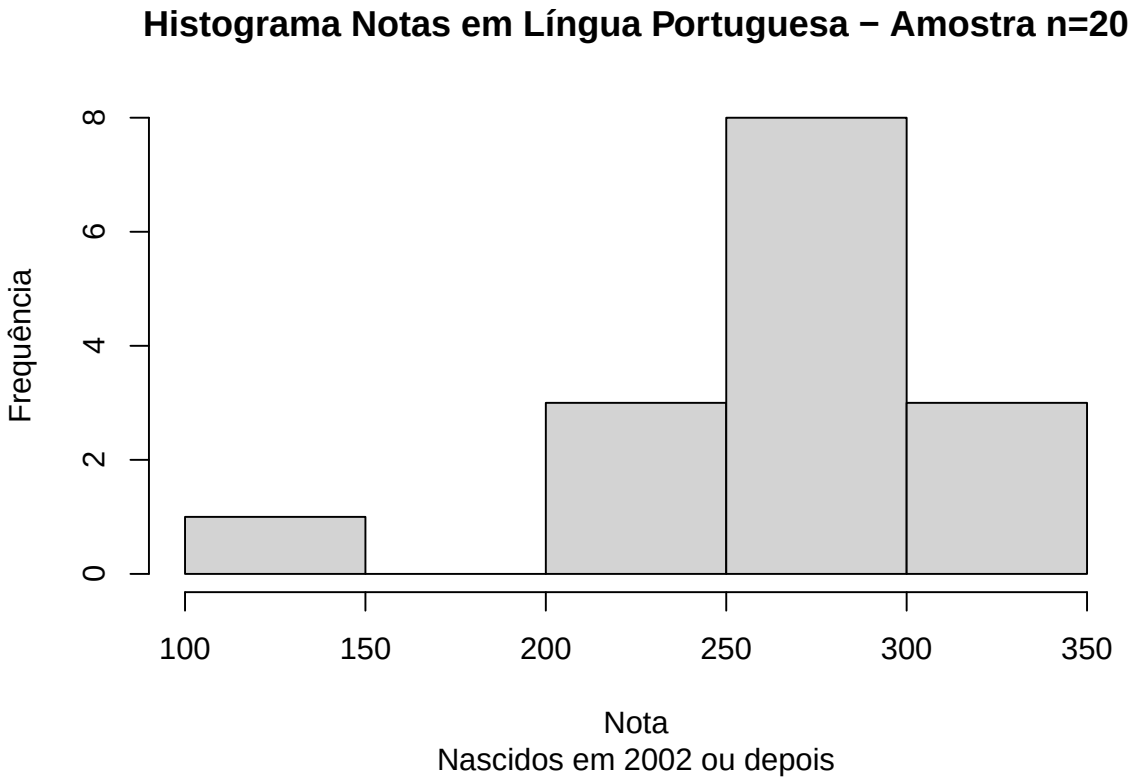


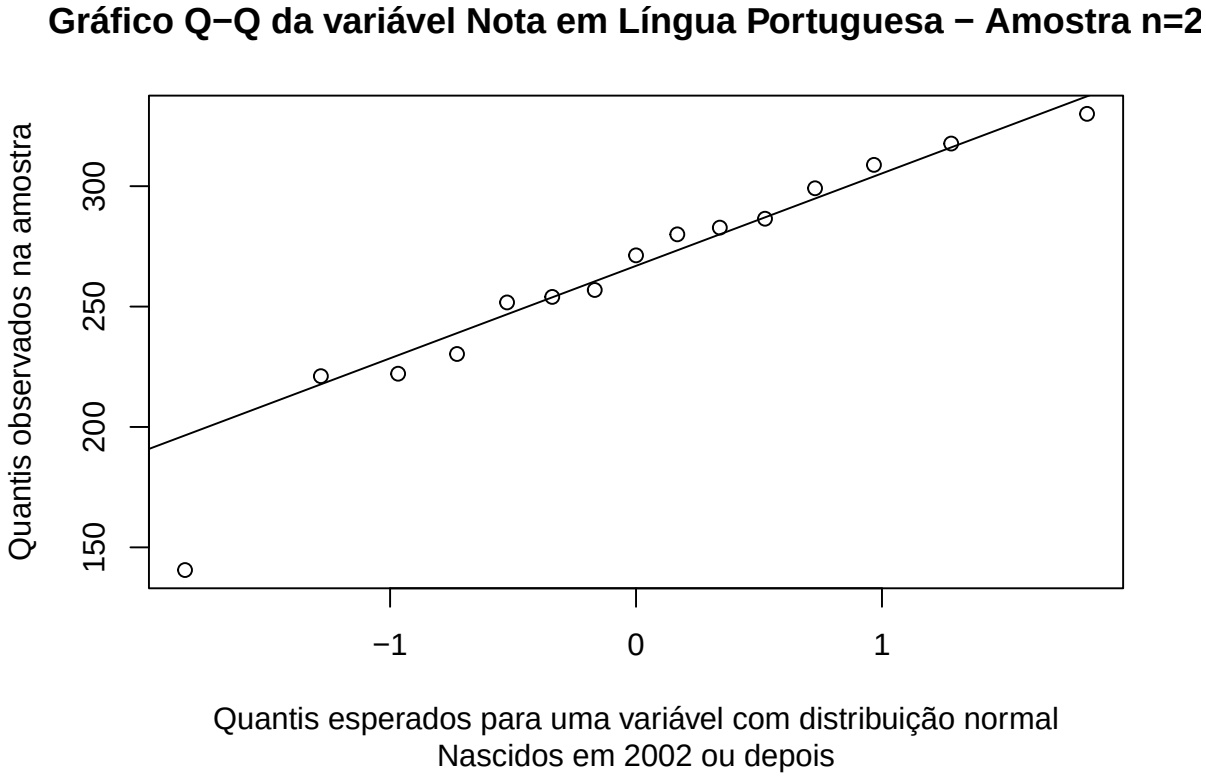
Nascidos em 2001 ou antes





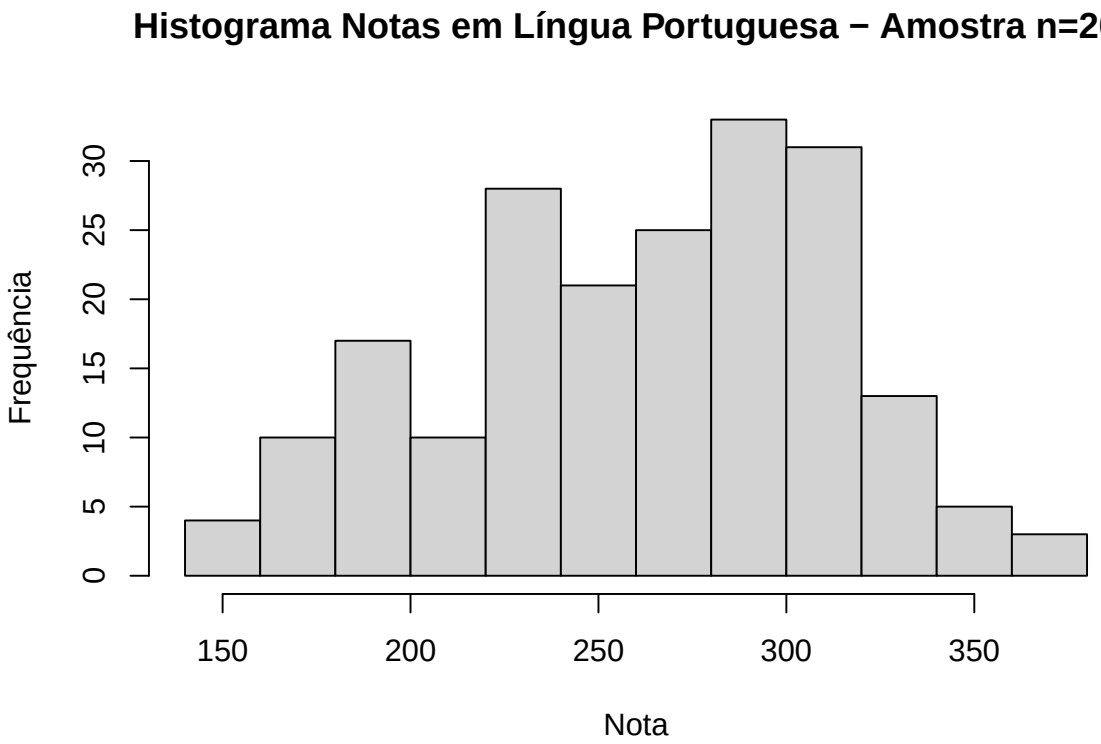
Nascidos em 2002 ou depois

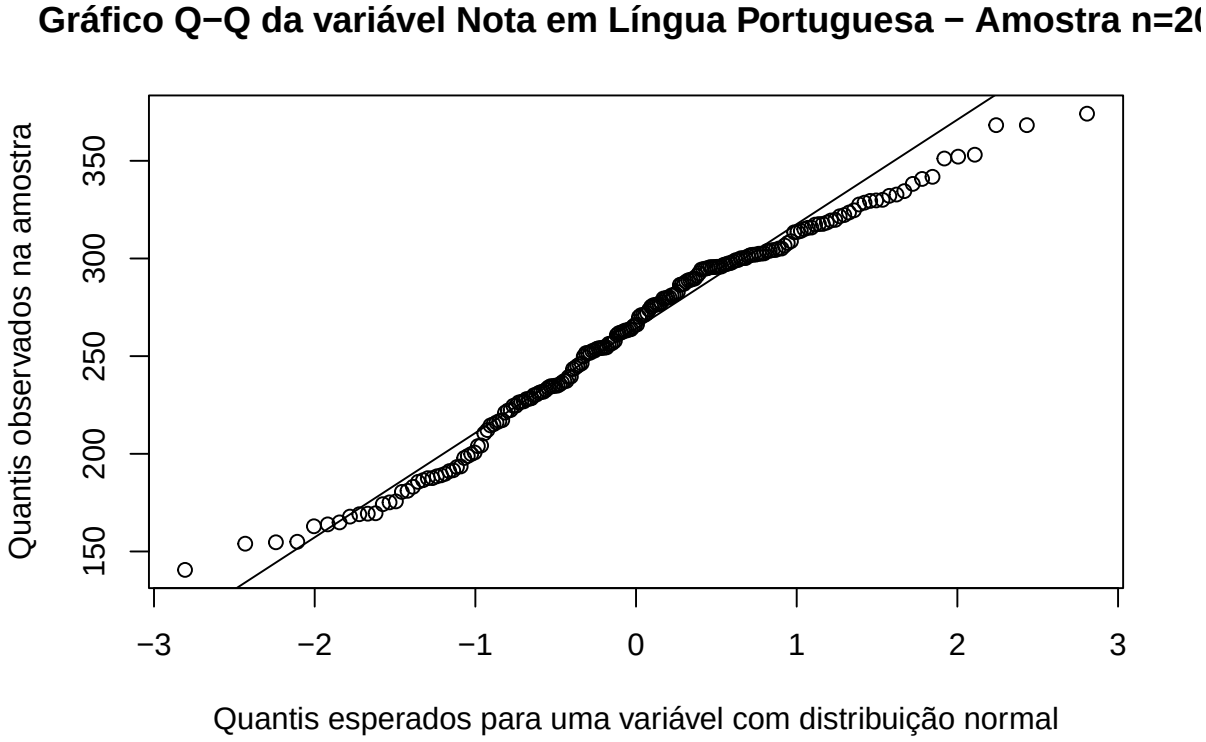




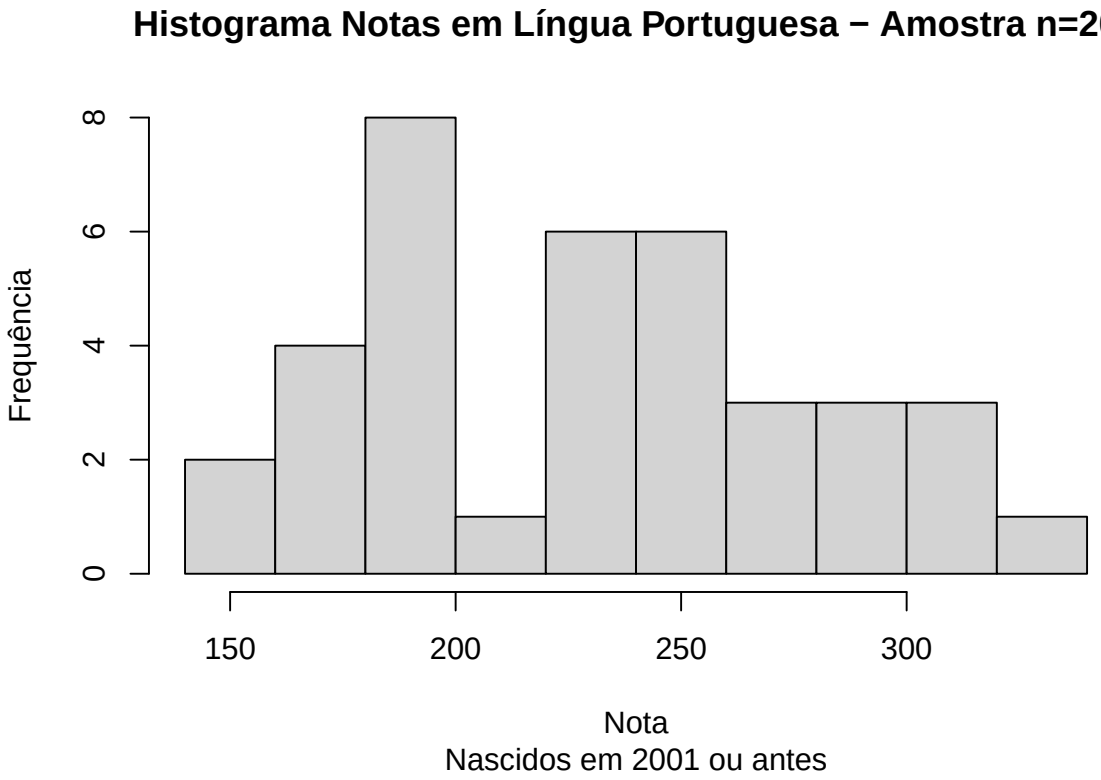
Amostra n=200, variável Nota em língua portuguesa

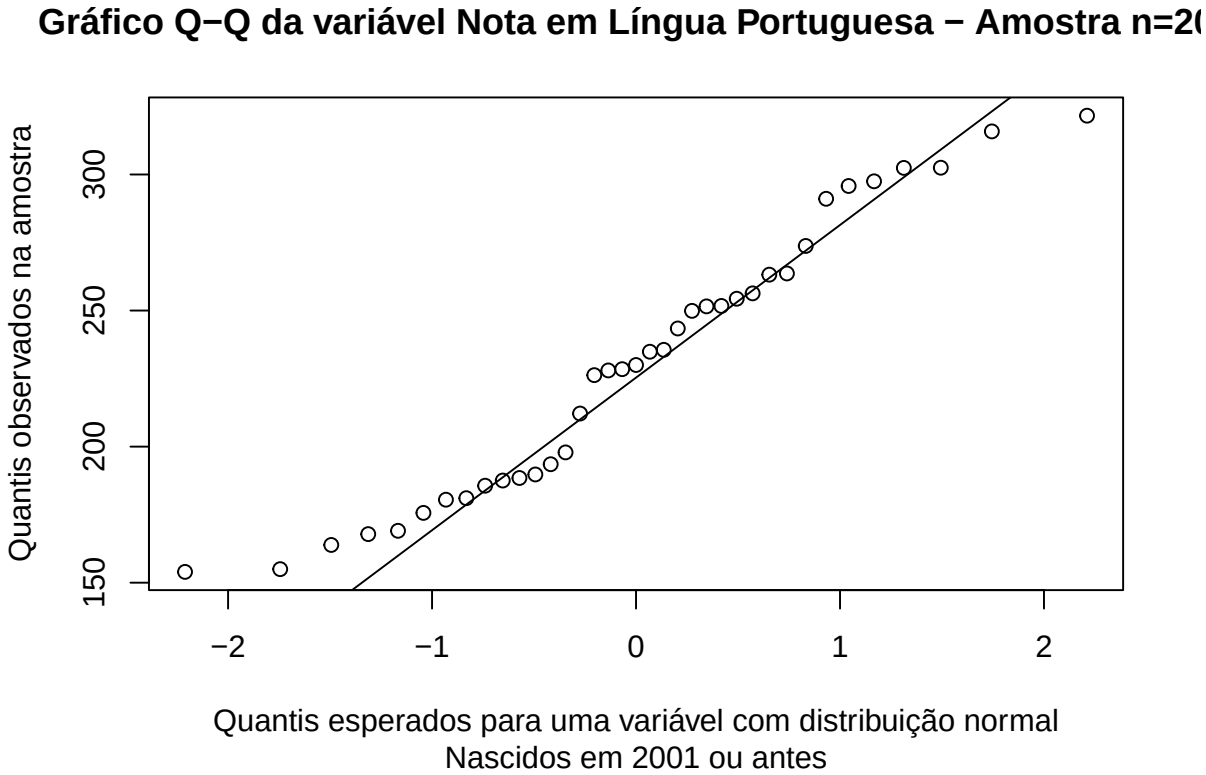
Ano de nascimento indiscriminado



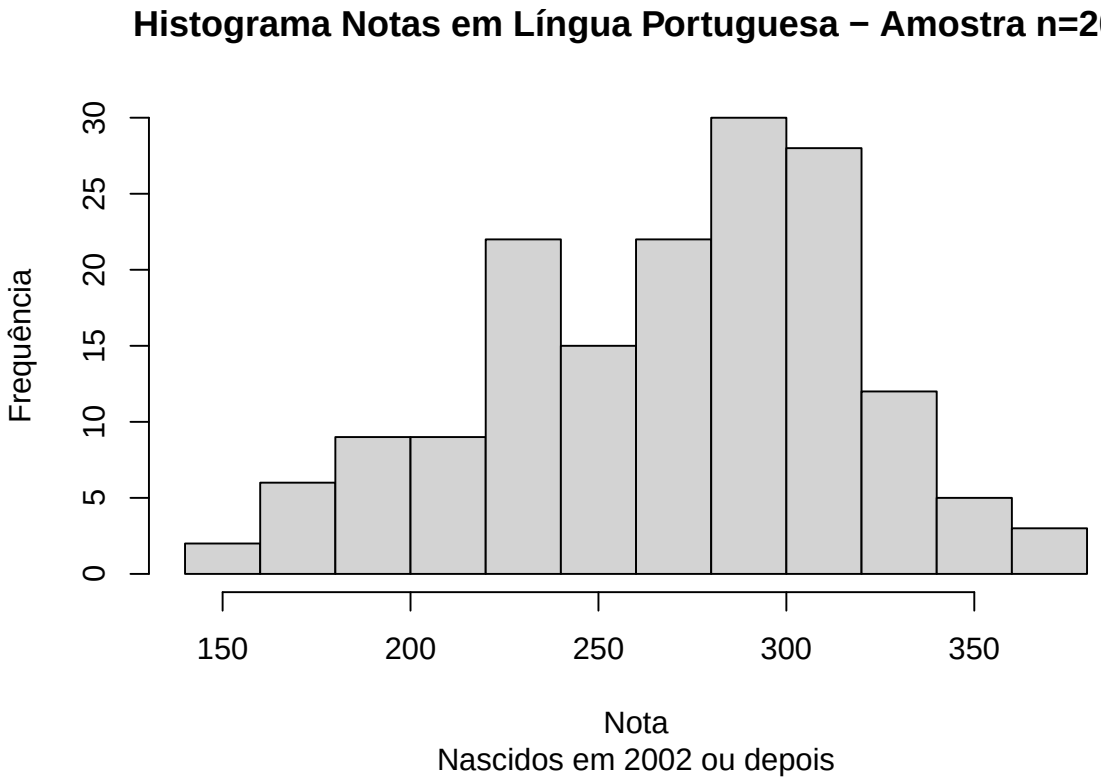


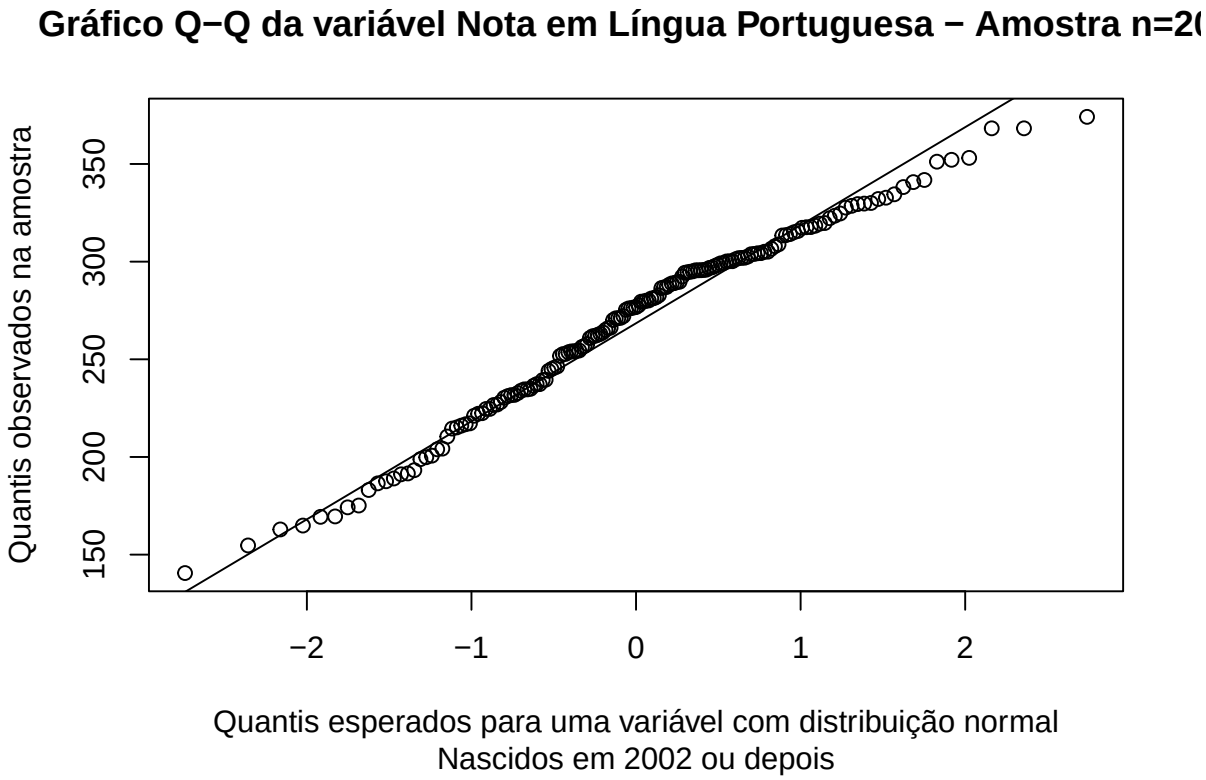
Nascidos em 2001 ou antes





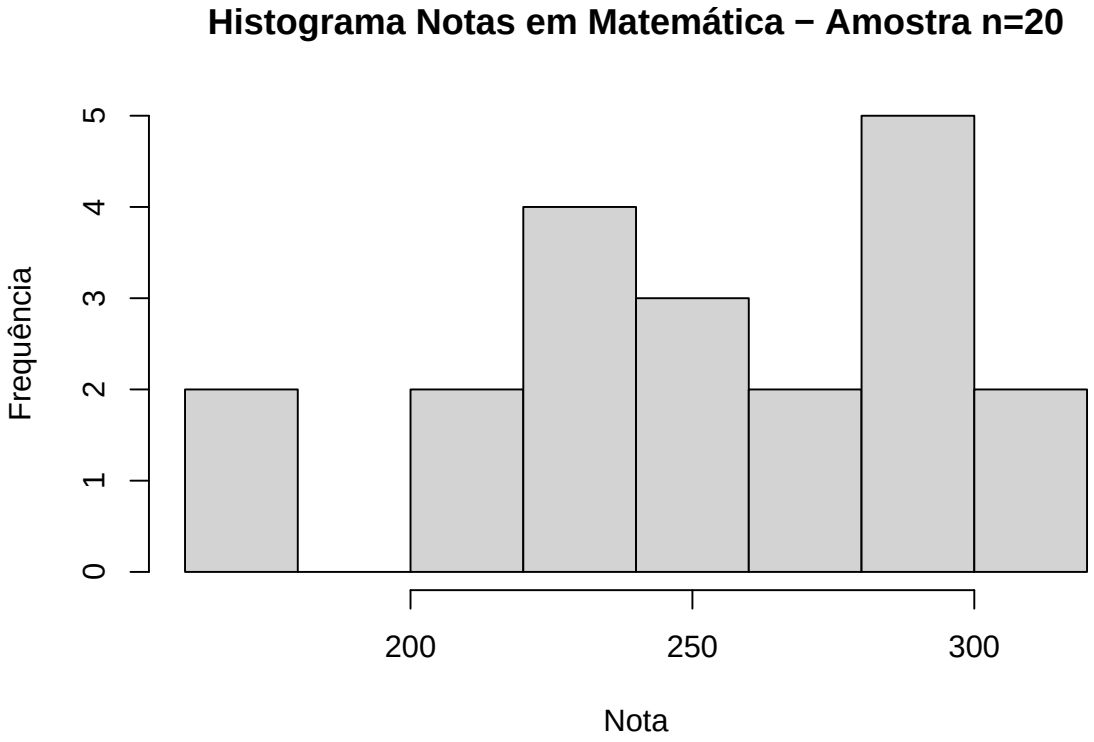
Nascidos em 2002 ou depois

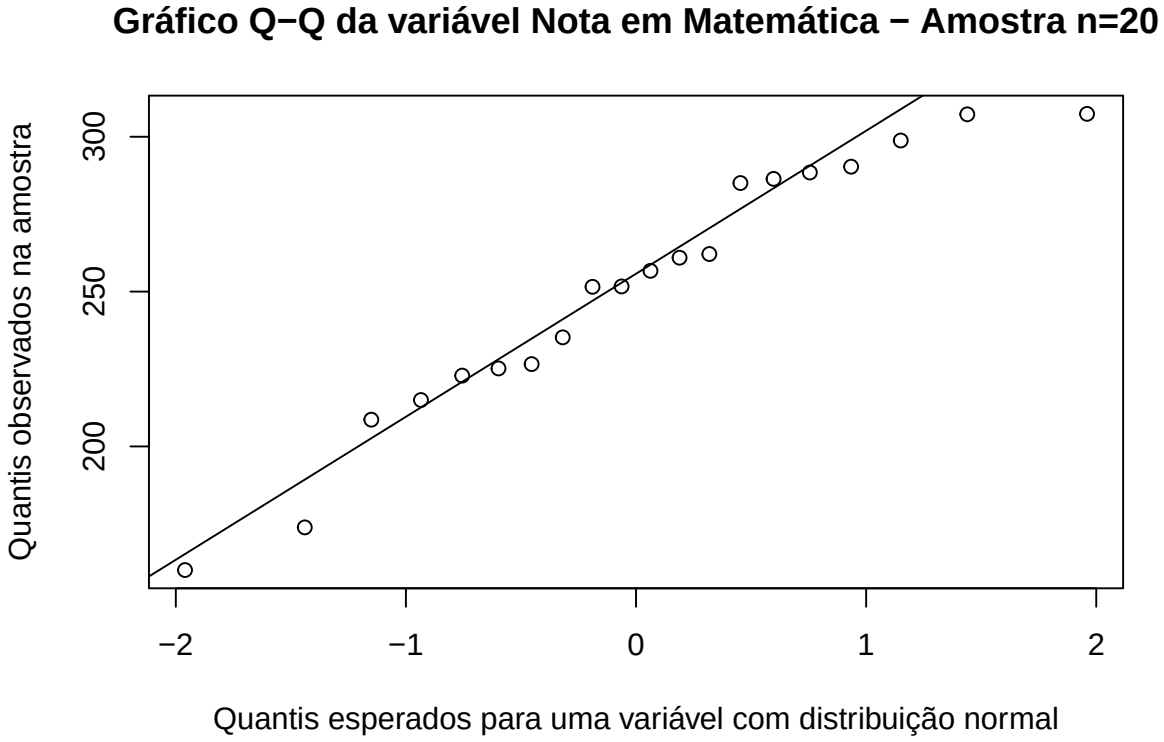




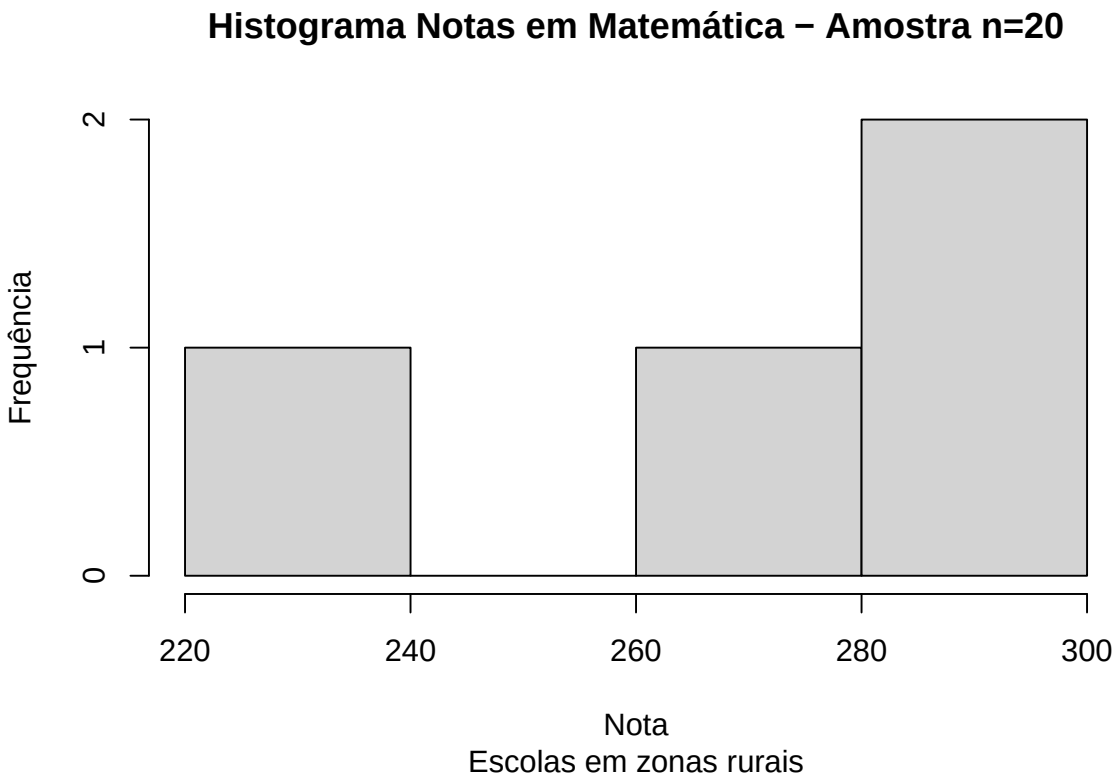
Amostra n=20, variável Nota em Matemática

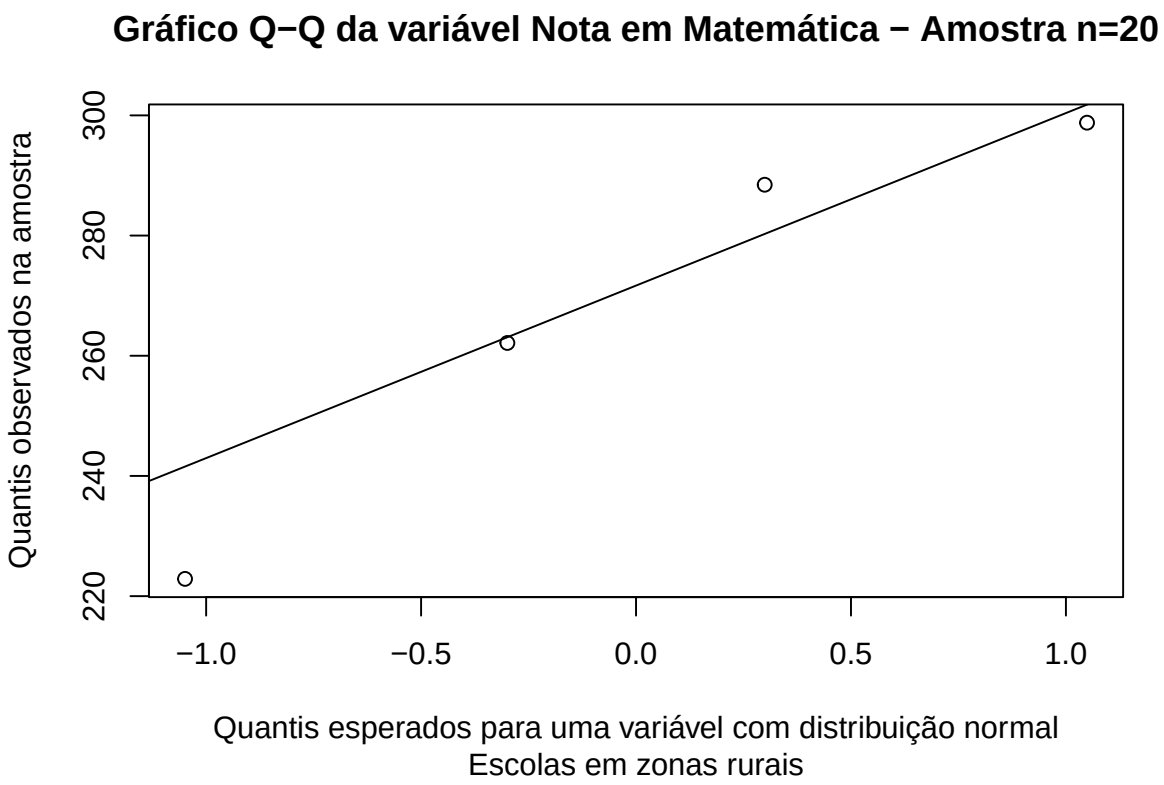
Localização da escola indiscriminada



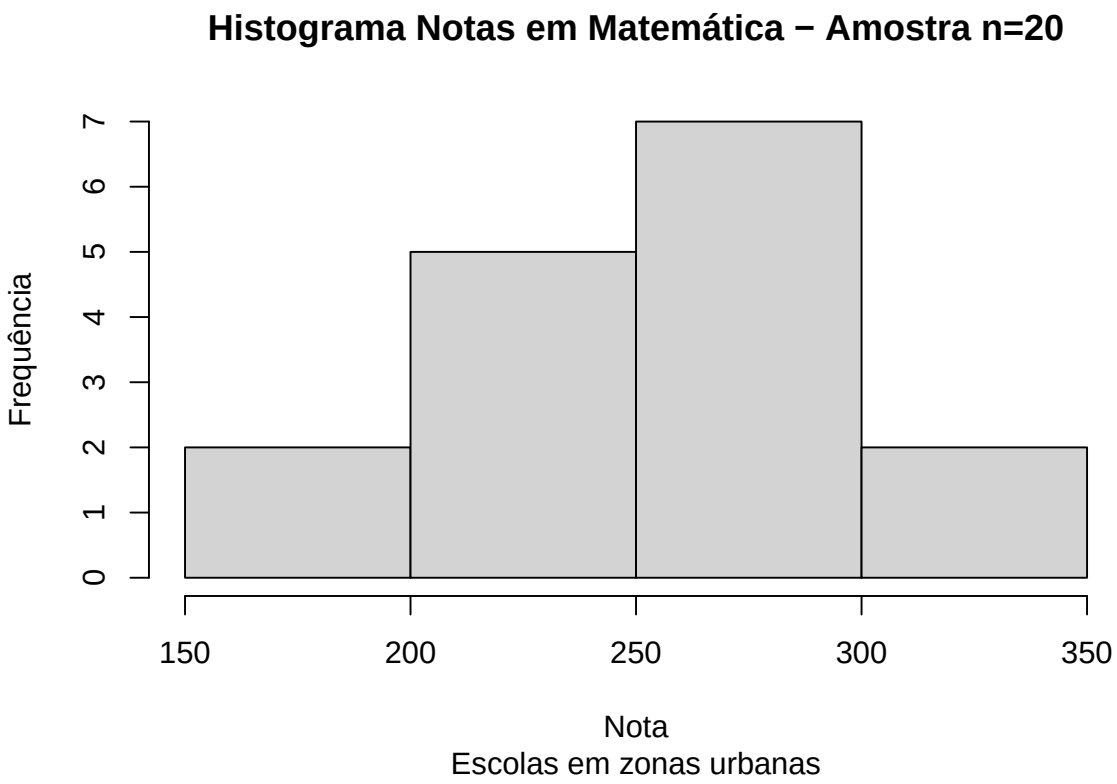


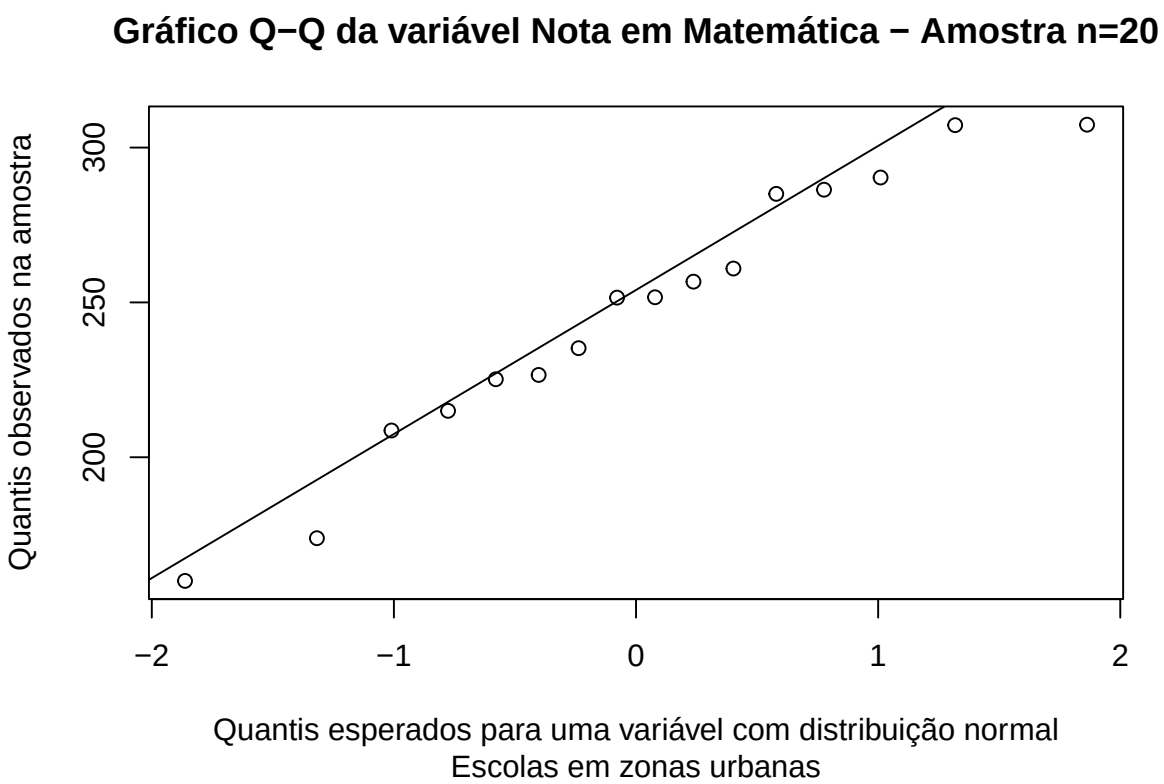
Escolas rurais





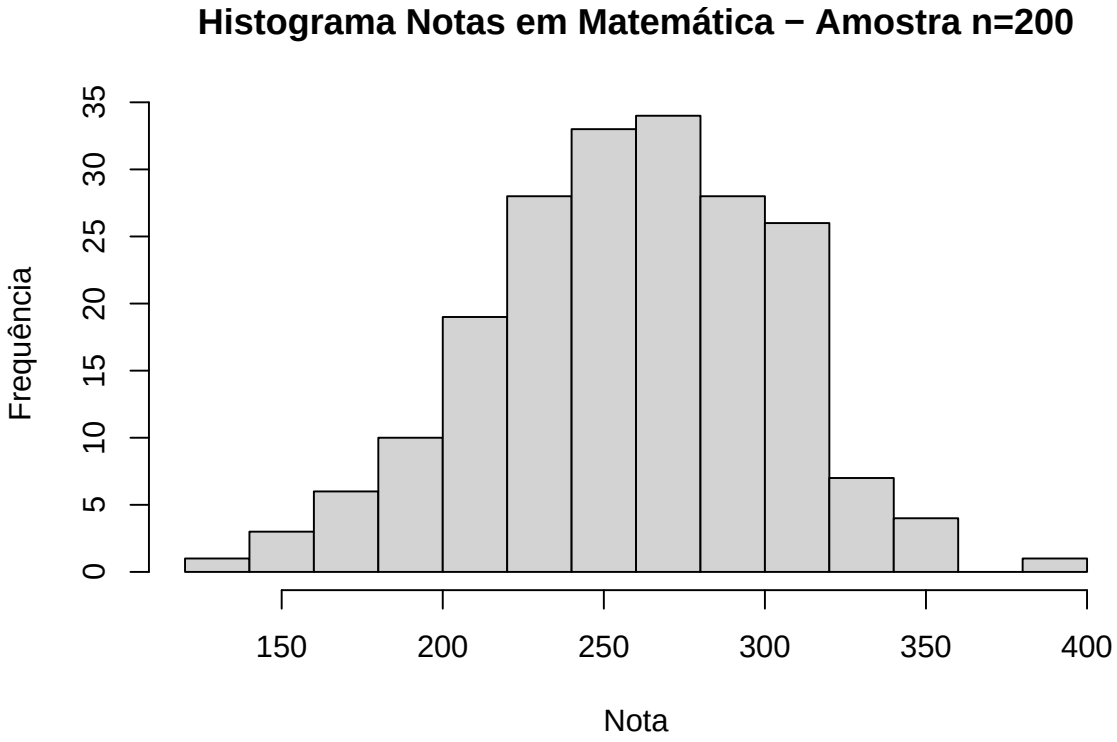
Escolas urbanas

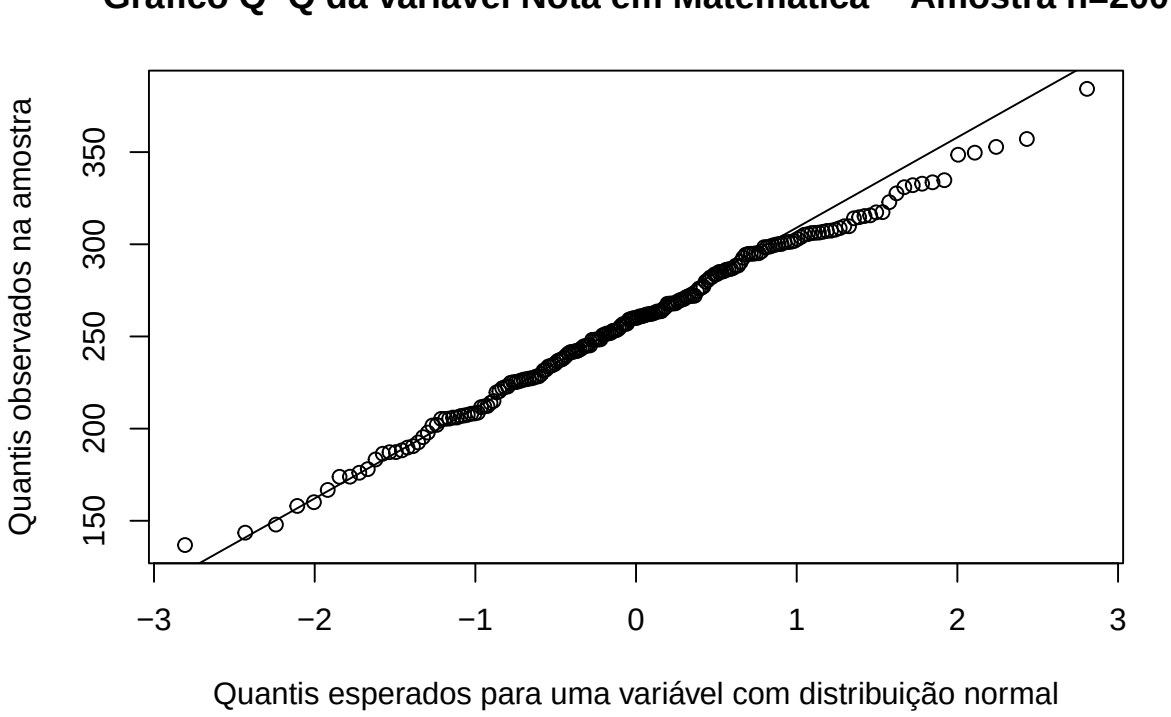




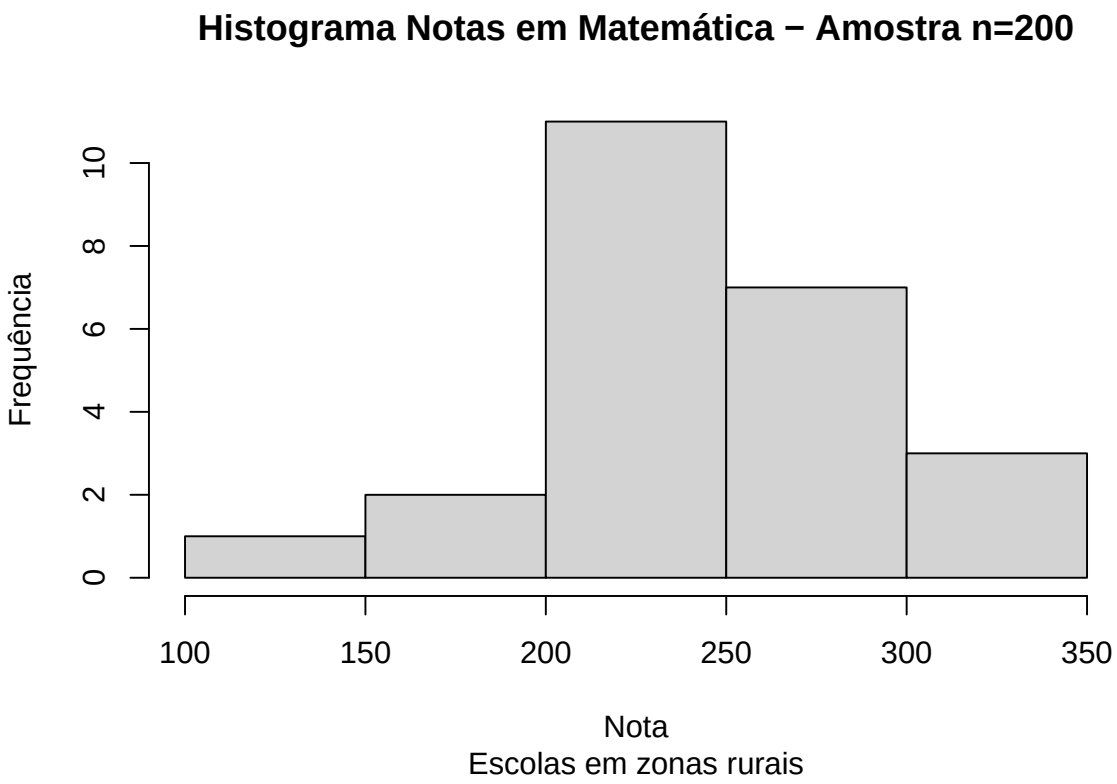
Amostra n=200, variável Nota em Matemática

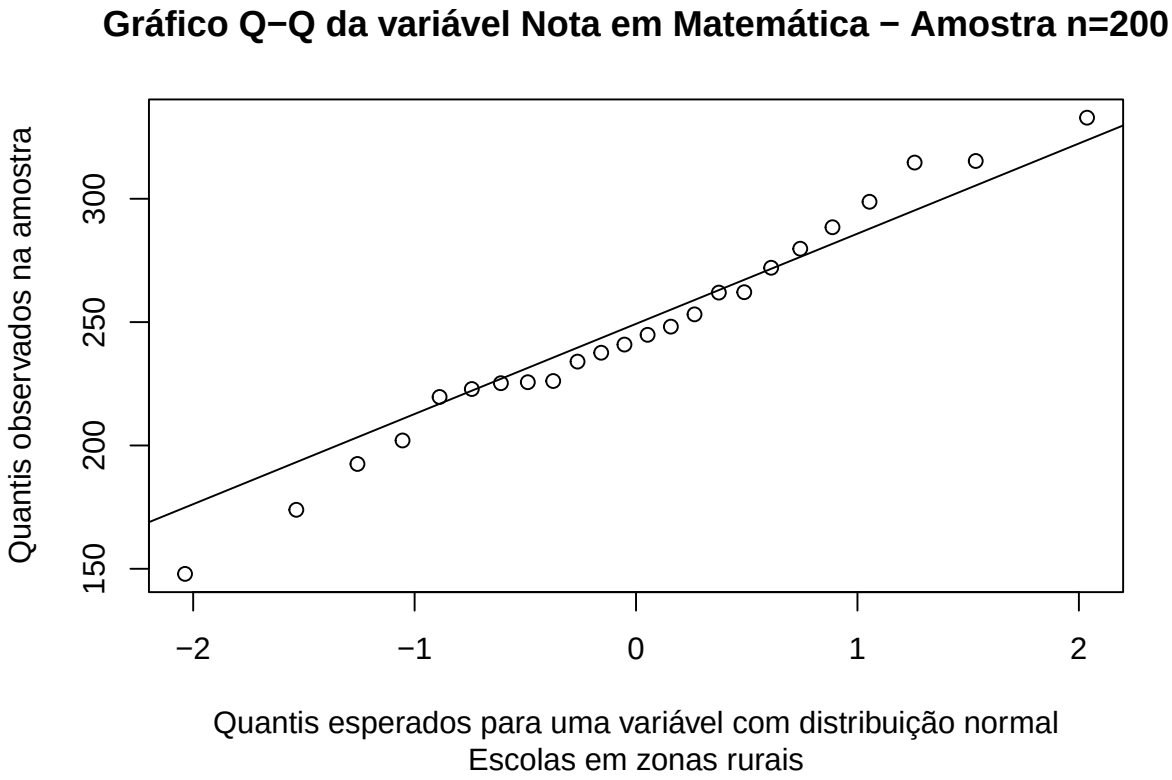
Localização da escola indiscriminada



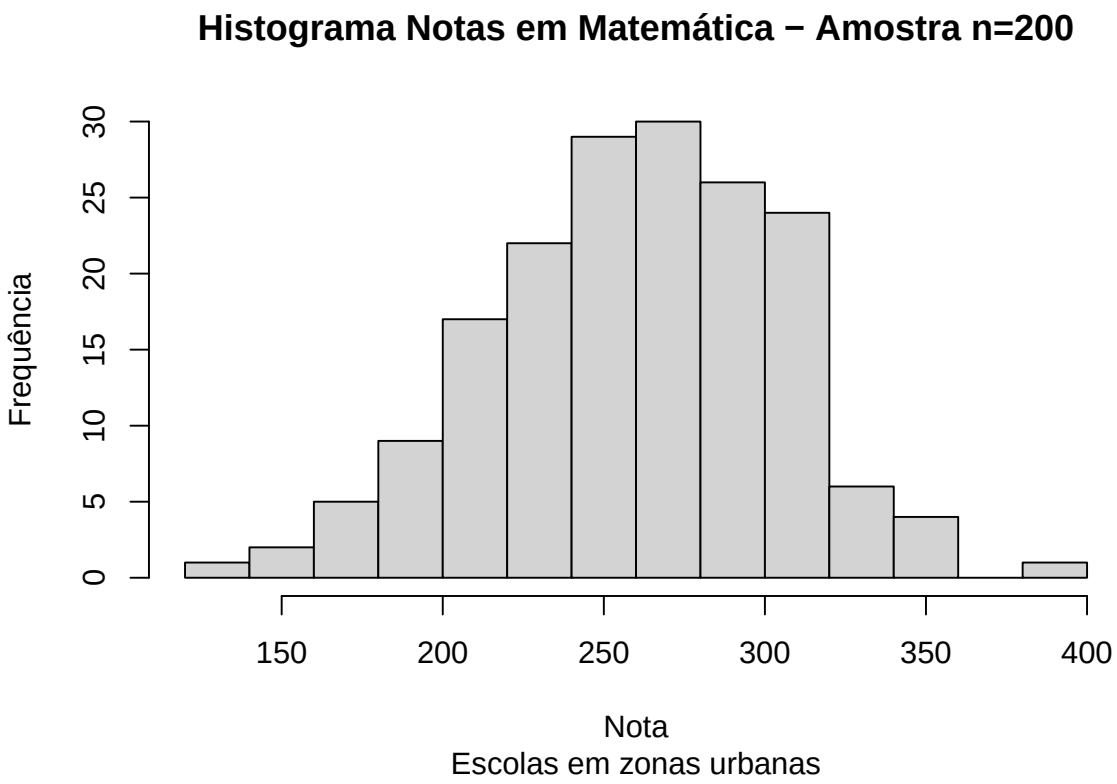


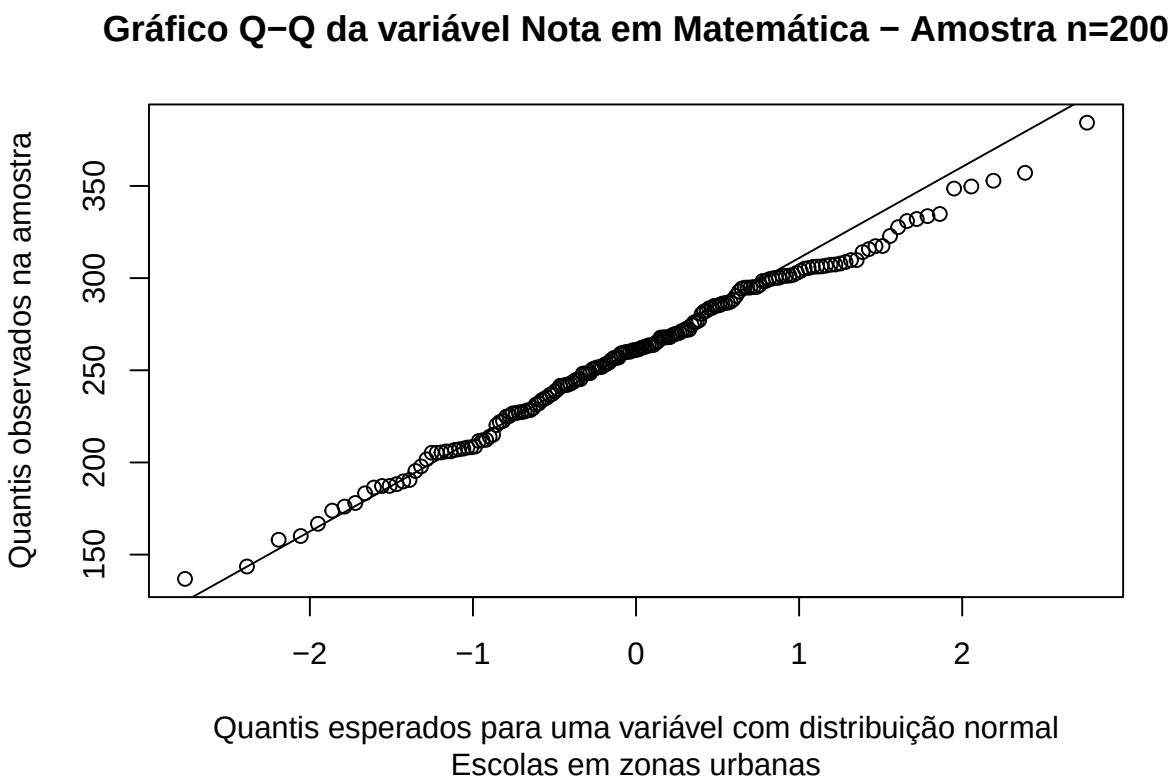
Escolas rurais





Escolas urbanas





Testando a normalidade das variáveis NOTA_MT e NOTA_LP

Testes:

1. **Qui-quadrado**
2. **Shapiro-Wilk**
3. **Anderson-Darling**

OBS: A normalidade desses dados foi melhor explorada nos exercícios anteriores. Aqui se encontra apenas um resumo, com o resultado dos testes de aderência.

Qui-quadrado: [OBS: Não é um bom teste para as amostras n=20]

```
##
##  Pearson chi-square normality test
##
## data:  amostra20$NOTA_LP
## P = 1.7, p-value = 0.7907
```

```
##
##  Pearson chi-square normality test
##
## data:  amostra200$NOTA_LP
## P = 29.67, p-value = 0.008471
```

```
##
##  Pearson chi-square normality test
##
## data:  amostra20$NOTA_MT
## P = 3.8, p-value = 0.4337
```

```
##
##  Pearson chi-square normality test
##
## data:  amostra200$NOTA_MT
## P = 20.32, p-value = 0.1204
```

Shapiro-Wilk:

```
##
##  Shapiro-Wilk normality test
##
## data:  amostra20$NOTA_LP
## W = 0.95157, p-value = 0.3915
```

```
##
##  Shapiro-Wilk normality test
##
## data:  amostra200$NOTA_LP
## W = 0.97833, p-value = 0.003468
```

```
##
##  Shapiro-Wilk normality test
##
## data:  amostra20$NOTA_MT
## W = 0.94585, p-value = 0.3085
```

```
##
##  Shapiro-Wilk normality test
##
## data:  amostra200$NOTA_MT
## W = 0.9947, p-value = 0.705
```

Anderson-Darling:

```
##
##  Anderson-Darling normality test
##
## data:  amostra20$NOTA_LP
## A = 0.29621, p-value = 0.5583
```

```
##
##  Anderson-Darling normality test
##
## data:  amostra200$NOTA_LP
## A = 1.451, p-value = 0.0009283
```

```
##
##  Anderson-Darling normality test
##
## data:  amostra20$NOTA_MT
## A = 0.34378, p-value = 0.4519
```

```
##
##  Anderson-Darling normality test
##
## data:  amostra200$NOTA_MT
## A = 0.35056, p-value = 0.4682
```

Para uma análise mais aprofundada acerca da normalidade, favor conferir minhas atividades 2.2, 3.2 e 3.3.

Testando a homocedasticidade das variáveis NOTA_MT e NOTA_LP

```
##
##  F test to compare two variances
##
## data:  rural20$NOTA_MT and urbana20$NOTA_MT
## F = 0.59448, num df = 3, denom df = 15, p-value = 0.7435
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1431514 8.4729486
## sample estimates:
## ratio of variances
##          0.5944798
```

```
##
##  F test to compare two variances
##
## data:  rural20$NOTA_MT and urbana20$NOTA_MT
## F = 0.59448, num df = 3, denom df = 15, p-value = 0.7435
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1431514 8.4729486
## sample estimates:
## ratio of variances
##          0.5944798
```

```
##
##  F test to compare two variances
##
## data:  antes200$NOTA_LP and depois200$NOTA_LP
## F = 1.0296, num df = 36, denom df = 162, p-value = 0.8671
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6406743 1.8023960
## sample estimates:
## ratio of variances
##           1.029612
```

```
##
##  F test to compare two variances
##
## data:  antes20$NOTA_LP and depois20$NOTA_LP
## F = 1.2538, num df = 4, denom df = 14, p-value = 0.6675
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.3221586 10.8878328
## sample estimates:
## ratio of variances
##           1.253814
```

Portando, os dados apresentam homocedasticidade.

Para a amostra n=20, Comparar as variáveis NOTA_MT e NOTA_LP pareadas.

Testes:

1. **t de Student**
2. **Wilcoxon**
3. **Sinais**

```
##
##  Paired t-test
##
## data:  amostra20$NOTA_MT and amostra20$NOTA_LP
## t = -1.5089, df = 19, p-value = 0.9261
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  -22.3399      Inf
## sample estimates:
## mean difference
##       -10.41031
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  amostra20$NOTA_MT and amostra20$NOTA_LP
## V = 72, p-value = 0.2305
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Dependent-samples Sign-Test
##
## data:  amostra20$NOTA_MT and amostra20$NOTA_LP
## S = 7, p-value = 0.9423
## alternative hypothesis: true median difference is greater than 0
## 95 percent confidence interval:
##  -21.31161       Inf
## sample estimates:
## median of x-y
##     -15.91809
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level   L.E.pt U.E.pt
## Lower Achieved CI     0.9423 -20.6601    Inf
## Interpolated CI       0.9500 -21.3116    Inf
## Upper Achieved CI     0.9793 -23.8045    Inf
```

Todos os testes divergem da região de aceitação, evidenciando que a nota em língua portuguesa não influencia na nota em matemática no mesmo aluno, e vice-versa.