Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

26 setembro 2022

Atividade 4.5 - Análise de dados - comparação entre várias populações

Profª. Ana Maria Nogales

Métodos Estatísticos 2

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

1. Considere uma amostra de tamanho 500. Relacione as seguintes variáveis:

- NOTA_MT e Região
- NOTA_LP e Uso do tempo de tela: considerar as seguintes categorias

1. **Não vê TV..... + Menos de 1 hora (juntar as duas categorias)**
2. **Entre 1 e 2 horas**
3. **Mais de 2, até 3 horas**
4. **Mais de 3 horas**

- Exclui os valores faltantes

Para avaliar essas relações construa os gráficos adequados e medidas de posição e variabilidade segundo categorias das variáveis qualitativas.

Tabela 1: Medidas de posição e variabilidade para a variável Nota em Matemática segundo região geográfica:

REGIAO	Média	Variância	Desvio padrão	Mediana
Norte	256.9442	2103.801	45.86721	253.7189
Nordeste	241.1483	2382.157	48.80735	234.6699
Centro-Oeste	263.3456	1862.276	43.15409	262.9818
Sudeste	258.0409	2276.503	47.71271	262.7613
Sul	276.0202	1711.415	41.36925	273.0607

Tabela 2: Medidas de posição e variabilidade para a variável Nota em língua portuguesa segundo categorias de tempo de uso de telas:

USO_TEMPO_TELAS	Média	Variância	Desvio padrão	Mediana
Até 1h	234.2482	2153.595	46.40684	235.5082
1 a 2 horas	262.2674	2408.030	49.07168	265.3901
2 a 3 horas	268.4701	1951.753	44.17865	268.7871
Mais de 3 horas	266.0293	2392.968	48.91797	267.4759

Você diria que existem diferenças entre as proficiências em matemática segundo a região geográfica da escola?

Pergunta: As média da nota em matemática das escolas são diferentes segundo região geográfica?

Seja:

μ_N = Média da nota em matemática na região Norte

μ_{NE} = Média da nota em matemática na região Nordeste

μ_{CO} = Média da nota em matemática na região Centro-Oeste

μ_{SE} = Média da nota em matemática na região Sudeste

μ_S = Média da nota em matemática na região Sul

Então:

$h_0) = \mu_N = \mu_{NE} = \mu_{CO} = \mu_{SE} = \mu_S$

$h_1)$ = Alguma das médias é diferente.

Primeiramente, devemos testar se as variáveis seguem uma distribuição normal

Seja:

X_N = Distribuição da variável nota em matemática na região Norte

X_{NE} = Distribuição da variável nota em matemática na região Nordeste

X_{CO} = Distribuição da variável nota em matemática na região Centro-Oeste

X_{SE} = Distribuição da variável nota em matemática na região Sudeste

X_S = Distribuição da variável nota em matemática na região Sul

Então:

$h_0) = X_i \sim N(\mu_i, \sigma_i^2); i = \{N, NE, CO, SE, S\}$

$h_1)$ = Alguma distribuição não é normal.

Testes para normalidade das variáveis:

Diversos testes podem ser utilizados para testar a normalidade das variáveis. Neste caso, para cada variável, farei os seguintes testes:

- Teste de *Kolmogorov-Smirnov*
- Teste de *Anderson-Darling*
- Teste de *Shapiro-Wilk*

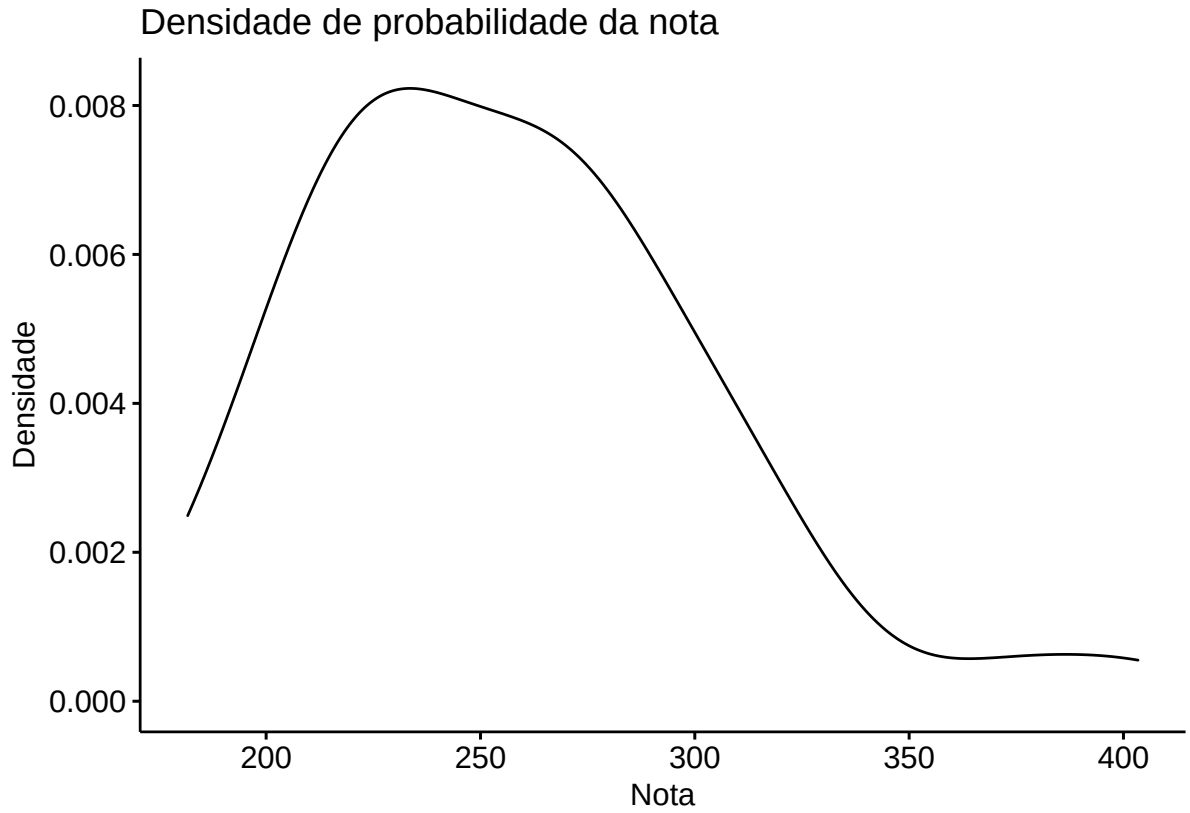
Além desses testes, também farei gráficos de densidade de probabilidade e quantil-quantil, para trazer uma abordagem visual acerca da normalidade de cada uma.

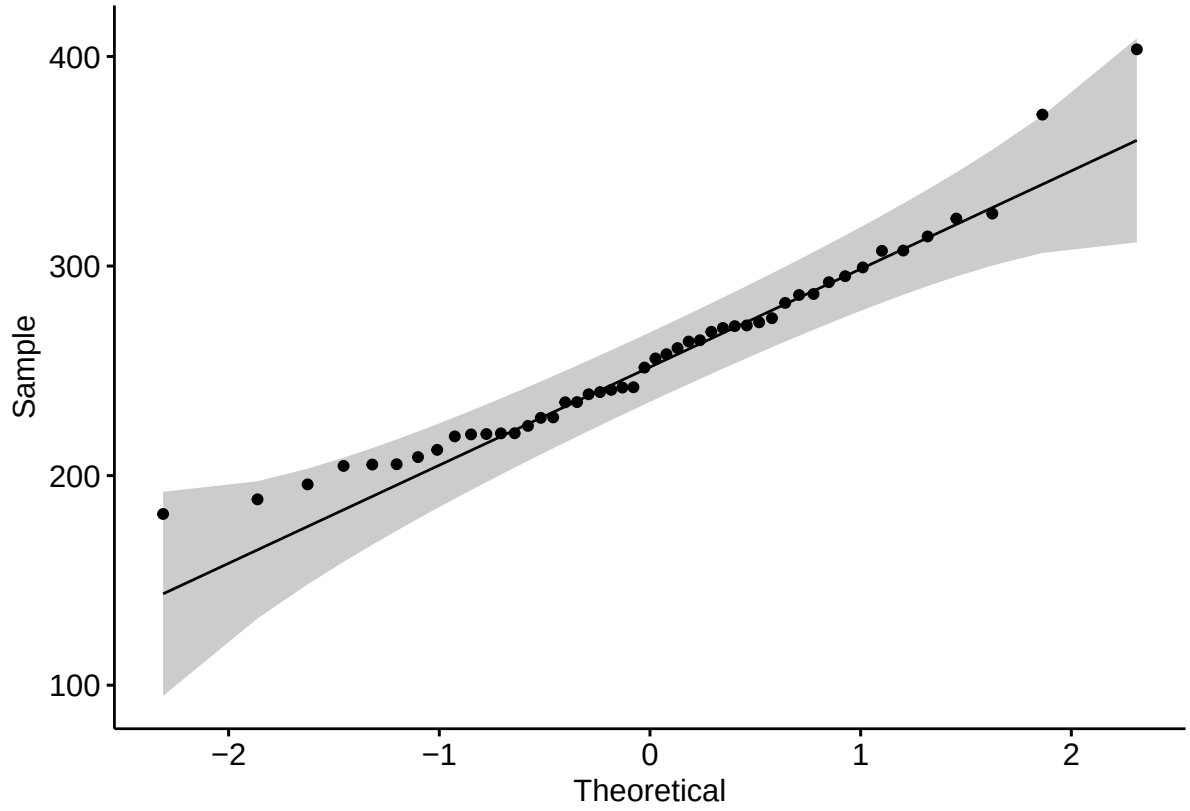
Norte:

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  NO$NOTA_MT
## D = 1, p-value = 8.882e-16
## alternative hypothesis: two-sided
```

```
##
##  Anderson-Darling normality test
##
## data:  NO$NOTA_MT
## A = 0.53247, p-value = 0.1648
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NO$NOTA_MT
## W = 0.94876, p-value = 0.0357
```





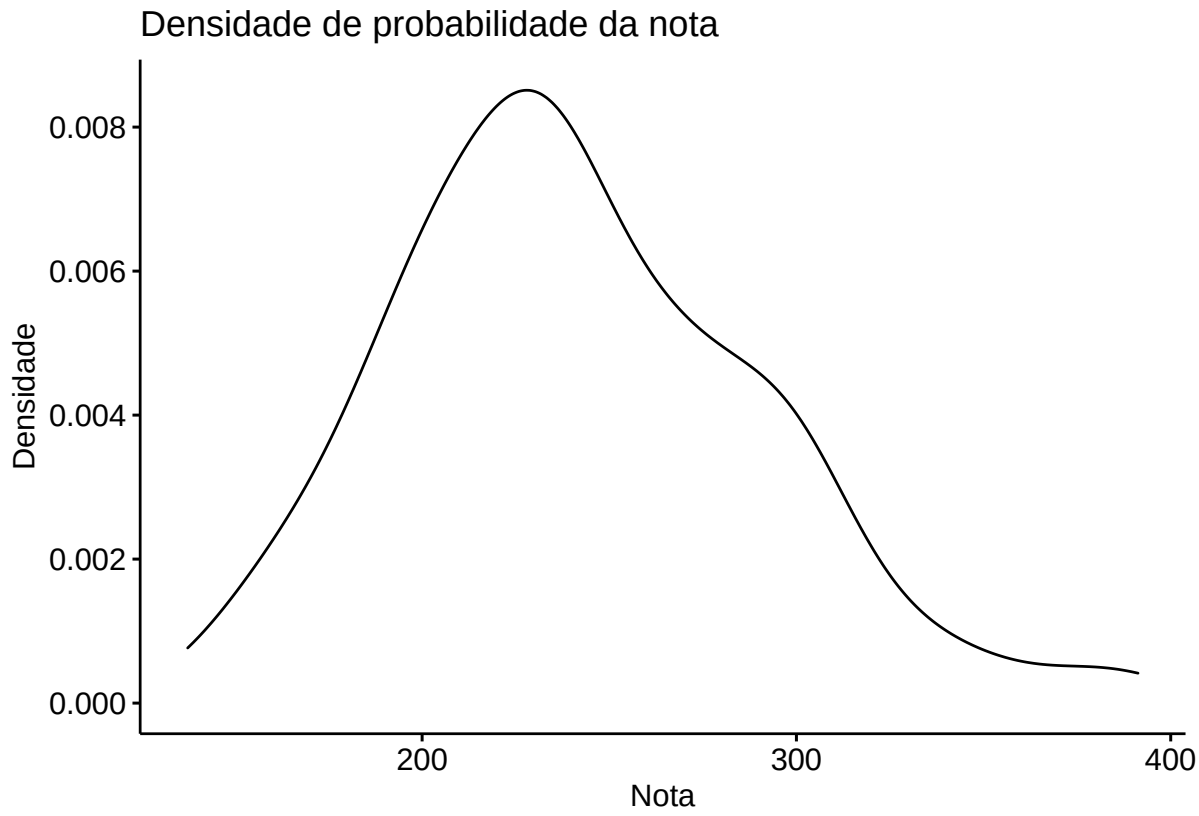
É Normal.

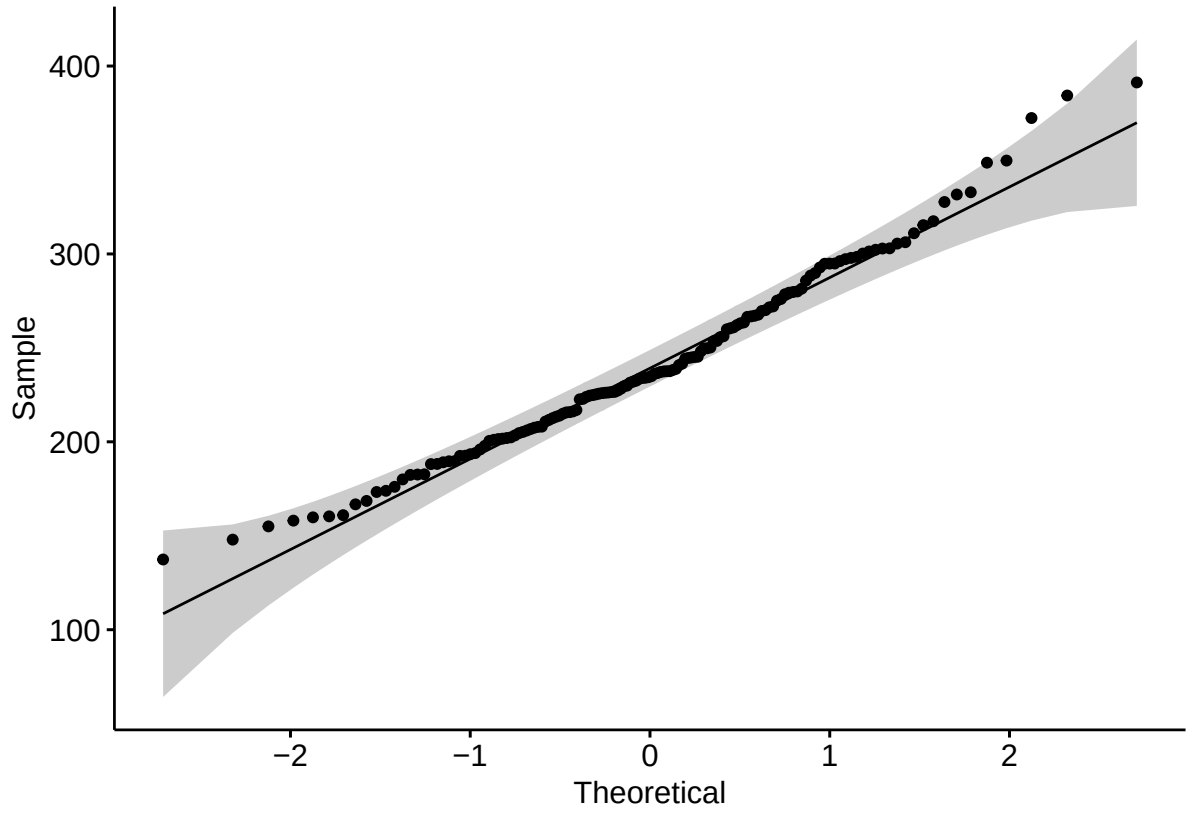
Nordeste:

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  NE$NOTA_MT
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Anderson-Darling normality test
##
## data:  NE$NOTA_MT
## A = 0.72379, p-value = 0.05785
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NE$NOTA_MT
## W = 0.98018, p-value = 0.03096
```





É Normal.

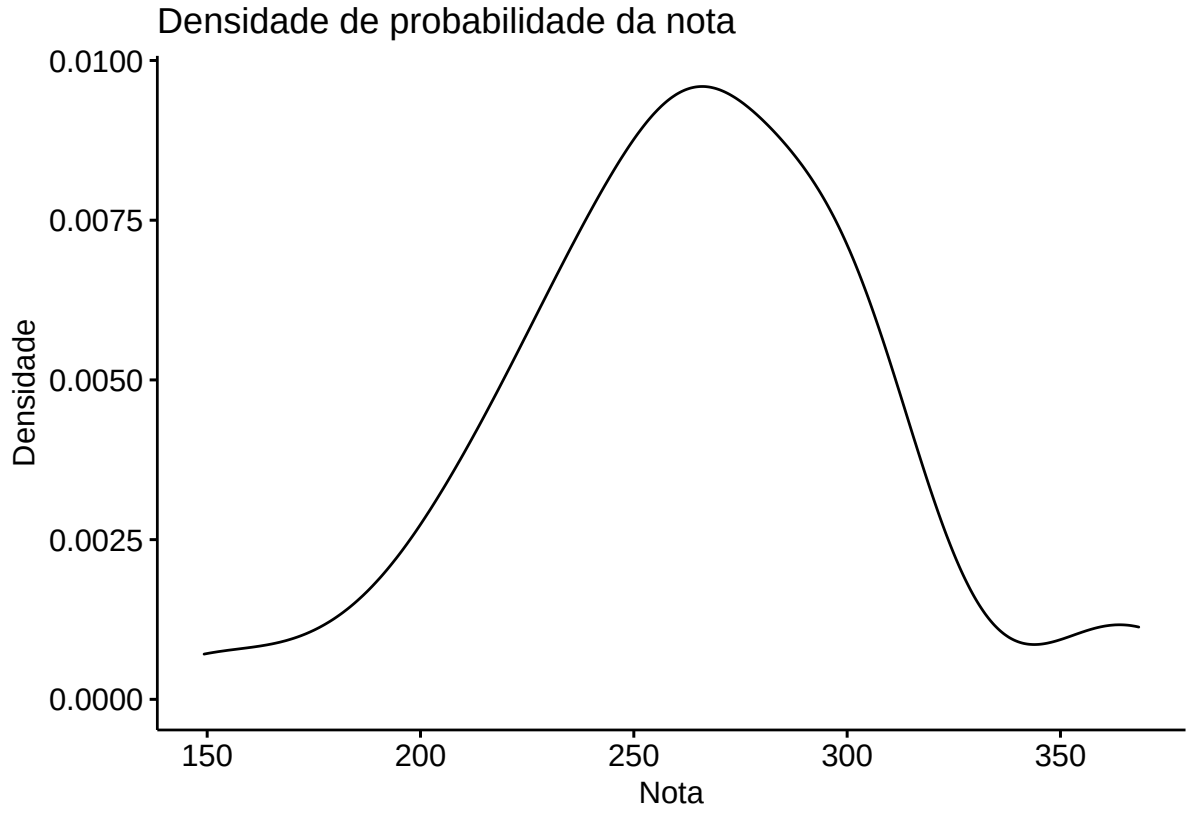
Centro-Oeste:

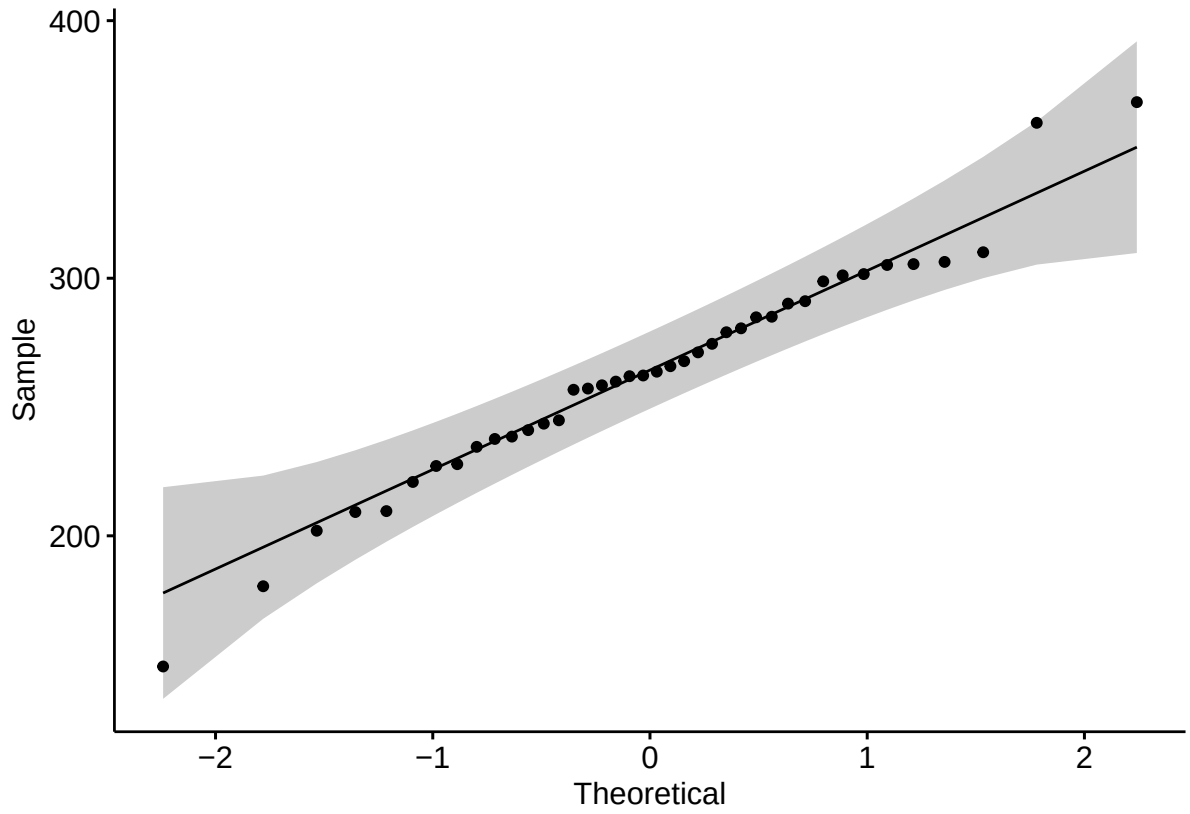
##

```
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  CO$NOTA_MT
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided

##
##  Anderson-Darling normality test
##
## data:  CO$NOTA_MT
## A = 0.31681, p-value = 0.5264

##
##  Shapiro-Wilk normality test
##
## data:  CO$NOTA_MT
## W = 0.97789, p-value = 0.6118
```





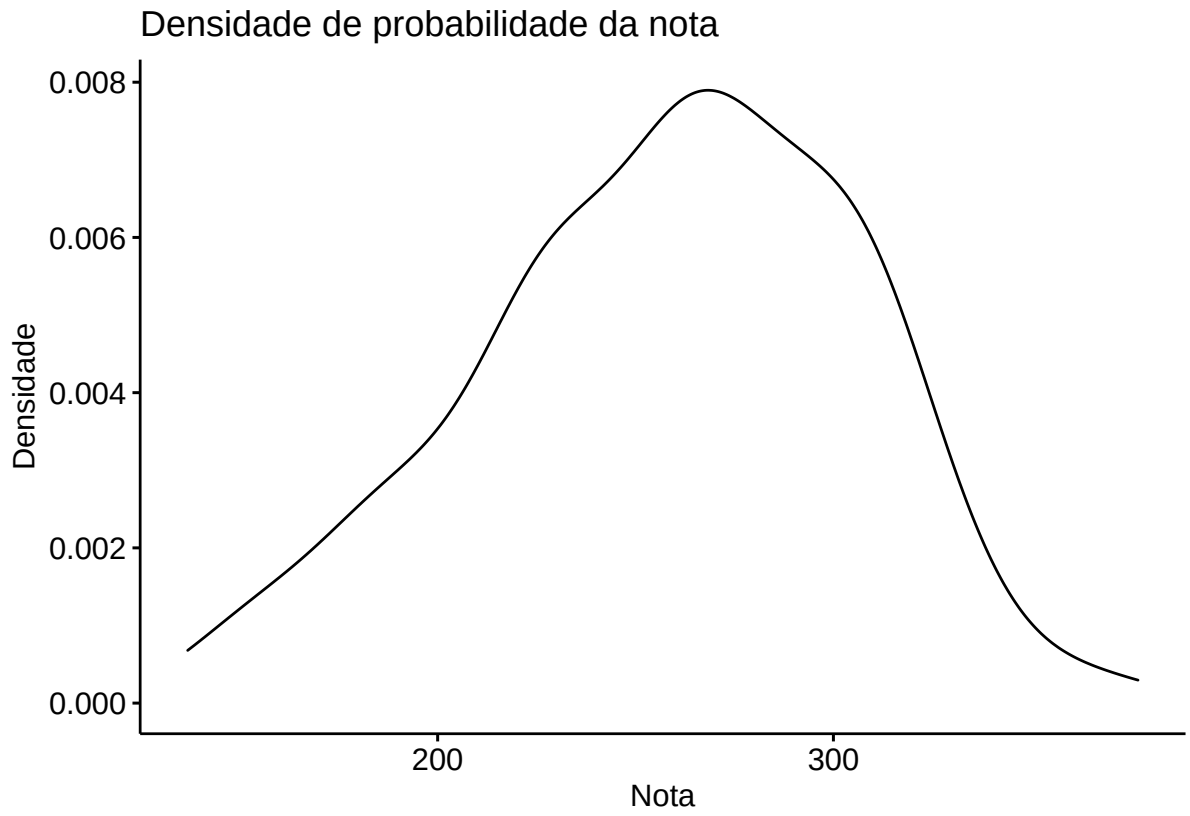
É Normal.

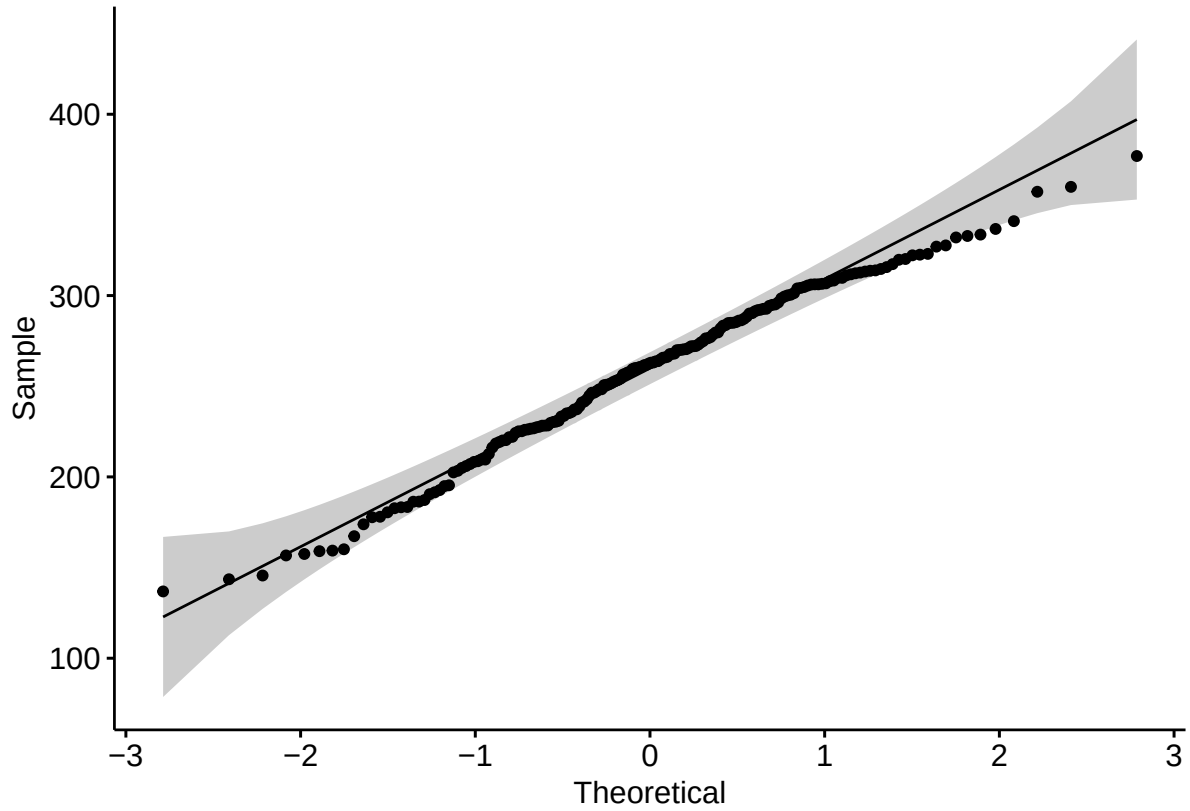
Sudeste:

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  SE$NOTA_MT
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Anderson-Darling normality test
##
## data:  SE$NOTA_MT
## A = 0.76054, p-value = 0.04712
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SE$NOTA_MT
## W = 0.98643, p-value = 0.06749
```


Densidade de probabilidade da nota
Densidade
0.008
0.006
0.004
0.002
0.000
200
300
Nota


Sample
400
300
200
100
−3
−2
−1
0
1
2
3
Theoretical

É Normal.

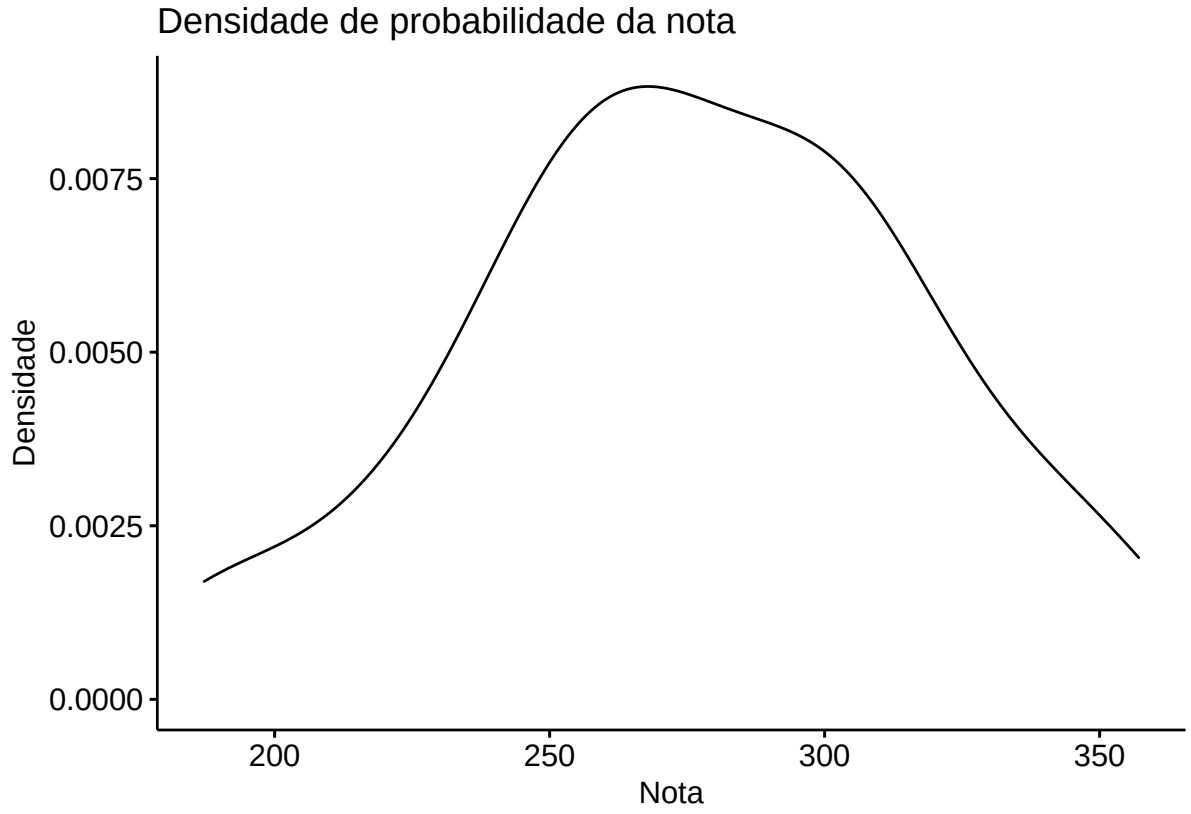
Sul:

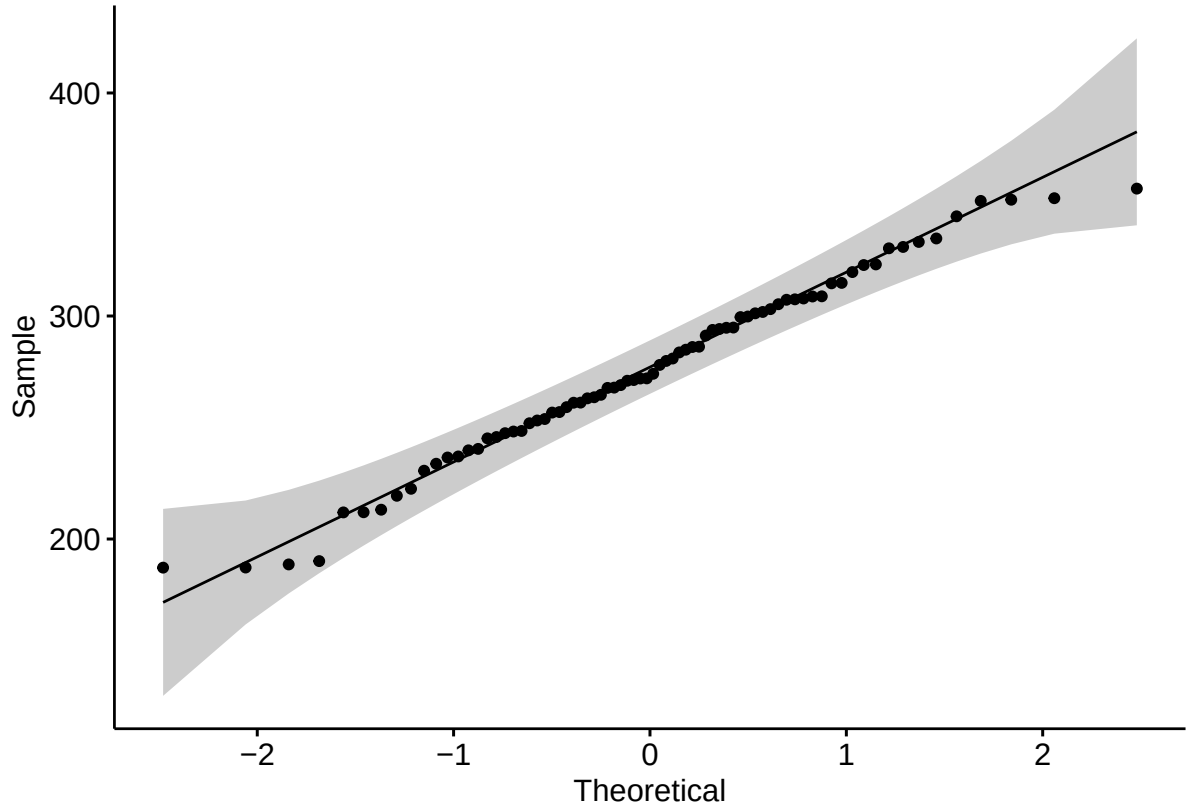
##

```
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  SUL$NOTA_MT
## D = 1, p-value = 2.22e-16
## alternative hypothesis: two-sided
```

```
##
##  Anderson-Darling normality test
##
## data:  SUL$NOTA_MT
## A = 0.18238, p-value = 0.9089
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SUL$NOTA_MT
## W = 0.98421, p-value = 0.4666
```





É Normal.

Portanto, a distribuição das notas de matemática segundo cada região geográfica respeita a distribuição normal, conforme esperado, e em concordância com os exercícios anteriormente realizados sob essa variável.

Voltamos, portanto, para as primeiras hipóteses h0 e h1, onde vamos testar se médias são iguais.

$h_0) = \mu_N = \mu_{NE} = \mu_{CO} = \mu_{SE} = \mu_S$

$h_1)$ = Alguma das médias é diferente.

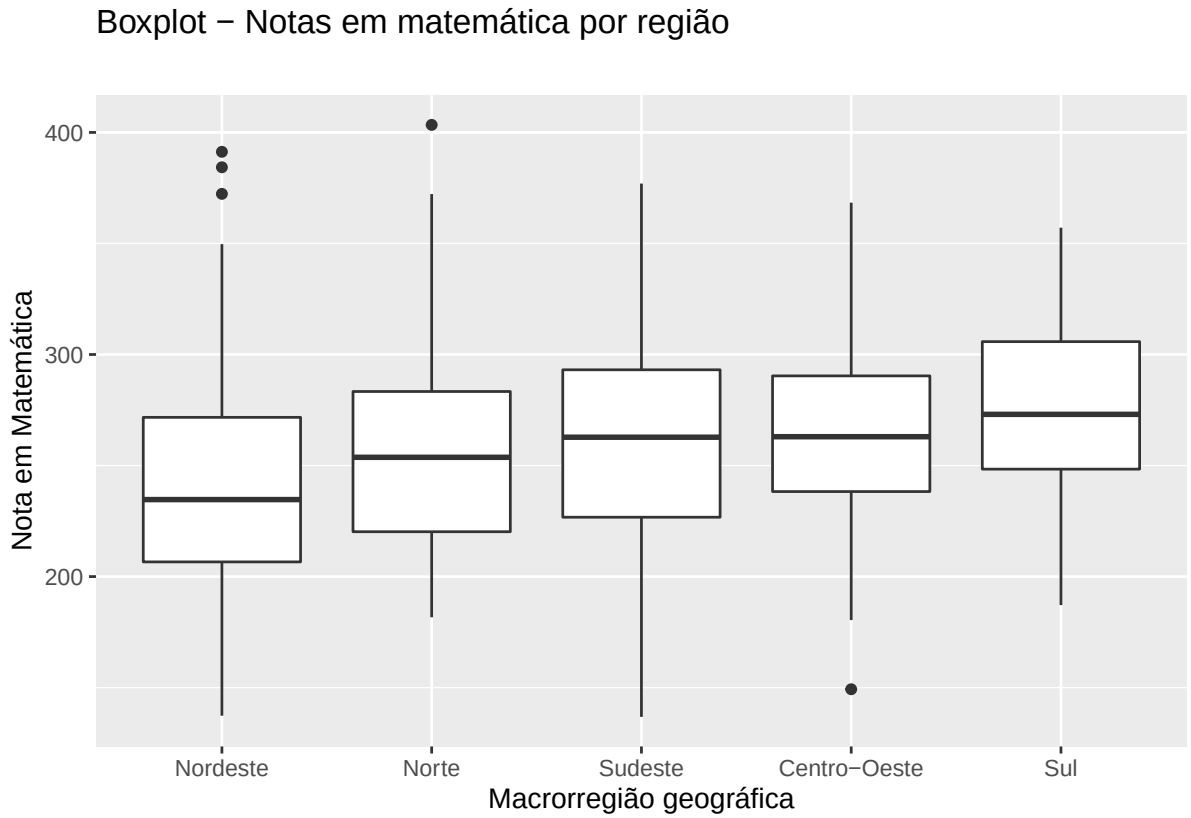
Nível de significância: $\alpha = 0,05$

Estatística de teste: Teste de hipótese para igualdade de médias de várias populações que seguem distribuição normal:

ANOVA:

```
##                 Df  Sum Sq Mean Sq F value   Pr(>F)
## amostra$REGIAO   4   66086   16522   7.602 5.98e-06 ***
## Residuals      495 1075747    2173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Boxplot: Média das notas segundo região geográfica:



Portanto, rejeitamos a hipótese h_0) de igualdade das médias. Ou seja, existe diferença na proeficiência segundo região geográfica.

Descobrindo aonde estão as desigualdades, utilizando o método de ajustamento de *Bonferroni*:

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  amostra$NOTA_MT and as.factor(amostra$REGIAO)
##
##           Centro-Oeste Nordeste Norte Sudeste
## Nordeste  0.078        -        -     -
## Norte     1.000        0.419    -     -
## Sudeste   1.000        0.010    1.000 -
## Sul       1.000        1.7e-06  0.269 0.047
##
## P value adjustment method: bonferroni
```

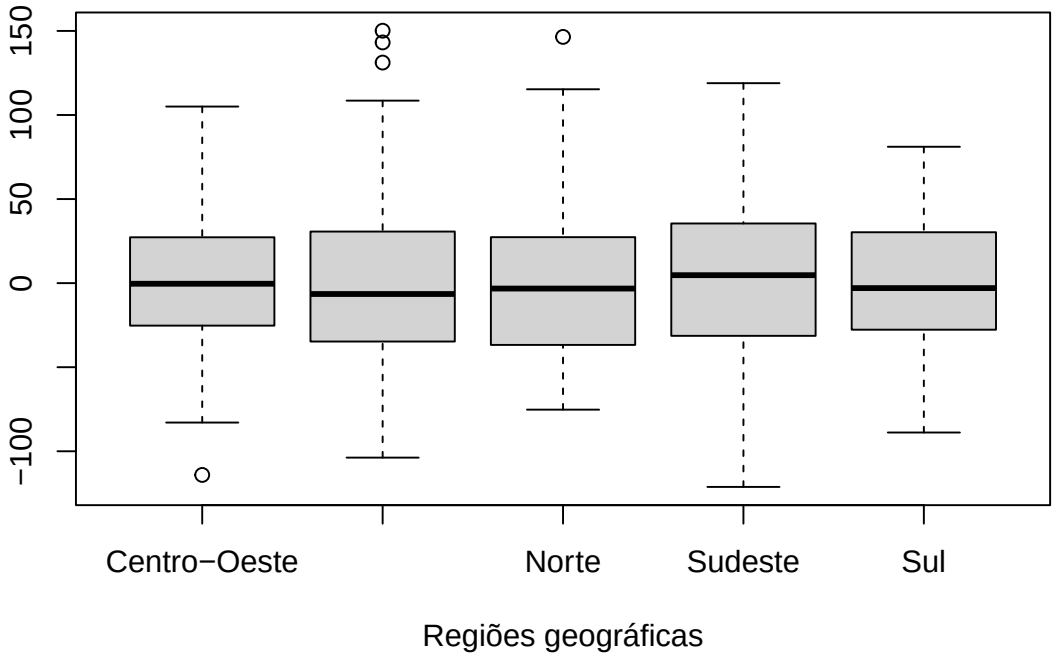
Concluímos portanto que:

$$\mu_{CO} = \mu_N = \mu_{SE} = \mu_S$$
$$\mu_{CO} \neq \mu_{NE}$$
$$\mu_{SE} \neq \mu_{NE}$$
$$\mu_S \neq \mu_{NE}$$
$$\mu_S \neq \mu_{SE}$$

A ANOVA tem como pressuposto a homocedasticidade das variâncias entre as variáveis. Verificaremos isto com um boxplot e com o teste de Bartlett.

As variâncias são iguais?

Boxplot das variâncias:



Teste de Bartlett:

```
##
##  Bartlett test of homogeneity of variances
##
## data:  mod1$res by amostra$REGIAO
## Bartlett's K-squared = 3.27, df = 4, p-value = 0.5137
```

Notamos que o boxplot e o teste nos sugerem a homocedasticidade das variâncias.

Existe diferença entre as proficiências em língua portuguesa segundo categoria de uso do tempo de tela?

Pergunta: As média da nota em língua portuguesa das escolas são diferentes segundo categoria de uso do tempo de tela?

Seja:

μ_1 = Média da nota em língua portuguesa na categoria uso de tela: Até 1h

μ_2 = Média da nota em língua portuguesa na categoria uso de tela: Entre 1 e 2 horas

μ_3 = Média da nota em língua portuguesa na categoria uso de tela: Entre 2 e 3 horas

μ_4 = Média da nota em língua portuguesa na categoria uso de tela: Mais de 3 horas

Então:

$h_0) = \mu_1 = \mu_2 = \mu_3 = \mu_4$

$h_1)$ = Alguma das médias é diferente.

Primeiramente, devemos testar se as variáveis seguem uma distribuição normal

Seja:

X_1 = Distribuição da variável nota em língua portuguesa na categoria uso de tela: Até 1h

X_2 = Distribuição da variável nota em língua portuguesa na categoria uso de tela: Entre 1 e 2 horas

X_3 = Distribuição da variável nota em língua portuguesa na categoria uso de tela: Entre 2 e 3 horas

X_4 = Distribuição da variável nota em língua portuguesa na categoria uso de tela: Mais de 3 horas

Então:

$h_0) = X_i \sim N(\mu_i, \sigma_i^2); i = \{1, 2, 3, 4\}$

$h_1)$ = Alguma distribuição não é normal.

Testes para normalidade das variáveis:

Diversos testes podem ser utilizados para testar a normalidade das variáveis. Neste caso, para cada variável, farei os seguintes testes:

- Teste de *Kolmogorov-Smirnov*
- Teste de *Anderson-Darling*
- Teste de *Shapiro-Wilk*

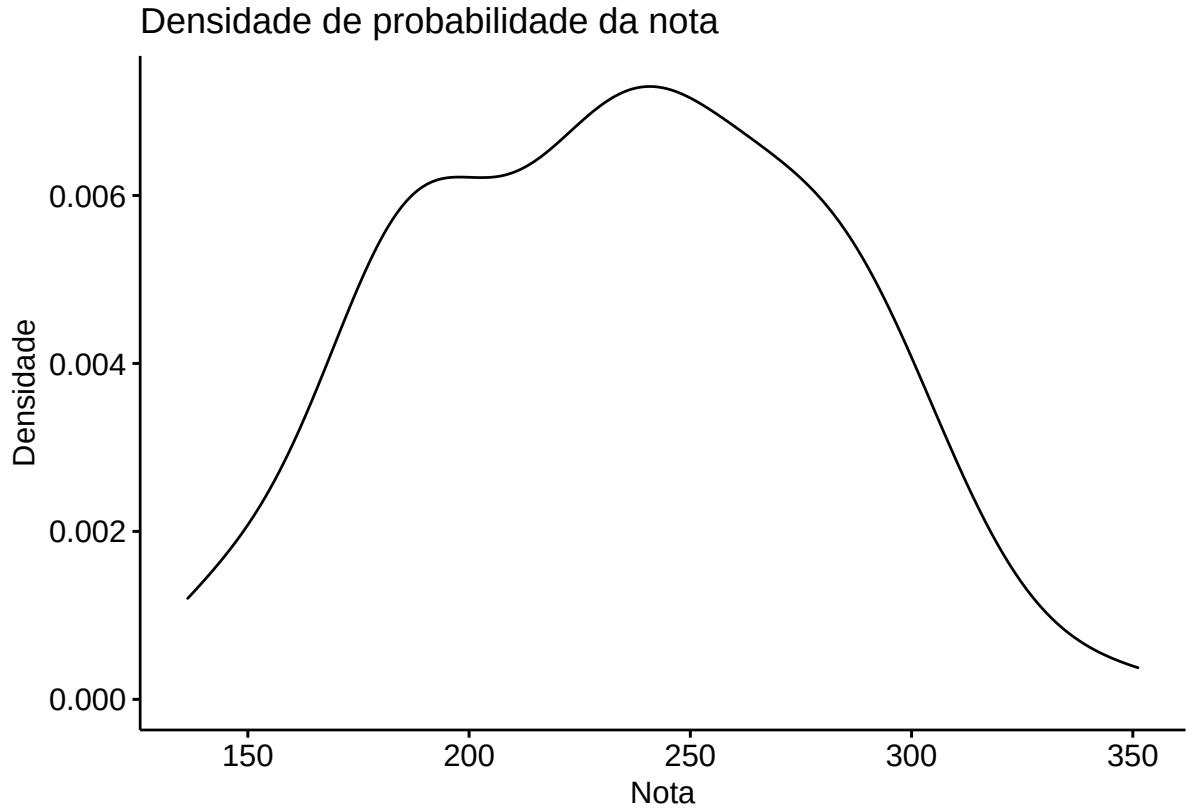
Além desses testes, também farei gráficos de densidade de probabilidade e quantil-quantil, para trazer uma abordagem visual acerca da normalidade de cada uma.

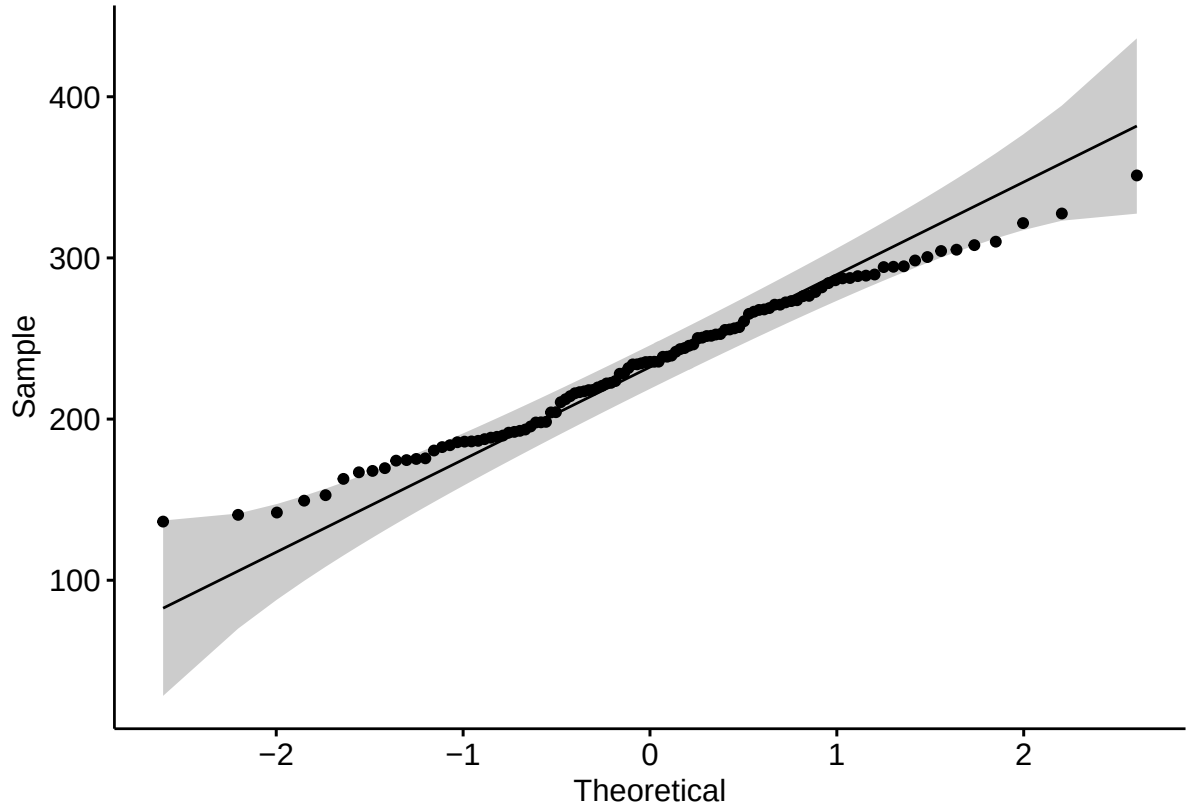
Até 1h:

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  menos1$NOTA_LP
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Anderson-Darling normality test
##
## data:  menos1$NOTA_LP
## A = 0.44426, p-value = 0.28

##
##  Shapiro-Wilk normality test
##
## data:  menos1$NOTA_LP
## W = 0.98709, p-value = 0.3811
```





É Normal

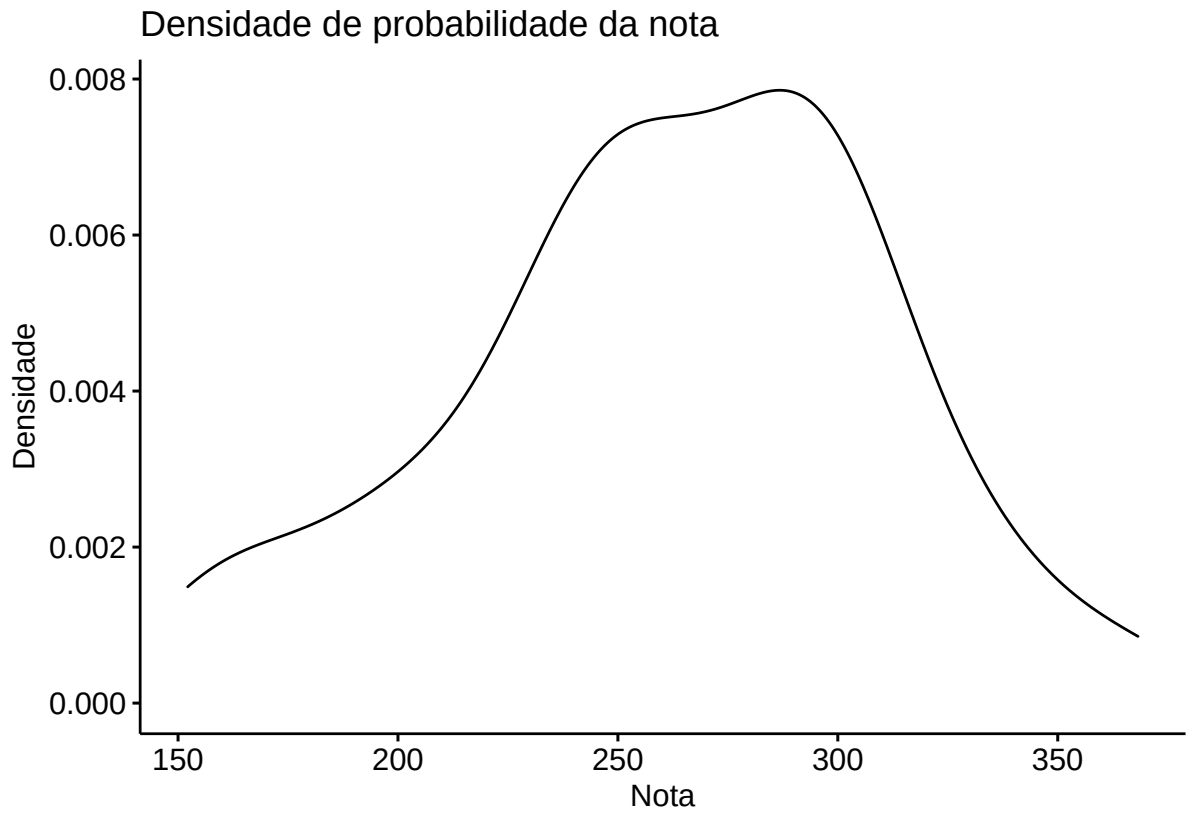
Entre 1 e 2 horas:

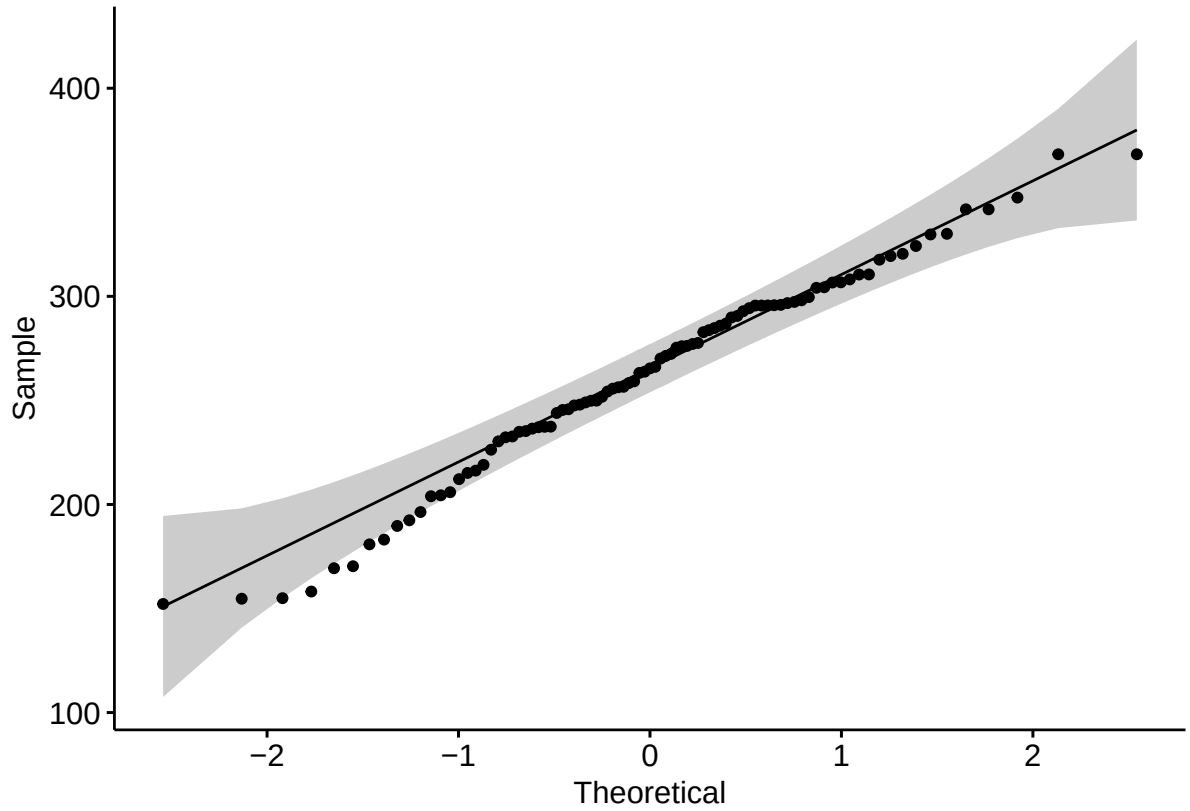
```
## Warning in ks.test.default(x = umaa2$NOTA_LP, y = pnorm): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  umaa2$NOTA_LP
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Anderson-Darling normality test
##
## data:  umaa2$NOTA_LP
## A = 0.47868, p-value = 0.23
```

```
##
##  Shapiro-Wilk normality test
##
## data:  umaa2$NOTA_LP
## W = 0.98123, p-value = 0.2138
```





É Normal

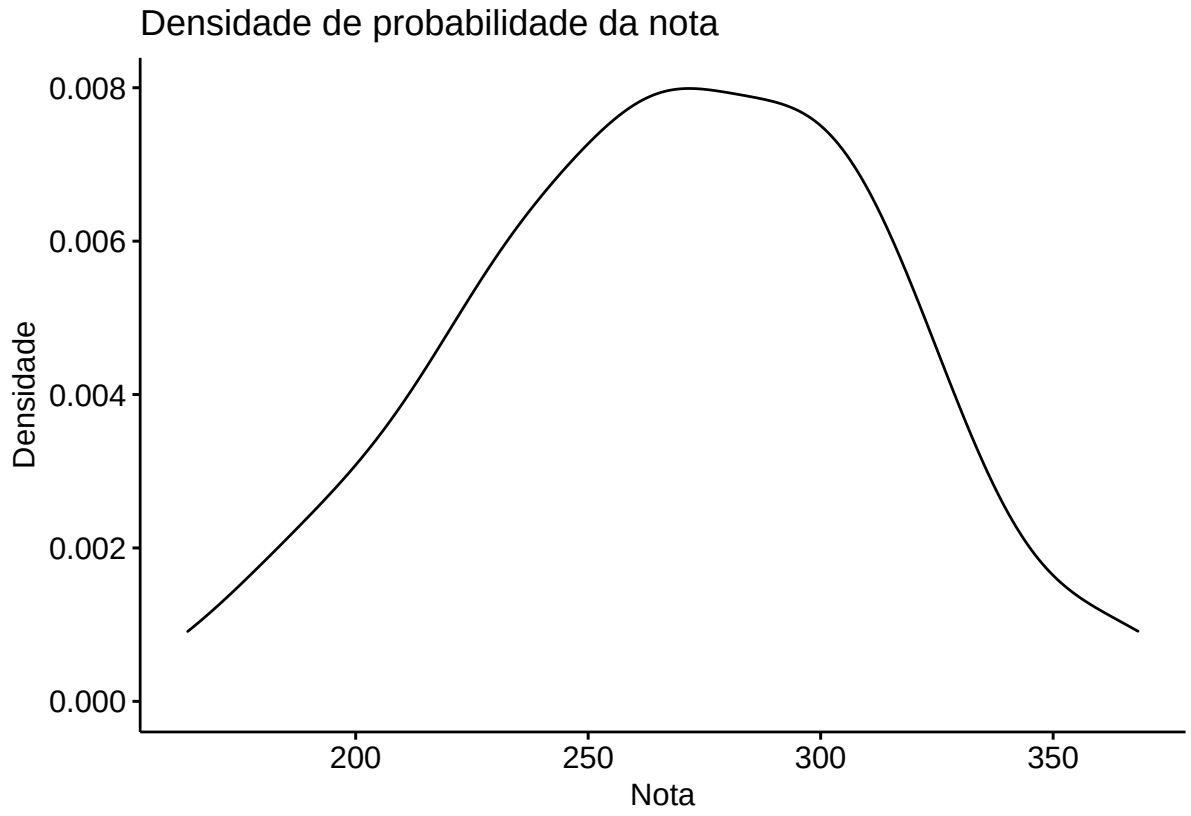
Entre 2 e 3 horas:

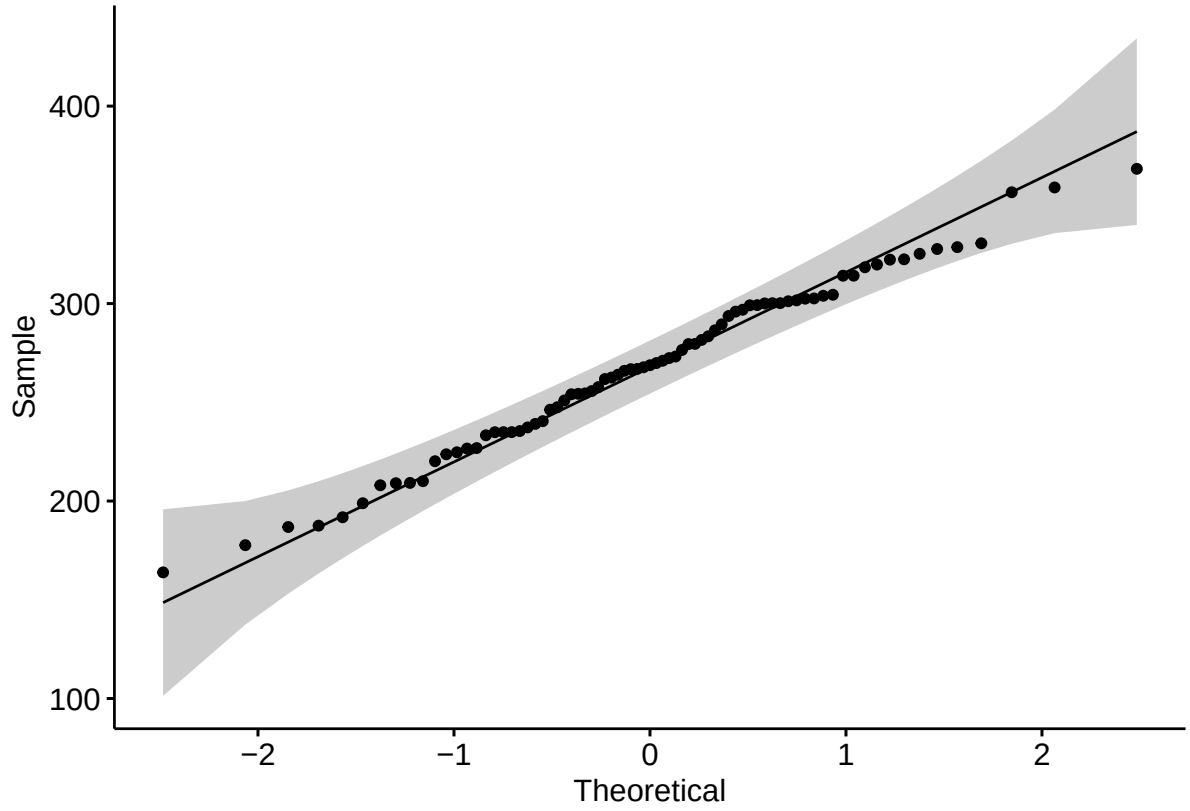
##

```
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  duasa3$NOTA_LP
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided

##
##  Anderson-Darling normality test
##
## data:  duasa3$NOTA_LP
## A = 0.26351, p-value = 0.6907

##
##  Shapiro-Wilk normality test
##
## data:  duasa3$NOTA_LP
## W = 0.9896, p-value = 0.79
```





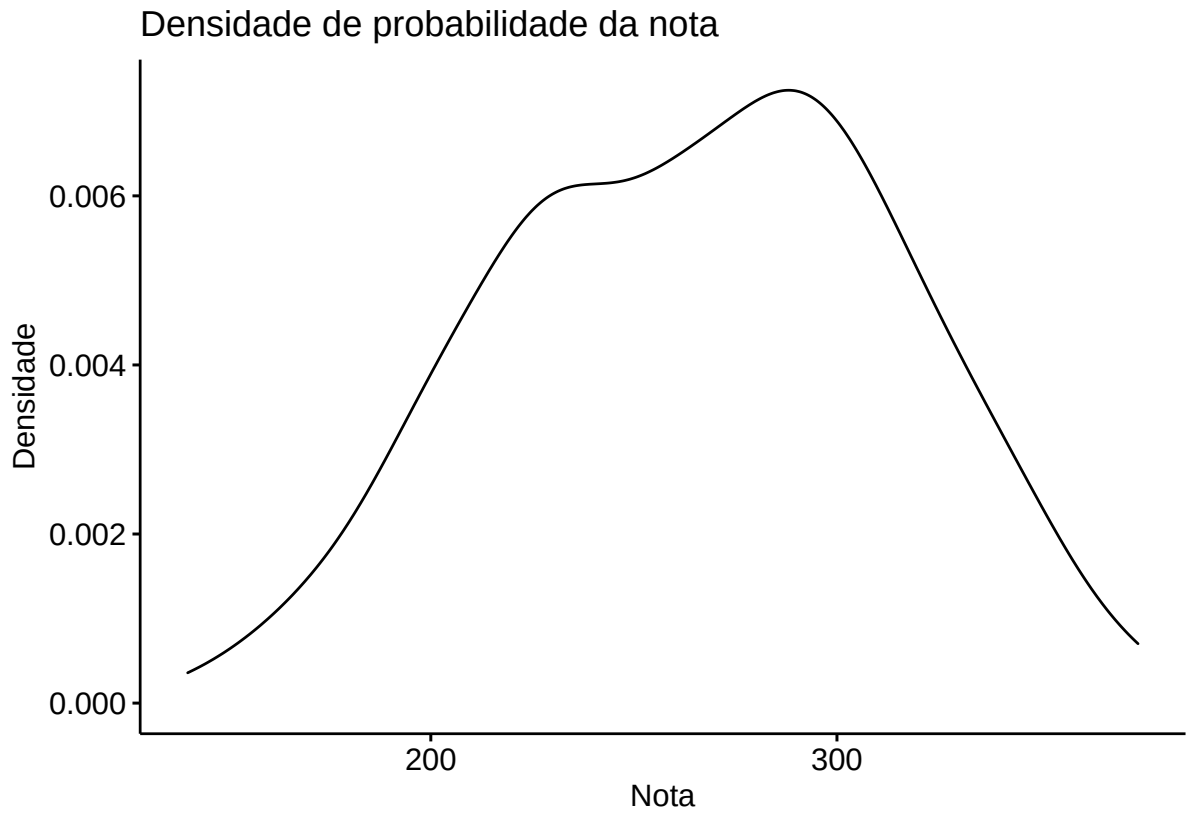
É Normal

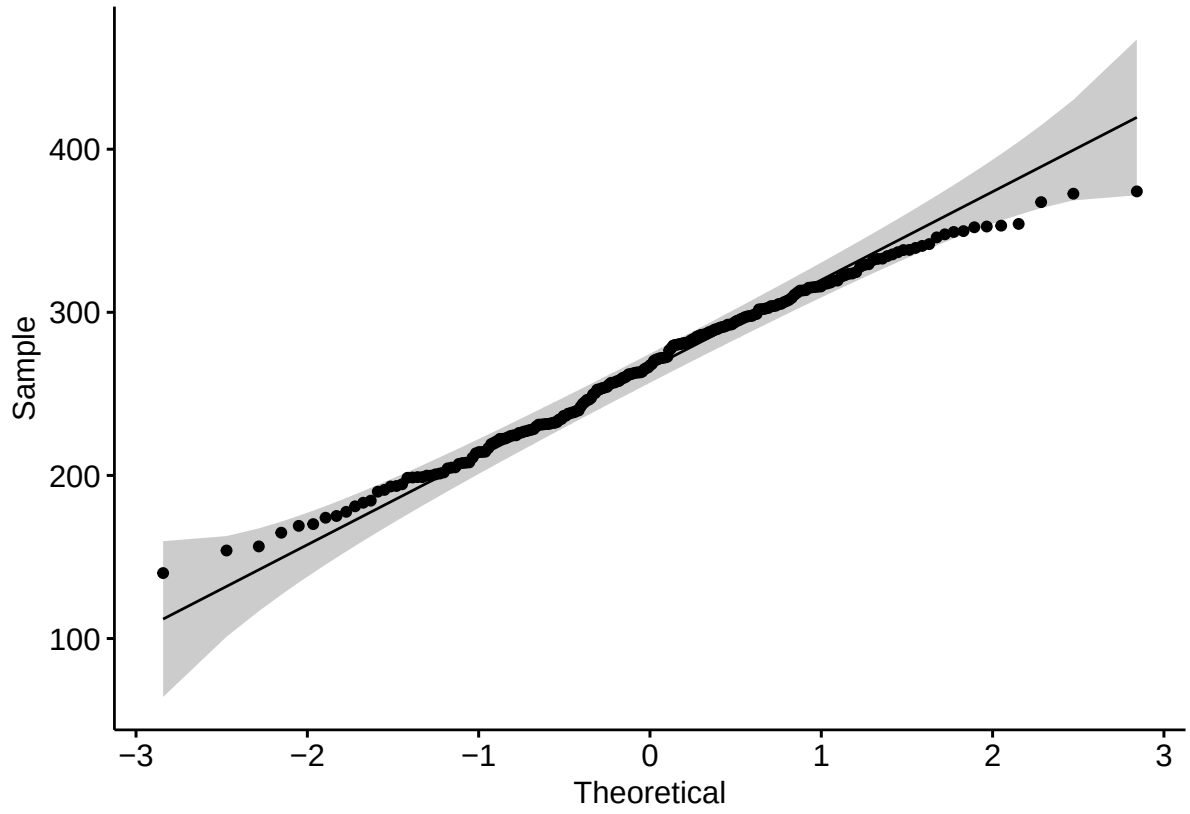
Mais de 3 horas:

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  mais3$NOTA_LP
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Anderson-Darling normality test
##
## data:  mais3$NOTA_LP
## A = 0.63159, p-value = 0.09858
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mais3$NOTA_LP
## W = 0.98991, p-value = 0.1219
```





É Normal

Portanto, a distribuição das notas de língua portuguesa segundo categorias de tempo de uso de tela respeita a distribuição normal, conforme esperado e em concordância com os exercícios anteriormente realizados sob essa variável.

Voltamos, portanto, para as primeiras hipóteses h0 e h1, onde vamos testar se médias são iguais.

h_0) = $\mu_1 = \mu_2 = \mu_3 = \mu_4$

h_1) = Alguma das médias é diferente.

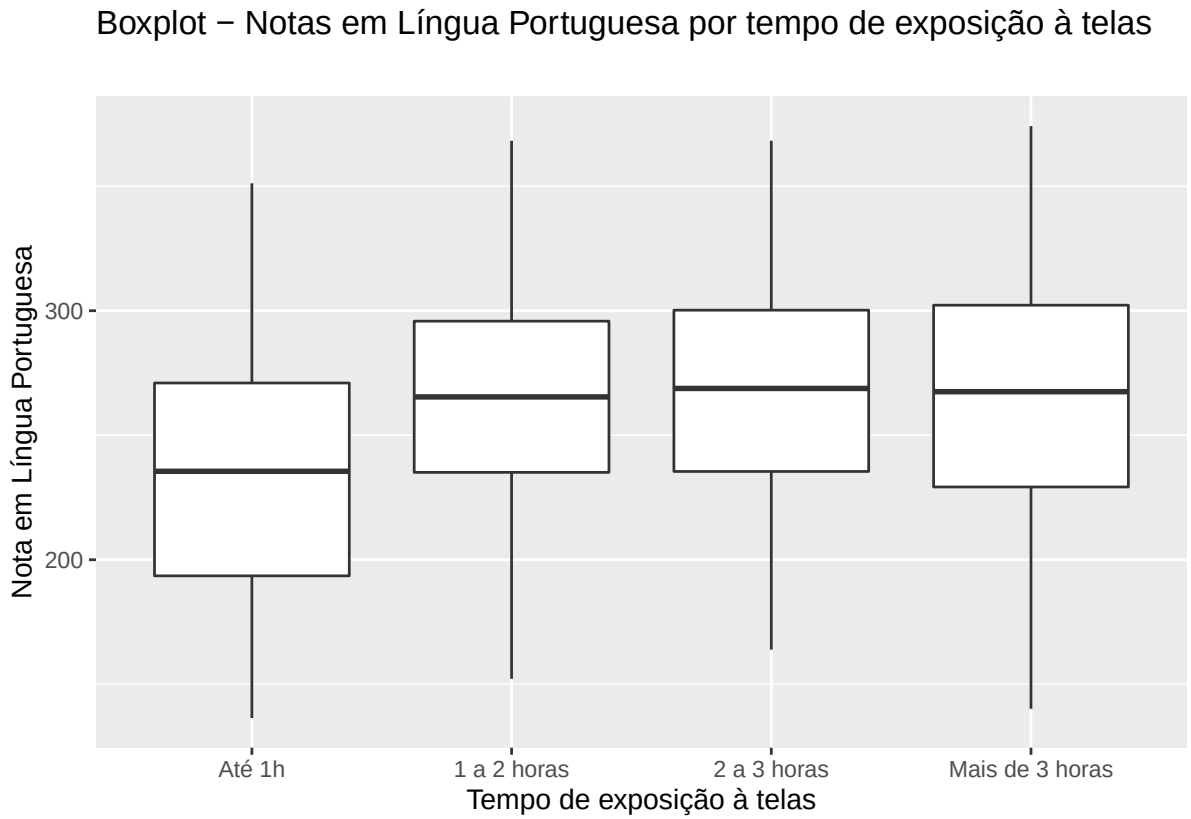
Nível de significância: $\alpha = 0,05$

Estatística de teste: Teste de hipótese para igualdade de médias de várias populações que seguem distribuição normal:

ANOVA:

```
##                           Df  Sum Sq Mean Sq F value   Pr(>F)
## amostra$USO_TEMPO_TELAS    3   85653   28551   12.54 6.44e-08 ***
## Residuals                496 1128883    2276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Boxplot: Média das notas segundo categoria de uso de tempo de tela:



Portanto, rejeitamos a hipótese h_0) de igualdade das médias. Ou seja, existe diferença na proeficiência segundo categoria de tempo de uso de telas.

Descobrindo aonde estão as desigualdades, utilizando o método de ajustamento de *Bonferroni*:

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  amostra$NOTA_LP and as.factor(amostra$USO_TEMPO_TELAS)
##
##                1 a 2 horas 2 a 3 horas Até 1h
```

```
## 2 a 3 horas     1.00000     -           -
## Até 1h          0.00025     1.2e-05     -
## Mais de 3 horas 1.00000     1.00000     1.2e-07
##
## P value adjustment method: bonferroni
```

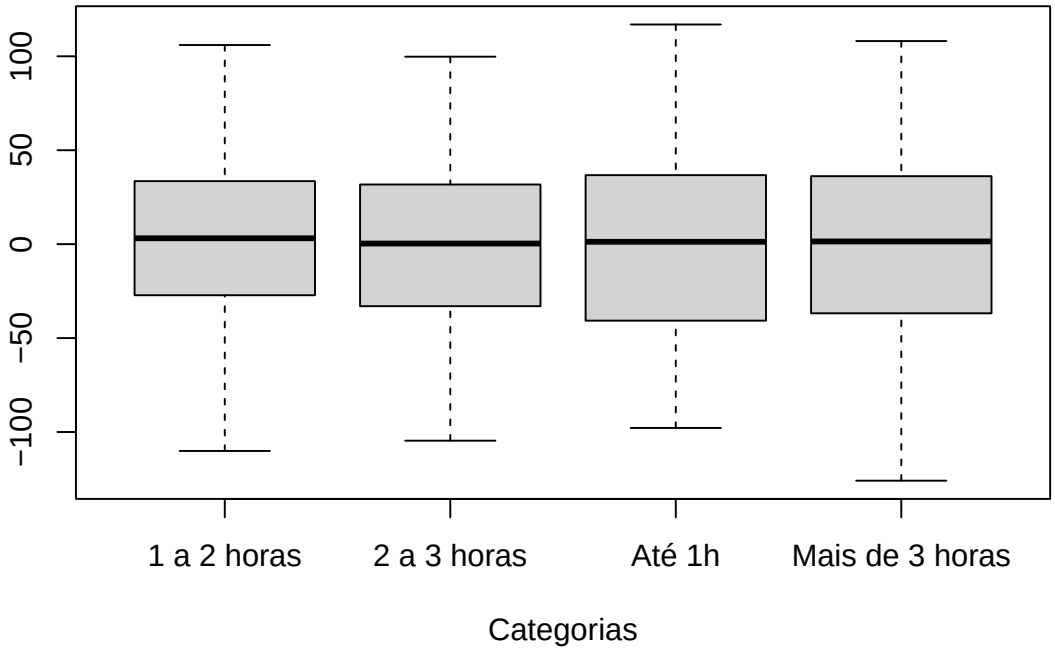
Concluímos portanto que:

$$\mu_2 = \mu_3 = \mu_4 \neq \mu_1$$

A ANOVA tem como pressuposto a homocedasticidade das variâncias entre as variáveis. Verificaremos isto com um boxplot e com o teste de Bartlett.

As variâncias são iguais?

Boxplot das variâncias:



Teste de Bartlett:

```
##
##  Bartlett test of homogeneity of variances
##
## data:  mod2$res by amostra$USO_TEMPO_TELAS
## Bartlett's K-squared = 1.4389, df = 3, p-value = 0.6964
```

Notamos que o boxplot e o teste nos sugerem a homocedasticidade das variâncias.