Estatística para Ciência de Dados
Resolução do trabalho 04: Ex4-teste-normalidade - três variáveis

Eduardo Façanha Dutra
2016473

Conteúdo

1 Enunciado

Objetivos: Analisar a distribuição da amostra onde o fenômeno Tempo agora diz respeito aos valores de 1 fator e 3 níveis Fazer os testes para análise da variância do fenômeno.

Considere o caso em que o tempo do usuário é a variável independente (calculada) e representando o tempo que o usuário passou em uma determinada conferência virtual, quando fez uso de um dos Meet virtuais, se usando Zoom, Hangout ou Skype. A hipótese é saber se existe diferença significativa entre os três Meet. O arquivo segue os seguintes princípios para a realização deste trabalho: - independência dos dados, quem usou um meet não usou o outro. - A variável tempo é mais próxima de uma log normalidade, porque a medida que o usuário usa um sistema, ele se torna mais especialista e o tempo, no eixo X tende a diminuir com o tempo; ou ainda tem poucas atividades que levam muito tempo e muitas que levam pouco tempo, afetadas pela experiência do usuário.

Pede-se: Fazer os testes para análise da variância do fenômeno, realizando as três técnicas dadas a seguir e considerando que:

1. a técnica Shapiro-wilk permite testar a normalidade, para uma amostra pequena
2. a técnica Kolmogorov-Smirnov permite testar a lognormalidade da amostra.
3. a técnica Levene
4. Visualize os dados usando Boxplot, histograma e qqplot.

Senão houver normalidade da amostra, então transforme os dados em uma log normal, depois, verifique como ficaram os dados, e repita os testes dos passos.

Obs: In a test statistic: the result expresses in a single number how much my data differ from my null hypothesis. So it indicates to what extent the observed scores deviate from a normal distribution. Now, if my null hypothesis is true, then this deviation percentage should probably be quite small. That is, a small deviation has a high probability value or p-value. Reversely, a huge deviation percentage is very unlikely and suggests that my reaction times don't follow a normal distribution in the entire population. So a large deviation has a low p-value. As a rule of thumb, we reject the null hypothesis if $p < 0.05$. So if $p < 0.05$, we don't believe that our variable follows a normal distribution in our population.

Resolução:

Resumo:

- Análises gráficas podem ser muito elucidativas para conhecer os dados que o pesquisador está trabalhando, entretanto testes estatísticos são necessários para obter evidências mais precisas;

- Curtose e assimetria podem indicar desvio da normalidade mas não podem ser utilizadas isoladamente para se obter conclusões;

- A amostra Zoom segue a normalidade para o teste de Komogorov-Smirnov(P= 0.3517). Para o teste de Shapiro-Wilk (P= 0.004191) a normalidade só é verificada se o teste for realizado excluindo-se o valor extremo (P= 0.9482). O teste sem o valor extremo só foi cogitado após a visualização gráfica dos dados;

- Quanto ao teste de lognormalidade da amostra Zoom (P= 0.8181), o resultado se mostrou positivo. Esse resultado pode ser devido ao baixo número de amostras, entretanto observou-se que o valor p da lognormalidade superou o da normalidade para o mesmo teste (0.8181 contra 0.3517);

- A amostra Hangout não pode ser considerada normal para o teste de Shapiro-Wilk (P= 0.01281), mas pode ser considerada para o teste de Komogorov-Smirnov (P= 0.375). Não há valores extremos que possam ser retirados para que se possa obter um teste positivo para normalidade;

- Quanto ao teste lognormalidade da amostra Hangout (P= 0.871), o resultado se mostrou positivo, a hipótese de lognormalidade não pode ser descartad.

- A amostra Skype pode não ser considerada normal para o teste de Shapiro-Wilk (P= 0.02294), entretanto pode ser considerada para o teste de Komogorov-Smirnov (P= 0.1186);

- Quanto ao teste lognormalidade da amostra Hangout (P= 0.4377), o resultado se mostrou positivo, a hipótese de lognormalidade não pode ser descartada;

- Considera-se então que as duas amostras seguem melhor uma distribuição lognormal do que uma distribuição normal;

- Mesmo que após os testes de normalidade as amostras tenham sido consideradas como provenientes de uma distribuição lognormal, os testes de Levene foram realizados utilizando como centro a média e a mediana para obter um comparativo entre os testes;

- Não houve discordância na conclusão entre os testes com média ou mediana para qualquer um dos casos testados;

- Após realizar os testes de Levene entre as 3 amostras ($P_{\text{mediana}} = 0.01388$, $P_{\text{média}} = 0.0004823$), conclui-se que pelo menos uma das 3 amostras veio de uma população diferente, ou há diferença na usabilidade de pelo menos uma plataforma para as demais;

- Após realizar o teste de Levene entre Zoom e Hangout ($P_{\text{mediana}} = 0.01984$, $P_{\text{média}} = 0.001356$), concluiu-se que elas podem ter sido originadas de população diferentes, ou há diferença na usabilidade entre as plataformas;

- Após realizar o teste de Levene entre Skype e Hangout ($P_{\text{mediana}} = 0.03061$, $P_{\text{média}} = 0.002625$), concluiu-se que elas podem ter sido originadas de população diferentes, ou há diferença na usabilidade entre as plataformas;

- Após realizar o teste de Levene entre Zoom e Skype($P_{\text{mediana}} = 0.7678$, $P_{\text{média}} = 0.7632$), concluiu-se que elas podem ter sido originadas de população de características semelhantes, ou não há diferença na usabilidade entre as plataformas.

A resolução da atividade seguirá as seguintes etapas:

1. Leitura dos dados, inicialização das variáveis e das funções para os gráficos;
2. Análise de normalidade e lognormalidade para as amostras que utilizaram a plataforma Zoom;
3. Análise de normalidade e lognormalidade para as amostras que utilizaram a plataforma Hangout;
4. Análise de normalidade e lognormalidade para as amostras que utilizaram a plataforma Skype;
5. Testes de Levene

2 Leitura dos dados

```r
library(readr)
#Leitura do arquivo CSV
meet_file <- read_csv("Dados/meet3-file.csv",
                      col_types = cols(Meet = col_factor(levels = c("Zoom",
                                                                    "Hangout",
                                                                    "Skype")),
                                       Subject = col_skip()))

meet_file$logTempo <- log(meet_file$Tempo)


#seleção dos dados que representam a plataforma Zoom
```

```r
Zoom = meet_file[meet_file$Meet == "Zoom","Tempo"]

Zoom$Tempo <- sort(Zoom$Tempo, FALSE)

#seleção dos dados que representam a plataforma Hangout
Hangout = meet_file[meet_file$Meet == "Hangout","Tempo"]

Hangout$Tempo <- sort(Hangout$Tempo, FALSE)

#seleção dos dados que representam a plataforma Hangout
Skype = meet_file[meet_file$Meet == "Skype","Tempo"]

Skype$Tempo <- sort(Skype$Tempo, FALSE)

#Função para configuração dos gráficos
library(ggplot2)
library(cowplot)
library(qqplotr)
library(car)

gera_histograma <- function(dados, bins=9){

  n            <- length(dados$Tempo)
  nome         <- deparse(substitute(dados))
  mediaAmostra <- mean(dados$Tempo)

  sd<-sqrt(var(dados$Tempo)*(n-1)/n)

  histograma <- ggplot(dados,aes(Tempo))
  histograma <- histograma + geom_histogram(bins = bins,
                                            aes(y=..density.., fill=..count..))
  histograma <- histograma + labs( x="",y="Frequência",
                                  title=paste("Tempo utilizado por usuário na plataforma",
                                              nome))
  histograma <- histograma + scale_fill_gradient("Amostra por caixa",
                                                low="#DCDCDC",
                                                high="#7C7C7C",)
  histograma <- histograma + stat_function(fun=dnorm,
                                          color="red",
                                          args=list(mean=mediaAmostra,
                                                    sd=sd))

  ylimHist= layer_scales(histograma)$x$range$range

  diagCaixa <- ggplot(dados, aes(y=Tempo))
  diagCaixa <- diagCaixa + geom_boxplot()
  diagCaixa <- diagCaixa + theme(axis.title.y=element_blank(),
                                axis.text.y=element_blank(),
                                axis.ticks.y=element_blank())
  diagCaixa <- diagCaixa + labs(y=paste("Tempo de uso do",nome))
  diagCaixa <- diagCaixa + coord_flip(ylim = ylimHist)

  plot_grid(histograma, diagCaixa,
```

```r
                     ncol = 1, rel_heights = c(2, 1),
                     align = 'v', axis = "rlbt")
}

gera_qqplot <- function(dados){



  nome            <- deparse(substitute(dados))

  diagramaQuartil <- ggplot(dados, mapping= aes(sample = Tempo))
  diagramaQuartil <- diagramaQuartil + stat_qq_band(bandType = "pointwise")
  diagramaQuartil <- diagramaQuartil + stat_qq_line()
  diagramaQuartil <- diagramaQuartil + stat_qq_point()
  diagramaQuartil <- diagramaQuartil + labs( x="Quantis teóricos de uma distribuição normal",
                                            y="Quantis amostrais",
                                            title=paste("Diagrama QQ para a plataforma", nome))
  diagramaQuartil}

gera_ksplot <- function(dados, distribuicao){


  media    <- mean(dados$Tempo)
  sd       <- sd(dados$Tempo)
  nome     <- deparse(substitute(dados))
  nomeDist <- deparse(substitute(distribuicao))
  group    <- c(rep(nome, length(dados$Tempo)),
             rep("Dist Normal", length(distribuicao)))
  dat      <- data.frame(KSD = c(dados$Tempo,distribuicao), group = group)

  cdf1 <- ecdf(dados$Tempo)
  cdf2 <- ecdf(distribuicao)

  minMax <- seq(min(dados$Tempo, distribuicao),
                max(dados$Tempo, distribuicao),
                length.out=length(dados$Tempo))
  x0 <- minMax[which(abs(cdf1(minMax) - cdf2(minMax)) ==
                       max(abs(cdf1(minMax) - cdf2(minMax))) )]
  y0 <- cdf1(x0)
  y1 <- cdf2(x0)
  ggplot(dat, aes(x = KSD, group = group, color = group))+
    stat_ecdf(size=1) +
      theme(legend.position ="top") +
      xlab("Amostra") +
      ylab("ECDF") +
      #geom_line(size=1) +
      geom_segment(aes(x = x0[1], y = y0[1], xend = x0[1], yend = y1[1]),
          linetype = "dashed", color = "red") +
      geom_point(aes(x = x0[1] , y= y0[1]), color="red", size=2) +
      geom_point(aes(x = x0[1] , y= y1[1]), color="red", size=2) +
      ggtitle(paste("K-S Test: Plataforma",nome,"/",nomeDist)) +
      theme(legend.title=element_blank())
}
```

3 Análise de normalidade para as amostras que utilizaram a plataforma Zoom

A análise de normalidade é importante para permitir ao pesquisador decidir que tipo de testes estatísticos são pertinentes nos dados gerados pelo objeto de estudo. Com essa análise pode-se concluir ou não se a mostra foi retirada de uma população que segue uma distribuição normal.

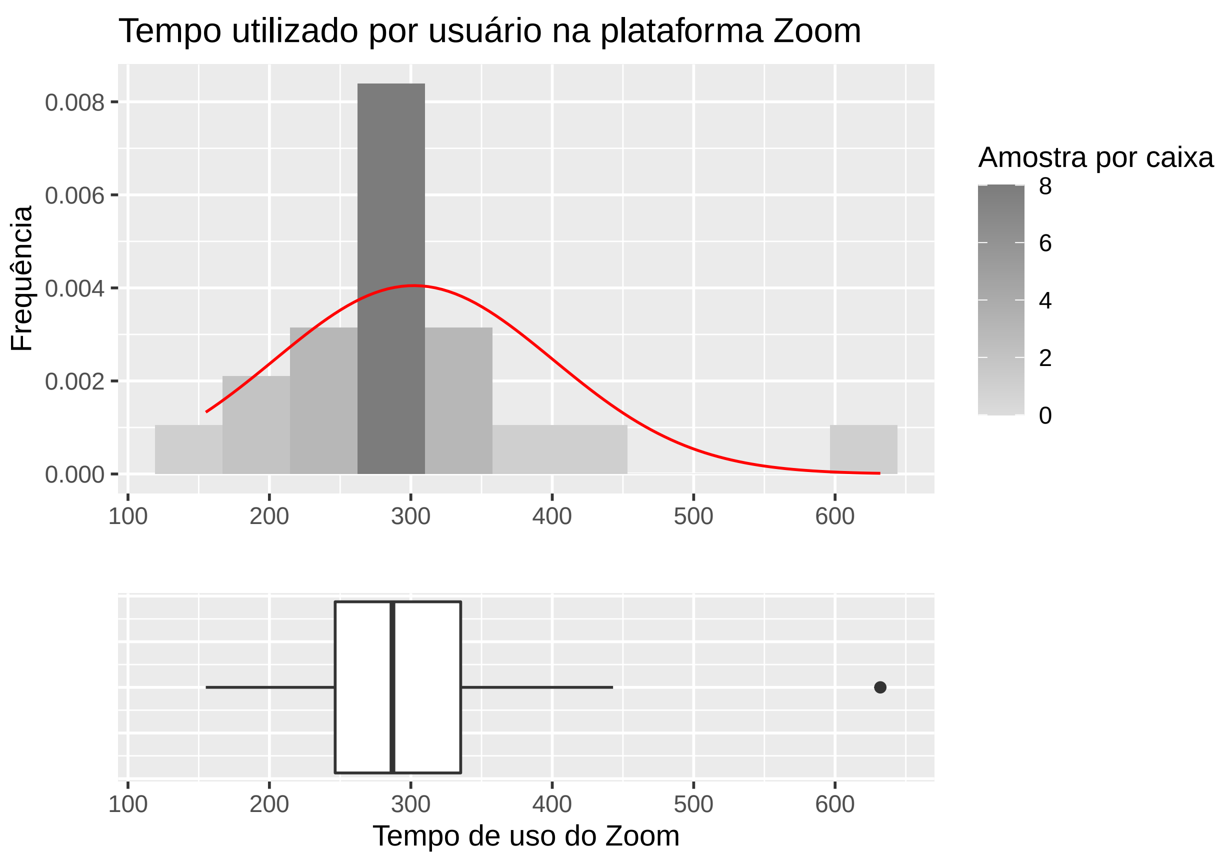
A seguir são realizados os testes de normalidade para as amostras do arquivo meet-file.csv que utilizaram a plataforma Zoom.

A análise da normalidade pode ser feita por métodos visuais, cálculo de parâmetros e/ou testes estatísticos.

Serão aplicados os 3 métodos isoladamente para a conclusão sobre a normalidade da amostra.

3.1 Análise dos gráficos

```
gera_histograma(Zoom,bins= 11)
```



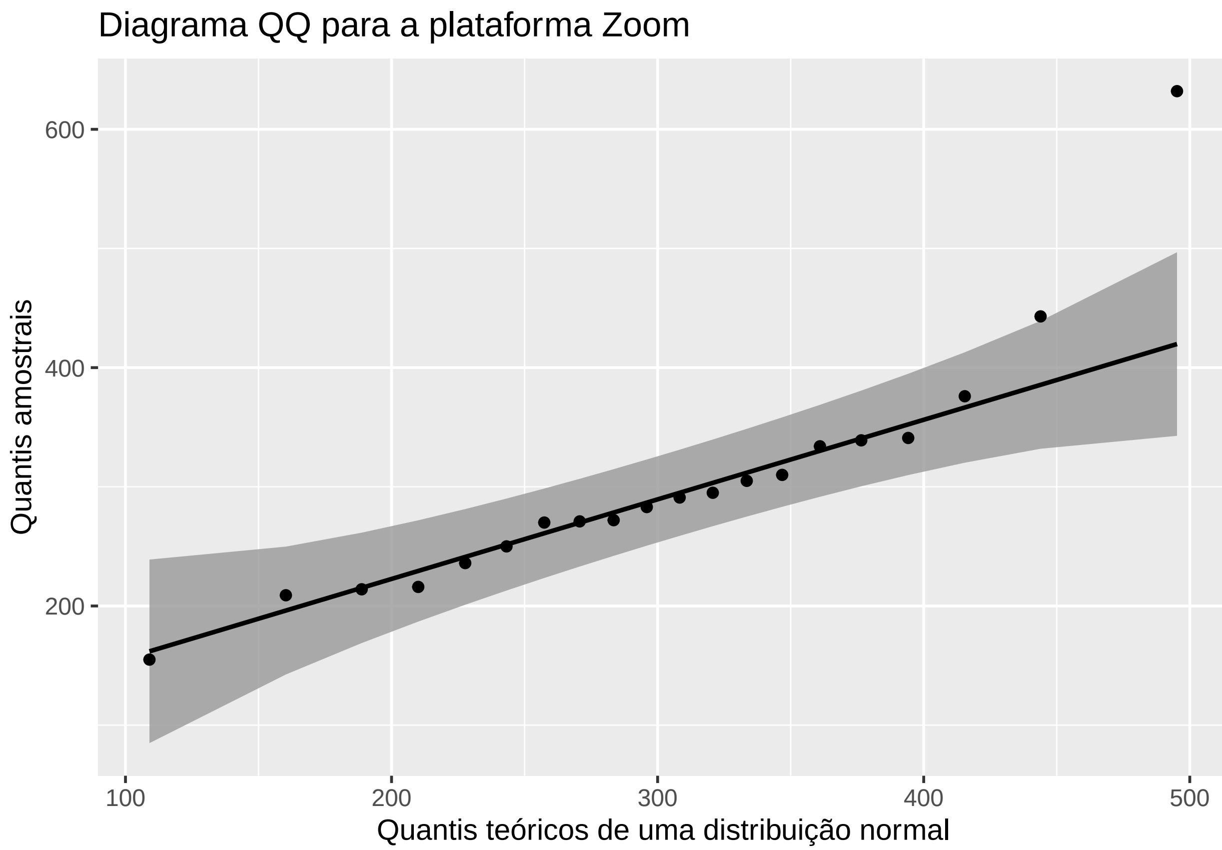
Foi plotado um diagrama de caixas, um histograma com 11 caixas (bins) e uma curva gaussiana com a média e desvio padrão iguais aos da amostra de tempo utilizado na ferramenta Zoom, para utilizar como referência visual.

O gráfico do histograma mostra que a amostra seguiria uma uma distribuição aparentemente normal se o valor extremo à direita fosse excluído, pois: observa-se que a média da gaussiana e a barra com maior frequência e número de observações aparentementam estar muito próximas, e o número de observações ao redor da média e sua frequência são similares.

O valor extremo à direita faz com que haja uma **assimetria positiva** na amostra.

Uma das ferramentas gráficas utilizadas para avaliar a normalidade é o gráfico de quantil-quantil, onde são representados os quantis de cada observação da amostra e comparado com uma linha que representa os quantis de uma distribuição normal.

```
gera_qqplot(Zoom)
```



Observa-se que os pontos são plotados ao longo da reta que representam os quantis de uma distribuição normal, à exceção do valor extremo visualizado no gráfico anterior.

Portanto, baseado na visualização dos gráficos pode-se inferir que a amostra foi retirada de uma população que segue uma distribuição normal.

3.2 Cálculo da curtose e assimetria

O cálculo da curtose e assimetria de uma amostra se dá utilizando o quarto e terceiro momentos centrais, respectivamente, ajustados para dados amostrais.

Os valores esperados de curtose e assimetria para uma curva normal são 0.0 e 0.0 respectivamente. No código abaixo são geradas 1.000.000 de amostras aleatórias de uma distribuição normal e calculados seus parâmetros de assimetria e curtose:

```
library(e1071)

distNormal<- rnorm(1000000)
```

```r
curtoseNormal <- kurtosis(distNormal)
assimetriaNormal <- skewness(distNormal)
cat(" Curtose de uma distribuição normal: ",
    curtoseNormal,"\n","Assimetria uma distribuição normal: ",
    assimetriaNormal)
```

```
##  Curtose de uma distribuição normal:  -0.002673145
##  Assimetria uma distribuição normal:  -0.0006117007
```

Abaixo são calculados os mesmos parâmetros para a amostra Zoom:

```r
curtoseZoom <- kurtosis(Zoom$Tempo)
assimetriaZoom <- skewness(Zoom$Tempo)
cat(" Curtose para as amostras que utilizaram a plataforma Zoom: ",
    curtoseZoom,"\n","Assimetria para as amostras que utilizaram a plataforma Zoom: ",
    assimetriaZoom)
```

```
##  Curtose para as amostras que utilizaram a plataforma Zoom:  3.208934
##  Assimetria para as amostras que utilizaram a plataforma Zoom:  1.617429
```

Percebe-se que os valores estão desviados do valor esperado para uma curva normal. Quanto à curtose pode-se classificar a amostra como **leptocúrtica**, ou seja, mais alongada que uma distribuição normal

A partir da assimetria calculada pode-se afirmar que a distribuição possui uma **assimetria positiva**, espera-se que a distribuição possua uma cauda mais longa à direita.

Portanto, a partir dos parâmetros calculados, conclui-se que a amostra não foi retirada de uma população que siga uma distribuição normal, pois seus parâmetros muitos se distanciam dos parâmetros para uma curva normal (0.0 e 0.0 para ambos).

3.3 Testes estatísticos

Chegou-se a conclusões distintas quanto à normalidade utilizando o método gráfico e o cálculo da assimetria e curtose.

É necessário portanto aplicar testes estatísticos de normalidade para a obtenção de resultados mais conclusivos.

3.3.1 Teste de Shapiro-Wilk:

O teste de Shapiro-Wilk apresenta a estatística W e o valor P para representar a significância estatística do teste. A hipótese nula é:

H0: A amostra foi retirada de uma população que segue uma distribuição normal.

A estatística W do teste, varia de entre 0 e 1, quanto mais alto for W mais a amostra se aproxima de uma distribuição normal.

O teste também apresenta o valor de significância estatística valor p para a amostra em questão.

Se o valor p para uma dada amostra for menor que um nível de significância designado pode-se rejeitar a hipótese nula e afirmar que a amostra não segue uma distribuição normal. Valores comuns para comparação de testes de hipóteses são: 0.1, 0.05, 0.01, a depender do que se está estudando e o nível de rigor requerido.

Abaixo a amostra Zoom é testada para normalidade seguindo o método de Shapiro-Wilk:

```
testeZoom <- shapiro.test(Zoom$Tempo)
testeZoom
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Zoom$Tempo
## W = 0.84372, p-value = 0.004191
```

A partir do teste aplicado nas amostras que utilizaram Zoom pode-se afirmar que:

A um nível de significância de 0.1, 0.05 ou 0.01 a hipótese nula pode ser rejeitada chegando-se a conclusão que a amostra não vem de uma população que segue uma distribuição normal.

O resultado do teste confirma o que foi visto através do desvio acentuado da assimetria e curtose da amostra e contraria a análise gráfica realizada.

O teste é aplicado novamente removendo o valor extremo:

```
testeZoom <- shapiro.test(Zoom$Tempo[1:19])
testeZoom
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Zoom$Tempo[1:19]
## W = 0.98052, p-value = 0.9482
```

O novo teste aplicado sem o valor extremo possui um valor p maior do que o maior valor normalmente utilizado de 0.1, portanto a distribuição segue uma distribuição normal se o valor extremo for excluído.

3.3.2 Teste de Kolmogorov-Smirnov

O teste de Kolmogorov pode ser utilizado para comparar duas amostras ou para comparar uma amostra com uma distribuição padrão.

O teste de Kolmogorov apresenta a estatíca D: Máxima diferença absoluta entre duas funções de distribuições cumulativas e possui um valor P para representar a significância estatística do teste. O teste de Kolmogorov possui as seguintes hipóteses nulas:

Comparação entre duas amostras: H0: As duas amostras foram retiradas de uma população com a mesma distribuição.

Comparação entre uma amostra e uma distribuição de referência: H0: A amostra foi retirada de uma população que segue a distribuição de referência.

Aplica-se então o teste para comparar a amostra Zoom a uma distribuição normal de média e desvio padrão iguais aos da amostra:

```
testeKSZoom <- ks.test(Zoom$Tempo, "pnorm", mean=mean(Zoom$Tempo), sd=sd(Zoom$Tempo))
testeKSZoom
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
```

```
## data:  Zoom$Tempo
## D = 0.20017, p-value = 0.3517
## alternative hypothesis: two-sided
```

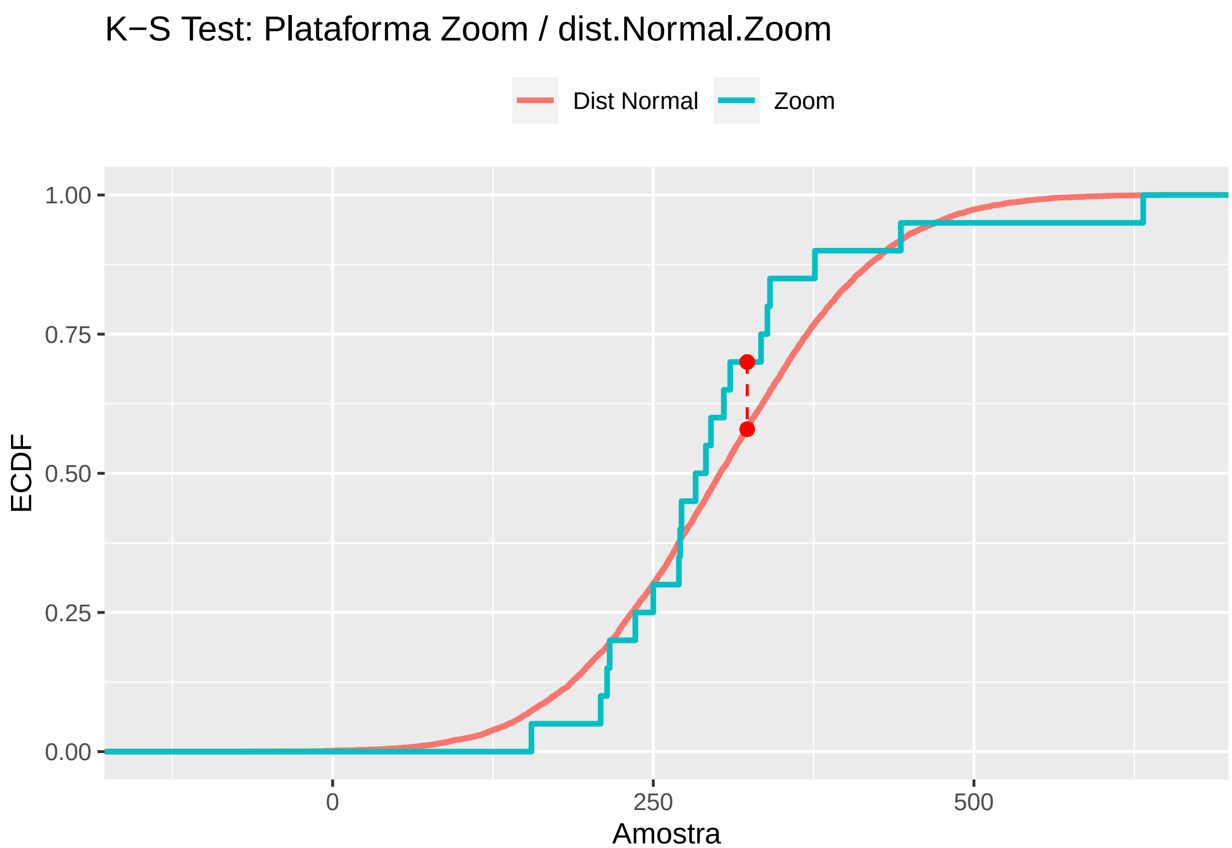
A partir das informações contidas no teste:

A um nível de significância de 0.1 a hipótese nula não pode ser rejeitada chegando-se a conclusão que a amostra segue uma distribuição normal.

A visualização da distribuição cumulativa comparada com a distribuição cumulativa da curva normal é mostrada em seguida, onde os dois pontos conectados entre as curvas demonstram a estatística D do teste de Komogorov.

```r
dist.Normal.Zoom<- rnorm(10000, mean(Zoom$Tempo), sd(Zoom$Tempo))

gera_ksplot(Zoom, dist.Normal.Zoom)
```



Teste de lognormalidade para a amostra Zoom:

```r
library(MASS)
#
fitlogZoom <- fitdistr(Zoom$Tempo, "lognormal")$estimate
meanlogZoom <- fitlogZoom[1]
sdlogZoom <- fitlogZoom[2]

testeKSlogZoom <- ks.test(Zoom$Tempo, "plnorm", meanlogZoom,sdlogZoom)
testeKSlogZoom
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  Zoom$Tempo
## D = 0.13421, p-value = 0.8181
## alternative hypothesis: two-sided
```

A partir das informações contidas no teste:

A um nível de significância de 0.1 a hipótese nula não pode ser rejeitada chegando-se a conclusão que a amostra segue uma distribuição lognormal.

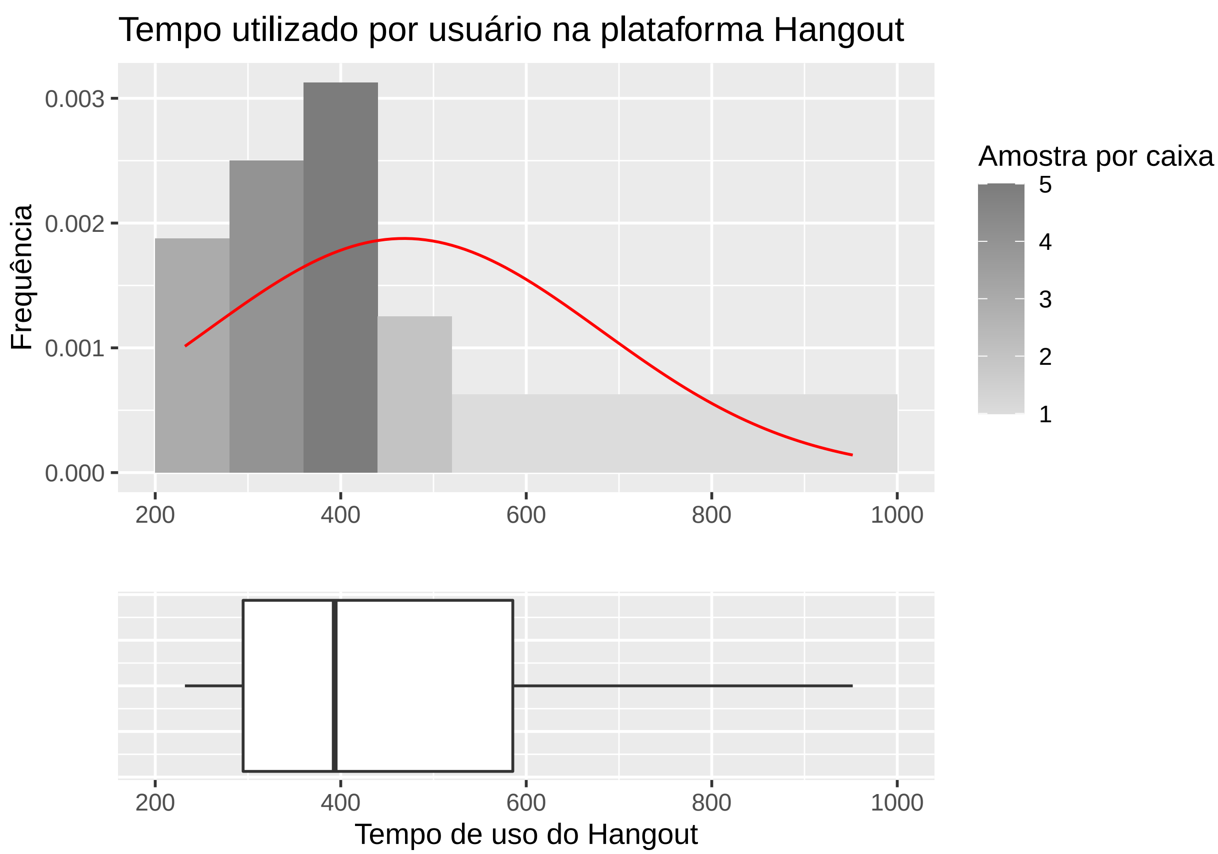
O resultado positivo tanto para normalidade quanto para normalidade pode ser devido ao baixo número de observações na amostra. Entretanto podemos observar que o valor-p para a lognormalidade é maior do que para normalidade.

4 Análise de normalidade para as amostras que utilizaram a plataforma Hangout

Serão aplicados os mesmos testes utilizados para a amostra Zoom.

4.1 Análise dos gráficos

```r
gera_histograma(Hangout,bins= 10)
```

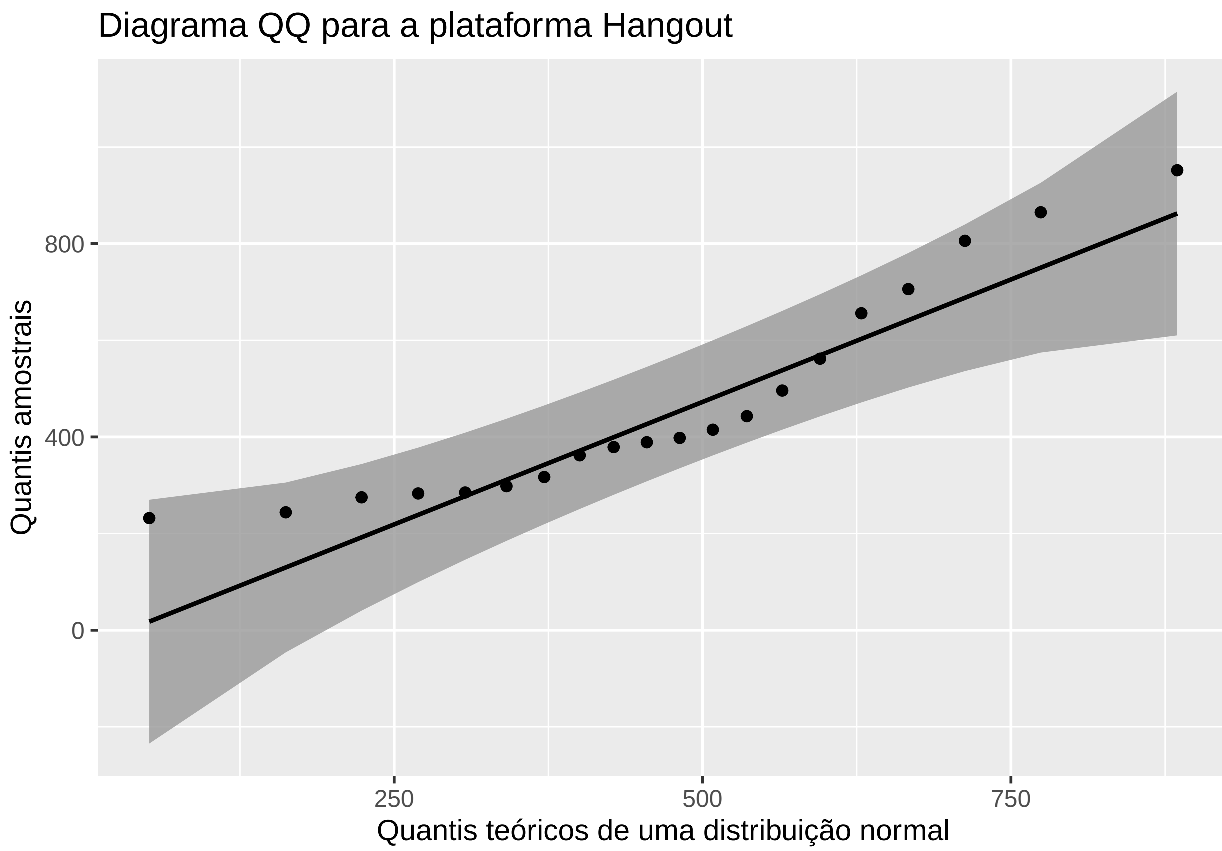


Foi plotado um diagrama de caixas, um histograma com 10 caixas (bins) e uma curva gaussiana com a média e desvio padrão iguais aos da amostra de tempo utilizado na ferramenta Hangout, para utilizar como referência visual.

Aparentemente a curva não segue uma distribuição normal devido aos valores maiores possuirem menor frequência na amostra.

A baixa frequência nos valores à direita faz com que haja **assimetria positiva** na amostra.

```
gera_qqplot(Hangout)
```



Observa-se que os pontos são plotados ao longo da reta que representam os quantis de uma distribuição normal, e não há valores extremos.

Portanto, baseado na visualização dos gráficos pode-se inferir que a amostra foi retirada de uma população que segue uma distribuição normal.

4.2 Cálculo da curtose e assimetria

Abaixo são calculados os mesmos parâmetros para a amostra Zoom:

```
curtoseZoom <- kurtosis(Hangout$Tempo)
assimetriaZoom <- skewness(Hangout$Tempo)
cat(" Curtose para as amostras que utilizaram a plataforma Hangout: ", curtoseZoom,"\n","Assimetria para
```

```
##  Curtose para as amostras que utilizaram a plataforma Hangout:  -0.6120307
##  Assimetria para as amostras que utilizaram a plataforma Hangout:  0.8665088
```

Percebe-se que os valores estão desviados do valor esperado para uma curva normal. Quanto à curtose pode-se classificar a amostra como **platicúrtica**, ou seja, mais achatada que uma distribuição normal, embora em baixa intensidade.

A partir da assimetria calculada pode-se afirmar que a distribuição possui uma **assimetria positiva**.

Portanto, a partir dos parâmetros calculados, conclui-se que a amostra pode ter sido retirada de uma população que segue uma distribuição normal, pois seus parâmetros pouco se distanciam dos parâmetros de uma curva normal (0.0 e 0.0 para assimetria e curtose).

4.3 Testes estatísticos

4.3.1 Teste de Shapiro-Wilk:

Abaixo a amostra Hangout é testada para normalidade seguindo o método de Shapiro-Wilk:

```
testeHangout <- shapiro.test(Hangout$Tempo)
testeHangout
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Hangout$Tempo
## W = 0.87213, p-value = 0.01281
```

A partir do teste aplicado nas amostras que utilizaram Hangout pode-se afirmar que:

A um nível de significância de 0.1 ou 0.05 a hipótese nula pode ser rejeitada chegando-se a conclusão que a amostra não vem de uma população que segue uma distribuição normal.

A um nível de significância de 0.01 a hipótese nula não pode ser rejeitada chegando-se a conclusão que a amostra vem de uma população que segue uma distribuição normal.

A depender do nível limiar de significância aplicado pelo pesquisador ambas as conclusões podem ser adotadas.

4.3.2 Teste de Kolmogorov-Smirnov:

O teste de Kolmogorov pode ser utilizado para comparar duas amostras ou para comparar uma amostra com uma distribuição padrão.

O teste de Kolmogorov apresenta a estatíca D: Máxima diferença absoluta entre duas funções de distribuições cumulativas e possui um valor P para representar a significância estatística do teste. O teste de Kolmogorov possui as seguintes hipóteses nulas:

Comparação entre duas amostras: H0: As duas amostras foram retiradas de uma população com a mesma distribuição.

Comparação entre uma amostra e uma distribuição de referência: H0: A amostra foi retirada de uma população que segue a distribuição de referência.

Aplica-se então o teste para comparar a amostra Hangout a uma distribuição normal de média e desvio padrão iguais aos da amostra:

```r
testeKSHangout <- ks.test(Hangout$Tempo, "pnorm", mean=mean(Hangout$Tempo), sd=sd(Hangout$Tempo))
testeKSHangout
```

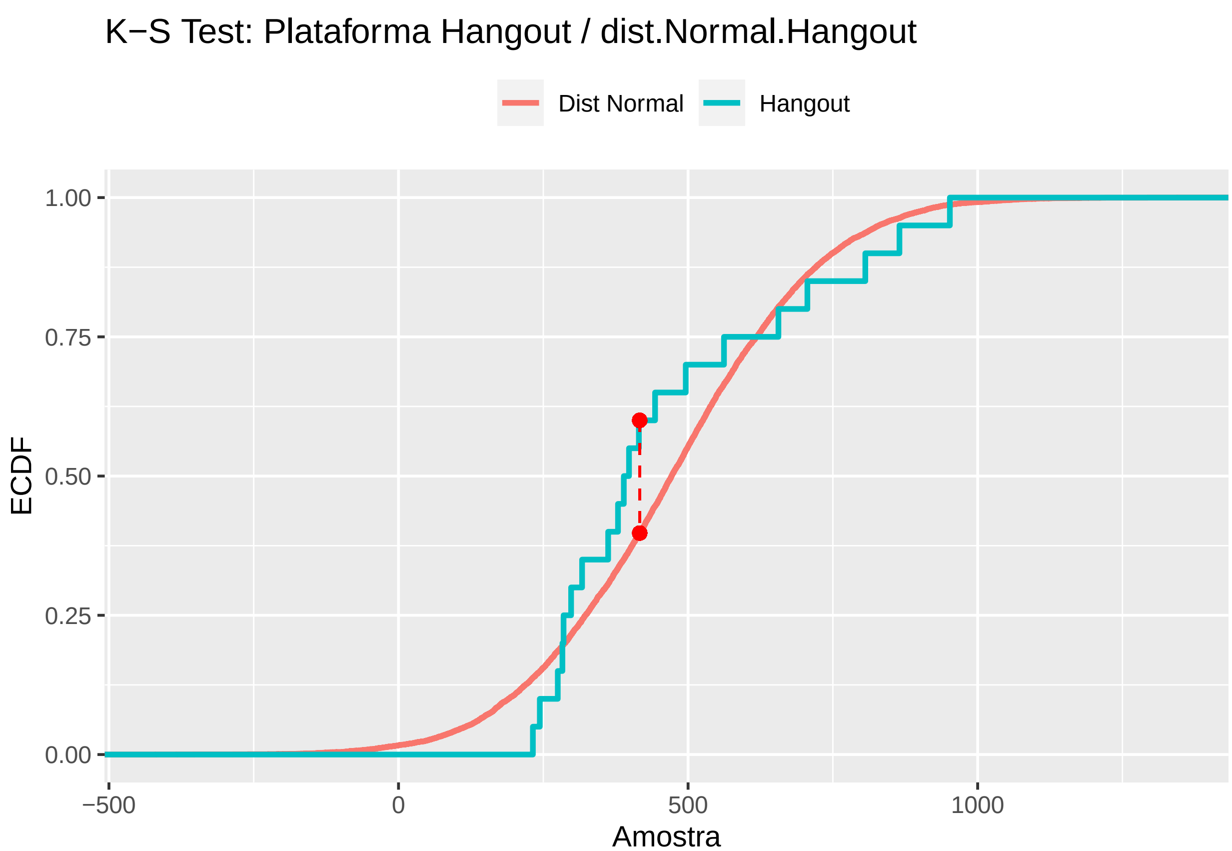
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  Hangout$Tempo
## D = 0.19626, p-value = 0.375
## alternative hypothesis: two-sided
```

A partir das informações contidas no teste:

A um nível de significância de 0.1 a hipótese nula não pode ser rejeitada chegando-se a conclusão de que a possibilidade que a amostra siga uma distribuição normal não pode ser descartada.

A visualização da distribuição cumulativa comparada com a distribuição cumulativa da curva normal é mostrada em seguida:

```r
dist.Normal.Hangout<- rnorm(10000, mean(Hangout$Tempo), sd(Hangout$Tempo))
gera_ksplot(Hangout, dist.Normal.Hangout)
```



Teste de lognormalidade para a amostra Zoom:

```r
fitlogHangout <- fitdistr(Hangout$Tempo, "lognormal")$estimate
meanlogHangout <- fitlogHangout[1]
sdlogHangout <- fitlogHangout[2]
testeKSlogHangout <- ks.test(Hangout$Tempo, "plnorm", meanlogHangout, sdlogHangout)
testeKSlogHangout
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  Hangout$Tempo
## D = 0.12583, p-value = 0.871
## alternative hypothesis: two-sided
```

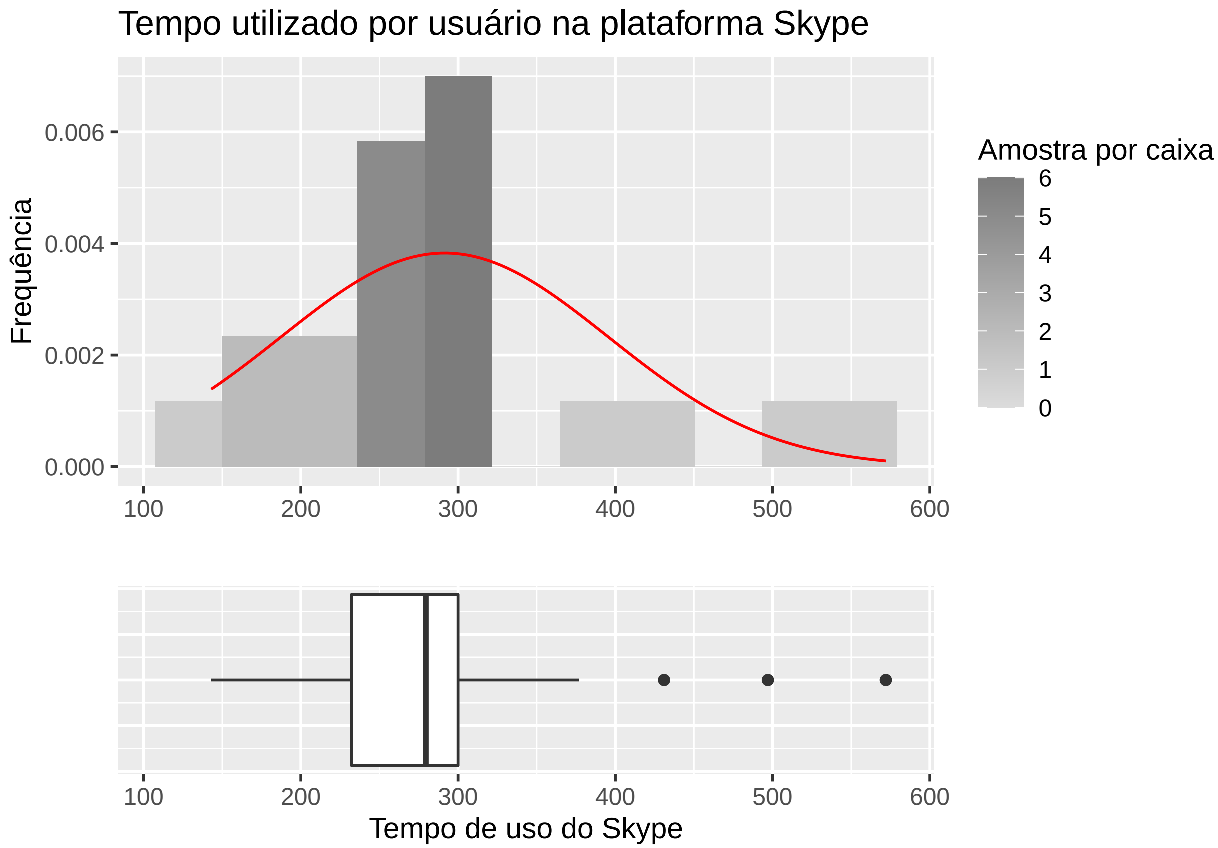
A partir das informações contidas no teste:

A qualquer nível de significância comumente utilizado a hipótese nula não pode ser rejeitada chegando-se a conclusão de que a possibilidade que a amostra siga uma distribuição lognormal não pode ser descartada.

5 Análise de normalidade para as amostras que utilizaram a plataforma Skype

5.1 Análise dos gráficos

```r
gera_histograma(Skype,bins= 11)
```

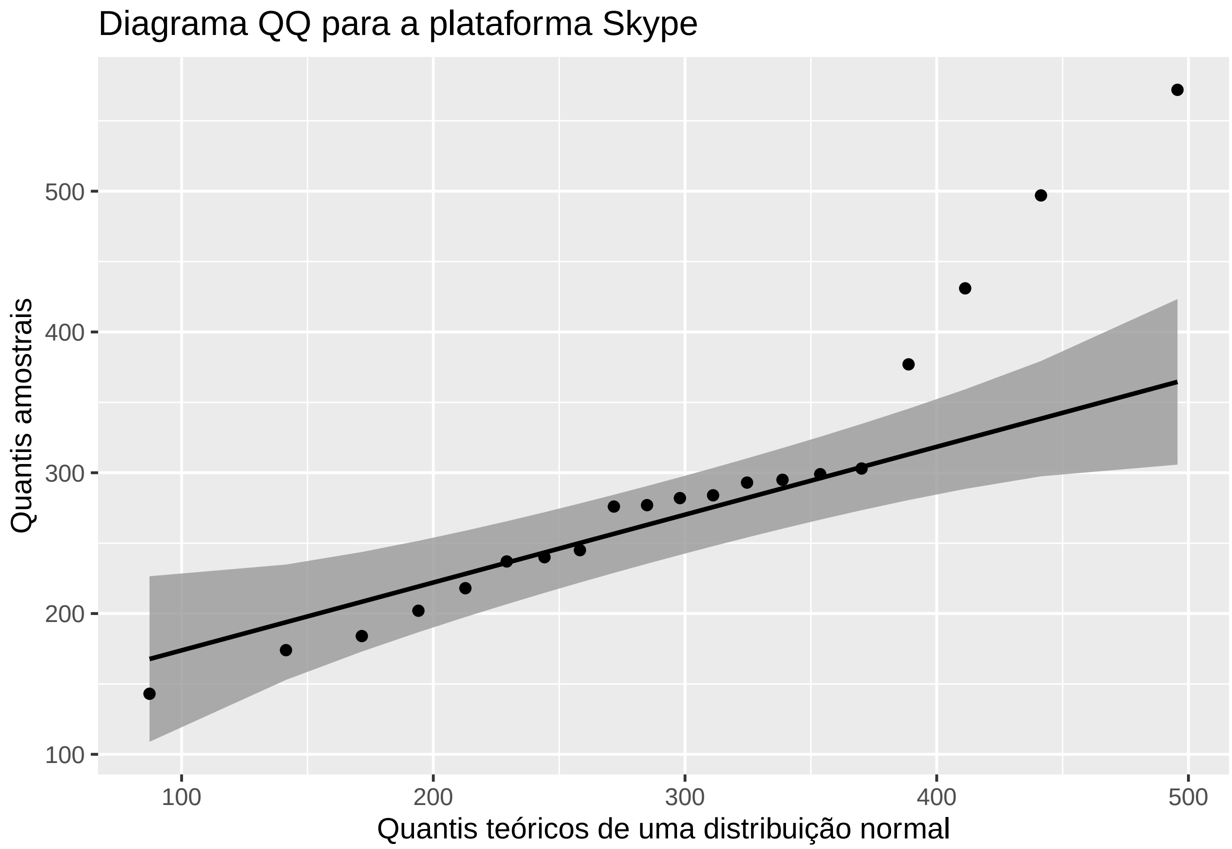


Foi plotado um diagrama de caixas, um histograma com 11 caixas (bins) e uma curva gaussiana com a média e desvio padrão iguais aos da amostra de tempo utilizado na ferramenta Skype, para utilizar como referência visual.

Aparentemente a curva não segue uma distribuição normal devido aos valores maiores possuirem menor frequência na amostra.

A baixa frequência nos valores à direita faz com que haja **assimetria positiva** na amostra.

```
gera_qqplot(Skype)
```



Observa-se que os pontos são plotados ao longo da reta que representam os quantis de uma distribuição normal, a não ser pelos valores extremos que estão em maior número que as amostras anteriores.

Portanto, baseado na visualização dos gráficos não se pode inferir que a amostra foi retirada de uma população que segue uma distribuição normal.

5.2 Cálculo da curtose e assimetria

Abaixo são calculados os mesmos parâmetros para a amostra Zoom:

```
curtoseZoom <- kurtosis(Skype$Tempo)
assimetriaZoom <- skewness(Skype$Tempo)
cat(" Curtose para as amostras que utilizaram a plataforma Skype: ", curtoseZoom,"\n","Assimetria para a
```

```
##  Curtose para as amostras que utilizaram a plataforma Skype:  0.5755356
##  Assimetria para as amostras que utilizaram a plataforma Skype:  1.083317
```

Percebe-se que os valores estão desviados do valor esperado para uma curva normal. Quanto à curtose pode-se classificar a amostra como **leptocúrtica**.

A partir da assimetria calculada pode-se afirmar que a distribuição possui uma **assimetria positiva**.

Portanto, a partir dos parâmetros calculados, conclui-se que a amostra pode NÃO ter sido retirada de uma população que segue uma distribuição normal.

5.3 Testes estatísticos

5.3.1 Teste de Shapiro-Wilk:

Abaixo a amostra Skype é testada para normalidade seguindo o método de Shapiro-Wilk:

```
testeSkype <- shapiro.test(Skype$Tempo)
testeSkype
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Skype$Tempo
## W = 0.88623, p-value = 0.02294
```

A partir do teste aplicado nas amostras que utilizaram Skype pode-se afirmar que:

A um nível de significância de 0.1 ou 0.05 a hipótese nula pode ser rejeitada chegando-se a conclusão que a amostra não vem de uma população que segue uma distribuição normal.

A um nível de significância de 0.01 a hipótese nula não pode ser rejeitada chegando-se a conclusão que a amostra vem de uma população que segue uma distribuição normal.

A depender do nível limiar de significância aplicado pelo pesquisador ambas as conclusões podem ser adotadas.

5.3.2 Teste de Kolmogorov-Smirnov:

Aplica-se então o teste para comparar a amostra Skype a uma distribuição normal de média e desvio padrão iguais aos da amostra:

```
testeKSSkype <- ks.test(Skype$Tempo, "pnorm", mean=mean(Skype$Tempo), sd=sd(Skype$Tempo))
testeKSSkype
```

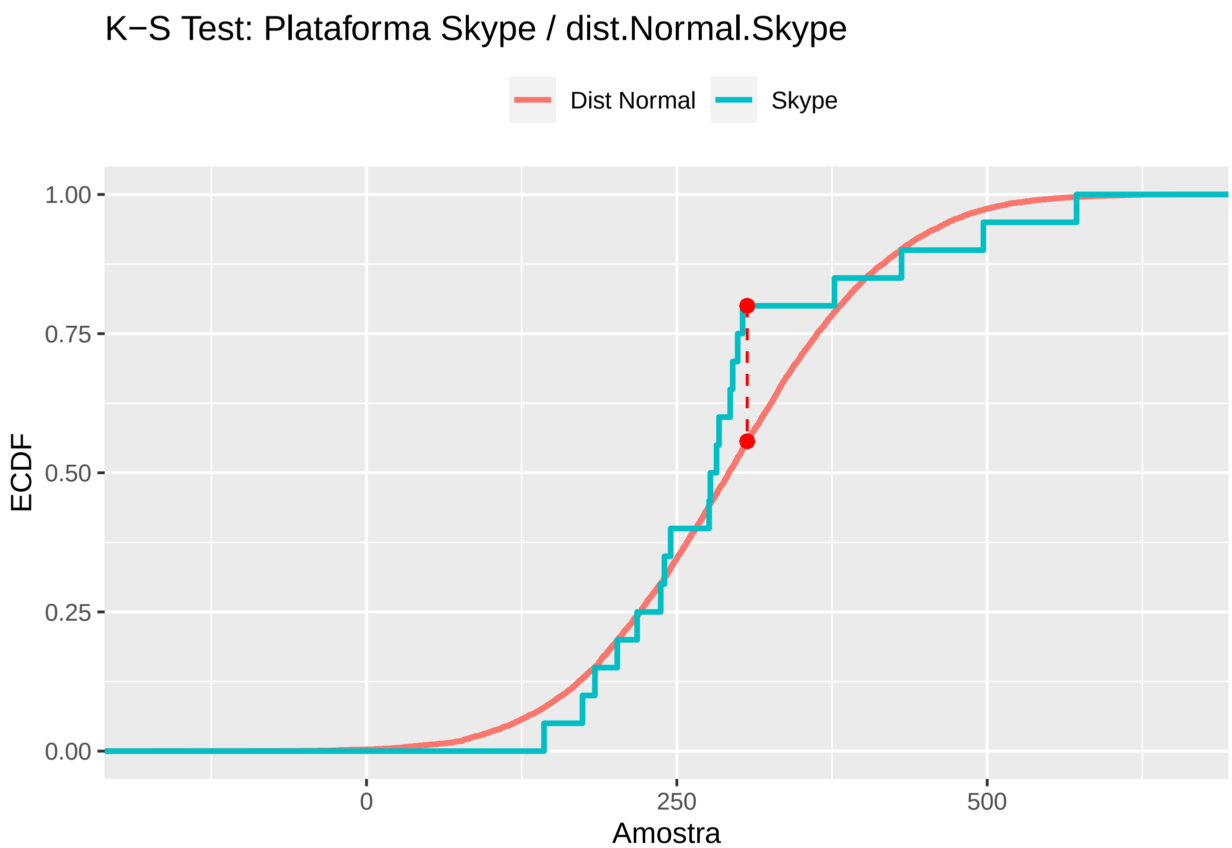
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  Skype$Tempo
## D = 0.25698, p-value = 0.1186
## alternative hypothesis: two-sided
```

A partir das informações contidas no teste:

A um nível de significância de 0.1 a hipótese nula não pode ser rejeitada chegando-se a conclusão que a hipótese de que a amostra segue uma distribuição normal não pode ser descartada.

A visualização da distribuição cumulativa comparada com a distribuição cumulativa da curva normal é mostrada em seguida:

```
dist.Normal.Skype<- rnorm(10000, mean(Skype$Tempo), sd(Skype$Tempo))
gera_ksplot(Skype, dist.Normal.Skype)
```



Teste de lognormalidade para a amostra Skype:

```
fitlogSkype <- fitdistr(Skype$Tempo, "lognormal")$estimate
meanlogSkype <- fitlogSkype[1]
sdlogSkype <- fitlogSkype[2]
testeKSlogSkype <- ks.test(Skype$Tempo, "plnorm", meanlogSkype, sdlogSkype)
testeKSlogSkype
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  Skype$Tempo
## D = 0.1864, p-value = 0.4377
## alternative hypothesis: two-sided
```

A partir das informações contidas no teste:

A qualquer nível de significância comumente utilizado a hipótese nula não pode ser rejeitada, chegando-se a conclusão que a hipótese de que a amostra segue uma distribuição lognormal não pode ser descartada.

6 Teste de Levene

O teste de Levene é utilizado para avaliar se a variância entre os grupos é homogênea. O teste calcula a estatística F como é mostrado na fórmula a seguir:

$$F = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_{i.})^2}$$

Entretanto, ao invés de calcular o quadrado da distância de cada ponto para a média de sua amostra, o teste de Levene transforma cada ponto da amostra para o módulo da diferença entre o ponto e a média ou mediana da amostra, como na fórmula a seguir:

$$Z_{ij} = \begin{cases} |Y_{ij} - \bar{Y}_{i.}|, & \bar{Y}_{i.} \text{ é a média do } i\text{-ésimo grupo}, \\ |Y_{ij} - \tilde{Y}_{i.}|, & \tilde{Y}_{i.} \text{ é a mediana do } i\text{-ésimo grupo}. \end{cases}$$

A escolha entre utilizar a média ou a mediana está relacionada com a normalidade da amostra. Para amostras que seguem uma distribuição normal é utilizada a média. Para amostras assimétricas o parâmetro indicado é a mediana.

O teste com centro na mediana também é chamado de teste de Brown-Forsythe

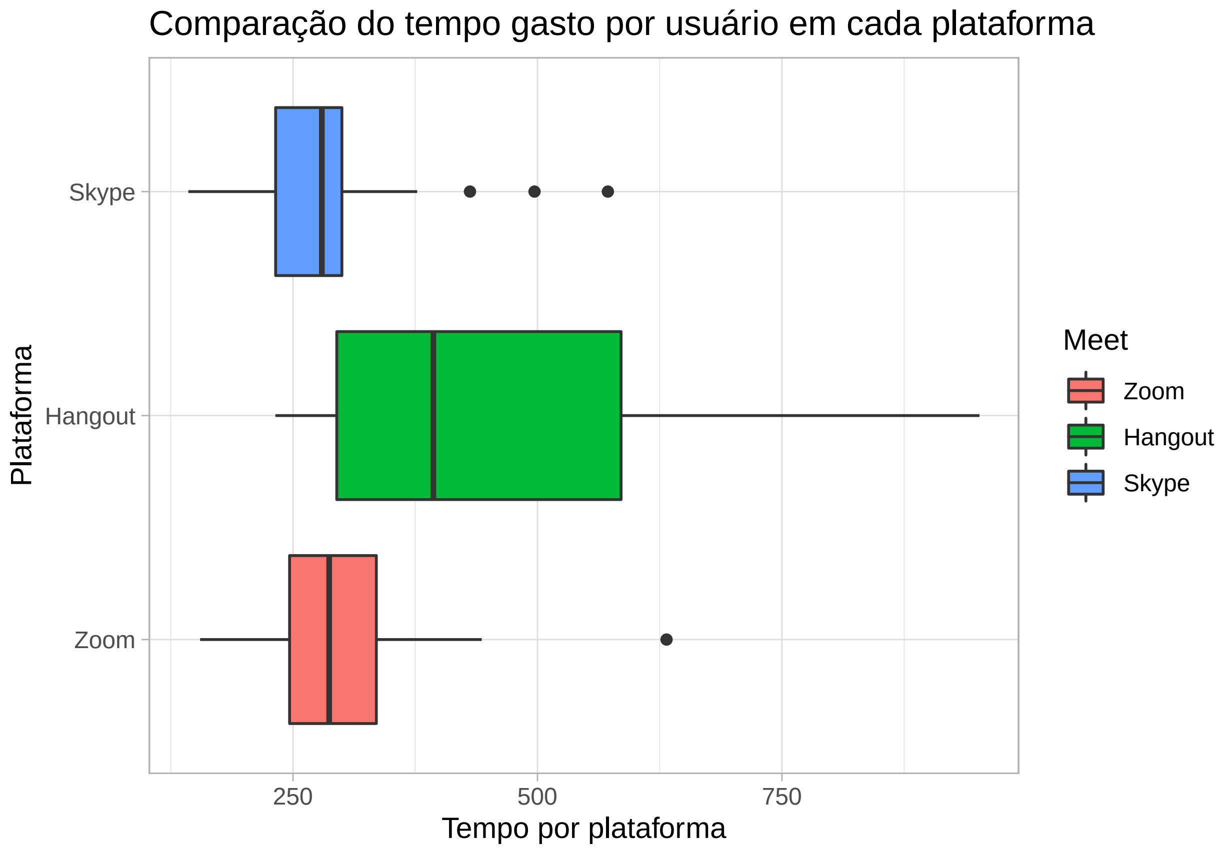
O teste possui como hipótese nula a seguinte afirmação: **H0: As amostras foram retiradas de populções com dispersões homogêneas**

6.1 Plotagem de diagrama de caixas

Antes de realizar os testes gera-se um gráfico para uma análise visual comparativa.

```
Boxplot3Variaveis <- ggplot(meet_file, aes(Meet,Tempo, fill=Meet))+
                      geom_boxplot() +
                      theme_light()+
                      labs(x= "Plataforma",
                           y="Tempo por plataforma",
                           title="Comparação do tempo gasto por usuário em cada plataforma")+
                      coord_flip()

Boxplot3Variaveis
```



6.2 Teste entre as 3 amostras

O teste de homogeneidade será realizado utilizando a média e a mediana como ajuste da amostra.

```
leveneTest(Tempo ~ Meet, data=meet_file, center= median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2  4.6149 0.01388 *
##       57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para uma siginificância de 0,05 a hipótese nula pode ser rejeitada demonstrando que pelo menos uma amostra pode vir de uma população com variância diferente das demais

```
leveneTest(Tempo ~ Meet, data=meet_file, center= mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value    Pr(>F)
## group  2   8.758 0.0004823 ***
##       57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para uma siginificância de 0,05 a hipótese nula pode ser rejeitada demonstrando que pelo menos uma amostra pode vir de uma população com variância diferente das demais

6.3 Testes 2 a 2

A partir da visualização do gráfico foi notado que a amostra Hangout parece ser a que mais se difere das demais. Para testar essa possibilidade é realizado o teste de cada combinação de amostras.

6.3.1 Testes entre Zoom e Hangout

```r
leveneTest(Tempo ~ Meet,
           data=meet_file[meet_file$Meet == "Zoom" | meet_file$Meet == "Hangout", ],
           center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  5.9144 0.01984 *
##       38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para uma siginificância de 0,05 a hipótese nula pode ser rejeitada demonstrando que as amostras Zoom e Hangout podem ser originadas de populações distintas

```r
leveneTest(Tempo ~ Meet,
           data=meet_file[meet_file$Meet == "Zoom" | meet_file$Meet == "Hangout", ],
           center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value   Pr(>F)
## group  1  11.959 0.001356 **
##       38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para uma siginificância de 0,05 a hipótese nula pode ser rejeitada demonstrando que as amostras Zoom e Hangout podem ser originadas de populações distintas

6.3.2 Testes entre Skype e Hangout

```r
leveneTest(Tempo ~ Meet,
           data=meet_file[meet_file$Meet == "Skype" | meet_file$Meet == "Hangout", ],
           center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1   5.044 0.03061 *
##       38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para uma siginificância de 0,05 a hipótese nula pode ser rejeitada demonstrando que as amostras Skype e Hangout podem ser originadas de populações distintas

```
leveneTest(Tempo ~ Meet,
           data=meet_file[meet_file$Meet == "Skype" | meet_file$Meet == "Hangout", ],
           center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value   Pr(>F)
## group  1  10.369 0.002625 **
##       38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para uma siginificância de 0,05 a hipótese nula pode ser rejeitada demonstrando que as amostras Skype e Hangout podem ser originadas de populações distintas

6.3.3 Testes entre Zoom e Skype

```
leveneTest(Tempo ~ Meet,
           data=meet_file[meet_file$Meet == "Zoom" | meet_file$Meet == "Skype", ],
           center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0884 0.7678
##       38
```

Para uma siginificância de 0,05 a hipótes nula não pode ser rejeitada demonstrando que as amostras Zoom e Skype podem ser originadas de populações de variâncias semelhantes ou da mesma população

```
leveneTest(Tempo ~ Meet,
           data=meet_file[meet_file$Meet == "Zoom" | meet_file$Meet == "Skype", ],
           center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value Pr(>F)
## group  1  0.0921 0.7632
##       38
```

Para uma siginificância de 0,05 a hipótese nula não pode ser rejeitada demonstrando que as amostras Zoom e Skype podem ser originadas de populações de variâncias semelhantes ou da mesma população