Universidade Federal do Rio de Janeiro
Observatório do Valongo

Primeira prova

Astroestatística

2025.1

Iago Lopes
DRE: 122077032

Professor: Hélio Jaques

31 de Maio de 2025

Questão 1 - O diâmetro de um asteroide do cinturão principal foi medido por técnicas de ocultação de estrelas. As observações de vários astrônomos forneceram as seguintes medidas de diâmetro em km: 320, 315, 327, 313, 318, 319, 330, 317, 328, 332.

a) **Estime qual é o diâmetro desse asteroide com um intervalo de confiança de 97.5 %.**

b) **Estime o intervalo de confiança de 95 % para a variância das medidas do diâmetro, sabendo que esse intervalo segue a fórmula:**

$$\frac{(n-1)S_X^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S_X^2}{\chi^2_{1-\alpha/2}}$$

Essa formula não está correta. Olhando no slide 19 do cap 2, vemos que a ordem está invertida e deveria ser:

$$\frac{(n-1)S_X^2}{\chi^2_{\alpha/2}} > \sigma^2 > \frac{(n-1)S_X^2}{\chi^2_{1-\alpha/2}}$$

Para estimar o tamanho desse asteroide com base em todas essas medidas, usarei o test de t-student nos dados, pois esse teste é adequado para a média de poucos dados. O teste resultou no intervalo de (316.18, 327.62) para o diâmetro do asteroide

```
d = c(320, 315, 327, 313, 318, 319, 330, 317, 328, 332)
t.test(d, conf.level = 0.975)
```

Seguindo a fórmula fornecida, estimei o intervalo de 95 % para a variância amostral:

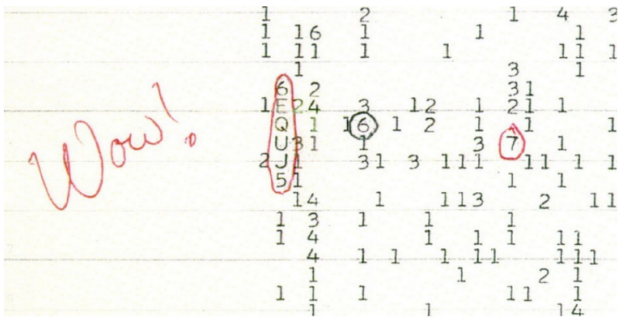
```
alpha = 0.05
n = length(d)
s_x = var(d)
x2_1 = qchisq(1-(alpha/2), df = n-1)
x2_2 = qchisq(alpha/2, df = n-1)

lower_interval = (n-1)*s_x/x2_1
upper_interval = (n-1)*s_x/x2_2

cat('Variance estimated:', s_x, '\nLower interval:', lower_interval,'\nUpper
    interval:', upper_interval)
```

```
Variance estimated: 45.43333
Lower interval: 21.49529
Upper interval: 151.4226
```

Questão 2 - A figura abaixo representa medições de rádio feitas pelo radiotelescópio Big Ear em 1977. As observações destacadas pelo círculo vermelho onde se lê 6EQUJ5 ficaram conhecidas como Sinal Uau (ou Wow, como quiser). As letras e números representam uma escala de intensidade, segundo a ordem: 123456789ABCDEFGHIJKLMNOPQRSTU. A cada 2 segundos, a intensidade do sinal medido seria representada por uma dessas letras. Suponha que essa escala seja linear, tal que 1 represente intensidade 1, 7 represente intensidade 7, A represente intensidade 10, e assim por diante, até a letra U, que representa a intensidade 30. Cada coluna composta por um caracter representa uma série de medidas. Com base nos dados dessa figura, estime a probabilidade do Sinal Uau ter sido um evento aleatório.



Este é um problema comum na astronomia e para resolvê-lo usarei a distribuição de Poisson, pois a intensidade é uma contagem de fótons, fazendo que essa distribuição seja a mais adequada para o problema. Para verificar se o evento Uau é um sinal aleatório, primeiro vou estimar a média das outras observações, ou seja, o λ da distribuição de Poisson.

Aqui vou assumir que os espaços vazios são intervalos de 2 segundos onde não houve nenhum fóton, logo sua intensidade será 0, e também vou desconsiderar a coluna do evento Uau nessa estimativa. Considerando 22 colunas e 17 linhas, temos:

```
values_no_wow_no_zero = c(rep(1,77), rep(2,8), rep(3,10), rep(4,6), rep(6,2),
    rep(7,1))

values_no_wow = c(values_no_wow_no_zero, rep(0,17*22-length(values_no_wow_no_
    zero)))

lambda = mean(values_no_wow)
```

Isso resulta em $\lambda \approx 0{,}444$, ou seja, em média temos 0,444 de intensidade a cada 2 segundos. Agora, podemos estimar a probabilidade de ocorrer 6EQUJ5 ou mais do que isso em 12 segundos:

$$I_{med} = \frac{6 + 14 + 26 + 30 + 19 + 5}{6} \approx 16{,}66$$

$$P(x > I_{med}) = \int_{I_{med}}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!} dx$$

A partir disso, podemos estimar a probabilidade de que isso seja um evento aleatório, usando:

$$P_{aleatorio} = 1 - P(x > I_{med}) = 1{,}862e - 21$$

```
wow = c(6, 14, 26, 30, 19, 5)
prob = ppois(mean(wow), lambda=lambda, lower.tail = F)
```

Logo, podemos dizer que a probabilidade de que isso seja um evento aleatório é praticamente nula e que é altamente provável que esse evento seja algum fenômeno físico a ser investigado.

Questão 3 - Um determinado modelo de formação estelar prevê que cerca de 38 % das estrelas de um aglomerado globular devem ser binárias. Seu colaborador estudou 568 estrelas de um aglomerado globular e encontrou que 200 delas devem ser binárias. Com base nesses dados, você pode refutar o modelo mencionado acima?

Para isso, precisamos utilizar um teste de hipótese baseado em proporções, que é o mais adequado ao problema:

```
prop.test(200, n = 568, p=0.38)
```

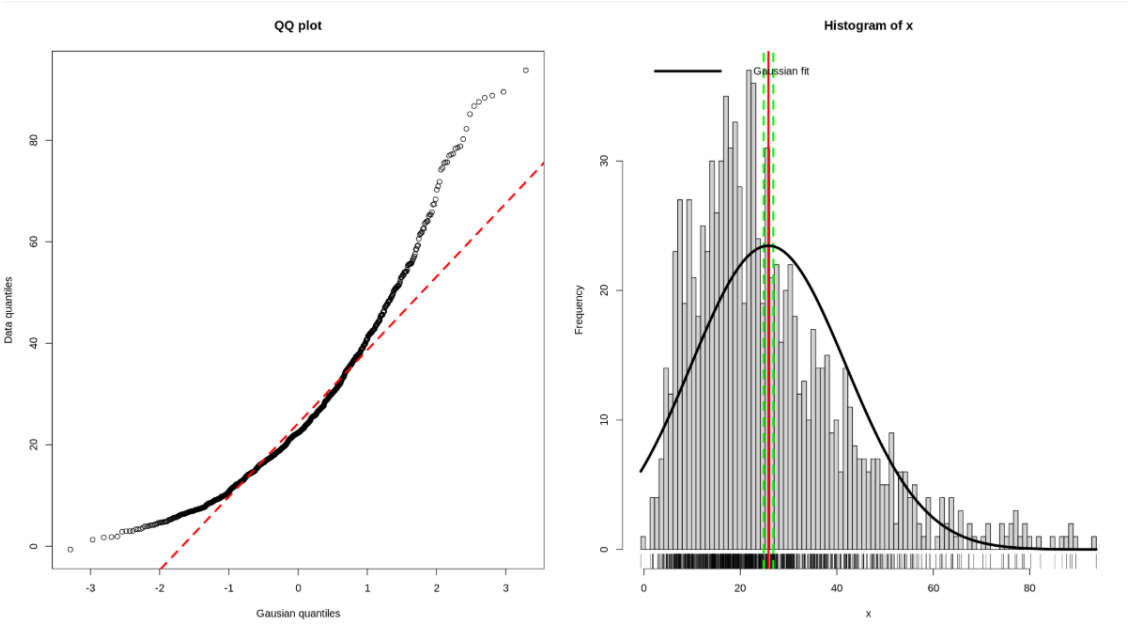
```
        1-sample proportions test with continuity correction

data:  200 out of 568, null probability 0.38
X-squared = 1.7584, df = 1, p-value = 0.1848
alternative hypothesis: true p is not equal to 0.38
95 percent confidence interval:
 0.3130942 0.3931624
sample estimates:
        p
0.3521127
```

De acordo com o teste de proporções, não podemos refutar o modelo mencionado, pois seu p-valor é de 0,38 e nessa análise estamos considerando como evidência forte valores p menores do que 0,05.

Questão 4 - Crie uma função no R que receba um vetor de dados qualquer e forneça como resultado ao usuário: na tela gráfica, uma figura de dois painéis, um ao lado do outro, nos quais o primeiro mostra um gráfico qq-plot para a amostra fornecida e uma linha qqline vermelha tracejada de uma população normal; e no segundo, um histograma do vetor fornecido, acompanhado por um "tapete"de valores individuais na abscissa (comando rug) e de três barras verticais, sendo a central em linha sólida vermelha indicando o valor da média amostral, e as duas linhas laterais tracejadas verdes indicando o erro da média amostral, isto é, $\mu \pm \epsilon\mu$. Lembre-se que o erro da média amostral costuma ser chamado de "erro padrão". Ainda, adicione nesse último painel a curva da melhor gaussiana ajustada aos dados fornecidos.

O plot de quantil-quantil é muito útil para uma análise exploratória da gaussianidade dos dados. Além dele, também podemos fazer uma investigação visual no histograma dos dados e comparar com um ajuste da melhor gaussiana. Para essa análise, utilizando o método de máxima verossimilhança, que seleciona os parâmetros que maximizam $\prod_{i=1}^{n} f(X_i; \theta)$ no ajuste. Para demostrar a função, criei um vetor com a soma de 3 distribuições diferentes e os resultados se encontram na imagem.



```
analysis = function(x){
    par(mfcol=c(1,2))
    options(repr.plot.width=18,repr.plot.height=10)

    # QQ plot
    qqnorm(x, xlab='Gausian quantiles', ylab = 'Data quantiles', main='QQ plot
        ')
    qqline(x, col='red', lwd=3, lty=2)

    # Histogram
    hist(x, breaks=seq(min(x), max(x), length.out = 100))
    rug(x)

    # Estimating mean and its error
    results = t.test(x)
    abline(v=results$estimate, col='red', lwd=3)
    abline(v=results$conf.int[1], col='green', lty=2, lwd=3)
    abline(v=results$conf.int[2], col='green', lty=2, lwd=3)

    # Fitting a gaussian dist
    fit = fitdist(x, 'norm', 'mle')
    x_plot = seq(min(x), max(x), length.out = 100)
    weight = (max(x)-min(x))/100 * length(x)
    y_plot = dnorm(x_plot, mean=fit$estimate[1], sd=fit$estimate[2])*weight
    lines(x_plot, y_plot, lwd=4)
    legend('topleft', legend='Gaussian fit', lwd=4, col='black', box.lwd=0)
}

vetor = rpois(1000,lambda = 3)+rnorm(1000) + 6*rchisq(1000,df=4)
analysis(vetor)
```

Questão 5 - Estime os parâmetros das leis de potência que descrevem:

1. a distribuição de massas das estrelas primárias (M1);
2. a distribuição de massa das estrelas secundárias (M2). Represente ambas as distribuições empíricas em um gráfico quantil-quantil.

A distribuição de Pareto é muito utilizada em astronomia devido às diferentes escalas envolvidas na área. Nessa amostra de dados, temos a massa das estrelas de vários sistemas binários em logaritmo. Para estimar os parâmetros dessas leis de potência, vou utilizar o método de mínimos quadrados:

```
    binaries = read.table('/content/belikov.dat', header=T, sep='|')

    pareto.MLE = function(X) {
      n = length(X)
      m = min(X)
      a = n/sum(log(X)-log(m))
      return(c(m,a))

    # Fitting
    params_m1 = pareto.MLE(10**binaries$M1)
    params_m2 = pareto.MLE(10**binaries$M2)
    cat('Fit for primary stars (location, shape):',params_m1[1],params_m1[2],'\n')
    cat('Fit for secundary stars (location, shape):',params_m2[1],params_m2[2])

    # Creating theoretical values from fitted distribution
    fit_m1 = log10(rpareto(length(binaries$M1), shape=params_m1[2], scale=params_
        m1[1]))
    fit_m2 = log10(rpareto(length(binaries$M2), shape=params_m2[2], scale=params_
        m2[1]))

    # QQ plot for Primaries
    qqplot(binaries$M1, fit_m1, xlab = "Primary mass (log)", ylab = "Primary mass
        from fit (log)")
    lines(binaries$M1, binaries$M1, col='red', lwd=4)
    legend('topleft', legend='Q_empirical = Q_fitted', lwd=4, col='red', box.lwd
        =0)

    # QQ plot for secundaries
    qqplot(binaries$M2, fit_m2, xlab = "Secundary mass (log)", ylab = "Secundary
        mass from fit (log)")
    lines(binaries$M2, binaries$M2, col='red', lwd=4)
    legend('topleft', legend='Q_empirical = Q_fitted', lwd=4, col='red', box.lwd
        =0)
}
```

```
    Fit for primary stars (location, shape): 1.16681 0.1031938
    Fit for secundary stars (location, shape): 1.135011 0.1446229
```

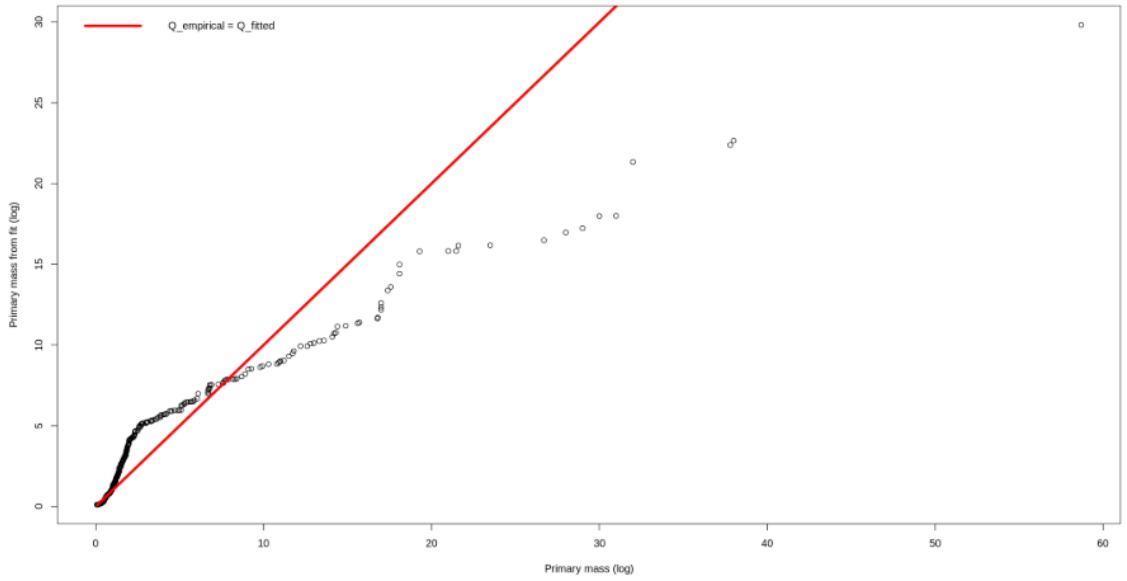


Figura 1: QQ plot com a massa primaria no eixo x e Pareto ajustada no eixo y. A linha vermelha representa o caso onde os quantis dos dados e da distribuição ajustada são iguais.

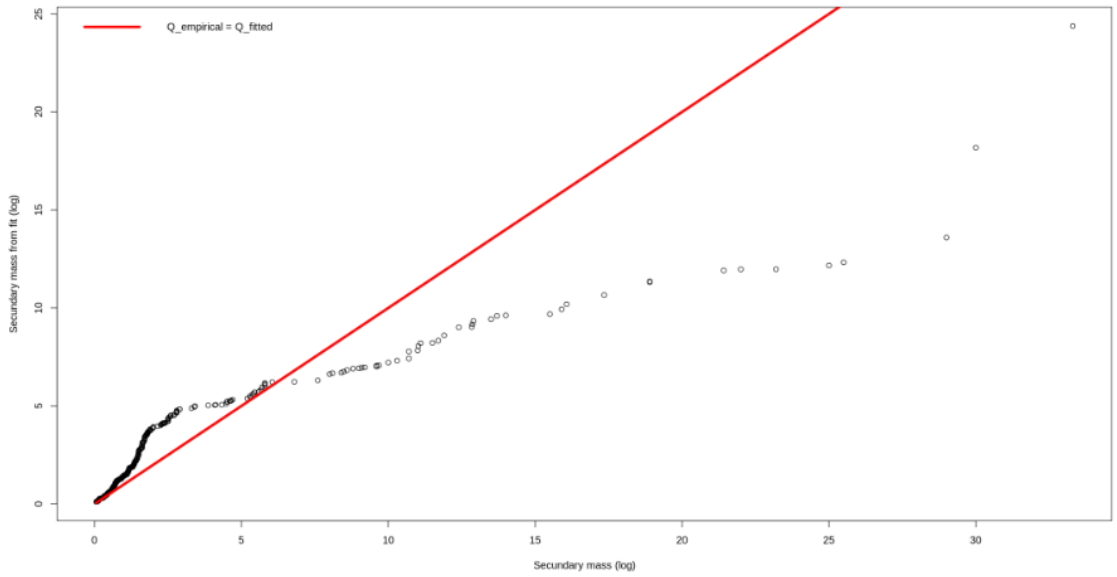


Figura 2: QQ plot com a massa secundária no eixo x e Pareto ajustada no eixo y. A linha vermelha representa o caso onde os quantis dos dados e da distribuição ajustada são iguais.

Questão 6 - Compare graficamente as distribuições de desvio para o vermelho (z) para cada classe (isto é, Type). A partir dessa comparação, descreva qualitativamente as diferenças encontradas.

Para comparar as distribuições, fiz um histograma de cada tipo, mantendo os bins iguais para facilitar a visualização.

```
data = read.table('/content/QSOBLC.dat', header=T, sep='|')

par(mfrow=c(3,1))
for (tip in c('QSO', 'SY1', 'BLZ')) {
  subset = data[data$Type == tip, ]
  hist(subset$z, breaks=seq(0,max(na.omit(data$z)+0.1), by=0.1), xlab = "
      Redshift z", xlim = c(0,max(na.omit(data$z))+0.1), main=tip, cex.main=2,
      cex.lab=1.5)
}
```

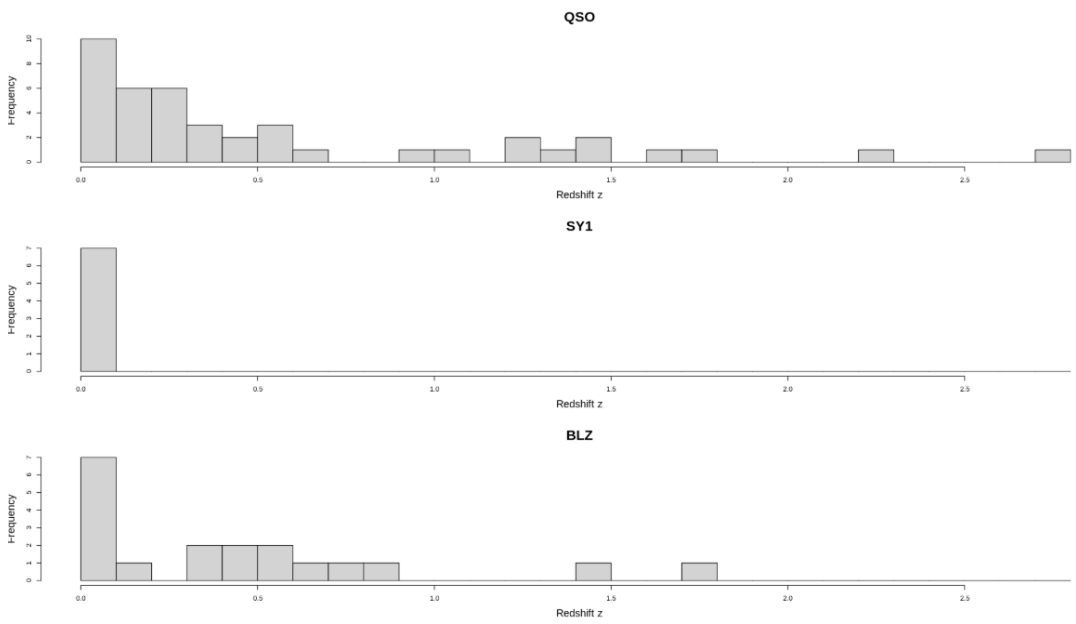


Figura 3: Distribuição de redshift para quasares, galáxias Seyfert 1 e blazares do catálogo dado.

Analisando os gráficos, podemos dizer que galáxias do tipo Seyfert 1 são objetos próximos à nossa galáxia, pois eles tem um baixo redshift. Além disso, pode-se observar semelhanças entre quasares e blazares.

Questão 7 - Teste a hipótese de que a distribuição de desvios para o vermelho é similar para os quasares (QSO) e blazares (BLZ). Use argumentos suficientes para justificar o(s) teste(s) escolhido(s)..

Para comparar as distribuições, podemos utilizar dois tipos de testes, paramétricos e não-paramétricos. Testes não-paramétricos são mais robustos e irei dar mais preferência a eles.

Nessa análise, escolhi usar o teste de Cramér-von Mises que baseia-se na distância entre a CDF das duas amostras. Apesar de haver o teste de Anderson-Darling e de Kolmogorov-Smirnov, decidi que o teste de Cramér-von Mises seria mais adequado por dois motivos:

1. Anderson-Darling dá mais peso para as pontas da distribuição, mas como temos poucos dados, não acho que seja uma boa ideia, pois a presença de um outlier faria muita diferença. Além disso, há muita incerteza em altos redshifts e também deve ser considerado.

2. O teste de Cramér-von Mises possui um melhor desempenho quando comparado ao teste de Kolmogorov-Smirnov, pois ele leva em conta todos os valores ao estimar a diferença das CDFs.

Outro teste escolhido foi o de Mann-Whitney-Wilcoxon, pois este baseia-se na soma de ranks das duas amostras. O motivo de sua escolha é que é um teste sólido na presença de outliers. Por fim, usei 2 testes paramétricos, um para a variância e outro para a média. Com isso, temos testes para diversas propriedades dos dados.

```
qso = data[data$Type=='QSO',]
blz = data[data$Type=='BLZ',]
sy1 = data[data$Type=='SY1',]

qso_z = na.omit(qso$z)
blz_z = na.omit(blz$z)
sy1_z = na.omit(sy1$z)

# Non-parametric tests
cvm = cvmts.test(qso_z, blz_z)
cat('p-value from CVM test: ',cvmts.pval(cvm, length(qso_z), length(blz_z)))
wilcox.test(qso_z,blz_z)

# Parametric tests
t.test(qso_z, blz_z)
var.test(qso_z, blz_z)
```

O valor p para o testes foram 0.2772023, 0.3664, 0.3668, 0.1654 para Cramér-von Mises, Wilcoxon, t-test, F test, respectivamente. Logo, não podemos afirmar nada sobre as semelhanças e diferenças entre as distribuições de redshift.

Questão 8 - Usando uma abordagem paramétrica e outra não paramétrica, calcule o coeficiente de correlação entre o desvio para o vermelho e a magnitude V, para as galáxias Seyfert 1 (SY1)? Qual a abordagem mais correta a ser considerada? Existe alguma base física para a relação encontrada?

A abordagem não-paramétrica geralmente é a mais adequada, pois não depende de conhecimentos prévios dos dados. Porém, como veremos na questão 10, não pode-se rejeitar a hipótese de que o redshift e a magnitude seguem uma distribuição Gaussiana e como o teste de Pearson assume Gaussianidade nas variáveis, então podemos usá-lo. Todavia, há poucos objetos na amostra e isso faz com que os testes falhem, incluindo o de Gaussianidade. Logo, eu considero mais correto utilizar o teste não-paramétrico.

```
cor.test(sy1$z, sy1$Vmag, method='pearson')
cor.test(sy1$z, sy1$Vmag, method='spearman')
```

```
        Pearson product-moment correlation

data:  sy1$z and sy1$Vmag
t = 0.92786, df = 5, p-value = 0.3961
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5198244  0.8818138
sample estimates:
      cor
0.3832664
```

```
        Spearman rank correlation rho

data:  sy1$z and sy1$Vmag
S = 36.828, p-value = 0.4523
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3423562
```

Desses valores, não podemos rejeitar a hipótese nula de que não existe correlação ($\rho = 0$). Existe sim uma relação física entre a magnitude e o redshift, pois com diferentes redshifts, teremos diferentes partes do espectro na banda V. Todavia, essa relação não é linear, já que o fluxo em diferentes partes do espectro de uma galáxia pode aumentar ou diminuir. Assim, é de se esperar que não tenha uma correlação entre as duas quantidades físicas. Por fim, devemos nos lembrar que há poucos dados e isso limita nosso poder de afirmações.

Questão 9 - Use 1000 sorteios de bootstrap para definir o intervalo de confiança de 97.5 % para o valor da mediana da distribuição de desvio para o vermelho (z) da amostra de quasares (QSO) e da amostra de Blazares (BLZ).

A técnica de bootstrap é baseada em amostragem com repetição. Dado um número de sorteios, cada sorteio seleciona N objetos dos dados com repetição e então estima a quantidade desejada para cada sorteio. No fim, obtém-se o uma distribuição da quantidade, onde tira-se o intervalo de confiança.

```
theta = function(x,i){
  median(x[i])
}
qso_boot = boot(qso_z, statistic=theta, R=1000)
boot.ci(qso_boot, conf=0.975)
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = qso_boot, conf = 0.975)

Intervals :
Level      Normal              Basic
97.5%   ( 0.1124,  0.4421 )   ( 0.0763,  0.4080 )

Level     Percentile            BCa
97.5%   ( 0.1770,  0.5087 )   ( 0.1710,  0.4860 )
```

```
blz_boot = boot(blz_z, statistic=theta, R=1000)
boot.ci(blz_boot, conf=0.975)
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = blz_boot, conf = 0.975)

Intervals :
Level      Normal              Basic
97.5%   ( 0.0370,  0.7308 )   ( 0.0640,  0.6638 )

Level     Percentile            BCa
97.5%   ( 0.0562,  0.6560 )   ( 0.0461,  0.5950 )
```

Levando o resultado do "Basic", temos que o intervalo de confiança da mediana quasares é menor do que para os blazares. Possivelmente isso é devido ao menor número de dados na amostra de blazares.

Questão 10 - As distribuições de z e Vmag de cada classe (Type) podem ser consideradas gaussianas? Justifique sua resposta com mais de um teste/análise.

```
for (tip in c('QSO', 'SY1', 'BLZ')) {
  subset = data[data$Type == tip, ]
  z = na.omit(subset$z)
  v = na.omit(subset$Vmag)
  cat('Starting', tip, 'type\n\n')

# Redshift
  if (length(z) > 7) {
  ad_z = ad.test(z)
  cvm_z = cvm.test(z)
  cat('p-value from AD test for redshift in',tip,':',ad_z$p.value,'\n')
  cat('p-value from CVM test for redshift in',tip,':',cvm_z$p.value,'\n')
  }

  dip = dip.test(z)
  shapiro = shapiro.test(z)
  lillie = lillie.test(z)
  cat('p-value from dip test for redshift in',tip,':',dip$p.value,'\n')
  cat('p-value from Shapiro test for redshift in',tip,':',shapiro$p.value,'\n'
      )
  cat('p-value from Lillie test for redshift in',tip,':',lillie$p.value,'\n','
      \n')

# V mag

  if (length(v) > 7) {
    ad_v = ad.test(v)
    cvm_v = cvm.test(v)
    cat('p-value from AD test for Vmag in', tip, ':', ad_v$p.value, '\n')
    cat('p-value from CVM test for Vmag in', tip, ':', cvm_v$p.value, '\n')
  }

  dip = dip.test(v)
  shapiro = shapiro.test(v)
  lillie = lillie.test(v)
  cat('p-value from dip test for Vmag in', tip, ':', dip$p.value, '\n')
  cat('p-value from Shapiro test for Vmag in', tip, ':', shapiro$p.value, '\n'
      )
  cat('p-value from Lillie test for Vmag in', tip, ':', lillie$p.value, '\n',
      '\n')
  cat('############################################################\n\n')
}
```

```
Starting QSO type

p-value from AD test for redshift in QSO : 7.681129e-09
p-value from CVM test for redshift in QSO : 1.121454e-07
p-value from dip test for redshift in QSO : 0.9870413
p-value from Shapiro test for redshift in QSO : 1.547669e-06
p-value from Lillie test for redshift in QSO : 1.573389e-05

p-value from AD test for Vmag in QSO : 0.03655097
p-value from CVM test for Vmag in QSO : 0.05203636
p-value from dip test for Vmag in QSO : 0.4685278
p-value from Shapiro test for Vmag in QSO : 0.03447447
p-value from Lillie test for Vmag in QSO : 0.06412205

############################################################

Starting SY1 type


p-value from dip test for redshift in SY1 : 0.1182018
p-value from Shapiro test for redshift in SY1 : 0.4400546
p-value from Lillie test for redshift in SY1 : 0.6378789

p-value from dip test for Vmag in SY1 : 0.9351645
p-value from Shapiro test for Vmag in SY1 : 0.2234065
p-value from Lillie test for Vmag in SY1 : 0.2478134

############################################################

Starting BLZ type

p-value from AD test for redshift in BLZ : 0.006366748
p-value from CVM test for redshift in BLZ : 0.01736082
p-value from dip test for redshift in BLZ : 0.7571165
p-value from Shapiro test for redshift in BLZ : 0.003163003
p-value from Lillie test for redshift in BLZ : 0.08508485

p-value from AD test for Vmag in BLZ : 0.06981848
p-value from CVM test for Vmag in BLZ : 0.06846265
p-value from dip test for Vmag in BLZ : 0.9826913
p-value from Shapiro test for Vmag in BLZ : 0.1373625
p-value from Lillie test for Vmag in BLZ : 0.119869

############################################################
```

Da análise de quasares, pode-se dizer que temos evidência para confirmar que as distribuições de redshift e de magnitude V **não** seguem uma normal. Apesar do p valor alto em alguns testes, como o dip, estou considerando as diversas características de uma distribuição normal, por isso faço tal afirmação.

Da análise de Seyfert 1, pode-se dizer que não há evidência para confirmar a refutar a hipótese nula de que as distribuições redshift e da magnitude V seguem uma normal.

Da análise de blazares, pode-se dizer que temos evidência para confirmar que as distribuições de redshift não segue uma normal. Para a magnitude V, não temos evidência para refutar a hipótese nula.