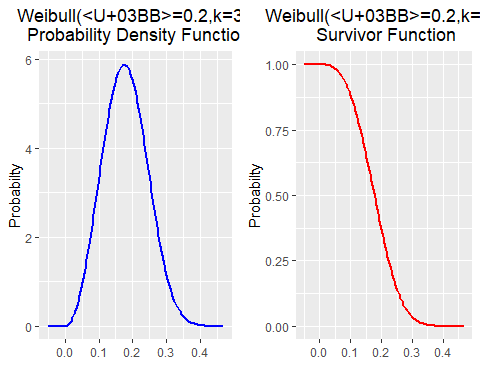
20150879 (replace with your own student number)

A2, B3, C5 (replace with your allocated questions)

# A2, 20150879

INSERT solution to question A0

#weibull parameters  
wshape = 3  
wscale = 0.2  
# create a variable name use as title on the plots  
fun\_name = sprintf("Weibull(λ=%s,k=%s)",wscale,wshape)  
  
# identify the boundaries of the distribution  
wmin = qweibull(0, shape=wshape, scale = wscale) # min as the quantile 0  
wmax = qweibull(0.9999, shape=wshape, scale = wscale) # max as the quantile 99.99%  
  
#Generate the density plot for he Weibull function   
DensityFig <- ggplot() +  
 stat\_function(fun = dweibull, args = list(shape=wshape, scale = wscale), # pass the function and the arguments for weibull  
 size = 1 , color = "blue" )   
SurvFig <- ggplot() +   
 stat\_function(  
 fun = function(x) 1-pweibull (x, shape=wshape, scale = wscale), # define a function as 1-cumulative weibull  
 size = 1 , colour = "red" )   
  
#Set format parameters for both graphs  
DensityFig <-  
 DensityFig + labs(title = paste(fun\_name, "\n Probability Density Function"))+  
 xlim(wmin-0.05, wmax+0.05) +   
 labs( x ="", y="Probability") + theme(plot.title = element\_text(hjust = 0.5))  
  
SurvFig<-SurvFig +  
 xlim(wmin-0.05, wmax+0.05) + labs(title = paste(fun\_name, "\n Survivor Function"))+  
 labs( x ="", y="Probability") + theme(plot.title = element\_text(hjust = 0.5))  
  
#Plot both figures in the same grid  
grid.arrange(DensityFig,SurvFig,ncol=2)



# B3, 20150879

The data file lights.dat contains data on the failure time of fluorescent strip lights in thousands of hours

# 1. read the data  
t\_fail<-scan("lights.dat")

# 2. summary of the data  
  
# each line enunciates what each functions does  
cat("Description of the data set","\n\n")

## Description of the data set

cat("Number of observations in the dataset :",length(t\_fail),"\n")

## Number of observations in the dataset : 100

cat("Mean of dataset :",mean(t\_fail),"\n")

## Mean of dataset : 4.10998

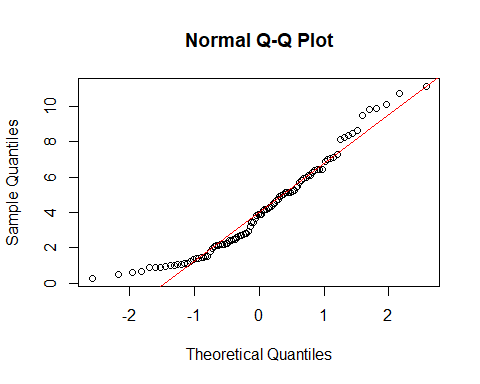
cat("Standard deviation of dataset :", sd(t\_fail) ,"\n")

## Standard deviation of dataset : 2.647777

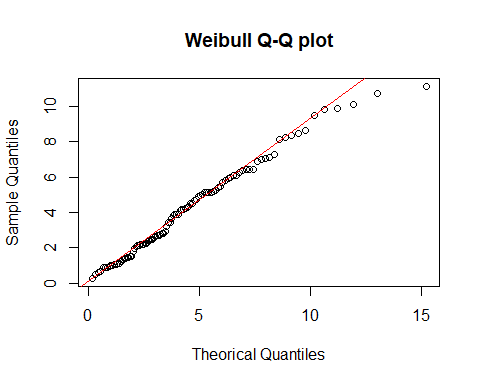
prop = sum(t\_fail>=5)/length(t\_fail) # numerator: sum vector of 1 and 0 in case where t\_fail is grater than 5 (data is in thousands of hours). divisor: total of observations  
cat("proportion of lights surviving beyond 5,000 hours : ", round(prop\*100,2),"%\n", sep = "")

## proportion of lights surviving beyond 5,000 hours : 35%

#3. Produce a Quantile-Quantile plot with reference to a normal distribution.   
  
# a. Plot Q-Q plot for a theorical normal distribution.   
# The first parameter is our data that we intend to compare it with.  
qqnorm(t\_fail)  
  
# b. plot the qqline. This function will create the appropriate reference line for our data against the normal distribution.  
qqline(t\_fail, col = 2)



#4. Produce a Quantile-Quantile plot with reference to a Weibull distribution with shape parameter 1.5 and scale parameter 5  
  
# a. set the parameters of the shape and scale  
wshape = 1.5  
wscale = 5  
  
# b. plot Q-Q plot. first parameter is the generated by the Weibull distribution our theorical distribution. The second parameters is our data that we intend to compare it with.  
qqplot(qweibull(ppoints(length(t\_fail)), shape = wshape, scale = wscale), t\_fail,  
 main = "Weibull Q-Q plot", xlab = "Theorical Quantiles",   
 ylab = "Sample Quantiles")  
  
# c. plot the qqline. This function will create the appropriate reference line for our data and the Weibull distribution given as first and second parameter respectively.  
qqline(t\_fail, distribution = function(p) qweibull(p, shape = wshape, scale = wscale), col = 2)



INSERT answer to question B0, including any necessary code chunks and outputs.

# C5, 20150879

### a. Goal

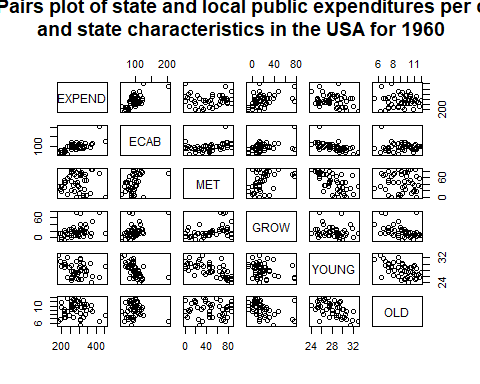
As summary we are asked to test the independency of two variables of the data in file spe.dat. To do this we are proposed to test in three different ways: 1. Perform a T-test using the statistic , where is the Pearson’s coefficient of correlation 2. Use the Fisher’s z-transform for , defined as to test the hypothesis and see the 95% confidence interval for 3. erform a T-test using the statistic , where is the Spearman’s coefficient of rank correlation

The data provided in file spe.dat contains data on per capita state and local public expenditures and associated state demographic and economic characteristics, in the USA, for 1960. It contains eight variables

### b. Select two variables to perform the analysis

First, is always a good idea to start by having an idea of the data around the goal searched for. The next chuck will load the data and plots scatters of each pairwise combination of variables, so it will be easy to identify for which pairs of variables we should expect high and low correlations.

#Load Data  
ecoUSA<-read.table("spe.dat", header = TRUE)  
  
#plot pair of varibles   
pairs(ecoUSA[,1:6], main = "Fig 1. Pairs plot of state and local public expenditures per capita \n and state characteristics in the USA for 1960")



Analysing the pair plot, it is no possible to find a clear perfect correlation between two variables (the ideal perfect correlated pair-plot would look as diagonal straight line of points). Nevertheless, visually, the most correlated variables probably are ECAB~EXPEND and ECAB~GROW. Conversely, the less correlated variable OLD~MET and MET~EXPEND as there is no correlation

We will analyse the relation between the variables  
EXPEND: Per capita state and local public expenditures ($) ECAB: Economic ability index, in which income, retail sales, and the value of output (manufactures, mineral, and agricultural) per capita are equally weighted.

## Modify the parameters that are manually needed to perform the analyses   
# select the variables to analyse   
x= ecoUSA$EXPEND  
y= ecoUSA$ECAB  
  
n= as.integer(length(x)) # number of observations. This will be use later.  
alpha =0.05 # state the alpha for the confidence interval. This will be use later.  
  
txtlabs= c("Test statistic", "P value") # a list that will be call later

### c. Pearson’s coefficient of correlation

cat("-----------------------","\n")

## -----------------------

cat("Analysis of no association between variables using Pearson's coefficient of correlation r", "\n", sep = "")

## Analisis of no association between variables using Pearson's coefficient of correlation r

#Pearson's coefficient of correlation r  
r = sum((x-mean(x))\*(y-mean(y)))/  
 (sum((x-mean(x))^2)\*sum((y-mean(y))^2))^(1/2)  
  
cat("Pearson's correlation (r): ", r ,"\n", sep = "")

## Pearson's correlation (r): 0.6558625

# T statistic  
Tp = r\*(n-2)^0.5/(1-r^2)^0.5  
### Eval "Tp" in t-student n-2 to get p value  
pval=2\*pt(-abs(Tp),df=n-2)  
  
  
# print outputs  
cat("\n", "Two side-test for ", sep = "")

##   
## Two side-test for

cat("H0: correlation coefficient ρ = 0","\n", sep = "")

## H0: correlation coefficient <U+03C1> = 0

cat(txtlabs[1]," : ", Tp ,"\n", sep = "")

## Test statistic : 5.89269

cat(txtlabs[2]," : ", pval ,"\n", sep = "")

## P value : 4.192755e-07

Therefore, there is statistical evidence to reject the null hypothesis of .

### d. Fisher’s z-transform

To compute the confidence interval, we need to calculate the inverse of the Fisher’s z-transform, so we are able have the probability in terms of . In the following lines is explained the mathematical step for this.

$$ Z=\dfrac{1}{2}~log \left( \dfrac{1+r}{1−r} \right) \\ \exp(2Z) = \dfrac{1+r}{1−r} \\ 1+r = \exp(2Z) - \exp(2Z) r \\ \exp(2Z) r+r = \exp(2Z) - 1 \\ r (\exp(2Z) +1) = \exp(2Z) - 1 \\ r = \dfrac{\exp(2Z) -1}{\exp(2Z) + 1} $$

Then we can use this last formula to change the values from to

cat("-----------------------","\n")

## -----------------------

cat("Analisis of no association between variables using Fisher's z-transform", "\n", sep = "")

## Analisis of no association between variables using Fisher's z-transform

cat("\n", "Approximate two-side-test for ", sep="")

##   
## Approximate two-side-test for

cat("H0: correlation coefficient ρ = 0","\n", sep = "")

## H0: correlation coefficient <U+03C1> = 0

# Fischers Z transform statistic  
Zfisher = 1/2 \* log((1+r)/(1-r))  
Zmean = 1/2 \* log((1+0)/(1-0))  
Zvar=1/(n-3)  
### Eval for H0 \ro=0 as N(Zmean,Zvar) / get p value   
pval=2\*pnorm(-abs(Zfisher),mean = Zmean, sd = sqrt(Zvar))  
  
# print outputs  
cat("Pearson's correlation (r): ", r ,"\n", sep = "")

## Pearson's correlation (r): 0.6558625

cat(txtlabs[1]," : ", Zfisher ,"\n", sep = "")

## Test statistic : 0.785518

cat(txtlabs[2]," : ", pval ,"\n", sep = "")

## P value : 1.368591e-07

### create a CI for with 95% confidence interval for \ro  
  
cat("\n", "95% confidence interval for ρ","\n", sep = "")

##   
## 95% confidence interval for <U+03C1>

CI\_inf <- Zfisher-qnorm(1-alpha/2)\*sqrt(Zvar) # calculate inferior bound in the statistc  
CI\_inf <- (exp(2\*CI\_inf)-1)/(exp(2\*CI\_inf)+1) # transform to ρ  
  
CI\_sup <- Zfisher+qnorm(1-alpha/2)\*sqrt(Zvar)# calculate superior bound in the statistic  
CI\_sup <- (exp(2\*CI\_sup)-1)/(exp(2\*CI\_sup)+1) # transform to ρ  
  
cat("[",CI\_inf,",",CI\_sup,"]", "\n") # print output

## [ 0.4568664 , 0.7923417 ]

Both, the two-side-test and the confidence interval gives evidence to reject with 95% confidence. The p-value is inferior to 5% and the interval does not contain the searched value 0

### e. Spearman’s coefficient of rank correlation

cat("-----------------------","\n")

## -----------------------

cat("Analysis of no association between variables using Spearman's coefficient of rank correlation (rs)", "\n", sep = "")

## Analysis of no association between variables using Spearman's coefficient of rank correlation (rs)

cat("\n", "Approximate two-side-test for ", sep="")

##   
## Approximate two-side-test for

cat("H0: X and Y are independent","\n", sep = "")

## H0: X and Y are independent

#Spearman's coefficient of rank correlation rs  
xs= rank(x)  
ys= rank(y)  
  
rs = sum((xs-mean(xs))\*(ys-mean(ys)))/  
 (sum((xs-mean(xs))^2)\*sum((ys-mean(ys))^2))^(1/2)  
  
cat("Spearman's correlation (rs): ", rs ,"\n", sep = "")

## Spearman's correlation (rs): 0.5850732

#Spearman statistic   
Ts = rs\*(n-2)^0.5/(1-rs^2)^0.5  
### Eval "Ts" in t-student n-2 / get p value  
pval=2\*pt(-abs(Ts),df=n-2)  
  
cat(txtlabs[1]," : ", Ts ,"\n", sep = "")

## Test statistic : 4.893039

cat(txtlabs[2]," : ", pval ,"\n", sep = "")

## P value : 1.257479e-05

Aligned with the previous results. we can reject the null hypothesis