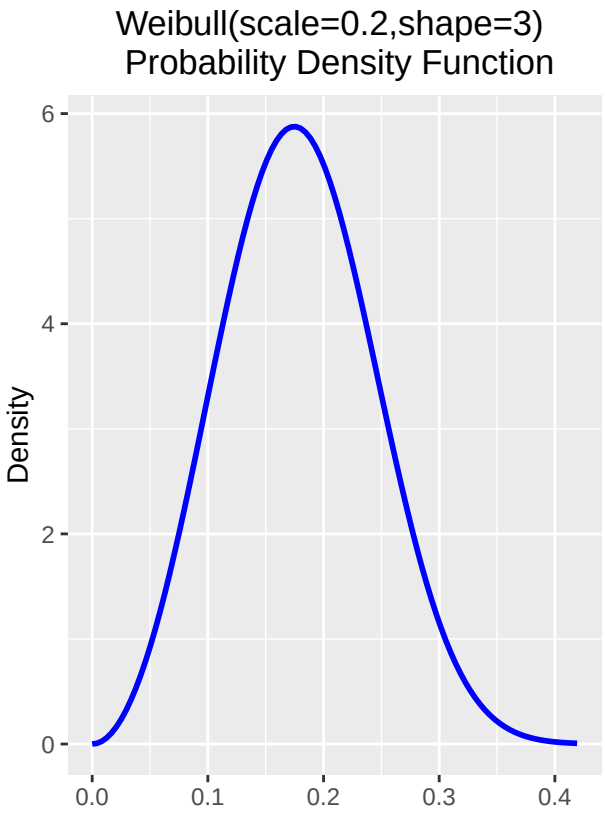
20150879

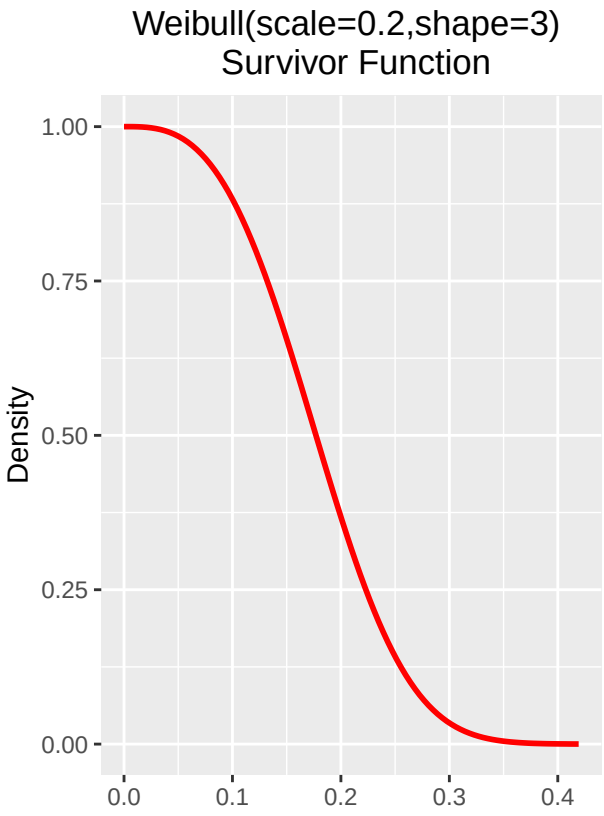
A2, B3, C5

```
#import packages
library(ggplot2)
library(gridExtra)
```

A2, 20150879

```
#a. weibull parameters
wshape = 3
wscale = 0.2

#b. 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%

#c. Generate plots for the Weibull function
#c.1 Generate the density plot for the Weibull function
DensityFig <- ggplot() +
  stat_function(fun = dweibull, args = list(shape=wshape, scale = wscale),
                size = 1 , color = "blue" )
#c.2 Generate the Survivot function plot for he Weibull function
SurvFig <- ggplot() +
  stat_function(
    fun = function(x) 1-pweibull (x, shape=wshape, scale = wscale),
    size = 1 , colour = "red" )

#c.3. Set format parameters for both graphs
#c.3.1. create a variable name use as title on the plots
fun_name = sprintf("Weibull(scale=%s,shape=%s)",wscale,wshape)

DensityFig <-
  DensityFig + labs(title = paste(fun_name, "\n Probability Density Function"))+
  xlim(wmin, wmax) +
  labs( x ="", y="Density") + theme(plot.title = element_text(hjust = 0.5))

SurvFig<-SurvFig +
  xlim(wmin, wmax)  + labs(title = paste(fun_name, "\n Survivor Function"))+
  labs( x ="", y="Density") + theme(plot.title = element_text(hjust = 0.5))

#d. 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

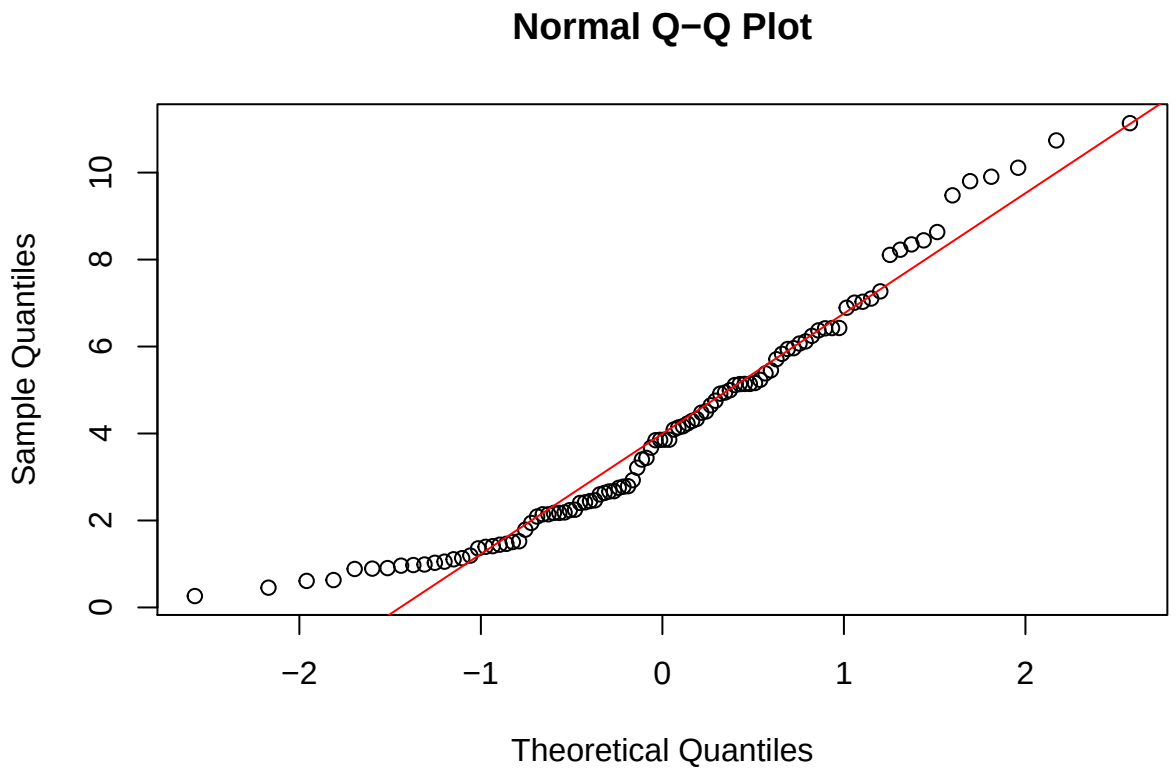
```
# 1. read the data
t_fail<-scan("lights.dat")
```

```
# 2. summary of the data

#calculate the proportion of obs. surviving beyond 5,000 hours
prop = sum(t_fail>=5)/# numerator: sum of cases where t_fail is >= 5 ( thousands of hours)
  length(t_fail) # divisor: total of observations

# print outputs. Each line enunciates  what each functions does
cat("Description of the data set","  \n  \n",
    "Number of observations in the dataset              : ",length(t_fail),"  \n",
    "Mean of dataset                                    : ",mean(t_fail),"  \n",
    "Standard deviation of dataset                      : ", sd(t_fail) ,"  \n",
    "proportion of lights surviving beyond 5,000 hours  : ", round(prop*100,2),"%",
    sep = "")
```

```
## Description of the data set
##
```

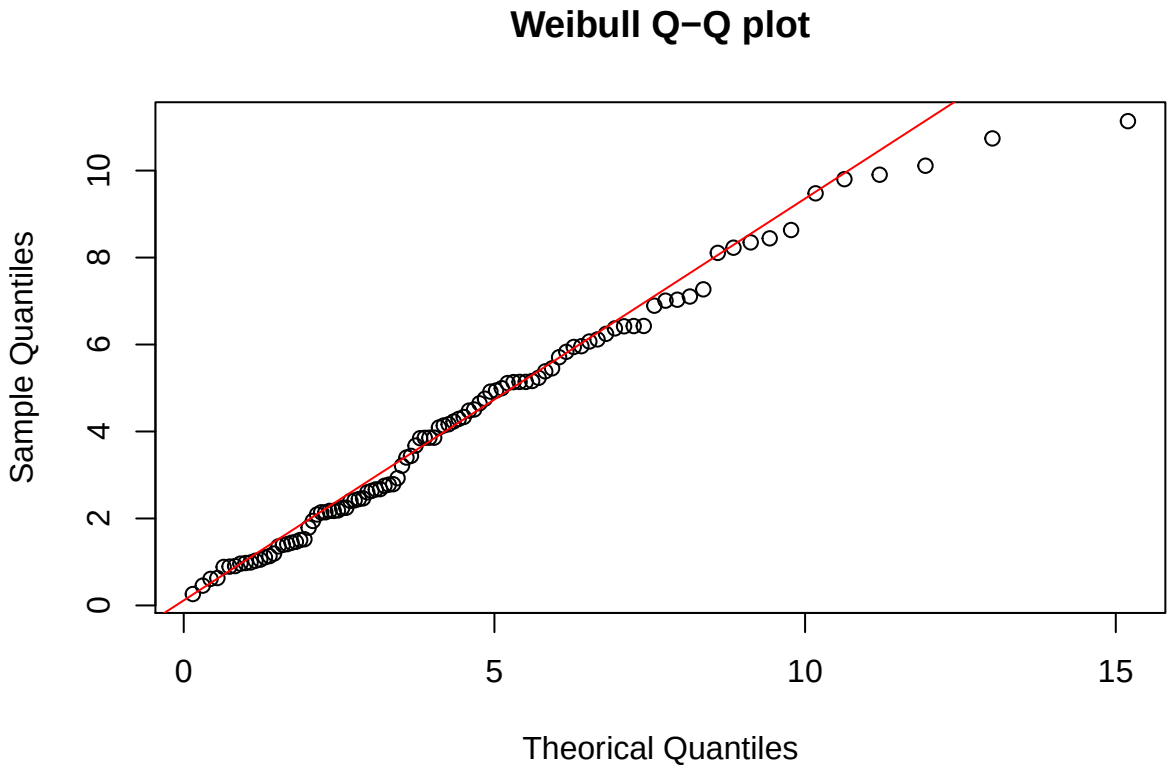
```
## Number of observations in the dataset              : 100
## Mean of dataset                                    : 4.10998
## Standard deviation of dataset                      : 2.647777
## proportion of lights surviving beyond 5,000 hours  : 35%
```

```r
#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.
#The appropriate reference line for our data against the normal distribution.
qqline(t_fail, col = 2)
```




```r
#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.
```

```r
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)
```



C5, 20150879

a. Goal

As summary we are asked to test the independency of two variables of the data in file spe.dat. Which is true if and only if the correlation between the two varailbes is close to zero. So we are proposed to test $H_0 : \rho = 0$ in three different ways:

1. Perform a T-test using the statistic $T = r\sqrt{n-2}/\sqrt{1-r^2}$, where r is the Pearson's coefficient of correlation.
2. Use the Fisher's z-transform for r, defined as $Z = 1/2\ log(1 + r/1 - r)$ to test the hypothesis and see the 95% confidence interval for ρ.

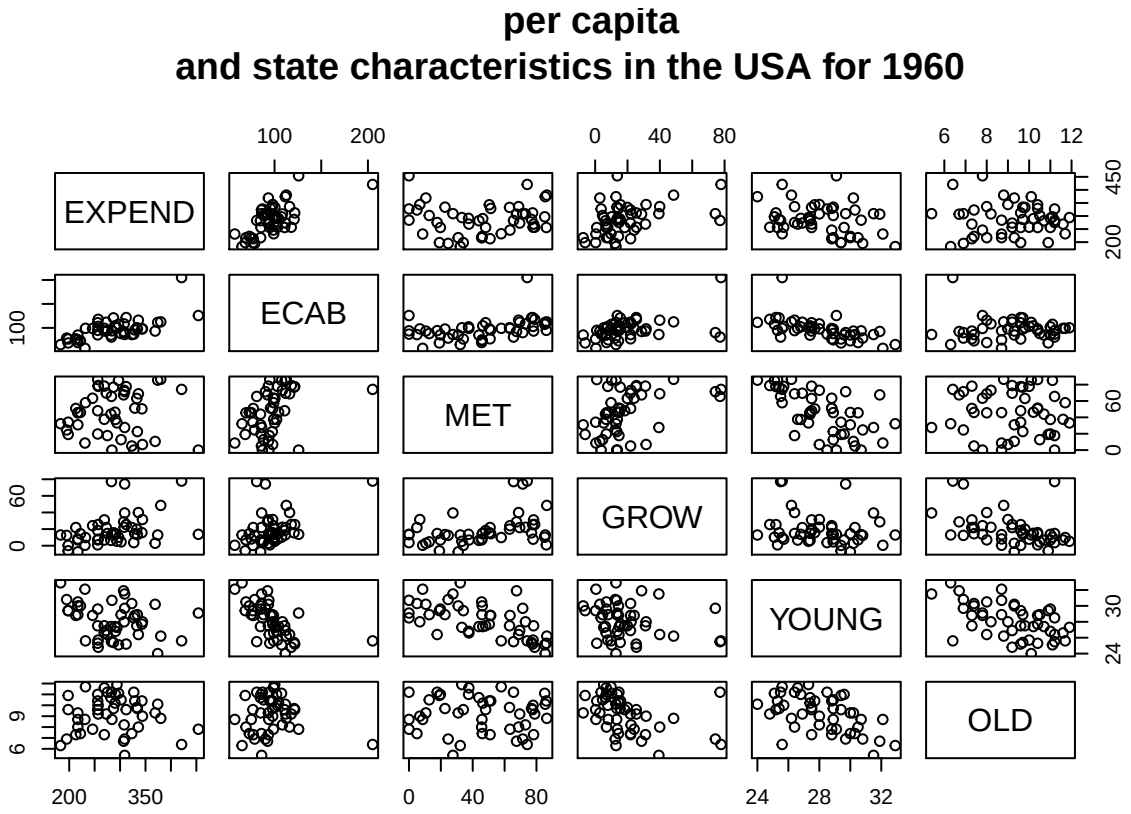
3. Perform a T-test using the statistic $T_s = r_s\sqrt{n-2}/\sqrt{1-r_S^2}$, where r_s is the Spearman's coefficient of rank correlation.

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 seeked. 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 next two variables, which looks correlated and therefore we expect to reject H_0:

- 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 use to print results
```

c. Pearson's coefficient of correlation

```
#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)

# 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("Analysis  of no association between variables using Pearson's
    coefficient of correlation r", "\n",
    "Pearson's correlation (r): ", r ,"\n",
    "\n", "Two side-test for ",
    "H0: correlation coefficient rho = 0","\n",
    txtlabs[1]," : ", Tp ,"\n",
    txtlabs[2],"        : ", pval ,"\n", sep = "")
```

```
## Analysis  of no association between variables using Pearson's
##     coefficient of correlation r
## Pearson's correlation (r): 0.6558625
##
## Two side-test for H0: correlation coefficient rho = 0
## Test statistic : 5.89269
## P value        : 4.192755e-07
```

Therefore, there is statistical evidence to reject the null hypothesis H_0 of $\rho = 0$.

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 ρ. In the following lines is explained the mathematical step for this.

$$Z = \frac{1}{2}\, log\left(\frac{1+r}{1-r}\right) \tag{1}$$

$$\exp(2Z) = \frac{1+r}{1-r} \tag{2}$$

$$1 + r = \exp(2Z) - \exp(2Z)r \tag{3}$$

$$\exp(2Z)r + r = \exp(2Z) - 1 \tag{4}$$

$$r(\exp(2Z) + 1) = \exp(2Z) - 1 \tag{5}$$

$$r = \frac{\exp(2Z) - 1}{\exp(2Z) + 1} \tag{6}$$

Then we can use this last formula to change the values from the distribution of the Fischers Z-transform statistic to ρ, the correlation coefficient.

```
# Fischers Z transform statistic using formulas given
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))

### create a CI for with  95% confidence interval for \ro
CI_inf <- Zfisher-qnorm(1-alpha/2)*sqrt(Zvar) # inferior  bound in the statistc dim.
CI_inf <- (exp(2*CI_inf)-1)/(exp(2*CI_inf)+1) # transform to \rho

CI_sup <- Zfisher+qnorm(1-alpha/2)*sqrt(Zvar)# superior  bound in the statistic dim.
CI_sup <- (exp(2*CI_sup)-1)/(exp(2*CI_sup)+1) # transform to \rho

# print outputs
cat("Analysis  of no association between variables using Fisher's z-transform", "\n",
    "\n", "Approximate two-side-test for ",
    "H0: correlation coefficient rho = 0","\n",
    "Pearson's correlation (r): ", r ,"\n",
    txtlabs[1]," : ", Zfisher ,"\n",
    txtlabs[2],"        : ", pval ,"\n",

    "\n", "95% confidence interval for rho","\n",
    "[",CI_inf,",",CI_sup,"]", "\n")
```

```
## Analysis  of no association between variables using Fisher's z-transform
##
##  Approximate two-side-test for  H0: correlation coefficient rho = 0
##  Pearson's correlation (r):  0.6558625
##  Test statistic  :  0.785518
##  P value         :  1.368591e-07
##
##  95% confidence interval for rho
##  [ 0.4568664 , 0.7923417 ]
```

Both, the two-side-test and the confidence interval gives evidence to reject H_0 with 95% confidence. The p-value is inferior to 5% and the interval does not contain the searched value for ρ: 0.

e. Spearman's coefficient of rank correlation

```
#a. calculate Spearman's coefficient of rank correlation rs
xs= rank(x) # rank values of variable x
ys= rank(y) # rank values of variable y

rs = sum((xs-mean(xs))*(ys-mean(ys)))/
  (sum((xs-mean(xs))^2)*sum((ys-mean(ys))^2))^(1/2) # compute Spearman coeficient

#b. Spearman statistic
Ts =  rs*(n-2)^0.5/(1-rs^2)^0.5 # calculate statistic
pval=2*pt(-abs(Ts),df=n-2) # Eval "Ts" in t-student n-2 / get p value

#print outputs
cat("Analysis  of no association between variables using Spearman's
    coefficient of rank correlation (rs)", "\n",
    "\n", "Approximate two-side-test for ",
    "H0: X and Y are independent","\n",
    "Spearman's correlation (rs): ", rs ,"\n",
    txtlabs[1]," : ", Ts ,"\n",
    txtlabs[2],"        : ", pval ,"\n", sep = "")
```

```
## Analysis  of no association between variables using Spearman's
##     coefficient of rank correlation (rs)
##
## Approximate two-side-test for H0: X and Y are independent
## Spearman's correlation (rs): 0.5850732
## Test statistic : 4.893039
## P value        : 1.257479e-05
```

Aligned with the previous results, we can reject the null hypothesis of $\rho = 0$