assignment3

2022-09-20

R Markdown

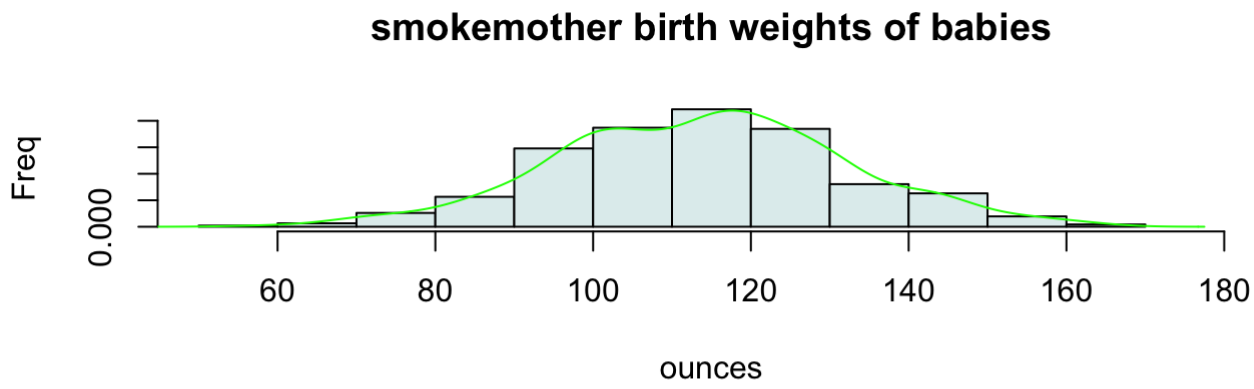
Access the data from url http://www.stat.berkeley.edu/users/statlabs/data/babies.data (http://www.stat.berkeley.edu/users/statlabs/data/babies.data) and store the infor- mation in an object named BABIES using the function read.table(). Use the option that reads the first line as header. A description of the variables can be found at http://www.stat.berkeley.edu/users/statlabs/labs.html (http://www.stat.berkeley.edu/users/statlabs/labs.html). Look for the data set Birth Weight II. These data are a subset from a much larger study dealing with child health and development.

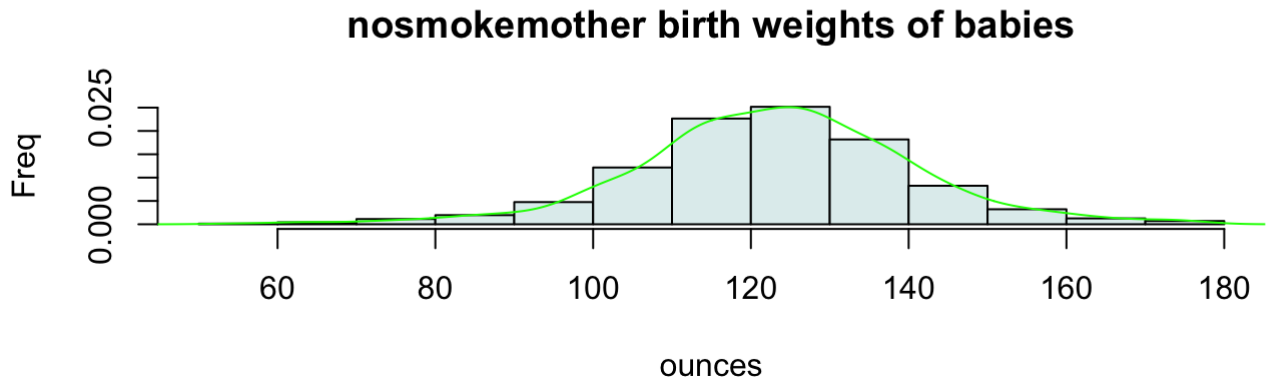
i. Create a “clean” data set that removes subjects if any observations on the subject are “unknown.” Note that bwt, gestation, parity, height, weight, and smoke use values of 999, 999, 9, 99, 999, and 9, respectively, to denote “unknown.” Store the modified data set in an object named CLEAN. The function subset may be useful here.

```
BABIES = read.table("http://www.stat.berkeley.edu/users/statlabs/data/babies.data",he
ader = TRUE)
CLEAN = subset(BABIES, bwt !=999 & gestation !=999 & parity != 9 & height != 99 & wei
ght !=999 & smoke != 9)
```

ii. Use the information in CLEAN to create a histogram of the birth weights of babies whose mothers have never smoked (smoke=0) and another histogram placed directly below the first in the same graphics device for the birth weights of babies whose mothers currently smoke (smoke=1). Use a common range of the x-axis for both histograms. Superimpose a density curve over each histogram. Use informative titles and labels for your graphs. Comment on what you observe.

```
smokemother = subset(CLEAN, smoke==1)
nosmokemother = subset(CLEAN, smoke==0)
par(mfcol=c(2,1))
hist(smokemother$bwt,col = 'azure2',xlab='ounces',xlim=c(50,180),  ylab='Freq',main =
"smokemother birth weights of babies", prob= TRUE) # must have prob parameter for des
ity
lines(density(smokemother$bwt),col = "green")
hist(nosmokemother$bwt,col = 'azure2',xlab='ounces', xlim=c(50,180), ylab='Freq',main
= "nosmokemother birth weights of babies",prob= TRUE)
lines(density(nosmokemother$bwt),col = "green")
```





The birth weights of babies whose mothers have never smoked have more big case than the birth weights of babies whose mothers currently smoke. The birth weights of babies whose mothers currently smoke have more outliers.

iii. The body weight index or body mass index (bmi) is defined as the weight of a person divided by the height squared and is measured in units of kg/m2. Compute the bmi for each mother in CLEAN. Observe that you have to convert the measurements in the data frame to metric (0.0254 m= 1 in., and 0.45359 kg= 1 lb.). Modify the variables weight and height so that they now appear in metric units (kg and m), and add bmi to CLEAN and store the result in CLEANP. Count how many subjects have bmi above 30.

```
CLEAN$weight = CLEAN$weight * 0.45359
CLEAN$height = CLEAN$height *0.0254
CLEAN <- within(CLEAN, bmi <- weight/ (height)^2 )
CLEANP = CLEAN$bmi
sum(CLEAN$bmi > 30,na.rm = TRUE)
```

```
## [1] 34
```

Question 2

The file data_q4.csv has four simulated samples of size 20 coming from the following distributions • Standard Cauchy, (rcauchy(20)) • Chi-square with 2 degrees of freedom, χ2, (rchisq(20,2)) • Lognormal with standard parameters, (rlnorm(20)) • Weibull with shape parameter 2 (rweibull(20,2)) You have to identify which is which using quantile plots. Since you will need to draw quantile plots with respect to distributions other than the normal, it will be convenient to use a new function named qqPlot in the package car. You will need to install

this package. If you are using RStudio, select the Packages tab on the panel on the right and then select the Install tab. Type car on the pop-up window and click install. After installing, you need to load the package using library(car). The function qqPlot has

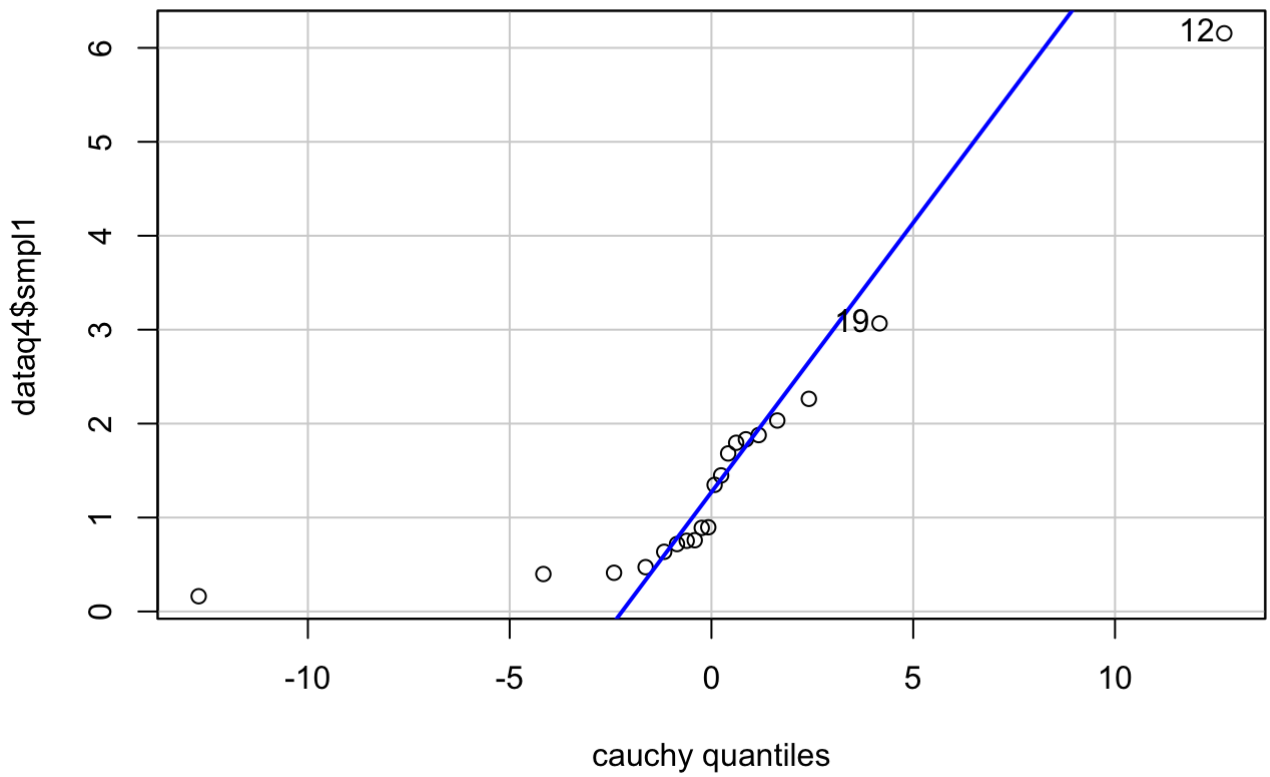
qqPlot(x, dist = 'rcauchy', shape = 2)

for plotting a quantile graph of vector x with respect to the Weibull distribution with shape parameter 2. The default distribution for qqPlot is the normal distribution. You can find more details in the help for qqPlot. By default, this function draws confidence bands which I find in many cases of little use, and in some cases misleading. If you don't want them in your graph, add envelope = FALSE in your call. Explain clearly the reasons for your choices.

```
library(car)
```

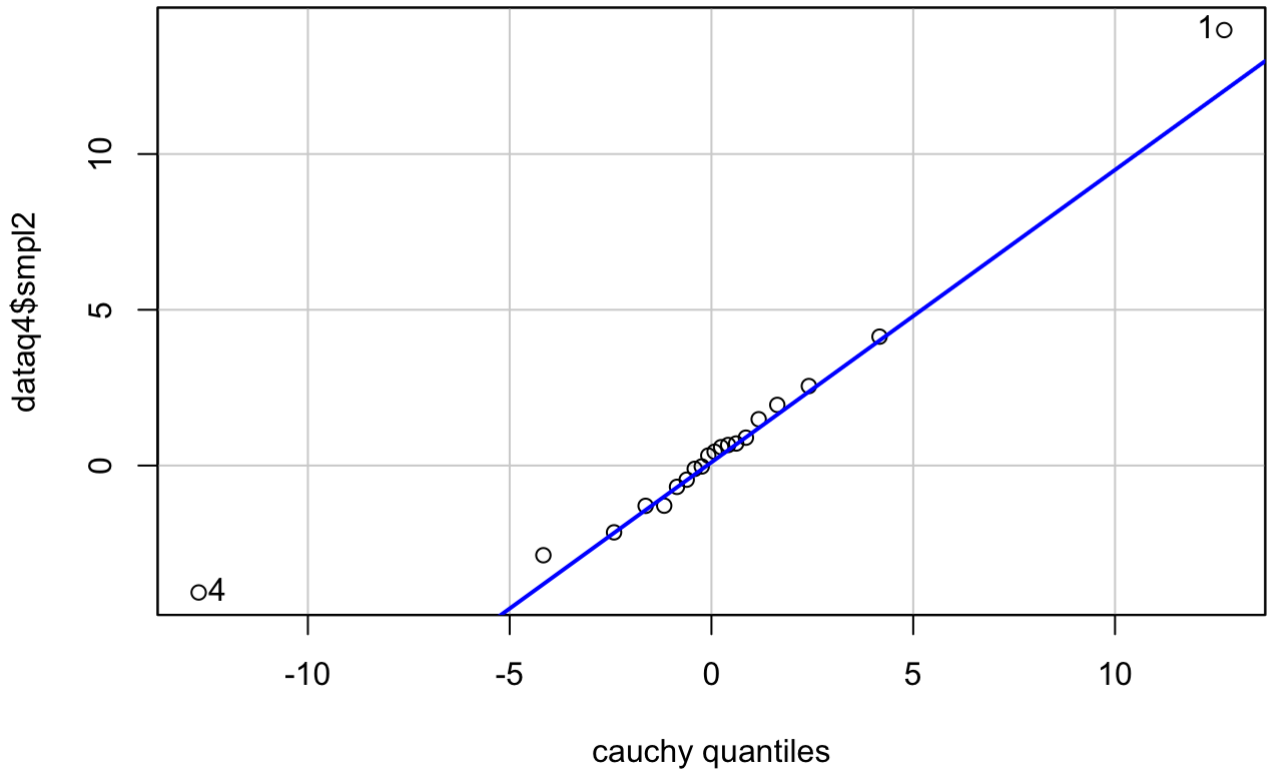
```
## Loading required package: carData
```

```
dataq4 = read.csv("data_q4.csv")
qqPlot(dataq4$smpl1, dist = 'cauchy',envelope = FALSE)
```



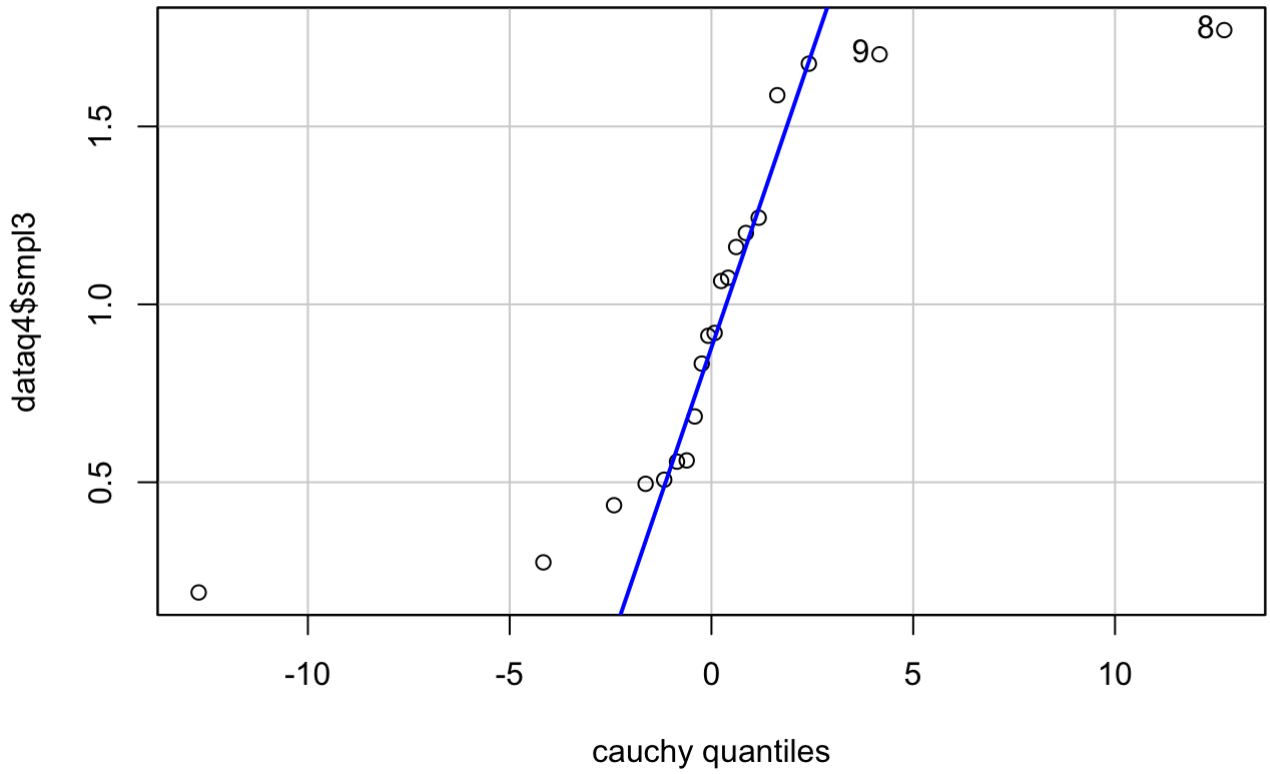

```
## [1] 12 19
```

```
qqPlot(dataq4$smpl2, dist = 'cauchy',envelope = FALSE)
```



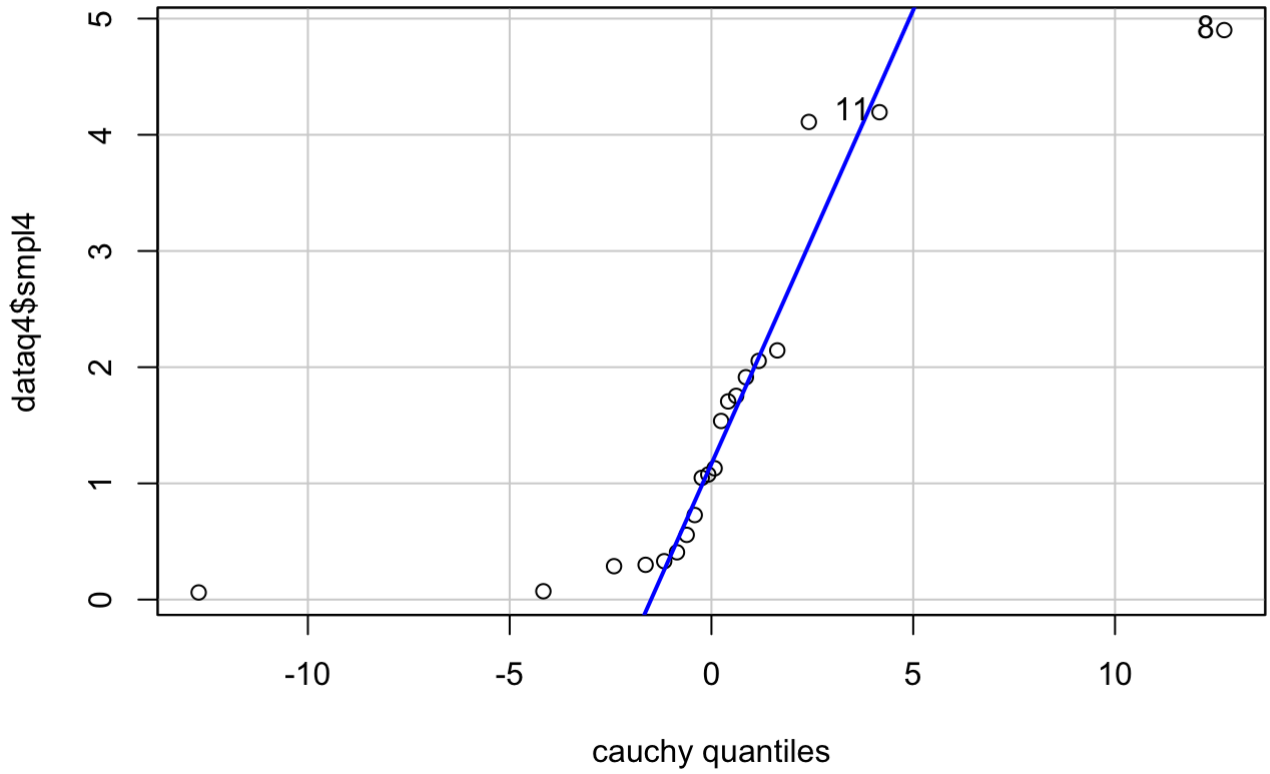

```
## [1] 1 4
```

```
qqPlot(dataq4$smpl3, dist = 'cauchy',envelope = FALSE)
```




```
## [1] 8 9
```

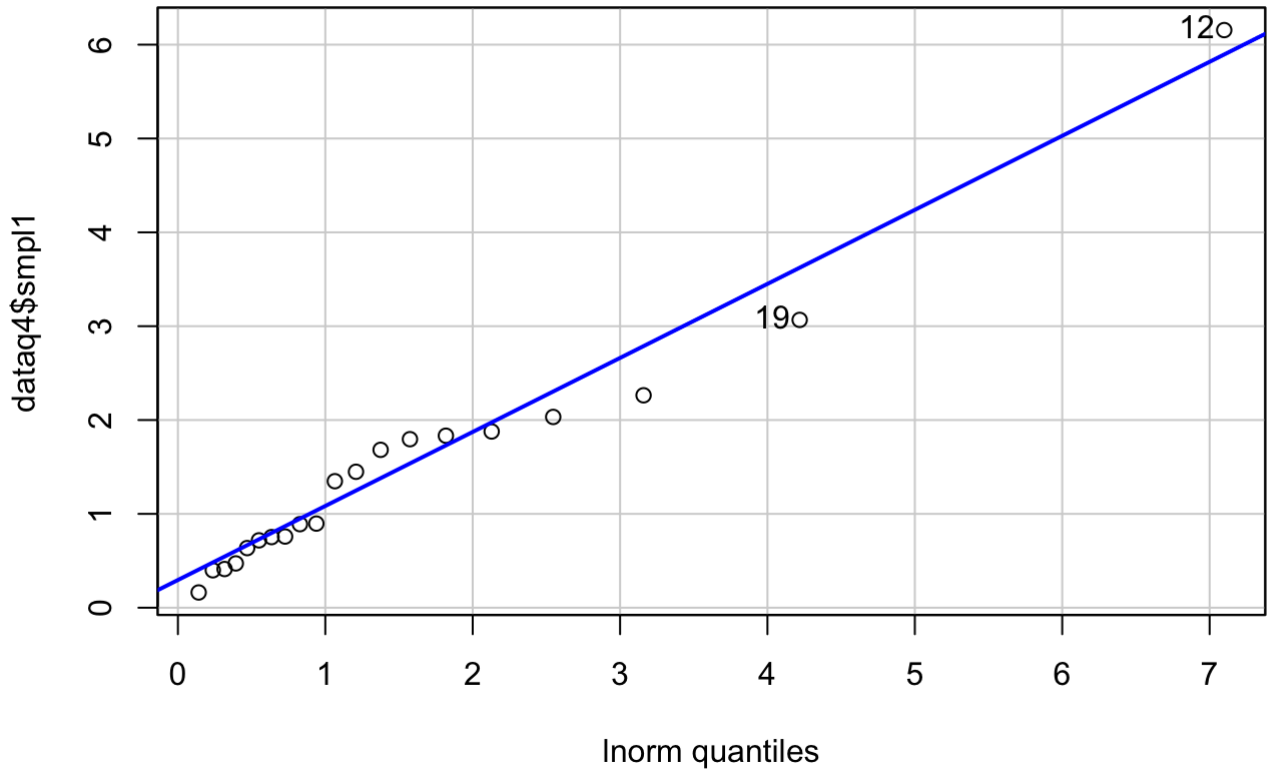
```
qqPlot(dataq4$smpl4, dist = 'cauchy',envelope = FALSE)
```




```
## [1]  8 11
```

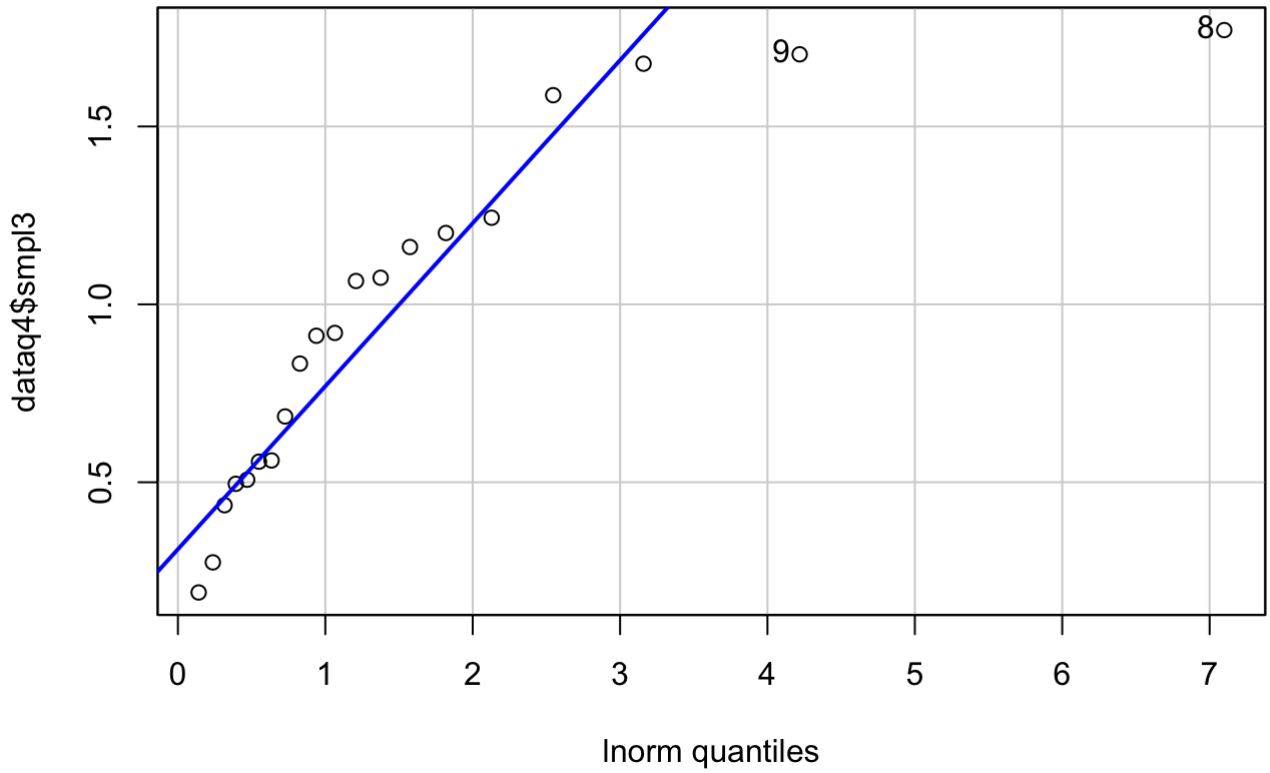
From above we can see Standard Cauchy fit sample2 best, the points appear to be on a straight line. Only 3 points not in line.

```
qqPlot(dataq4$smpl1, dist = 'lnorm',envelope = FALSE)
```



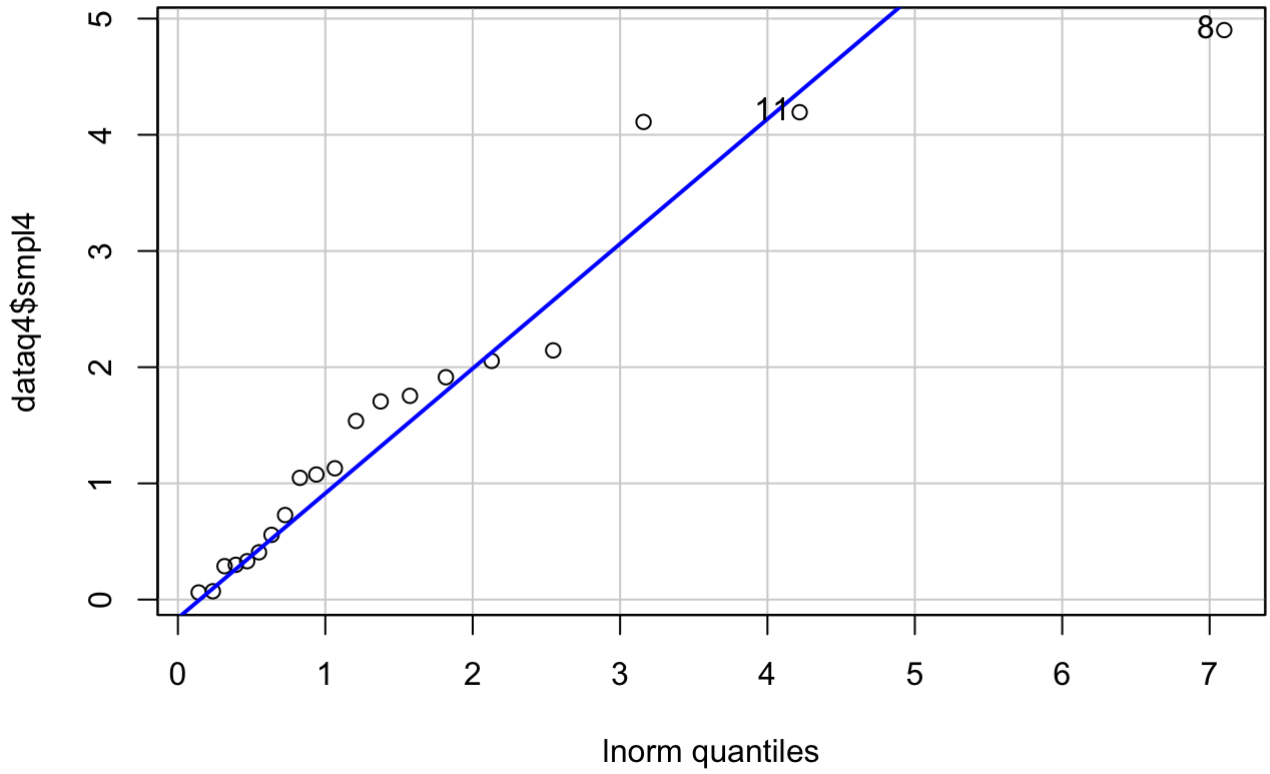

```
## [1] 12 19
```

```
qqPlot(dataq4$smpl3, dist = 'lnorm',envelope = FALSE)
```




```
## [1] 8 9
```

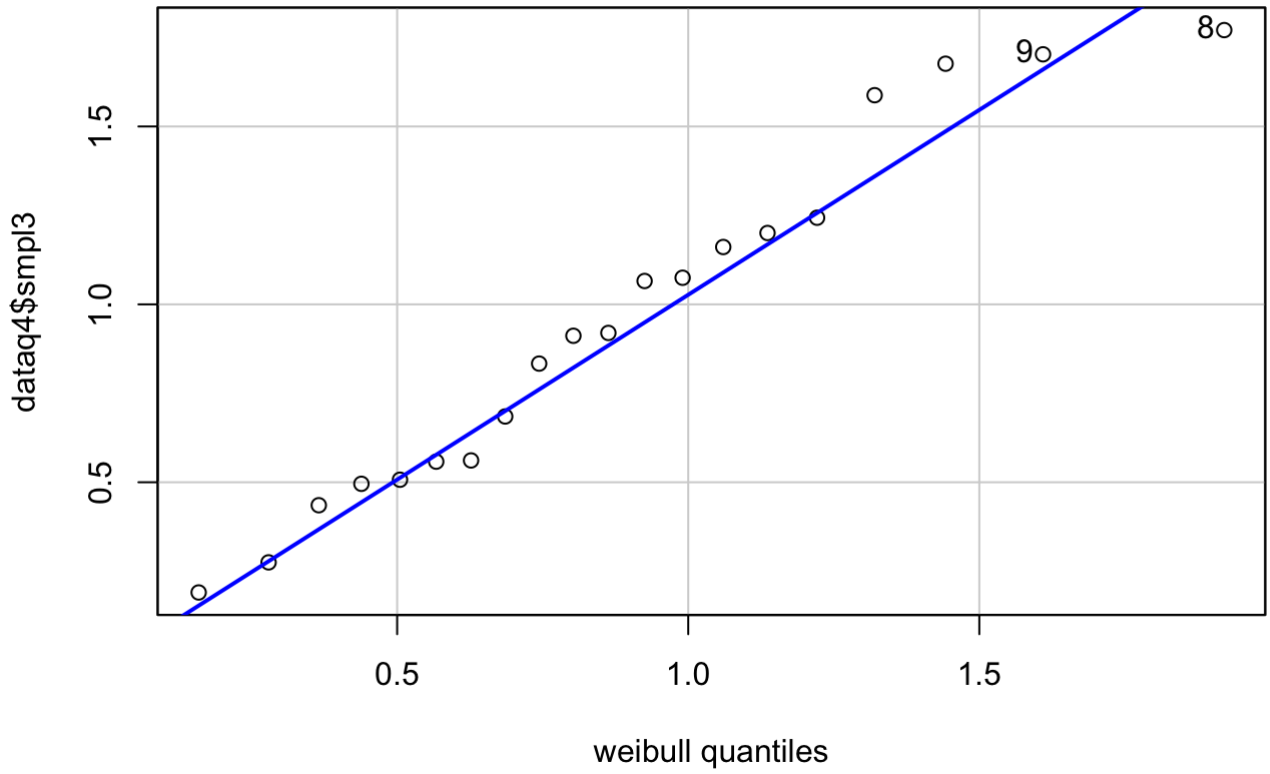
```
qqPlot(dataq4$smpl4, dist = 'lnorm',envelope = FALSE)
```




```
## [1]  8 11
```

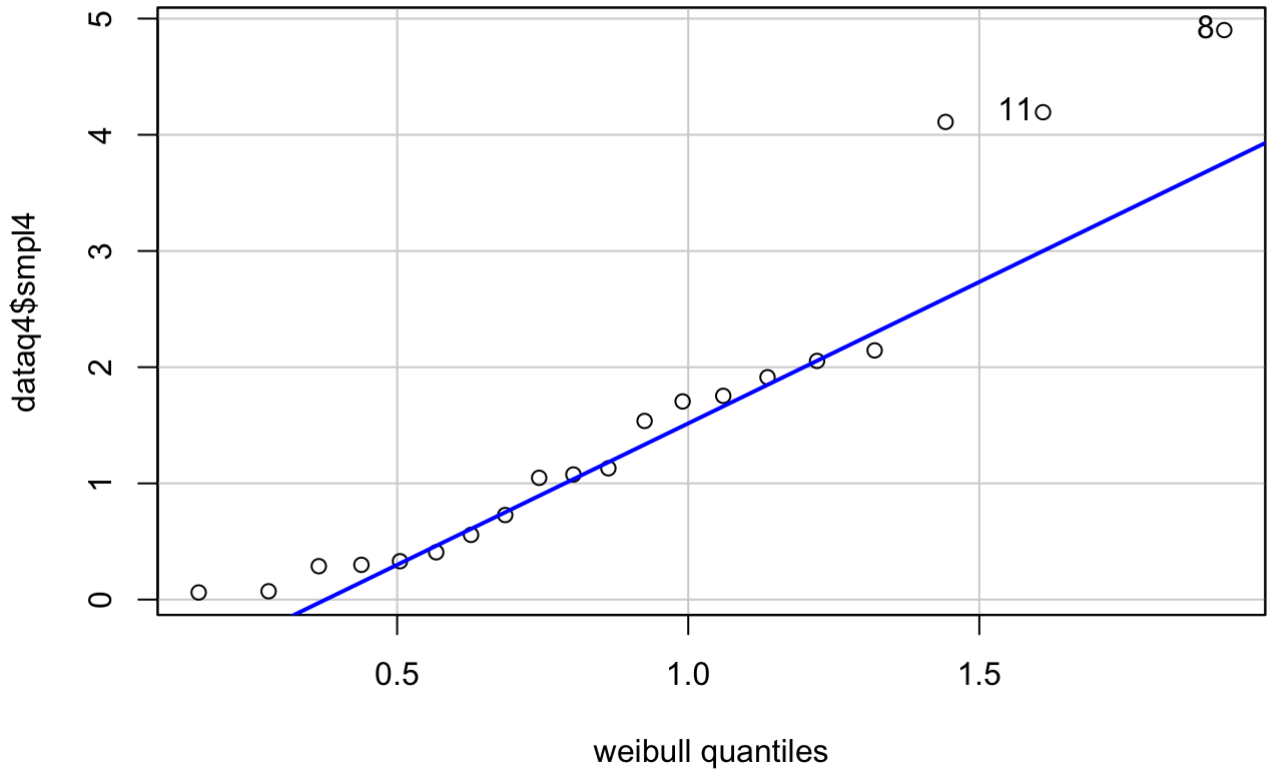
From above we can see Lognormal with standard parameters fit sample1 best, the points appear to be on a straight line.

```
qqPlot(dataq4$smpl3, dist = 'weibull',shape=2, envelope = FALSE)
```




```
## [1] 8 9
```

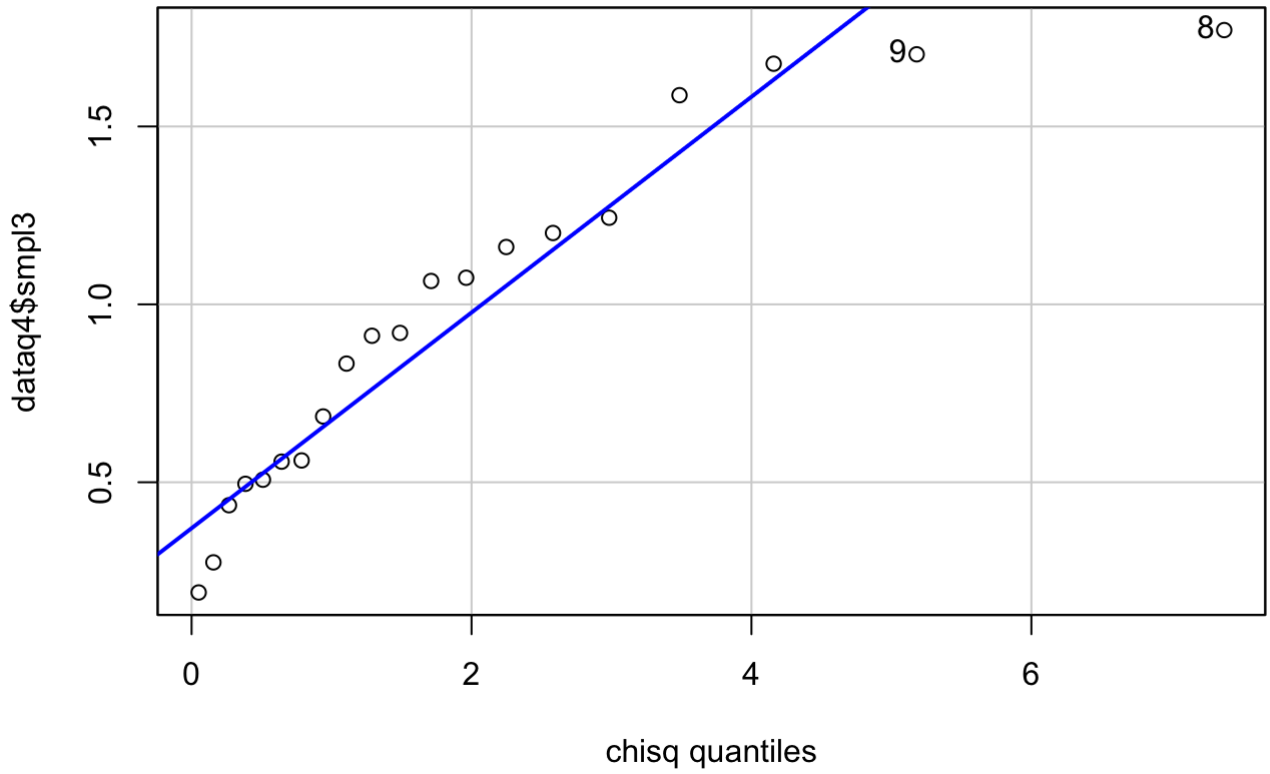
```
qqPlot(dataq4$smpl4, dist = 'weibull',shape=2, envelope = FALSE)
```




```
## [1]  8 11
```

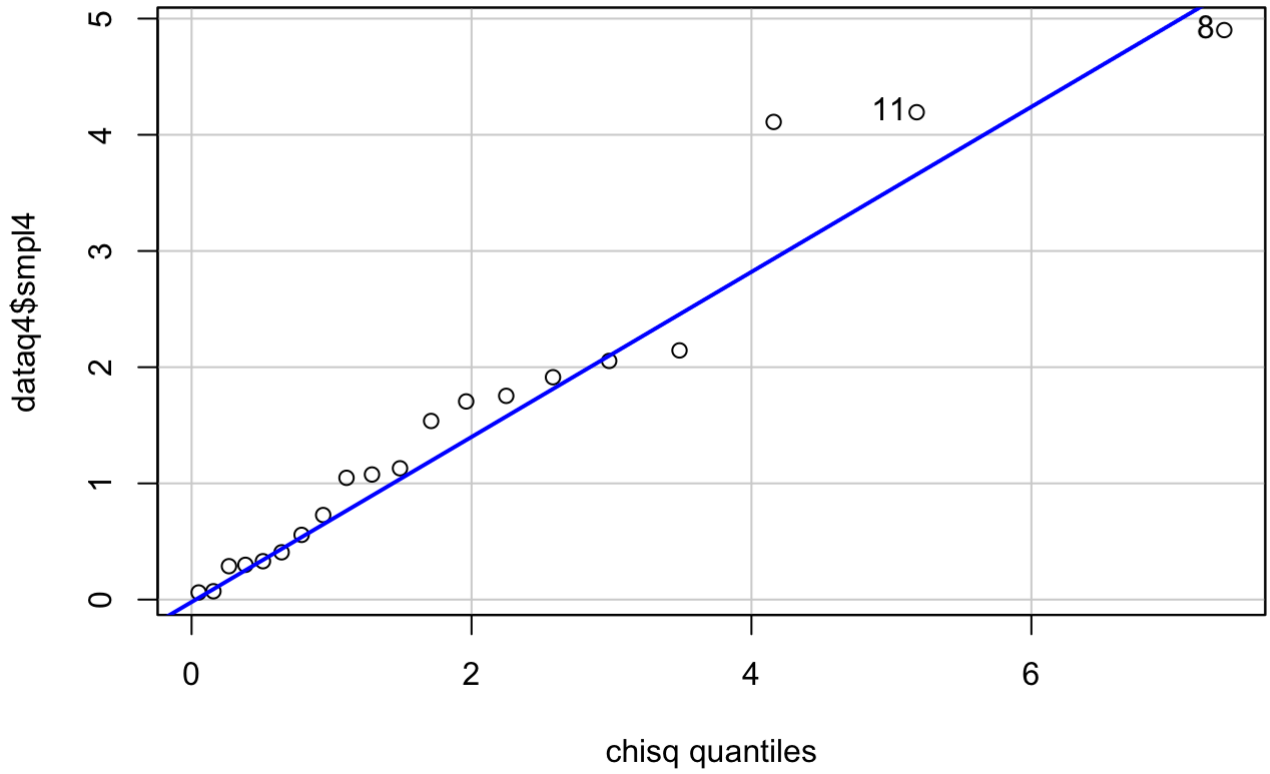
From above we can see Weibull with shape parameter 2 fit sample3 best, the points appear to be on a straight line.

```
qqPlot(dataq4$smpl3, dist = 'chisq',df=2, envelope = FALSE)
```




```
## [1] 8 9
```

```
qqPlot(dataq4$smpl4, dist = 'chisq',df=2, envelope = FALSE)
```




```
## [1]  8 11
```

From above we can see Chi-square with 2 degrees of freedom, χ2, (rchisq(20,2)) fit sample4 best, the points appear to be on a straight line.