STAT 210
Applied Statistics and Data Analysis:
Homework 3

Due on Oct. 5/2025

Question 1

For this question use the data set `penguins`, which is available in the `palmerpenguins` library. This data set was introduced in the second problem list and has four physical measurements for three species of penguins studied in Antartica. The species are `Adelie`, `Chinstrap` and `Gentoo`, and there are a total of 344 subjects, some with missing values. The penguins were observed in three different islands, `Biscoe`, `Dream` and `Togersen`. You can get more information looking at the help for this data set.

a) Find out

- how many subjects belong to each species,
- how many were observed in each island,
- how many missing values are there.

b) On a single plotting window, create boxplots of body mass grouped by both sex **and** species. This should result in six boxplots, one for each combination of sex and species. You can use the `boxplot` function, which allows you to specify a formula to define the variables for the plot. In the formula, use either `sex:species` or `sex + species` on the right-hand side to indicate grouping. Color the boxes according to species to distinguish them visually. Then, comment on what you observe in the resulting plot. Repeat the same process for bill depth, and again provide comments based on your observations.

c) Create a scatterplot matrix with plots of the four numerical variables in penguins. Color the plots according to `sex`. Comment on what you observe.

d) Reproduce the plot in Figure 1 below. The colors used for the dots are `dodgerblue1` and `darkblue` for Adelie, `green3` and `darkgreen` for Chinstrap, and `tomato1` and `tomato4` for Gentoo.

e) Finally, we want to assess whether the body weight measurements (`body_mass_g`) can reasonably be assumed to follow a normal distribution. To do this, use quantile-quantile (Q-Q) plots. Divide the plotting window into four panels, and create normal Q-Q plots for each of the three species (Adelie, Chinstrap, and Gentoo), and for the entire dataset Use the species name as the title for each plot. Also, include a reference line in each plot to help assess deviations from normality. After generating the plots, comment on your observations regarding the normality of body weight distributions across species and for the dataset as a whole.

Solution

Start by loading the library and attaching the data set

```
library(palmerpenguins)
data("penguins")
attach(penguins)
```

(a) We can use the function `tapply` for this

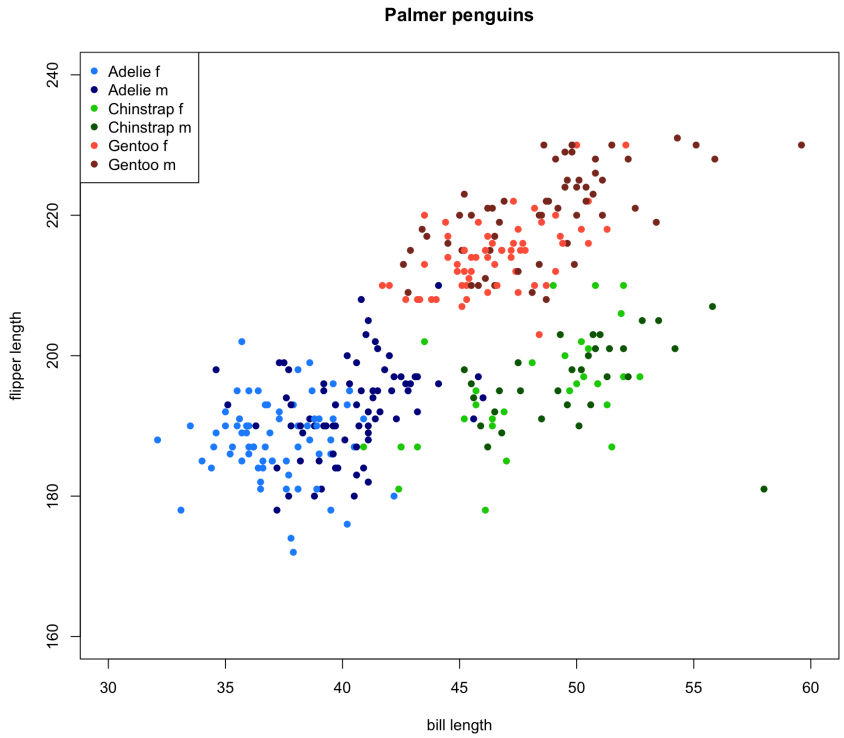


Figure 1: Figure for question 1(d)

```
tapply(sex, species, length)
```

```
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

Observe that in the previous command we could have used any variable instead of `sex` as the first component, because we are only interested in the length of the output, not in the content. For instance

```
tapply(island, species, length)
```

```
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

gives the same result, and even

```
tapply(species, species, length)
```

```
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

works. Similarly, for the islands

```
tapply(island, island, length)
```

```
##    Biscoe     Dream Torgersen
##       168       124        52
```

Finally, we use `apply` to find the missing values:

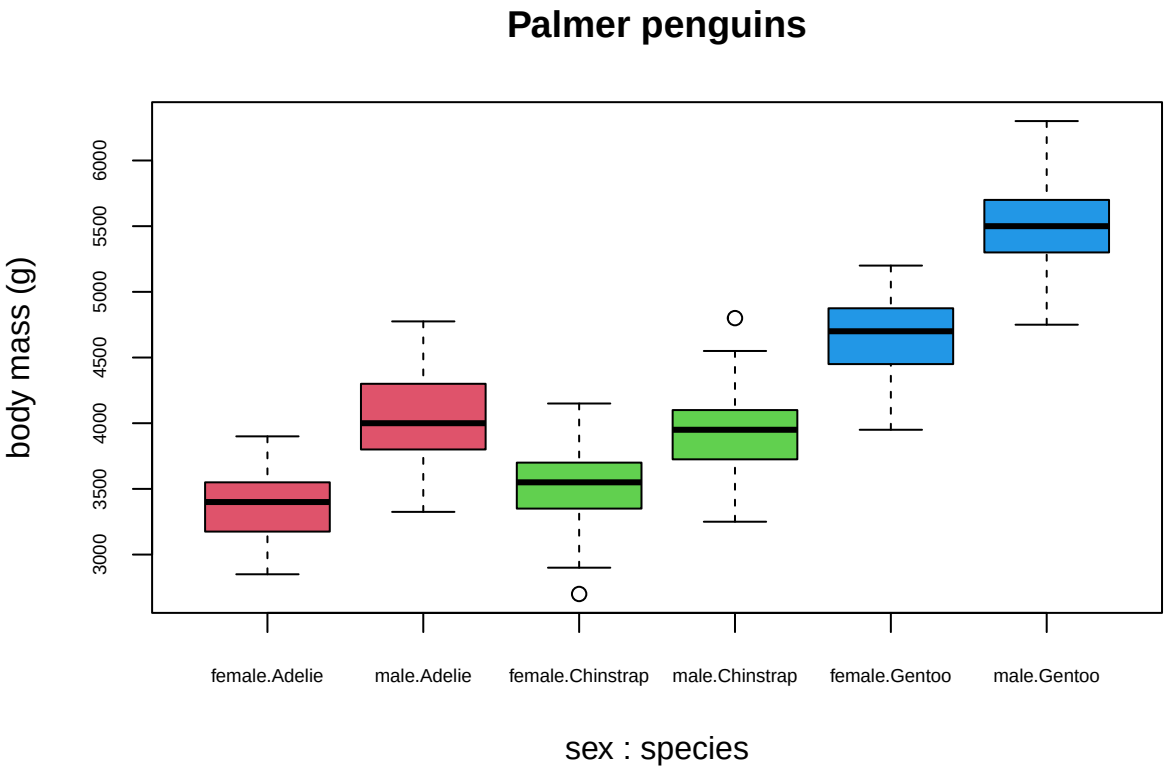
```r
apply(is.na(penguins), 2, sum)
```

```
##           species            island    bill_length_mm     bill_depth_mm
##                 0                 0                 2                 2
## flipper_length_mm       body_mass_g               sex              year
##                 2                 2                11                 0
```

There are two values missing for each of the four numerical variables, and eleven missing for `sex`.

(b) To get the boxplots we use the followiing commands:

```r
boxplot(body_mass_g ~ sex + species, col = rep(2:4, each = 2), cex.axis = 0.6,
        main = 'Palmer penguins', ylab = 'body mass (g)')
```

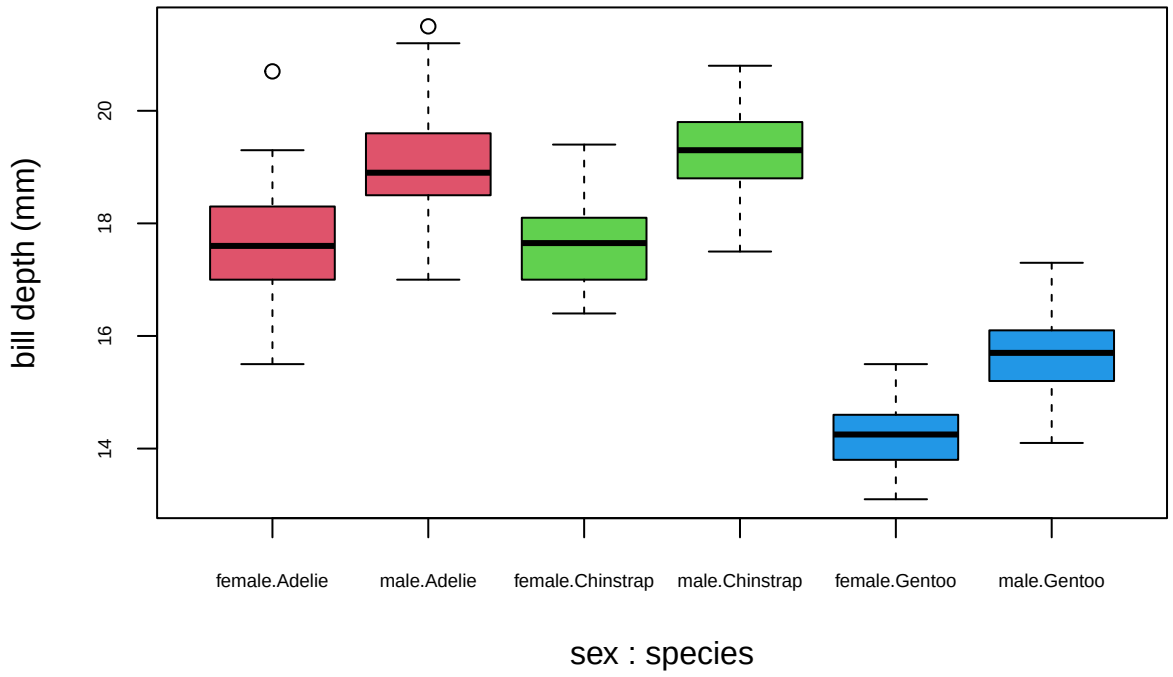


Comments:

- For the three species, males weight more than females.
- The interquartile range (IQR) for all the boxplots are similar, indicating that the dispersion observed for each combination of sex and species is similar.
- Species Adelie and Chinstrap have similar values for both genders, while Gentoo penguins are bigger.

For bill depth:

```r
boxplot(bill_depth_mm ~ sex + species, col = rep(2:4, each = 2), cex.axis = 0.6,
        main = 'Palmer penguins', ylab = 'bill depth (mm)')
```

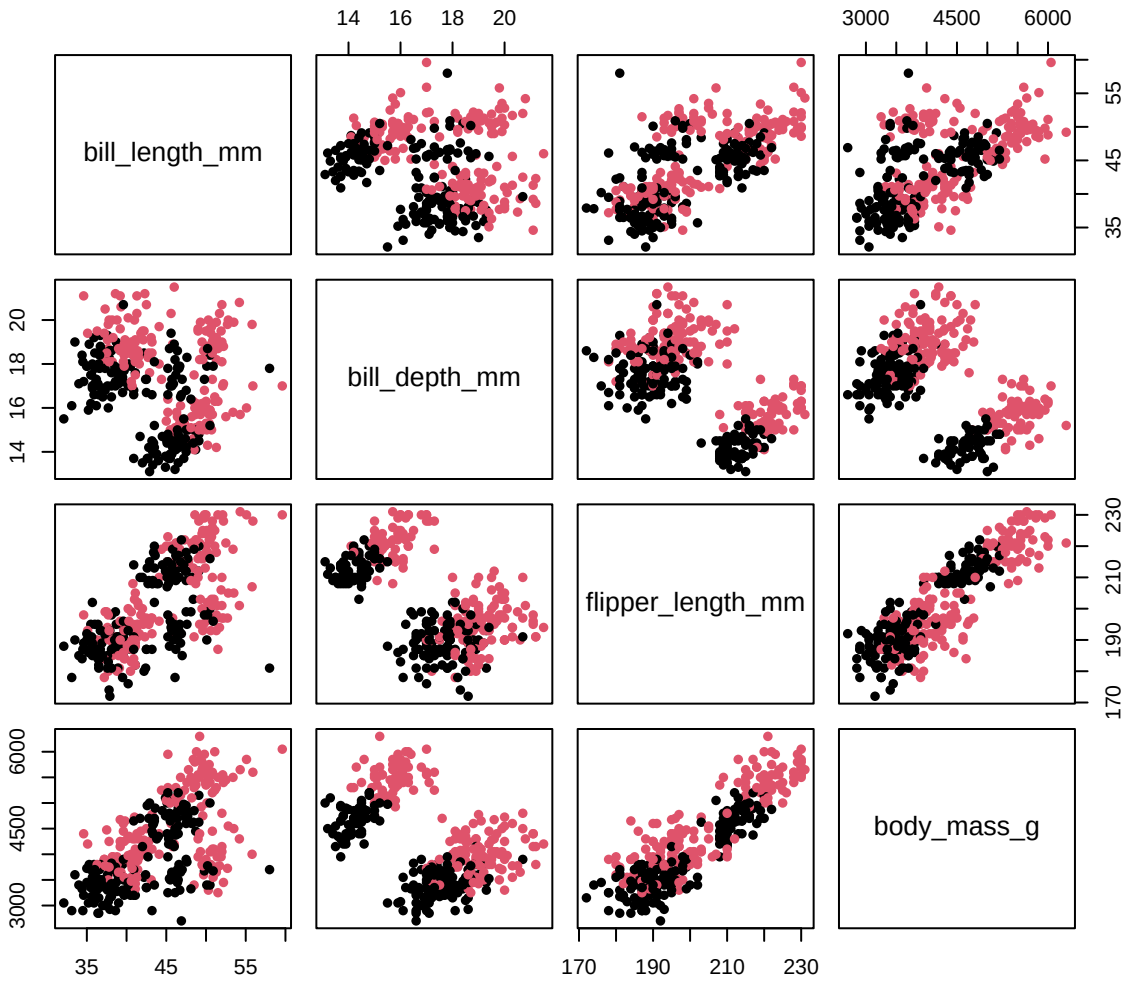


Comments:

- For the three species, bill depth for males is higher than for females.
- The interquartile range (IQR) for all the boxplots are similar, indicating that the dispersion observed for each combination of sex and species is similar.
- Species Adelie and Chinstrap have similar values for both genders, while Gentoo penguins are smaller.

(c) We create the scatterplot matrix with `plot`:

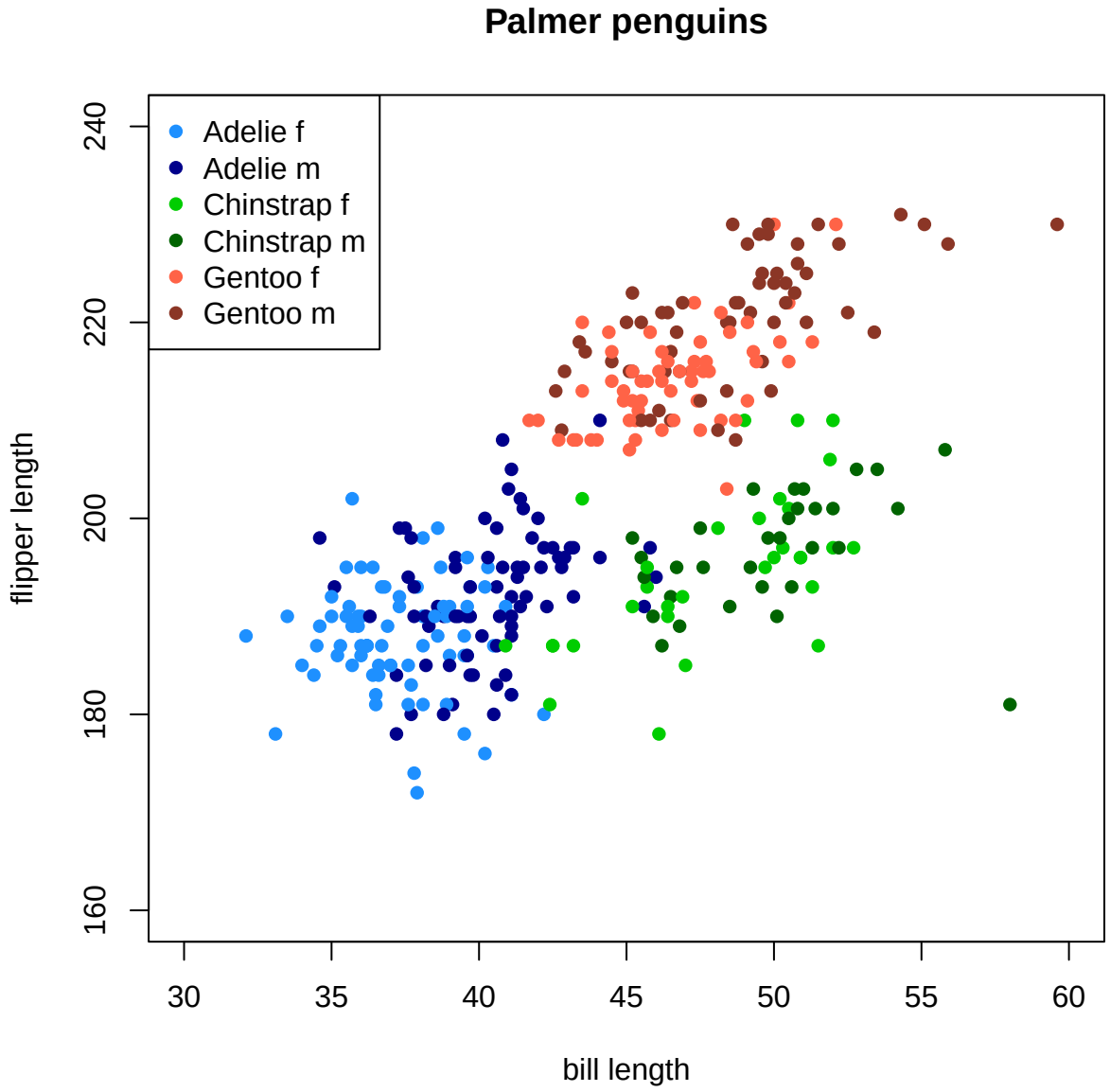
```
plot(penguins[3:6],pch = 16, col = sex, main = 'Palmer penguins')
```



In this plot, females are black and males are red. We see that, in each panel, red dots tend to be above black dots, indicating generally larger values for males. The plots in the first row show three clusters, which we know from the exercise in problem list 2 correspond to the different species. For the other three plots –those involving bill depth, flipper length and body mass– there seem to be only two clusters. In all cases, there is indication of a linear relation between the variables that seems to be valid for both sexes, but they cover different ranges of value.

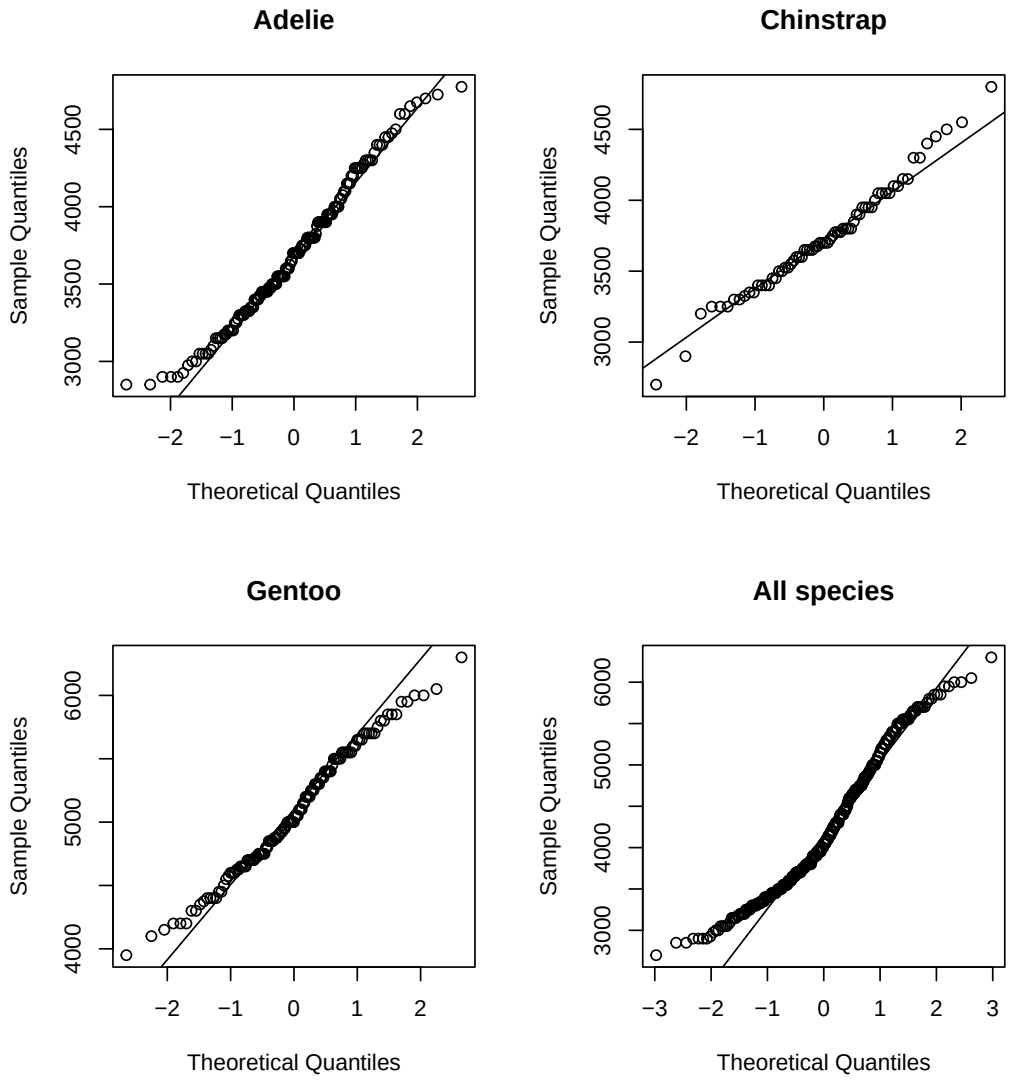
(d) The commands to reproduce this plot are

```
plot(bill_length_mm[species == 'Adelie'], flipper_length_mm[species == 'Adelie'],
     col = c('dodgerblue1','darkblue')[sex], pch = 16, xlim = c(30,60),
     ylim = c(160,240),  ylab = 'flipper length', xlab = 'bill length',
     main = 'Palmer penguins')
points(bill_length_mm[species == 'Chinstrap'], flipper_length_mm[species == 'Chinstrap'],
     col = c('green3','darkgreen')[sex], pch = 16)
points(bill_length_mm[species == 'Gentoo'], flipper_length_mm[species == 'Gentoo'],
     col = c('tomato1','tomato4')[sex], pch = 16)
legend('topleft',
       c('Adelie f','Adelie m', 'Chinstrap f', 'Chinstrap m', 'Gentoo f', 'Gentoo m'),
       col = c('dodgerblue1','darkblue','green3','darkgreen','tomato1','tomato4'),
       pch = 16)
```



(e) Quantile plots

```
par(mfrow = c(2,2))
## Adelie
qqnorm(body_mass_g[species=='Adelie'], main = 'Adelie')
qqline(body_mass_g[species=='Adelie'])
## Chinstrap
qqnorm(body_mass_g[species=='Chinstrap'], main = 'Chinstrap')
qqline(body_mass_g[species=='Chinstrap'])
## Gentoo
qqnorm(body_mass_g[species=='Gentoo'], main = 'Gentoo')
qqline(body_mass_g[species=='Gentoo'])
## All
qqnorm(body_mass_g, main ='All species')
qqline(body_mass_g)
```




```
par(mfrow = c(1,1))
detach(penguins)
```

We see that the fit for Adelie and Chinstrap is good, and for Gentoo the reference line is not helpful, but the points look to be reasonably aligned. However, for all the species together the fit is not good.

Question 2

The file `25Fhw3q2` has four simulated samples of size 30 obtained from the following distributions

- Standard Logistic, (`rlogis(30)`)
- Exponential with default parameter, (`rexp(30)`)
- Uniform in (0,10), (`runif(30, min - 0, max = 10)`)
- Cauchy with default parameter (`rcauchy(30)`)

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 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 syntax

```
qqPlot(x, dist = 'weibull', 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.

Solution:

Start by reading the data and looking at the structure of the data set.

```
dataQ2 <- read.table('25Fhw3q2')
str(dataQ2)
```

```
## 'data.frame':    30 obs. of  4 variables:
##  $ sp1: num  8.08 6.01 1.68 9.28 9.66 ...
##  $ sp2: num  -1.873 -0.107 0.776 1.221 -0.273 ...
##  $ sp3: num  1.479 1.029 1.265 0.586 0.272 ...
##  $ sp4: num  -1.202 -0.411 -0.933 -0.374 7.269 ...
```

We do quantile plots for all combinations of distributions and simulated samples. Since in this problem you know that each sample comes from a different distribution, you can either choose a distribution for each sample or a sample for each distribution. I will do this choosing a distribution for each sample.

Attach the data set and load the `car` library:

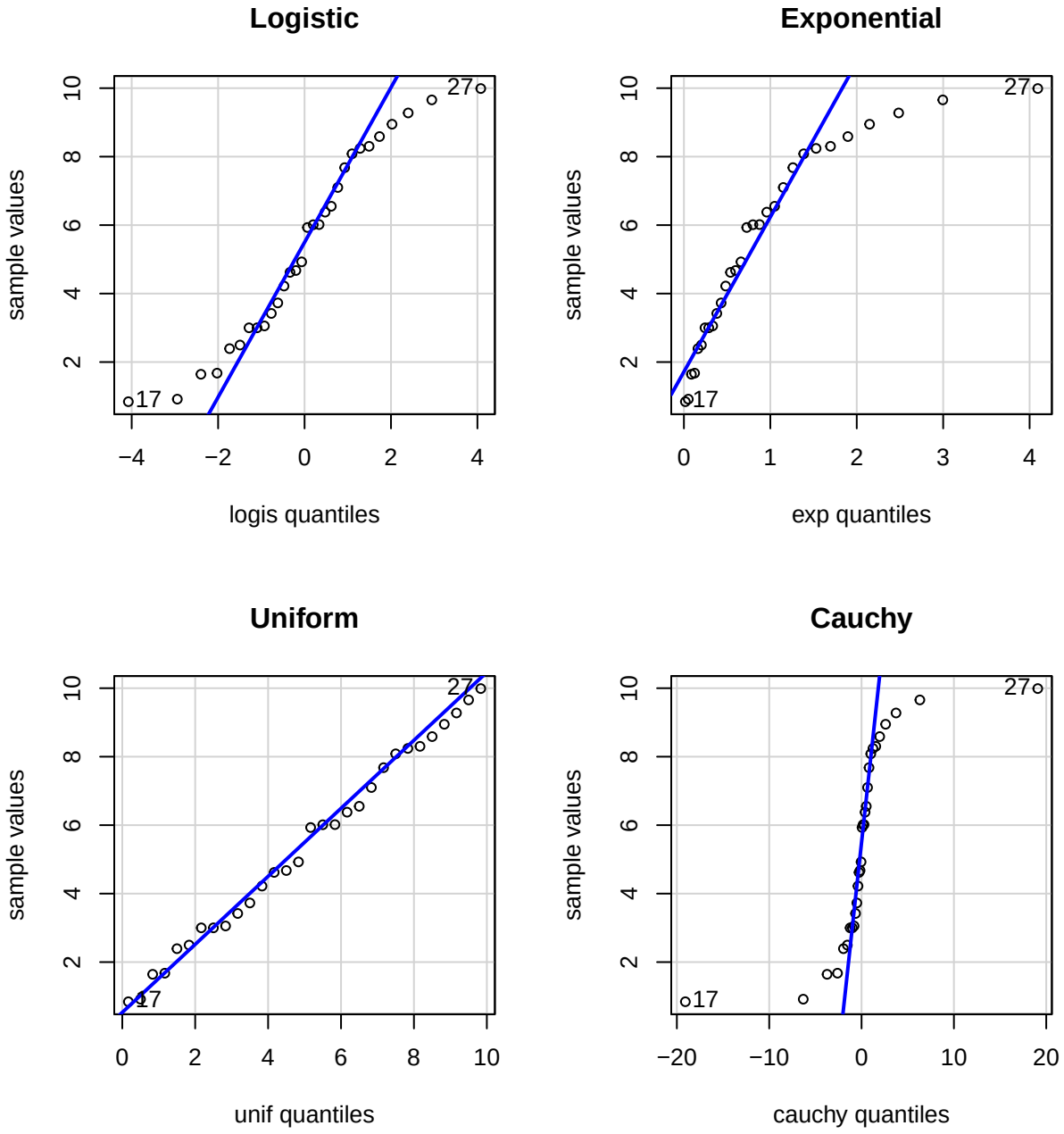
```
attach(dataQ2)
library(car)
```

Quantile plots for the first sample

The function `qqPlot` outputs a plot and the labels of identified points, which is not of interest in this problem. One way to eliminate this output is to assign the function's output to a variable, `xyz` in my case. `R` automatically prints the return value of an expression if it's not assigned to a variable. By assigning the output, you capture the numerical results without having them automatically printed. The plotting side-effect, however, will still occur.

```
## Sample 1
par(mfrow = c(2,2))
xyz <- qqPlot(sp1, dist = 'logis', envelope = F, main = 'Logistic',
              ylab = 'sample values')
```

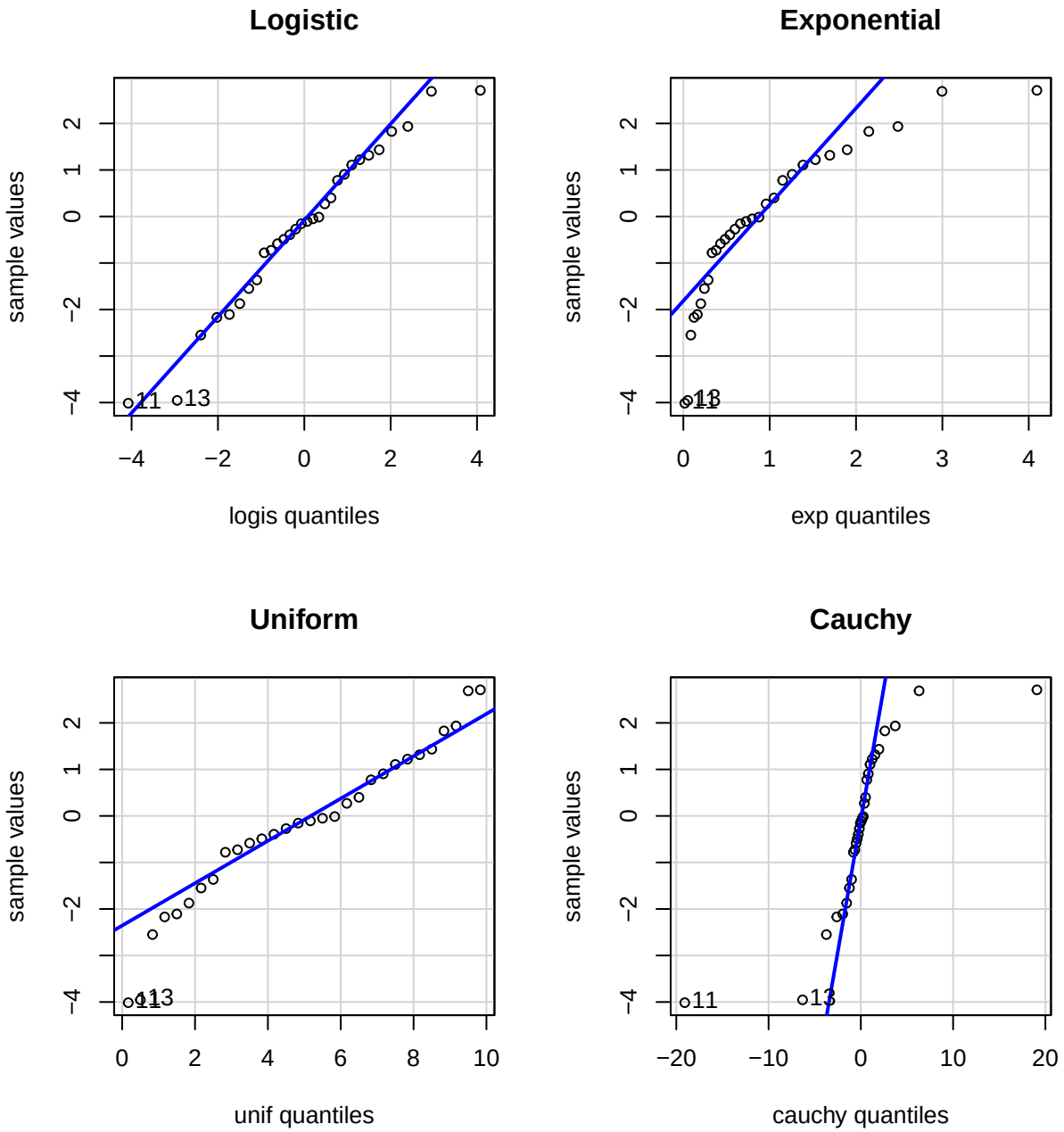
```
xyz <- qqPlot(sp1, dist = 'exp', envelope = F, main = 'Exponential',
              ylab = 'sample values')
xyz <- qqPlot(sp1, dist = 'unif', min = 0, max = 10, envelope = F,
              main = 'Uniform',
              ylab = 'sample values')
xyz <- qqPlot(sp1, dist = 'cauchy', envelope = F, main = 'Cauchy',
              ylab = 'sample values')
```



In this case the best fit corresponds to the uniform distribution. Observe also that, in this sample, all the values are positive (look at the y-axis in all the plots) and there are two distributions in the list that have only positive values: exponential, and uniform.

Quantile plots for the second sample

```
## Sample 2
par(mfrow = c(2,2))
xyz <- qqPlot(sp2, dist = 'logis', envelope = F, main = 'Logistic',
              ylab = 'sample values')
xyz <- qqPlot(sp2, dist = 'exp', envelope = F, main = 'Exponential',
              ylab = 'sample values')
xyz <- qqPlot(sp2, dist = 'unif', min = 0, max = 10, envelope = F,
              main = 'Uniform', ylab = 'sample values')
xyz <- qqPlot(sp2, dist = 'cauchy', envelope = F, main = 'Cauchy',
              ylab = 'sample values')
```

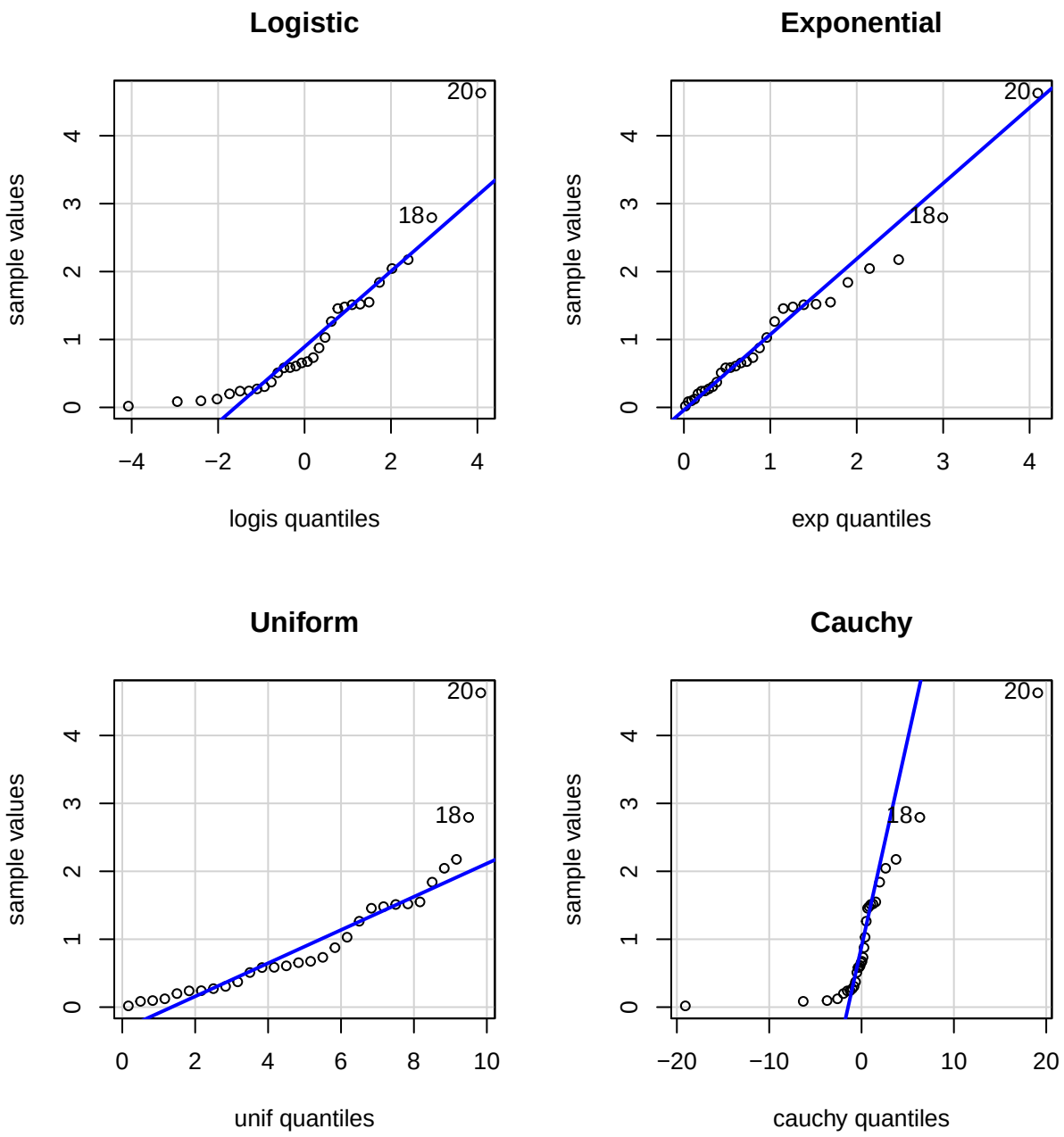


In this case the fit to the logistic distribution is very good so we identify this as the logistic sample. Observe

that the sample has both positive and negative values.

Quantile plots for the third sample

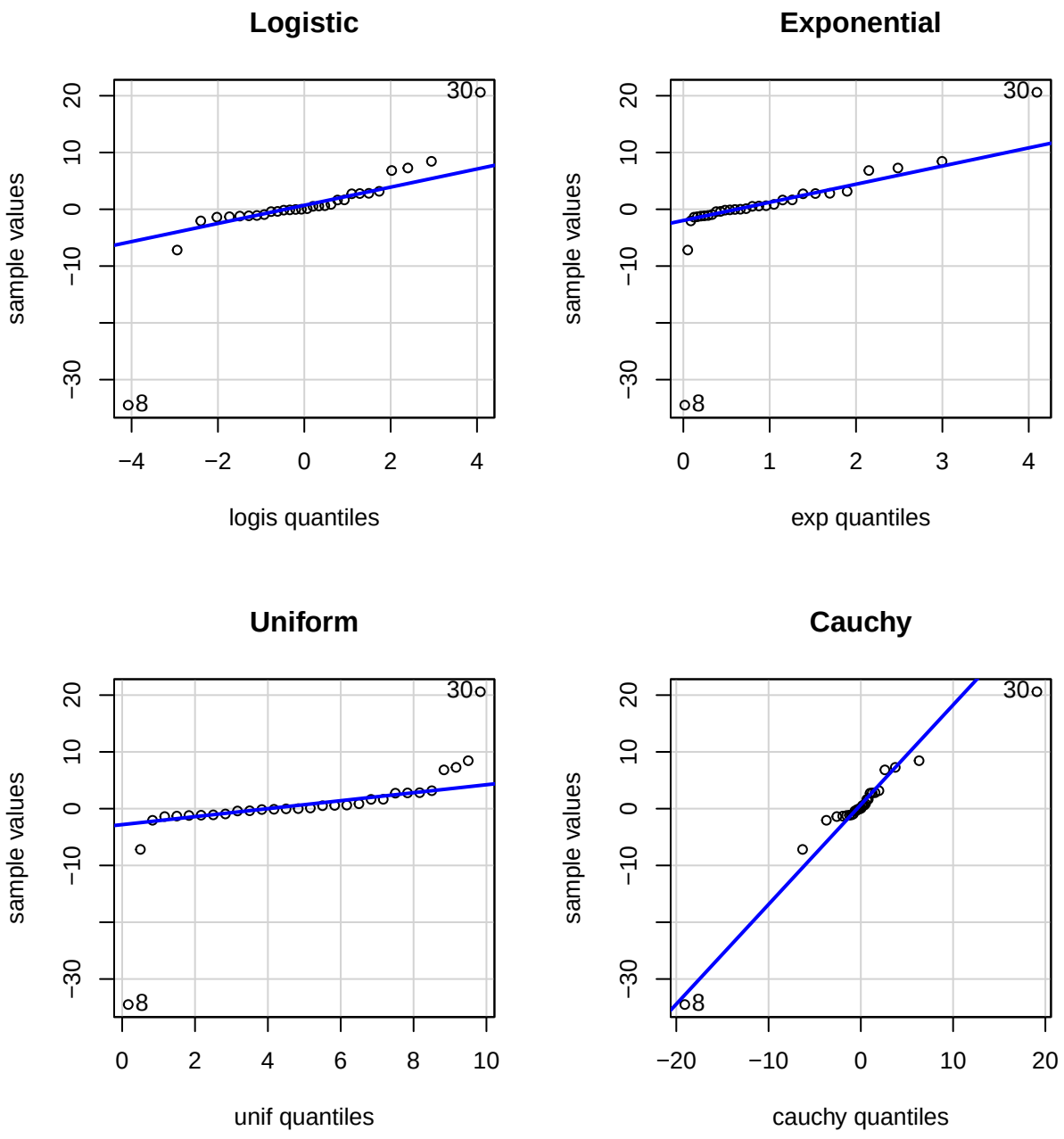
```
## Sample 3
par(mfrow = c(2,2))
xyz <- qqPlot(sp3, dist = 'logis', envelope = F, main = 'Logistic',
              ylab = 'sample values')
xyz <- qqPlot(sp3, dist = 'exp', envelope = F, main = 'Exponential',
              ylab = 'sample values')
xyz <- qqPlot(sp3, dist = 'unif', min = 0, max = 10, envelope = F,
              main = 'Uniform', ylab = 'sample values')
xyz <- qqPlot(sp3, dist = 'cauchy', envelope = F, main = 'Cauchy',
              ylab = 'sample values')
```



In this case the best fit corresponds to the exponential distribution

Quantile plots for the fourth sample

```
## Sample 4
par(mfrow = c(2,2))
xyz <- qqPlot(sp4, dist = 'logis', envelope = F, main = 'Logistic',
              ylab = 'sample values')
xyz <- qqPlot(sp4, dist = 'exp', envelope = F, main = 'Exponential',
              ylab = 'sample values')
xyz <- qqPlot(sp4, dist = 'unif', min = 0, max = 10, envelope = F,
              main = 'Uniform', ylab = 'sample values')
xyz <- qqPlot(sp4, dist = 'cauchy', envelope = F, main = 'Cauchy',
              ylab = 'sample values')
```




```
detach(dataQ2)
```

For this remaining sample the best fit corresponds to the Cauchy distribution which, fortunately, is the only remaining distribution.

Our classification is

Sample	Distribution
sp1	Uniform
sp2	Logistic
sp3	Exponential
sp4	Cauchy