Diabetes data: comparing distributions

13 marks (undergrads) plus potential *8 marks* bonus

21 marks (grads)

Comparing distributions

Download the `diabetes` data from the course website. In that file, there is a dataset on various measurements of 145 patients. Once you load this file into your R session (or equivalently, execute its contents there) there will be a data set called `diabetes`.

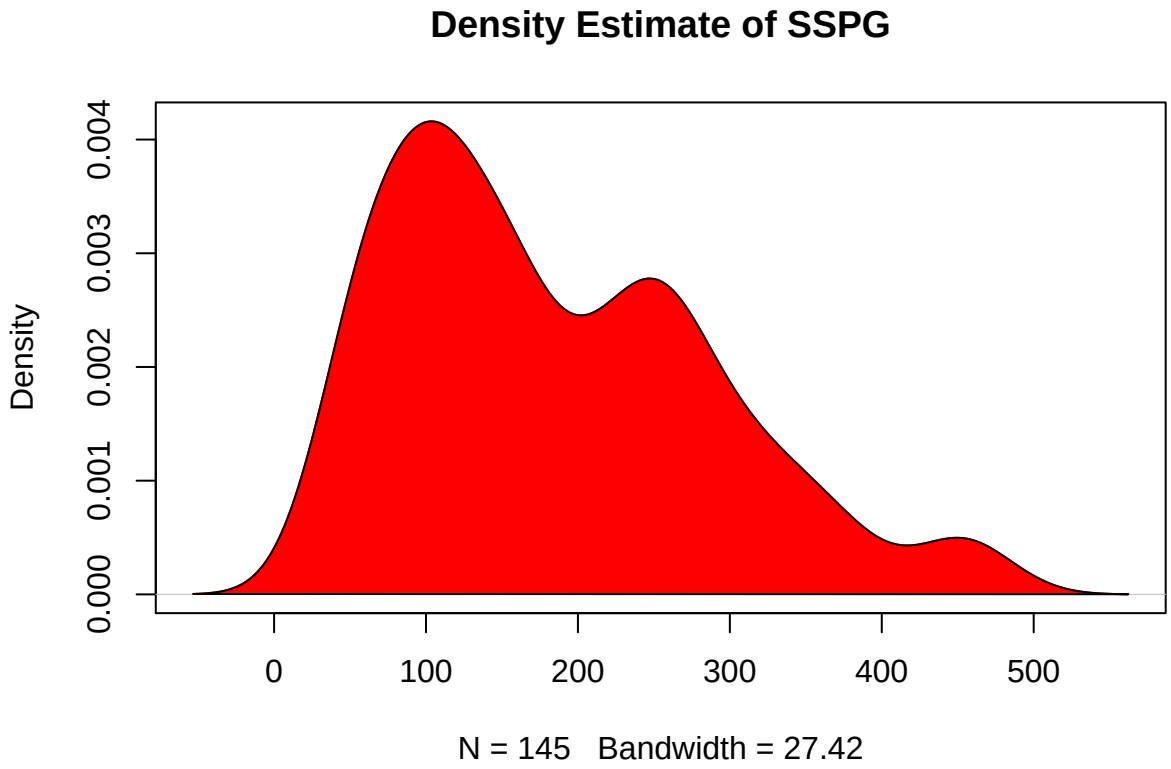
```
# For example, you could use the source command.
# Here the file is stored in the current directory
load("diabetes.Rda")
# Once loaded the data is available as the data frame 'diabetes'
head(diabetes)
```

```
##   PatientNumber RelativeWeight FastingPlasmaGlucose GlucoseArea InsulinArea
## 1             1           0.81                   80         356         124
## 2             2           0.95                   97         289         117
## 3             3           0.94                  105         319         143
## 4             4           1.04                   90         356         199
## 5             5           1.00                   90         323         240
## 6             6           0.76                   86         381         157
##   SSPG ClinClass
## 1   55         3
## 2   76         3
## 3  105         3
## 4  108         3
## 5  143         3
## 6  165         3
```

The variate `SSPG` stands for steady state plasma glucose which measures the patient's insulin resistance, a pathological condition where the body's cells fail to respond to the hormone insulin.

a. **(3 marks)** Produce a plot of a density estimate of `SSPG` and comment on what you see.

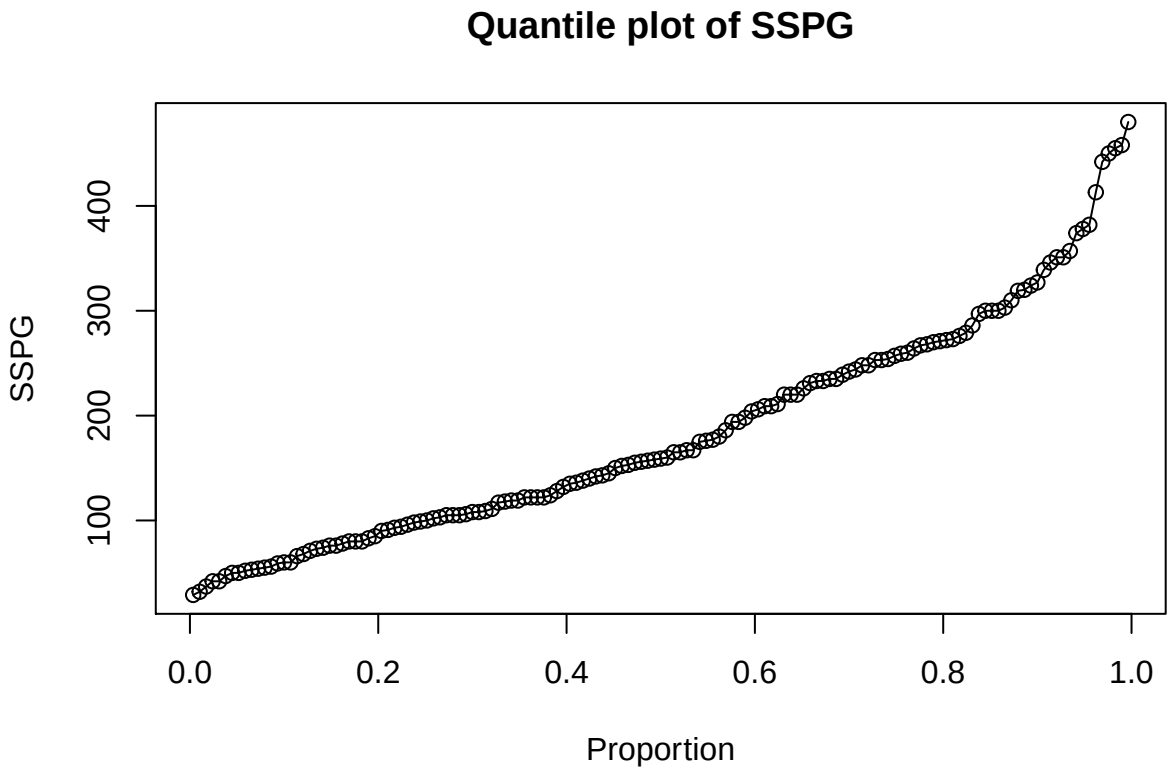
```
plot(density(diabetes$SSPG,bw='SJ'), col='red',
     main='Density Estimate of SSPG')
polygon(density(diabetes$SSPG,bw='SJ'), col='red')
```



Comment: The distribution is right skewed. There are three modes existing in the plot.

b. **(3 marks)** Construct a quantile plot of `SSPG` and comment on the shape of its distribution.

```
n <- length(diabetes$SSPG)
plot(ppoints(n), sort(diabetes$SSPG), type='o', xlab='Proportion',
     ylab='SSPG', main='Quantile plot of SSPG')
```

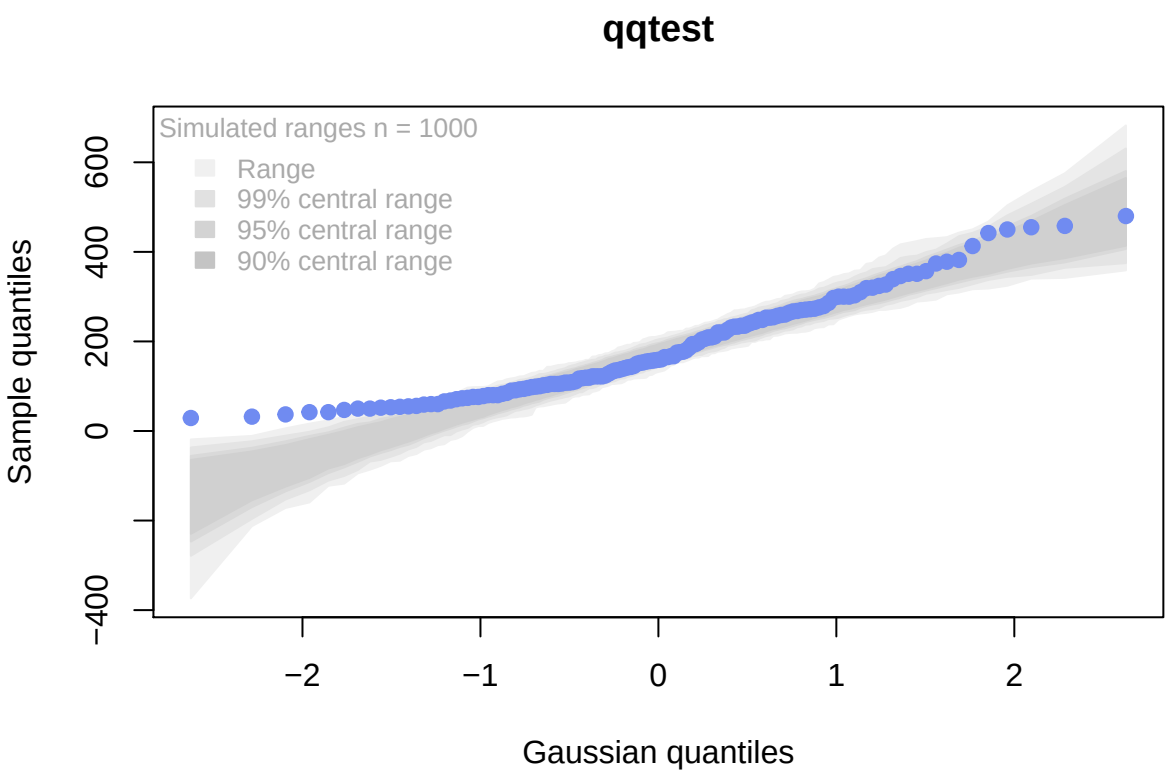


Comment: Since the top of the plot is steeper, the distribution has a relatively large right tail.

c. **(3 marks)** Use `qqtest` to construct a qqplot that compares `SSPG` to a standard normal distribution. Include envelopes in the plot. Comment on the distribution of `SSPG` and whether it might reasonably be regarded as a sample from some normal distribution. Explain your reasoning

Important: Before every `qqtest` execute `set.seed(3124159)` so that we are all seeing the same plots.

```
library(qqtest)
set.seed(3124159)
qqtest(diabetes$SSPG)
```



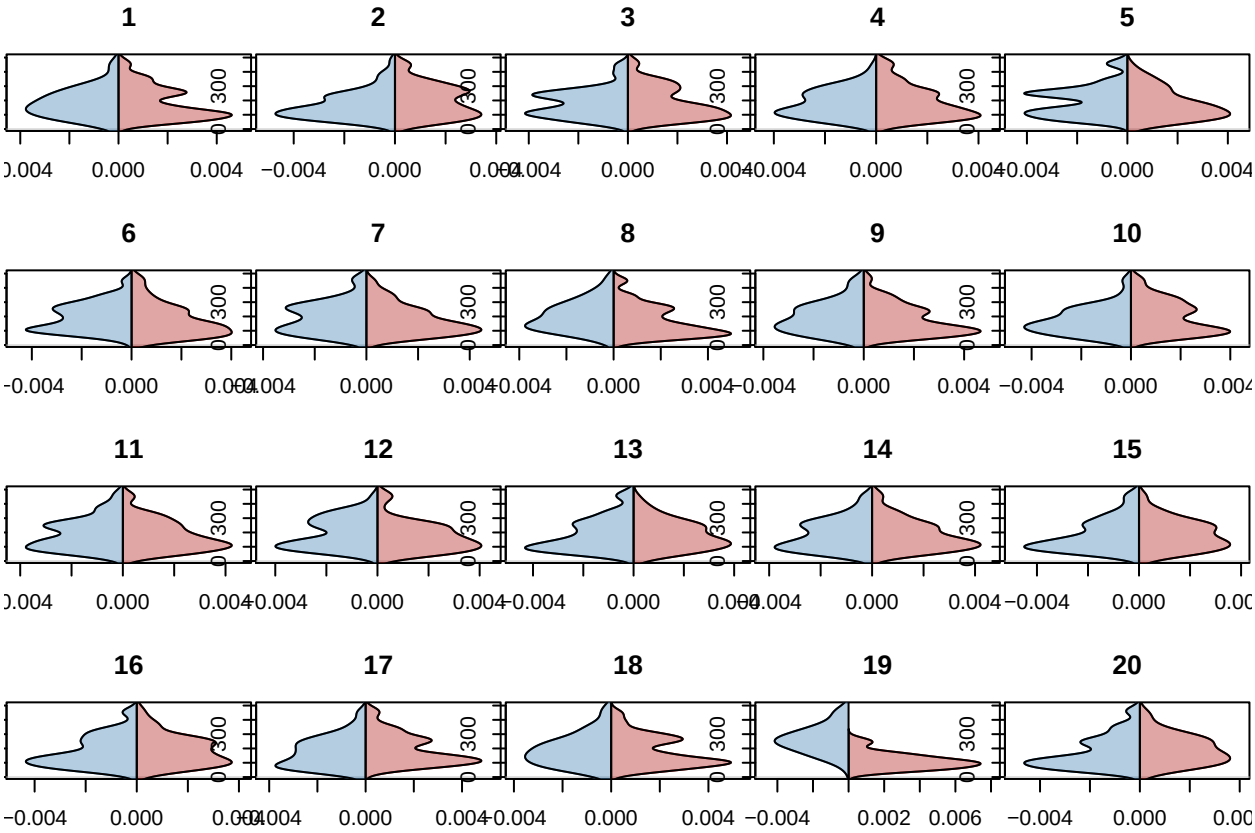
Comment: There are only a few points outside the envelope. All other points are well-placed in the envelope. We don't have a strong evidence against the null hypothesis. Therefore, we can conclude that the distribution of SSPG can be normal.

d. The last variate, `ClinClass`, represents the classification of each patient according to the 1979 medical criteria into one of three groups: 1 = "Overt Diabetic", 2 = "Chemical Diabetic", and 3 = "Normal".

i. **(4 marks)** Construct a back to back density line-up plot to assess whether the normal and diabetic (chemical and overt combined) `SSPG` values come from the same distribution. Use `set.seed(3124159)` and show your code. What conclusions do you draw?

```
data <- list(x = diabetes$SSPG[diabetes$ClinClass==3],
             y = diabetes$SSPG[diabetes$ClinClass!=3])
```

```
set.seed(3124159)
lineup(data,
       generateSubject = mixRandomly,
       showSubject = back2back,
       layout=c(4,5))
```




```
## $trueLoc
## [1] "log(7.93531459841716e+47, base=13) - 24"
```

Comment: After mixed randomly, we see that there are huge variations in shapes of the distributions. This gives an observation that is different from what we expected for the null hypothesis. Therefore, we have an evidence against the null hypothesis.

ii. **Grad students, bonus undergraduates (8 marks)** Consider the following code:

```
data <- list(x=x, y=y, z=z)
lineup(data,
       generateSuspect = mixRandomly,
       showSuspect = myQuantilePlot,
       layout=c(5,4))
```

The function `mixRandomly` will need to be rewritten to handle `data` being a list of three samples. Write the function `myQuantilePlot` so that it overlays the sample quantile functions of each of `x`, `y`, and `z` in the same display using different colours. Hand in your code for these two functions and illustrate the outcome (using `set.seed(314159)`) on `SSPG` for the three different clinical classes. Comment on your findings.

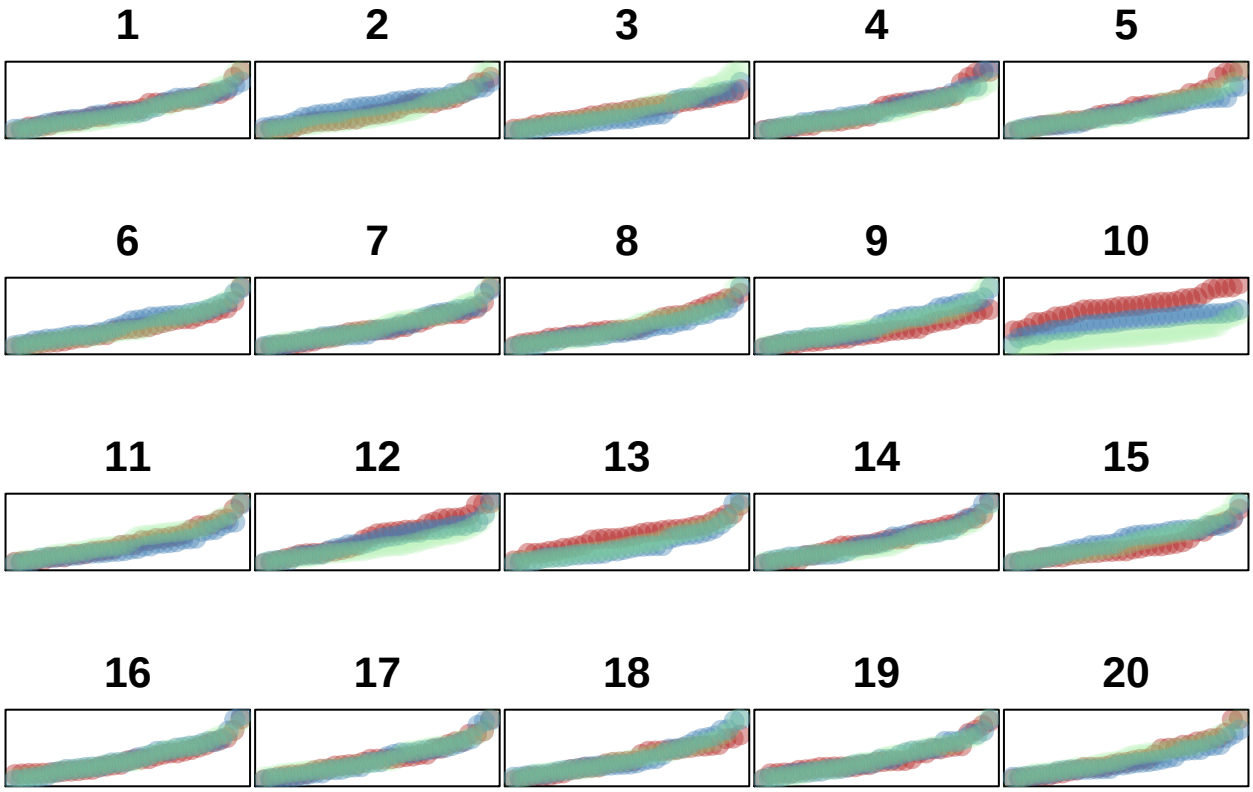
```
mixRandomly <- function(data) {
  x <- data$x
  y <- data$y
  z <- data$z
  m <- length(x)
  n <- length(y)
  l <- length(z)
  mix <- c(x, y, z)
```

```
  selectm <- sample(1:(m+n+l), m, replace = FALSE)
  new_x <- mix[selectm]
  remaining <- mix[-selectm]
  selectn <- sample(1:(n+l), n, replace = FALSE)
  new_y <- remaining[selectn]
  new_z <- remaining[-selectn]
  list(x=new_x, y=new_y, z=new_z)
}
```

```
myQuantilePlot <- function(data, subjectNo) {
  ylim <- extendrange(c(data$x, data$y, data$z))
  n_x <- length(data$x)
  n_y <- length(data$y)
  n_z <- length(data$z)
  p_x <- ppoints(n_x)
  p_y <- ppoints(n_y)
  p_z <- ppoints(n_z)
  plot(p_x, sort(data$x), type="b", col=adjustcolor("firebrick", 0.4),
       pch=19, cex=2, ylim = ylim, lwd=1,
       main=paste( subjectNo),
       cex.main = 2,
       ylab="", xlab="", xaxt="n", yaxt="n")
  points(p_y, sort(data$y), type="b", col=adjustcolor("steelblue",
                                                      0.4),
         pch=19, cex=2, lwd=1)
  points(p_z, sort(data$z), type="b", col=adjustcolor("lightgreen",
                                                      0.1),
         pch=19, cex=2, lwd=1)
}
```

```
set.seed(314159)
data <- list(x=diabetes$SSPG[diabetes$ClinClass==1],
             y=diabetes$SSPG[diabetes$ClinClass==2],
             z=diabetes$SSPG[diabetes$ClinClass==3])

lineup(data,
       generateSubject = mixRandomly,
       showSubject = myQuantilePlot,
       layout=c(4,5))
```




```
## $trueLoc
## [1] "log(1.32922799578492e+36, base=16) - 20"
```

Comment: From plot 9, 10 and 13, we see that they are not overlap too much. Therefore, we can say that we have an evidence against the null hypothesis.