Effect of increasing sample size

20 marks

In R there are functions that allow calculation of the density (or probability mass) function $f(x)$, the cumulative distribution function $F(x)$, and the quantile function $Q_X(p)$; there are also functions that will generate pseudo-random observations for each distribution. For example for a $N(0,1)$ distribution, the functions are `dnorm(...)`, `pnorm(...)`, `qnorm(...)`, and `rnorm(...)` respectively. To see all of the distributions for which these functions are built-in see `help("distributions")`.

In this question, you will be generating pseudo-random numbers from three different distributions, and four different sample sizes n:

- Gaussian or $N(0,1)$, Student (3) or t_3, and the χ^2_3 distribution.
- $n \in \{50, 100, 1000, 10000\}$

The goal is to compare different visualizations across distributions and to assess the effect of increasing sample size.

Note: So that we will all be looking at the same pictures, we will set a "seed" for the pseudo-random number generation. Be sure to set the seed as shown in each case below.

a. **(3 marks)** Complete (and hand in) the following code to generate the data that we will be considering

```
set.seed(314159)
# The normal data
z50 <- rnorm(50,0,1)
z100 <- rnorm(100,0,1)
z1000 <- rnorm(1000,0,1)
z10000 <- rnorm(10000,0,1)
zlims <- extendrange(c(z50, z100, z1000, z10000))

# The student t (3) data
t50 <- rt(50,3)
t100 <- rt(100,3)
t1000 <- rt(1000,3)
t10000 <- rt(10000,3)
tlims <- extendrange(c(t50, t100, t1000, t10000))

# The Chi-squared (3) data
c50 <- rchisq(50,3)
c100 <- rchisq(100,3)
c1000 <- rchisq(1000,3)
c10000 <- rchisq(10000,3)
clims <- extendrange(c(c50, c100, c1000, c10000))
```

You will be using these data to answer the remaining parts of this question.

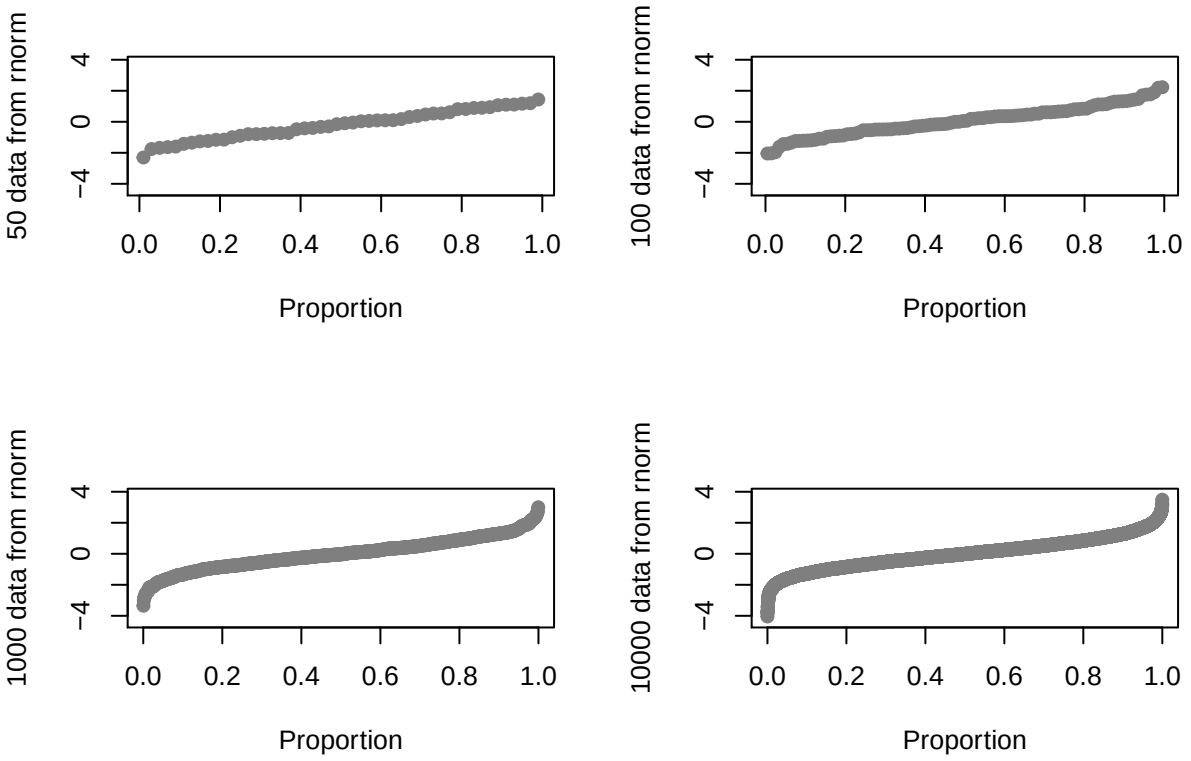
b. For each of the following arrange the corresponding visualizations of the underlying densities in a 2×2 array (e.g. via `savePar <- par(mfrow=c(2,2))`. Each plot in any given array should share the same data limits, the same underlying distribution, and be labelled according to the distribution that generated the sample, and the size of that sample. For each display type (i.e. quantile plot, boxplot, etc.) there should be three arrays (one for each generating distribution) where only the sample size n varies within array.

Fill all regions with "grey50".

For each array, comment on how the quality of the display changes as n increases.

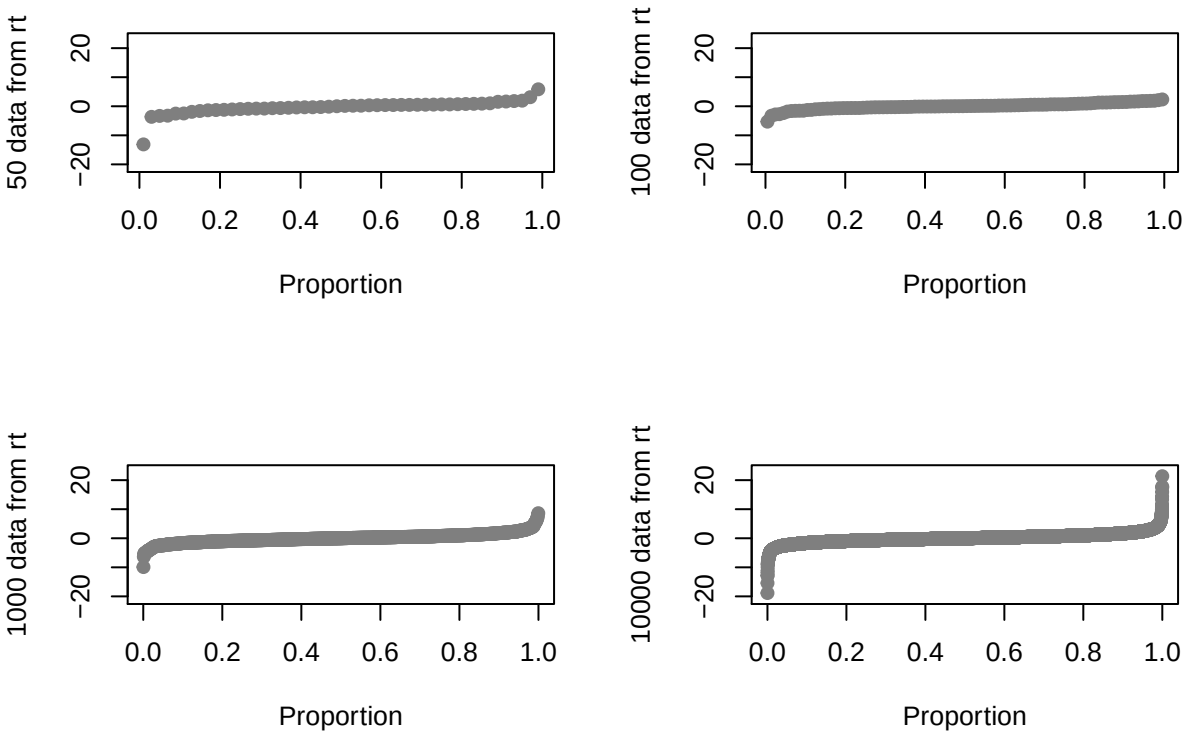
i. **(4 marks)** *quantile plots.* Produce the three arrays of changing n, one for each distribution ($N(0,1)$, t_3, and χ^2_3). Submit each arrangement of the four displayed plots and comment on how the quality of the display changes as n increases.

```
savePar <- par(mfrow=c(2,2))
plot(ppoints(50), sort(z50), ylim = zlims, xlab = "Proportion",
     ylab = "50 data from rnorm", type = "b", pch = 19, col = "grey50")
plot(ppoints(100), sort(z100), ylim = zlims, xlab = "Proportion",
     ylab = "100 data from rnorm", type = "b", pch = 19, col = "grey50")
plot(ppoints(1000), sort(z1000), ylim = zlims, xlab = "Proportion",
     ylab = "1000 data from rnorm", type = "b", pch = 19, col = "grey50")
plot(ppoints(10000), sort(z10000), ylim = zlims, xlab = "Proportion",
     ylab = "10000 data from rnorm", type = "b", pch = 19, col = "grey50")
```



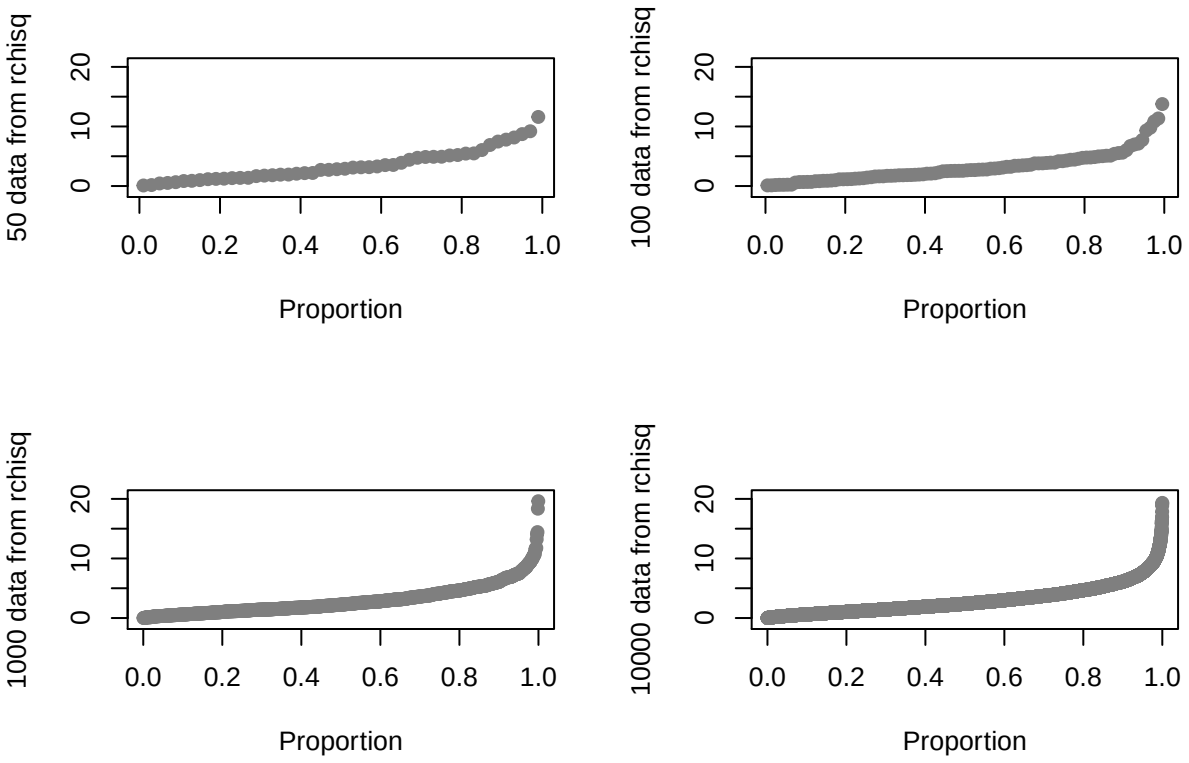
As the number of data increases, there is more dramatic increase in the beginning and end of the graph, which means that there is more data seems to have extreme value as the number of data increases. Compare the line of the 50 data, the larger data size (such as the 10000 data), the data is distributed more evenly, which means the line become much more smooth. It is not like the actual quantile plot of the normal distribution compare to the graph with large sample size. (ex, 10000 data) Therefore as the sample size increase, it perform better in quantile plot for normal distribution.

```
savePar <- par(mfrow=c(2,2))
plot(ppoints(50), sort(t50), ylim = tlims, xlab = "Proportion",
     ylab = "50 data from rt", type = "b", pch = 19, col = "grey50")
plot(ppoints(100), sort(t100), ylim = tlims, xlab = "Proportion",
     ylab = "100 data from rt", type = "b", pch = 19, col = "grey50")
plot(ppoints(1000), sort(t1000), ylim = tlims, xlab = "Proportion",
     ylab = "1000 data from rt",type = "b", pch = 19, col = "grey50")
plot(ppoints(10000), sort(t10000), ylim = tlims, xlab = "Proportion",
     ylab = "10000 data from rt", type = "b", pch = 19, col = "grey50")
```



There seems to have only small number of data in the beginning and end for the small data group(such as 50 data and 100 data). The disibution does not seems to be clear. As the number of the data increases, we can observe the distribution clearly especially for the extrme data on the ends. That is a significant feature of the quantile plot of the t=3 distribution. That makes the larger the sample size, the accuracy of the performance of the graph seems to increase.

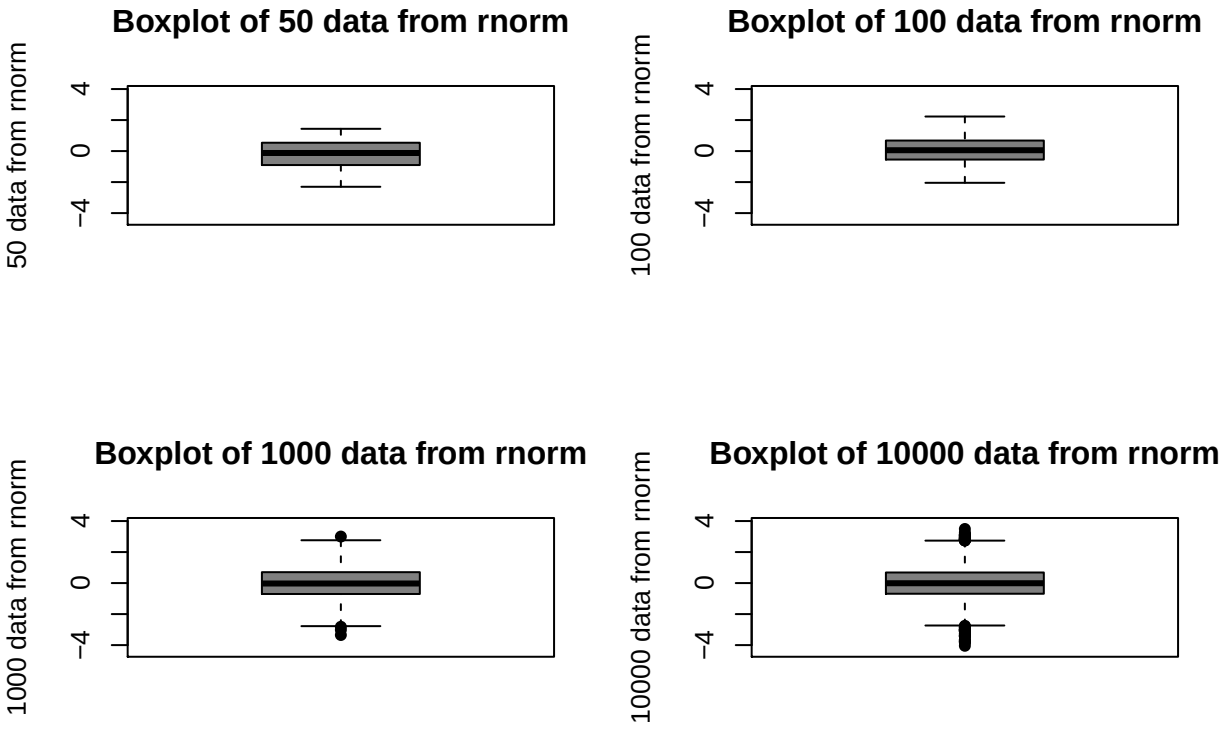
```r
savePar <- par(mfrow=c(2,2))
plot(ppoints(50), sort(c50), ylim = clims, xlab = "Proportion",
     ylab = "50 data from rchisq", type = "b", pch = 19, col = "grey50")
plot(ppoints(100), sort(c100), ylim = clims, xlab = "Proportion",
     ylab = "100 data from rchisq", type = "b", pch = 19, col = "grey50")
plot(ppoints(1000), sort(c1000), ylim = clims, xlab = "Proportion",
     ylab = "1000 data from rchisq", type = "b", pch = 19, col = "grey50")
plot(ppoints(10000), sort(c10000), ylim = clims, xlab = "Proportion",
     ylab = "10000 data from rchisq", type = "b", pch = 19, col = "grey50")
```



As discussed in the up 2 graphs, as the number of data increases, the line seems to be smoother. Also the extreme data in the end, in rchisq case, only the end of the graph(when the proportion close to 1), the extreme increase appear in the date group with large number of data. As the sample size increases, the graph seems more like to the real quantile plot for chisq.

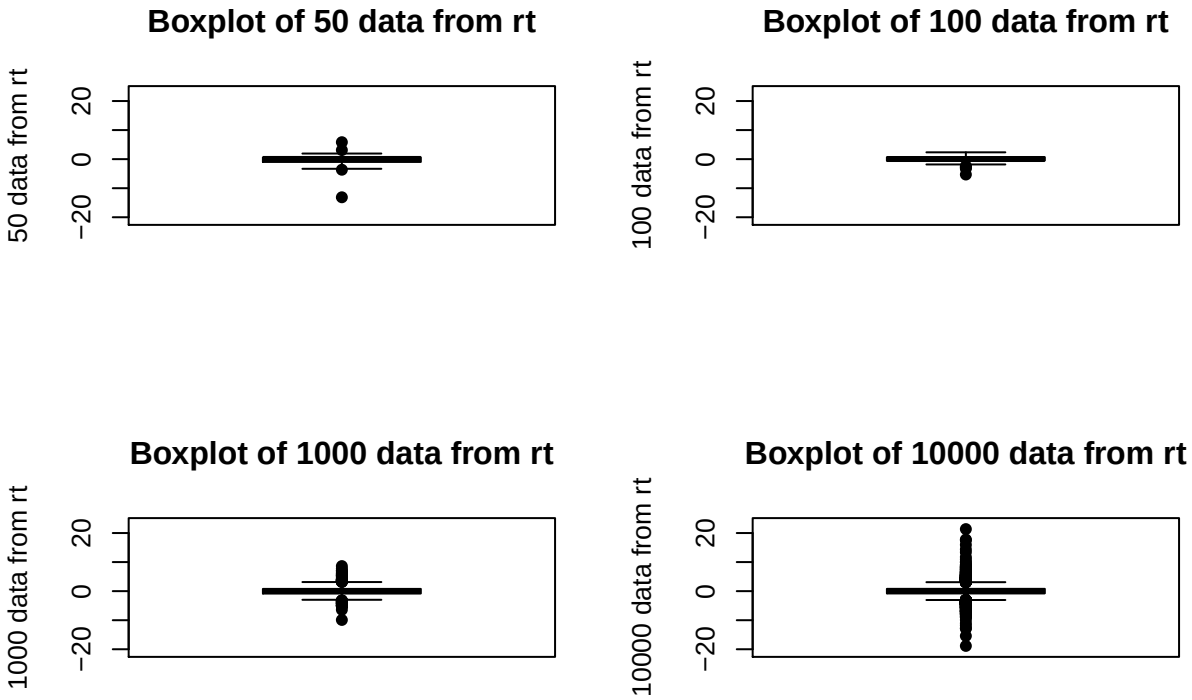
ii. **(4 marks)** *boxplots*. Produce the three arrays of changing n, one for each distribution ($N(

```
savePar <- par(mfrow=c(2,2))

boxplot(z50, ylim = zlims, main = "Boxplot of 50 data from rnorm",
        ylab = "50 data from rnorm", type = "b", pch = 19, col = "grey50")
boxplot(z100, ylim = zlims, main = "Boxplot of 100 data from rnorm",
        ylab = "100 data from rnorm", type = "b", pch = 19, col = "grey50")
boxplot(z1000, ylim = zlims, main = "Boxplot of 1000 data from rnorm",
        ylab = "1000 data from rnorm", type = "b", pch = 19, col = "grey50")
boxplot(z10000, ylim = zlims, main = "Boxplot of 10000 data from rnorm",
        ylab = "10000 data from rnorm", type = "b", pch = 19, col = "grey50")
```



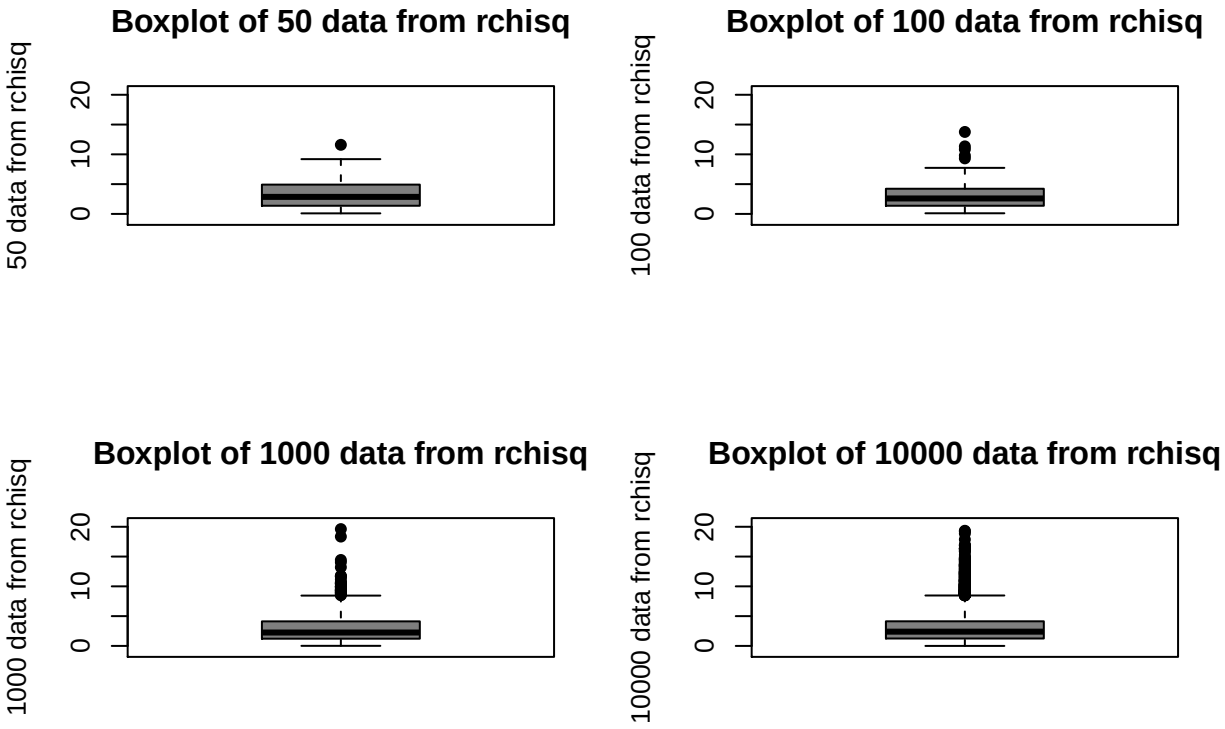
As the number of data increase, the range of the data increases, data is no longer centred around zero. What is more, as the number of data increases, the IQR increases, but the outlier also increases. As the data size is small, almost all the data we observed is in the IQR, which means the data is more concentrated in the center. As the data size increase, the data is more spread out. There should not be so many outlier for the plot, but the number of outliers increases dramatically compare to the increase in sample size as the sample size increase. That is not a good feature of the normal distribution, thus the increase in sample size does not improve the quality of the graph.

```r
savePar <- par(mfrow=c(2,2))

boxplot(t50, ylim = tlims, main = "Boxplot of 50 data from rt",
        ylab = "50 data from rt", type = "b", pch = 19, col = "grey50")
boxplot(t100, ylim = tlims, main = "Boxplot of 100 data from rt",
        ylab = "100 data from rt", type = "b", pch = 19, col = "grey50")
boxplot(t1000, ylim = tlims, main = "Boxplot of 1000 data from rt",
        ylab = "1000 data from rt", type = "b", pch = 19, col = "grey50")
boxplot(t10000, ylim = tlims, main = "Boxplot of 10000 data from rt",
        ylab = "10000 data from rt", type = "b", pch = 19, col = "grey50")
```



As the tdisterbitution is more concentrated in the center, the IQR is quite small. The data is more concetrated around 0. As the data size increases, the number of extreme data located out of the IQR increases. Same problem here, as the sample size increase, the number of outliers increase dramatically. That is a not siginficat feature fot the t distribution, therefore increasing the sample size may not be a good choice to estimate the boxplot for the t distribution.

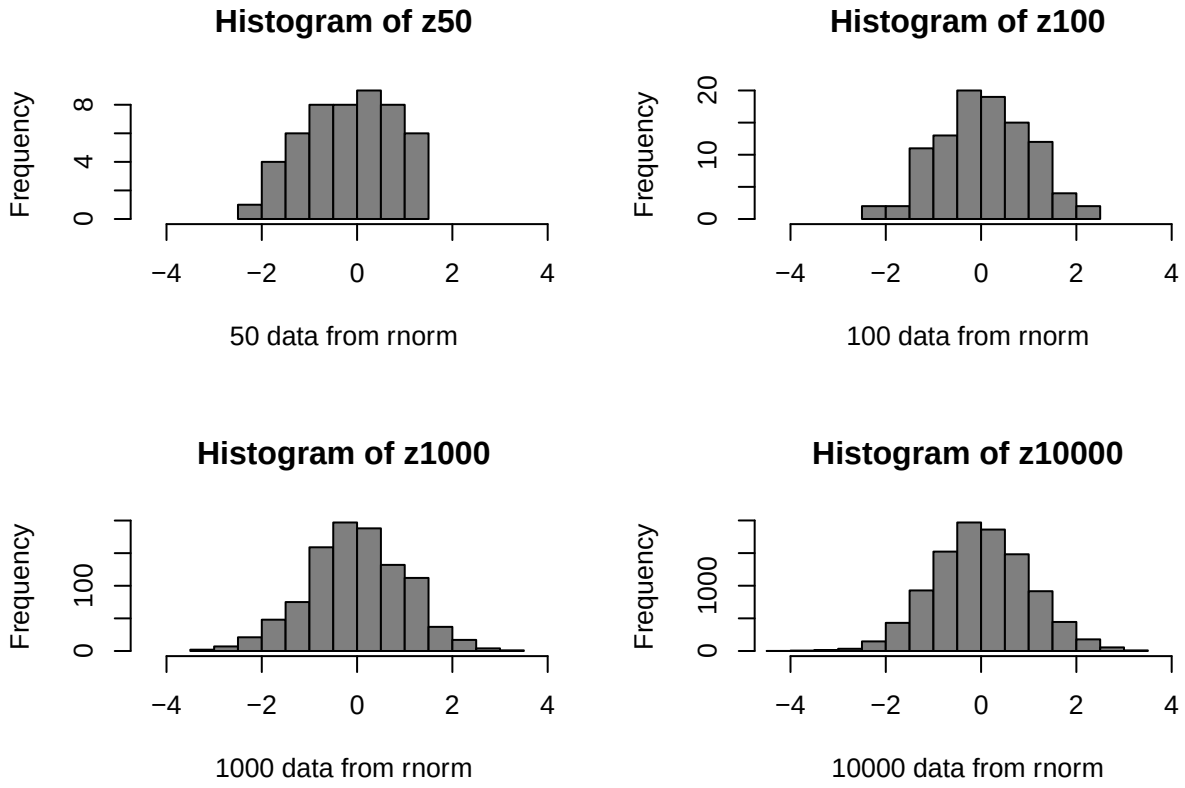
```r
savePar <- par(mfrow=c(2,2))

boxplot(c50, ylim = clims, main = "Boxplot of 50 data from rchisq",
        ylab = "50 data from rchisq", type = "b", pch = 19, col = "grey50")
boxplot(c100, ylim = clims, main = "Boxplot of 100 data from rchisq",
        ylab = "100 data from rchisq", type = "b", pch = 19, col = "grey50")
boxplot(c1000, ylim = clims, main = "Boxplot of 1000 data from rchisq",
        ylab = "1000 data from rchisq", type = "b", pch = 19, col = "grey50")
boxplot(c10000, ylim = clims, main = "Boxplot of 10000 data from rchisq",
        ylab = "10000 data from rchisq", type = "b", pch = 19, col = "grey50")
```



All the observed extreme value is located at the upper side in the graph. the IQR is quite small for all data size. As the data size increase, there is more extrme data observed. That is not a signature of the chisquare distribution. Thus increasing the sample size may bot be a good estimate for chisquare distribution.

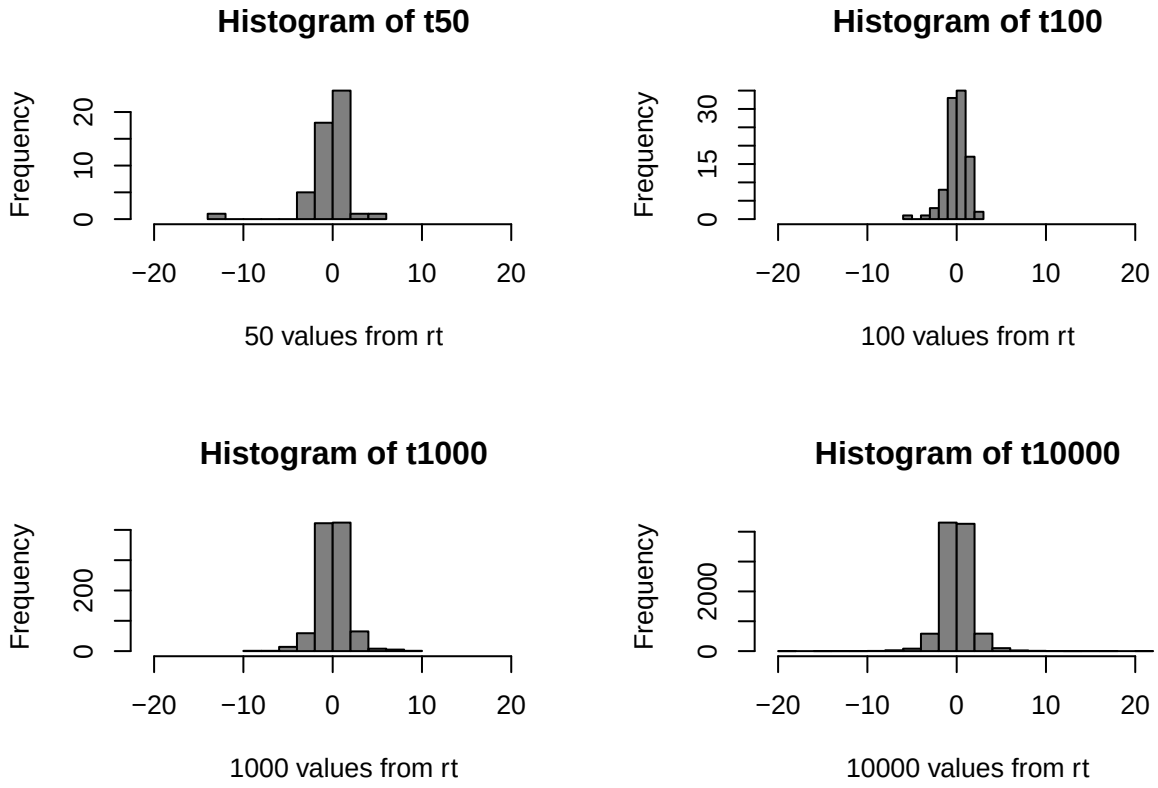
iii. **(4 marks)** *histograms*. Produce the three arrays of changing n, one for each distribution ($

```
savePar <- par(mfrow=c(2,2))


hist(z50, xlim = zlims, xlab = "50 data from rnorm", col = "grey50")
hist(z100, xlim = zlims, xlab = "100 data from rnorm", col = "grey50")
hist(z1000, xlim = zlims, xlab = "1000 data from rnorm", col = "grey50")
hist(z10000, xlim = zlims, xlab = "10000 data from rnorm", col = "grey50")
```



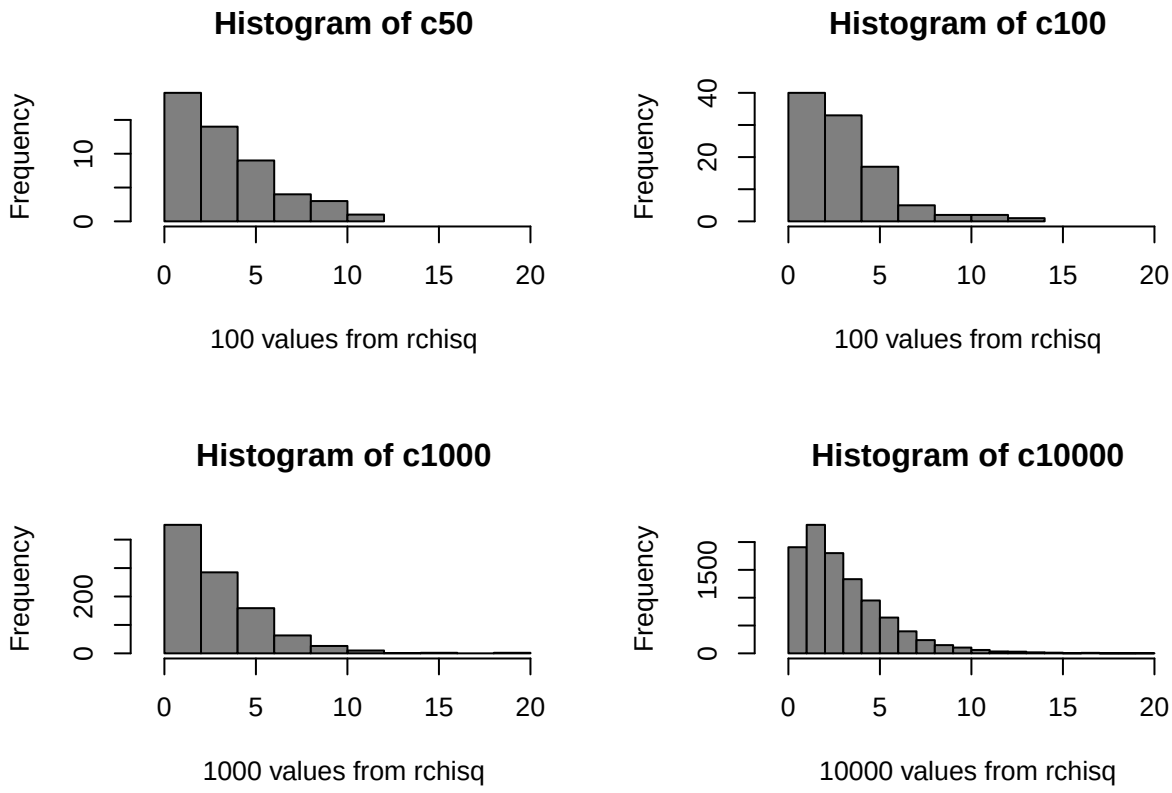
As the data size increases, the frequency increases. That make sense because there are more repeated data shown in the each data group. What is more, as the sample size increases, the data is more spread out, there is more extreme value shown in te graph even the number of centerd value also increases. When the data size is small, the distribution is not claear shown, some features of the distribution such as symmetric and the tails are not clearly shown. Thus as the data size increases, the normal distribution is more clearly shown.

```r
savePar <- par(mfrow=c(2,2))


hist(t50, xlim = tlims, xlab = "50 values from rt", col = "grey50")
hist(t100, xlim = tlims, xlab = "100 values from rt", col = "grey50")
hist(t1000, xlim = tlims, xlab = "1000 values from rt", col = "grey50")
hist(t10000, xlim = tlims, xlab = "10000 values from rt", col = "grey50")
```



As the data size increases, the data is more concentrated in the center.As the bin area increases in 100000, it is clear shown the data concentration. What is more, the spread decrease as the data size increases. Some feature such as symmetric and small tails are shown clearly for large sample size such as 1000 and 10000. As a result, increasing sample size may be a good way to estimate the histogram of the t distrubution.

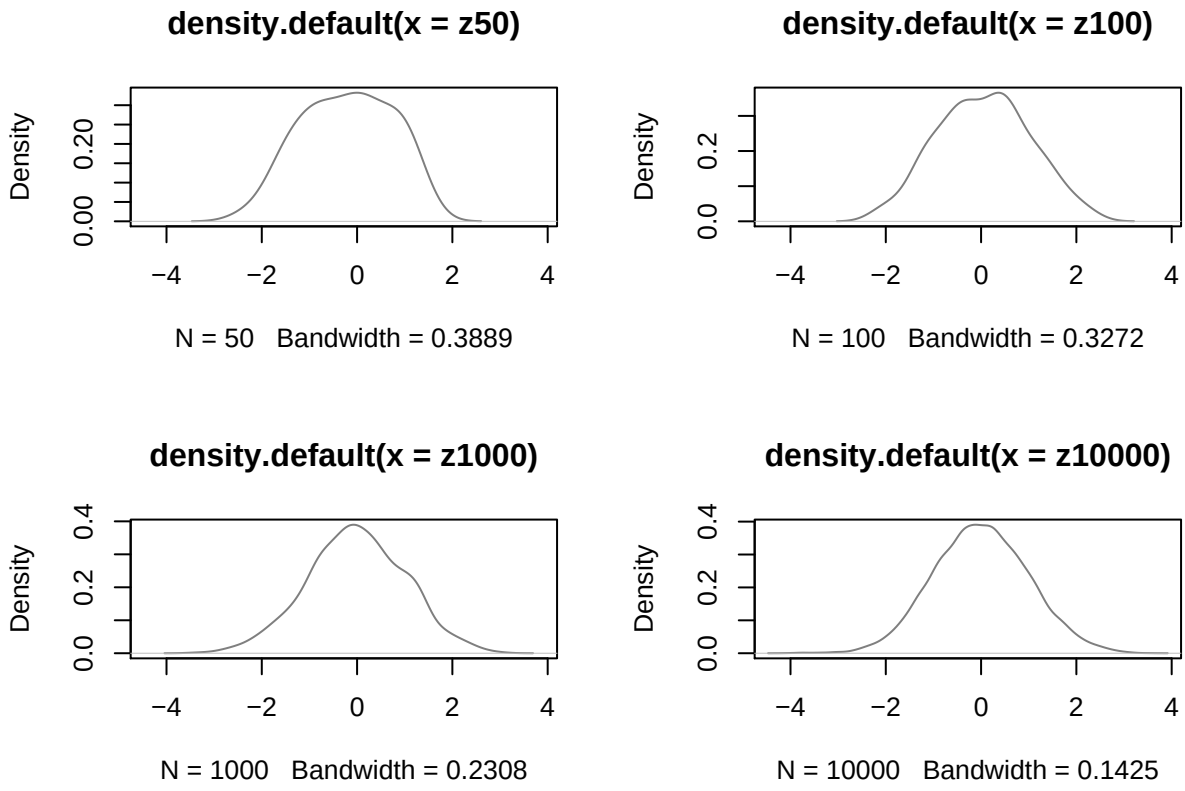
```
savePar <- par(mfrow=c(2,2))
hist(c50, xlim = clims, xlab = "100 values from rchisq", col = "grey50")
hist(c100, xlim = clims, xlab = "100 values from rchisq", col = "grey50")
hist(c1000, xlim = clims, xlab = "1000 values from rchisq", col = "grey50")
hist(c10000, xlim = clims, xlab = "10000 values from rchisq", col = "grey50")
```



As the sample size increase, the bin width decrease as there are more data in the orginal bin width makes the original bin width can not satisfy the new need. Also, only the sample size 10000 shows that the second bin has the largest frequency, which the increase from origin to the maximum is a signature feature of the chisq, thus increasing the sample size can increase the accuaracy of the graph for chisquare distribution.

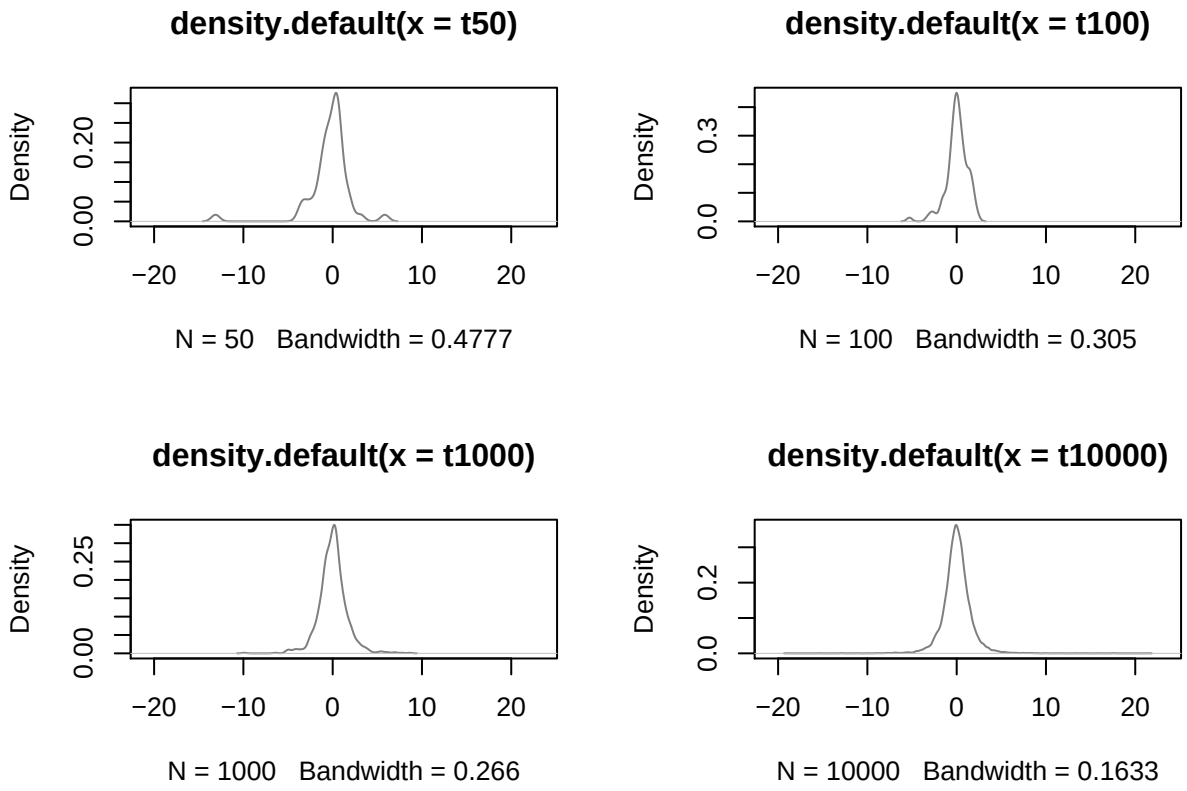
iv. **(5 marks)** *density plots*. Produce the three arrays of changing n, one for each distribution

```
savePar <- par(mfrow=c(2,2))
plot(density(z50), xlim = zlims, col = "grey50")
plot(density(z100), xlim = zlims, col = "grey50")
plot(density(z1000), xlim = zlims, col = "grey50")
plot(density(z10000), xlim = zlims, col = "grey50")
```



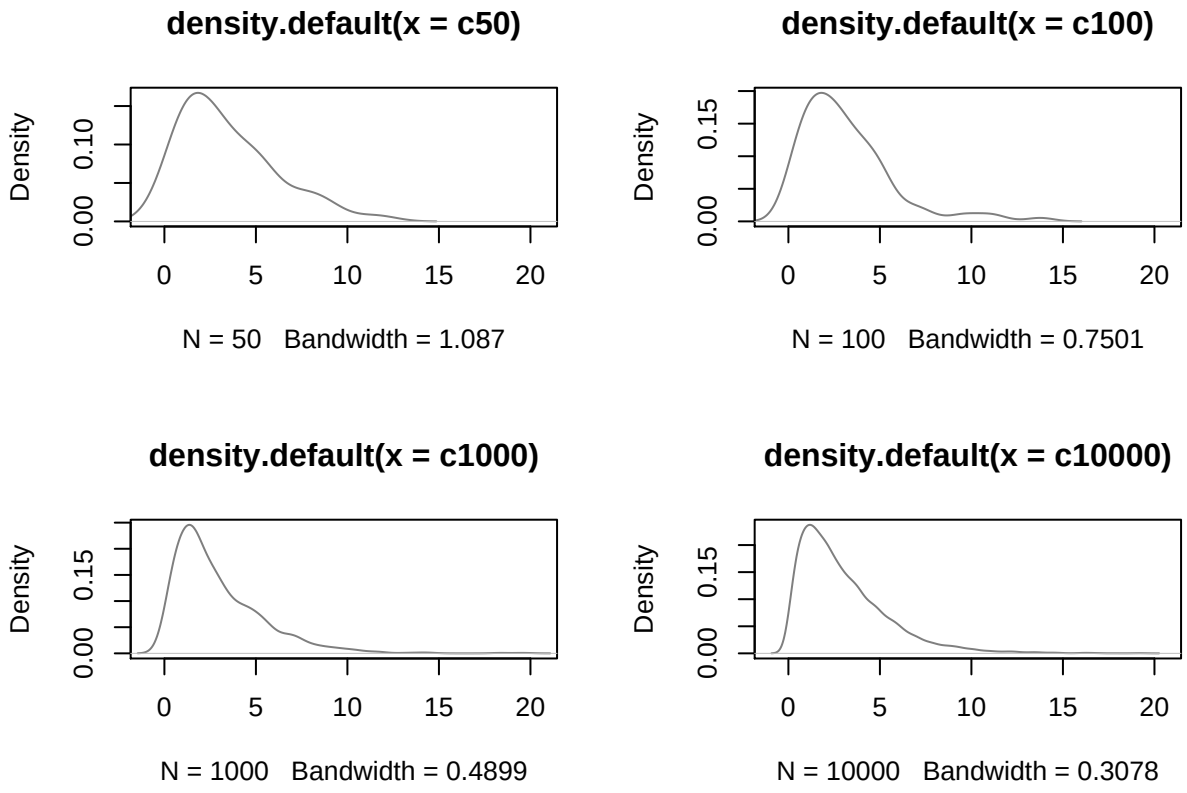
As the data size increases, the bandwidth continue decrease. As shown, the graph become much more smooth curve, which makes the graph more like a bell curve, which is a significant feature of the normal distribution. As the size increases, the normal features such as bell shape and symmtric are more clear, therefore increasing sample size is good for estimating normal distribution in density graph.

```r
savePar <- par(mfrow=c(2,2))
plot(density(t50), xlim = tlims, col = "grey50")
plot(density(t100), xlim = tlims, col = "grey50")
plot(density(t1000), xlim = tlims, col = "grey50")
plot(density(t10000), xlim = tlims, col = "grey50")
```



As the data size increases, the bandwidth continue decrease. As shown, the graph become much more smooth curve and centrted around 0. Also as the sample size increases, the density of extreme values decreases. The features of t distribution such as symmetric around 0 are more clearly shown. As a result, increasing the sample size is a good way for increasing accuaracy of the desity function for t distribution.

```r
savePar <- par(mfrow=c(2,2))
plot(density(c50), xlim = clims, col = "grey50")
plot(density(c100), xlim = clims, col = "grey50")
plot(density(c1000), xlim = clims, col = "grey50")
plot(density(c10000), xlim = clims, col = "grey50")
```



As the data size increases, the bandwidth continue decrease. As shown, the graph become much more smooth curve. What is more, as the sample size increases, the graph become much more likely to the chisquare distribution graph.