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

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

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

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

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

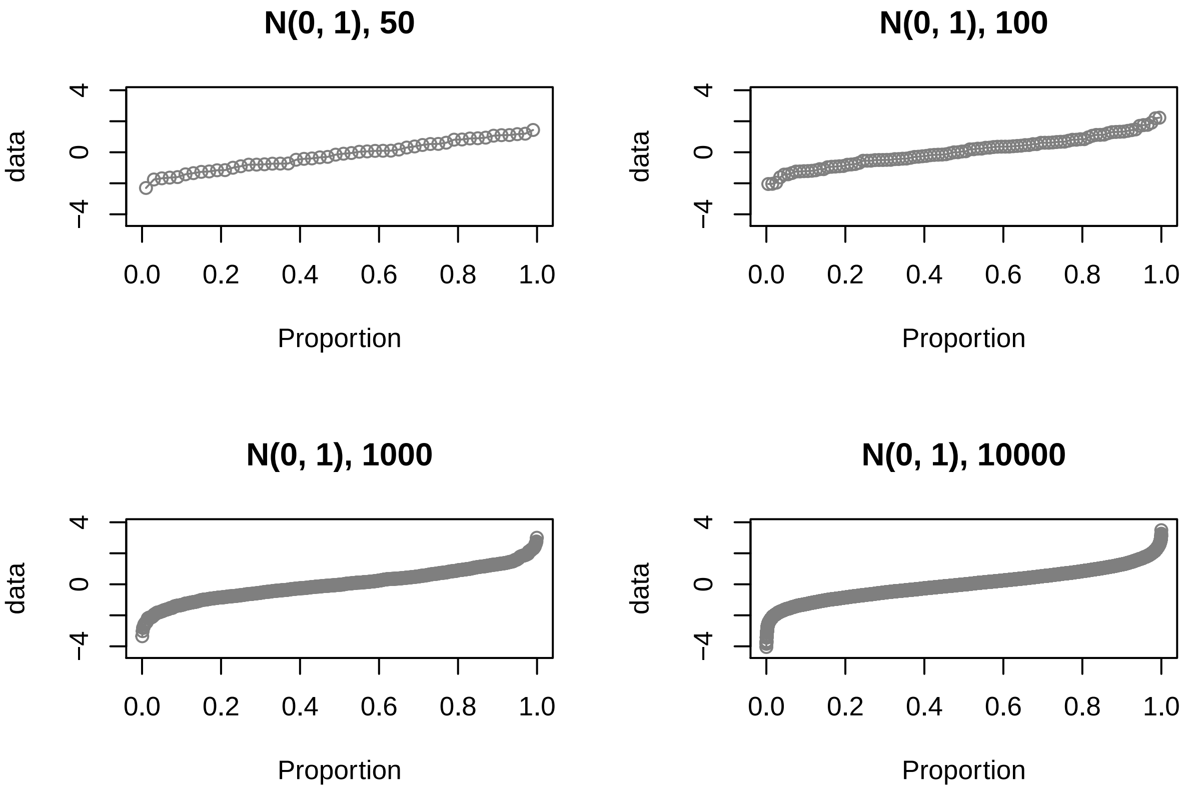
b. For each of the following arrange the corresponding visualizations of the underlying densities in a 2×2 array (e.g. via `savePar <- par(mfrow=c(2,2))`. Each plot in any given array should share the same data limits, the same underlying distribution, and be labelled according to the distribution that 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(x=ppoints(50), y=sort(z50), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=zlims,
     main=’N(0, 1), 50’)
plot(x=ppoints(100), y=sort(z100), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=zlims,
     main=’N(0, 1), 100’)
plot(x=ppoints(1000), y=sort(z1000), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=zlims,
     main=’N(0, 1), 1000’)
plot(x=ppoints(10000), y=sort(z10000), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=zlims,
     main=’N(0, 1), 10000’)
```



N(0, 1), 50

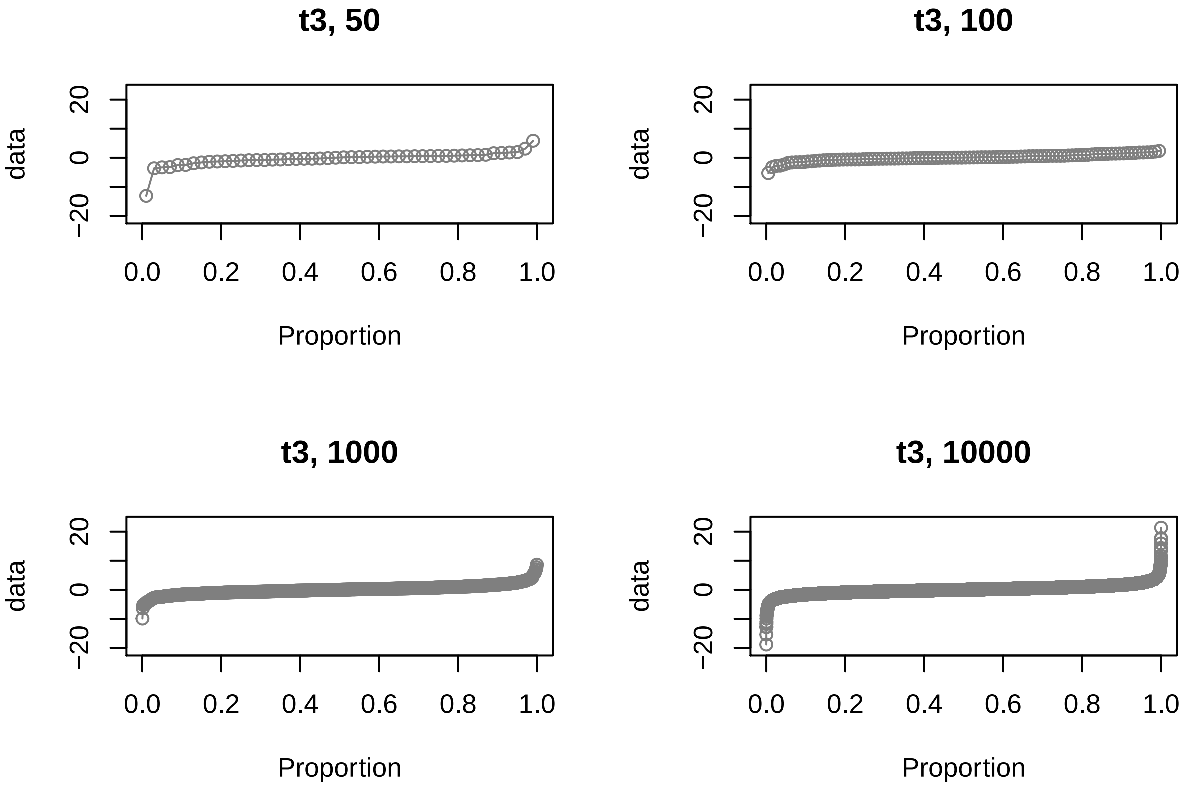
N(0, 1), 100

N(0, 1), 1000

N(0, 1), 10000

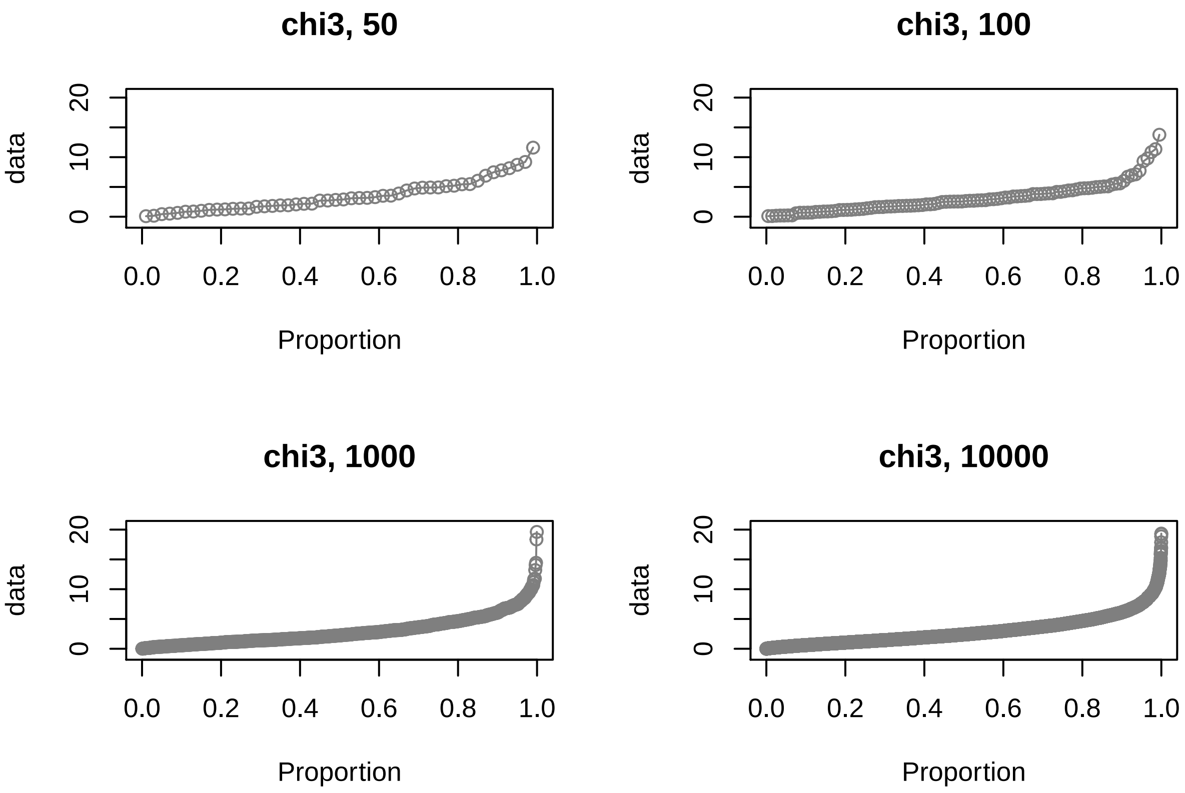
Comment: For $N(0, 1)$, with sample size increases, the graph shows a S-shape, which implies a unimodal symmetric shape.

```
savePar <- par(mfrow=c(2,2))
plot(x=ppoints(50), y=sort(t50), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=tlims,
     main=’t3, 50’)
plot(x=ppoints(100), y=sort(t100), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=tlims,
     main=’t3, 100’)
plot(x=ppoints(1000), y=sort(t1000), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=tlims,
     main=’t3, 1000’)
plot(x=ppoints(10000), y=sort(t10000), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=tlims,
     main=’t3, 10000’)
```



Comment: For student t, with sample size increases, the graph shows a S-shape, which implies a unimodal symmetric shape. However, comparing to normal distribution, vertical tails implies that student t distribution has low concentrations at both tails.

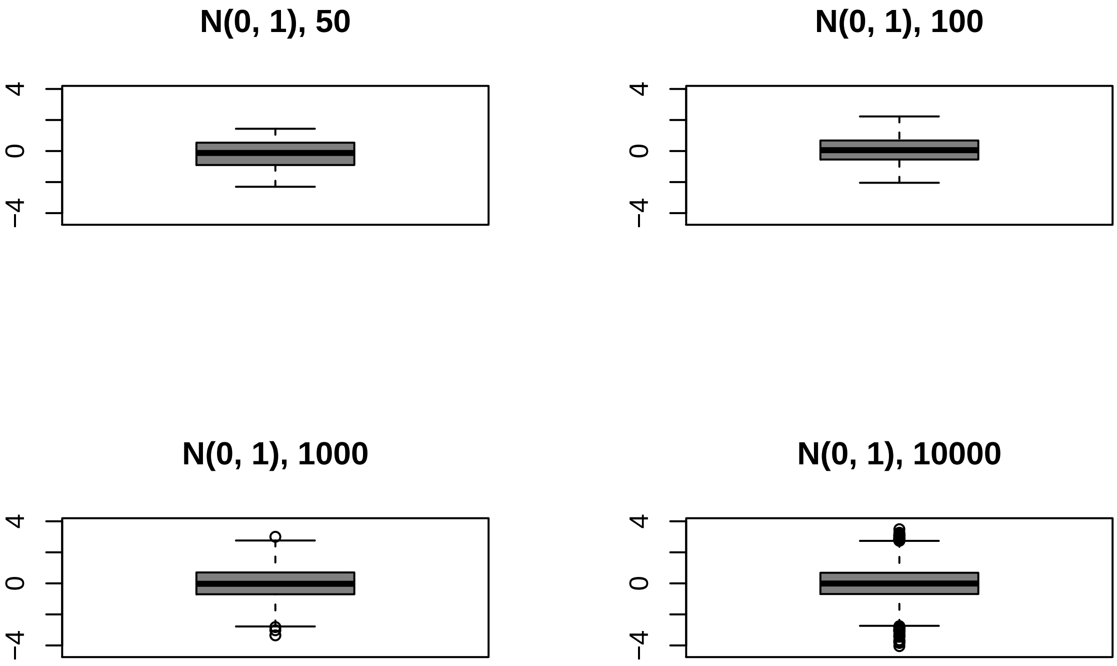
```
savePar <- par(mfrow=c(2,2))
plot(x=ppoints(50), y=sort(c50), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=clims,
     main=’chi3, 50’)
plot(x=ppoints(100), y=sort(c100), type=’o’, lwd=1, col=’grey50’,
     xlab=’Proportion’, ylab=’data’, xlim=c(0, 1), ylim=clims,
     main=’chi3, 100’)
plot(x=ppoints(1000), y=sort(c1000), type=’o’, lwd=1, col=’grey50’,
     xlab='Proportion', ylab='data', xlim=c(0, 1), ylim=clims,
     main='chi3, 1000')
plot(x=ppoints(10000), y=sort(c10000), type='o', lwd=1, col='grey50',
     xlab='Proportion', ylab='data', xlim=c(0, 1), ylim=clims,
     main='chi3, 10000')
```



Comment: For Chi-squared, the distribution is asymmetric and has a long right tail as the plot gets more vertical at the top.

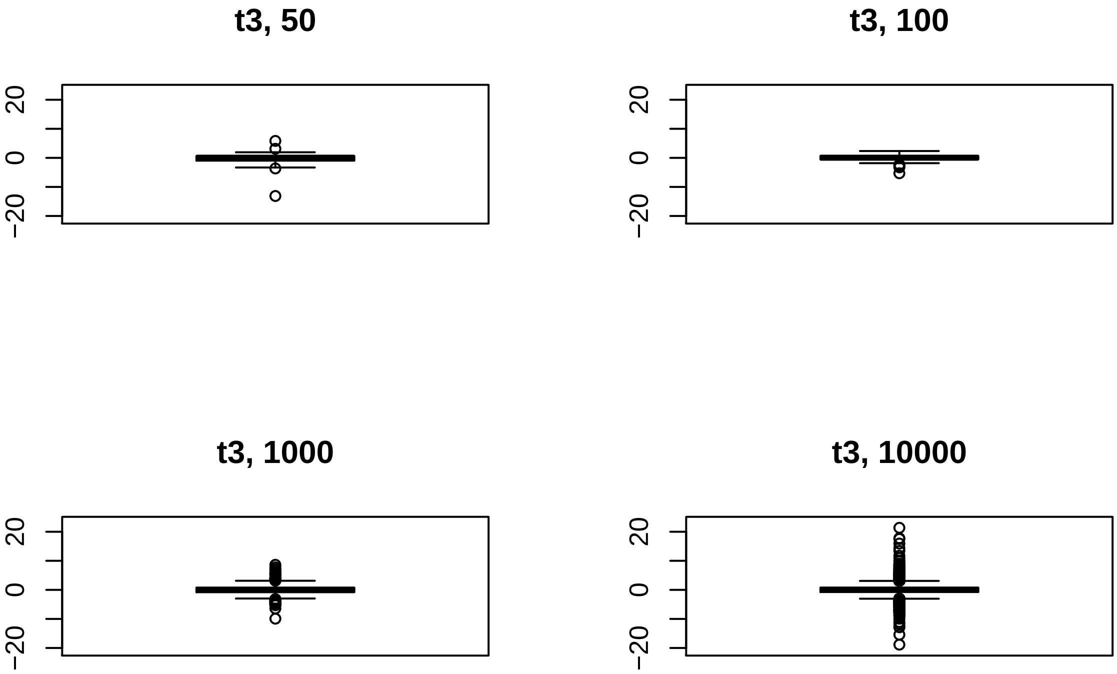
ii. **(4 marks)** *boxplots.* Produce the three arrays of changing n, one for each distribution ($N(0,1)$, t_3, and χ^2_3). Submit each arrangement of the four displayed plots and comment on how the quality of the display changes as n increases.

```
savePar <- par(mfrow=c(2,2))
boxplot(z50, ylim=zlims, col='grey50', main='N(0, 1), 50')
boxplot(z100, ylim=zlims, col='grey50', main='N(0, 1), 100')
boxplot(z1000, ylim=zlims, col='grey50', main='N(0, 1), 1000')
boxplot(z10000, ylim=zlims, col='grey50', main='N(0, 1), 10000')
```



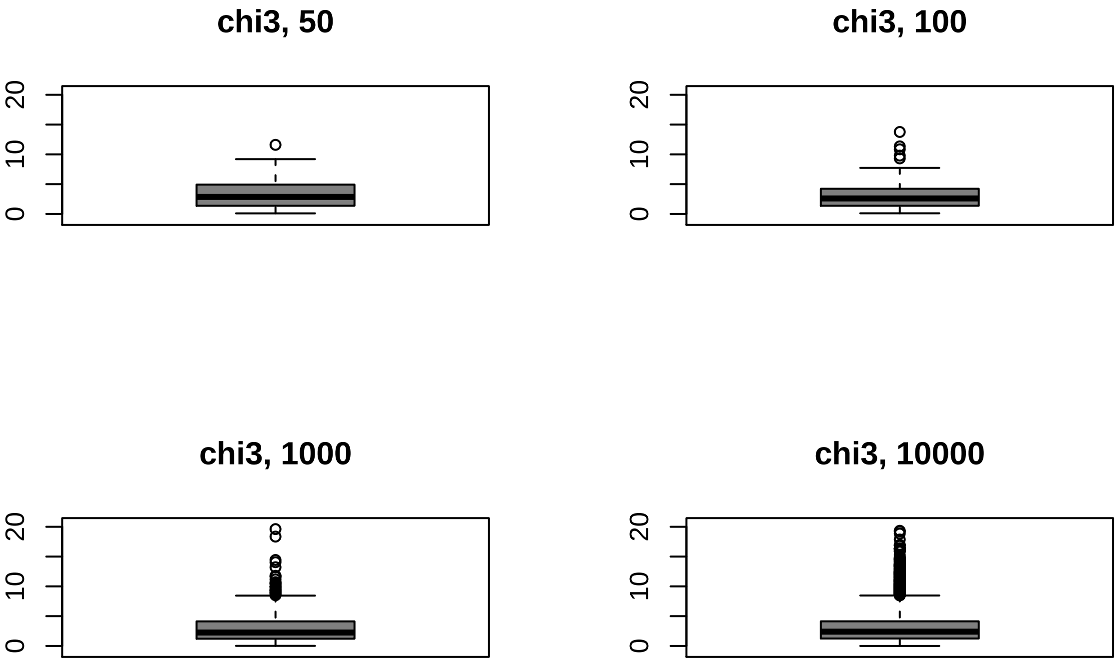
Comment: For normal distribution, outliers are not easily detected when n=50, with sample size increases, there are more outliers detected.

```
savePar <- par(mfrow=c(2,2))
boxplot(t50, ylim=tlims, col='grey50', main='t3, 50')
boxplot(t100, ylim=tlims, col='grey50', main='t3, 100')
boxplot(t1000, ylim=tlims, col='grey50', main='t3, 1000')
boxplot(t10000, ylim=tlims, col='grey50', main='t3, 10000')
```



Comment: For student t, we see that outliers are easier to be detected than normal distribution. With sample size increases, there are great amount of outliers shown in the boxplot. This implies that student t distribution has longer tails.

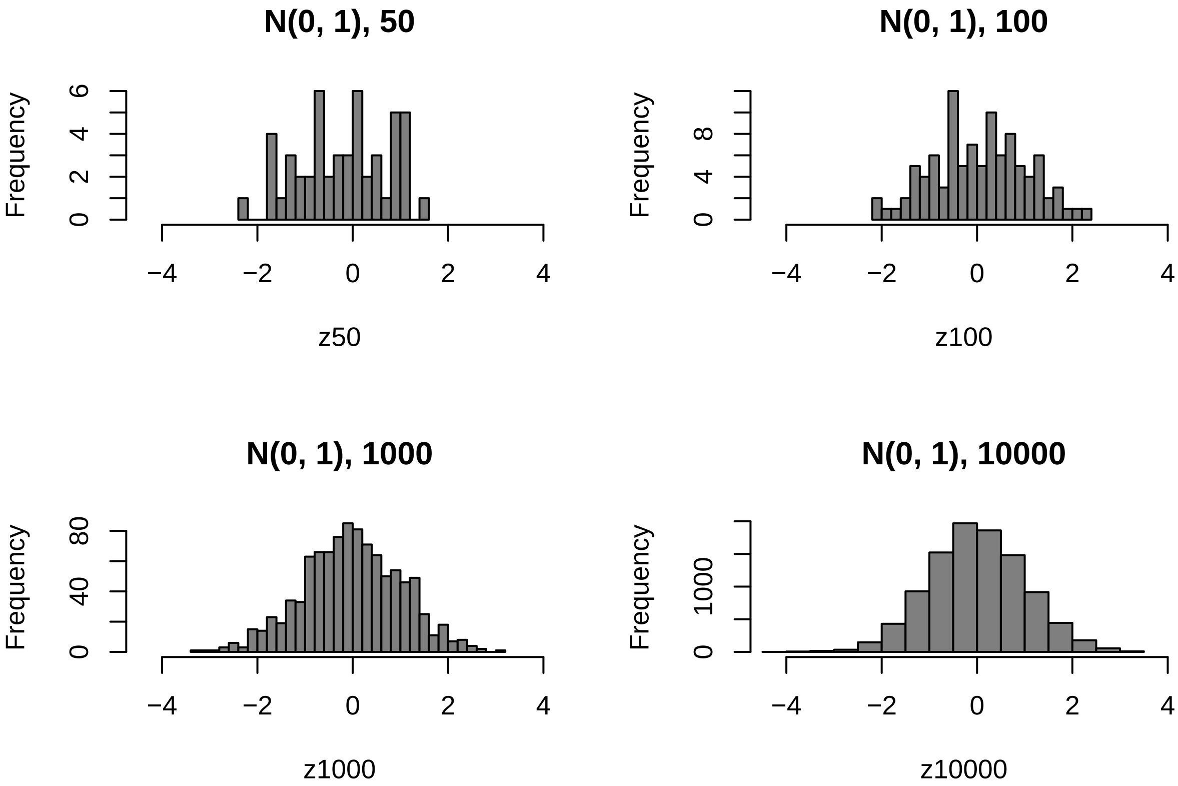
```
savePar <- par(mfrow=c(2,2))
boxplot(c50, ylim=clims, col='grey50', main='chi3, 50')
boxplot(c100, ylim=clims, col='grey50', main='chi3, 100')
boxplot(c1000, ylim=clims, col='grey50', main='chi3, 1000')
boxplot(c10000, ylim=clims, col='grey50', main='chi3, 10000')
```



Comment: For Chi-squared, we see that all outliers exist at the the upper area of the boxplot (or the right tail), also with sample size increases, outliers at the right tail get more and more. From the box itself, we can conclude that the distribution has asymmetric shape.

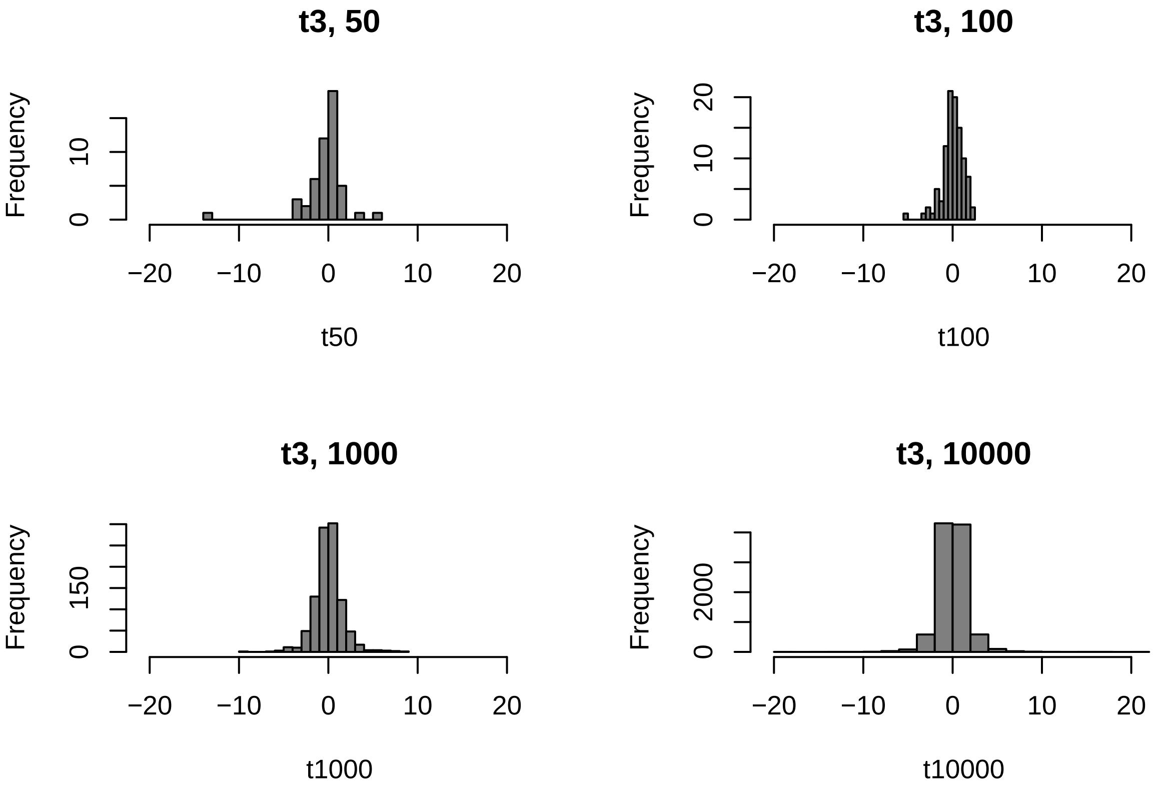
iii. **(4 marks)** *histograms.* Produce the three arrays of changing n, one for each distribution ($N(0,1)$, t_3, and χ^2_3). Submit each arrangement of the four displayed plots and comment on how the quality of the display changes as n increases.

```
savePar <- par(mfrow=c(2,2))
hist(z50, xlim=zlims, breaks = 25, col='grey50',
     main='N(0, 1), 50')
hist(z100, xlim=zlims, breaks = 25, col='grey50',
        main='N(0, 1), 100')
hist(z1000, xlim=zlims, breaks = 25, col='grey50',
        main='N(0, 1), 1000')
hist(z10000, xlim=zlims, breaks = 25, col='grey50',
        main='N(0, 1), 10000')
```



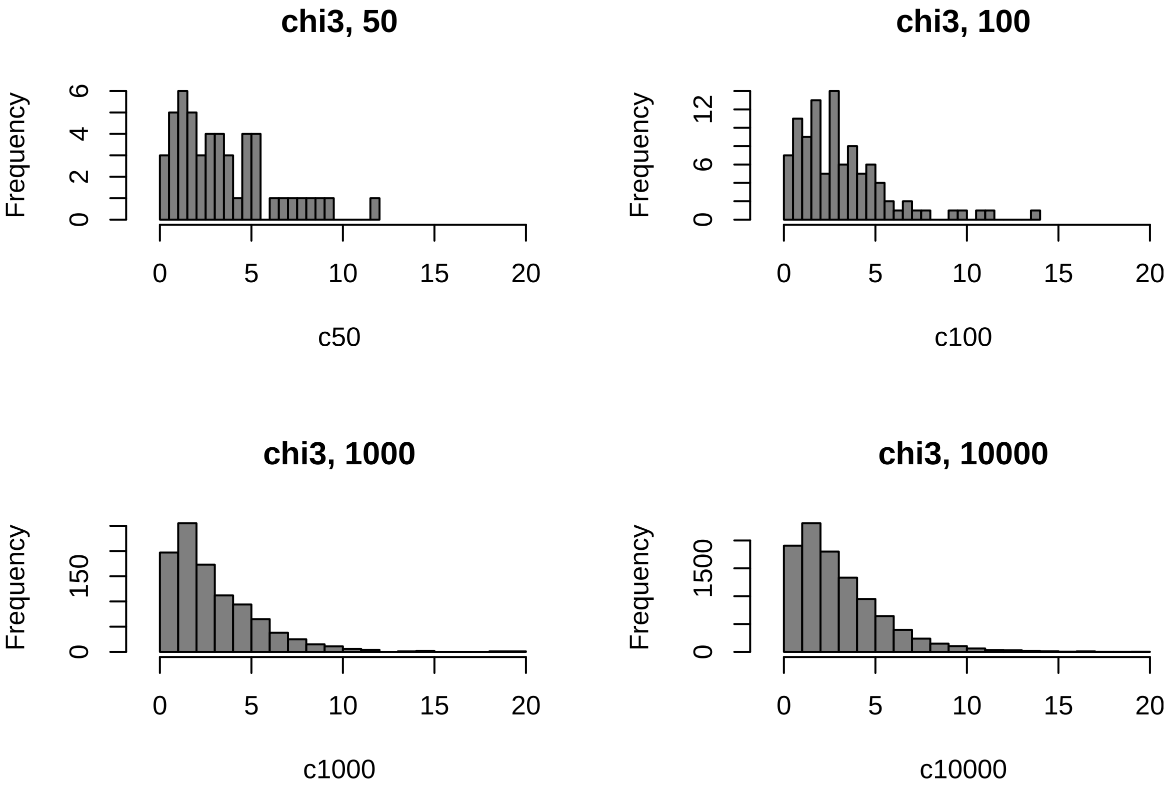
Comment: with sample size increases, the histogram becomes more and more symmetric.

```
savePar <- par(mfrow=c(2,2))
hist(t50, xlim=tlims, breaks = 25, col=’grey50’,
     main=’t3, 50’)
hist(t100, xlim=tlims, breaks = 25, col=’grey50’,
        main=’t3, 100’)
hist(t1000, xlim=tlims, breaks = 25, col=’grey50’,
        main=’t3, 1000’)
hist(t10000, xlim=tlims, breaks = 25, col=’grey50’,
        main=’t3, 10000’)
```



Comment: with sample size increases, the distribution becomes more and more symmetric but both tails become heavier and heavier.

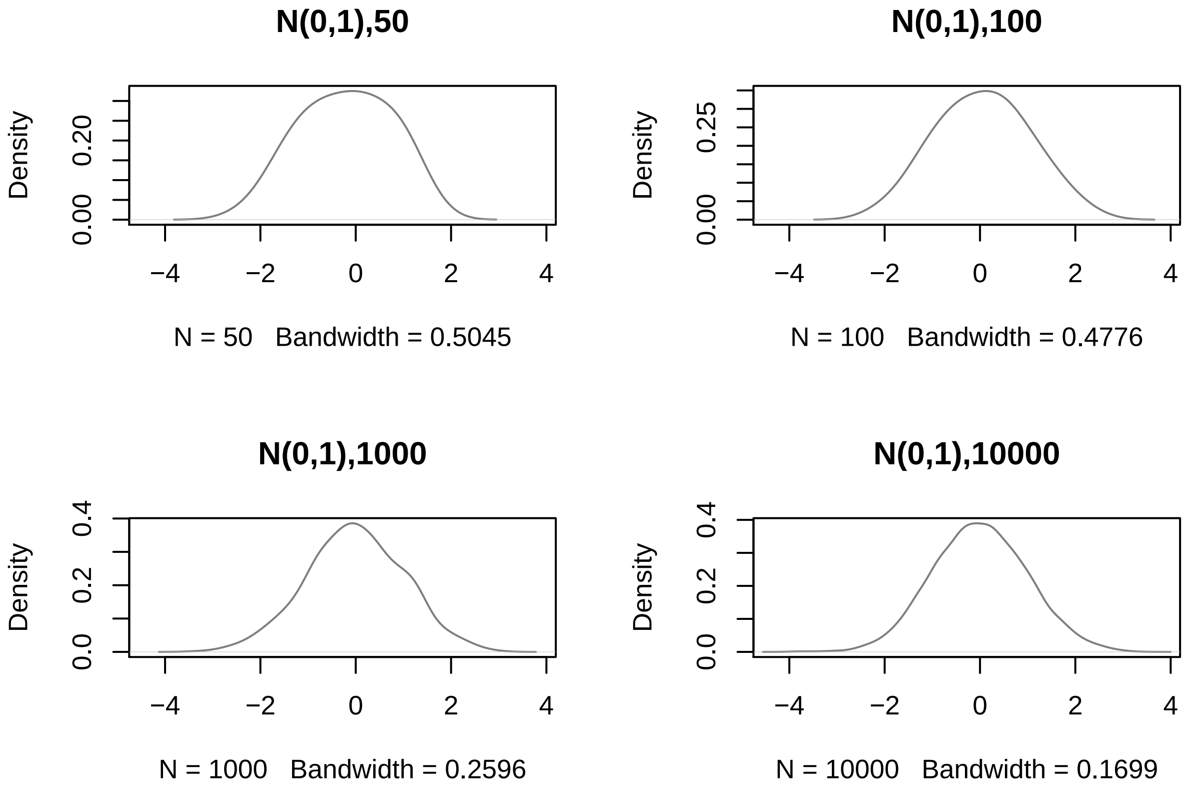
```
savePar <- par(mfrow=c(2,2))
hist(c50, xlim=clims, breaks = 25, col=’grey50’,
     main=’chi3, 50’)
hist(c100, xlim=clims, breaks = 25, col=’grey50’,
        main=’chi3, 100’)
hist(c1000, xlim=clims, breaks = 25, col=’grey50’,
        main=’chi3, 1000’)
hist(c10000, xlim=clims, breaks = 25, col=’grey50’,
        main=’chi3, 10000’)
```



Comment: all four graphs can conclude an asymmetric shape as most data are distributed at the left. However, with sample size increases, the right tail becomes heavier and heavier.

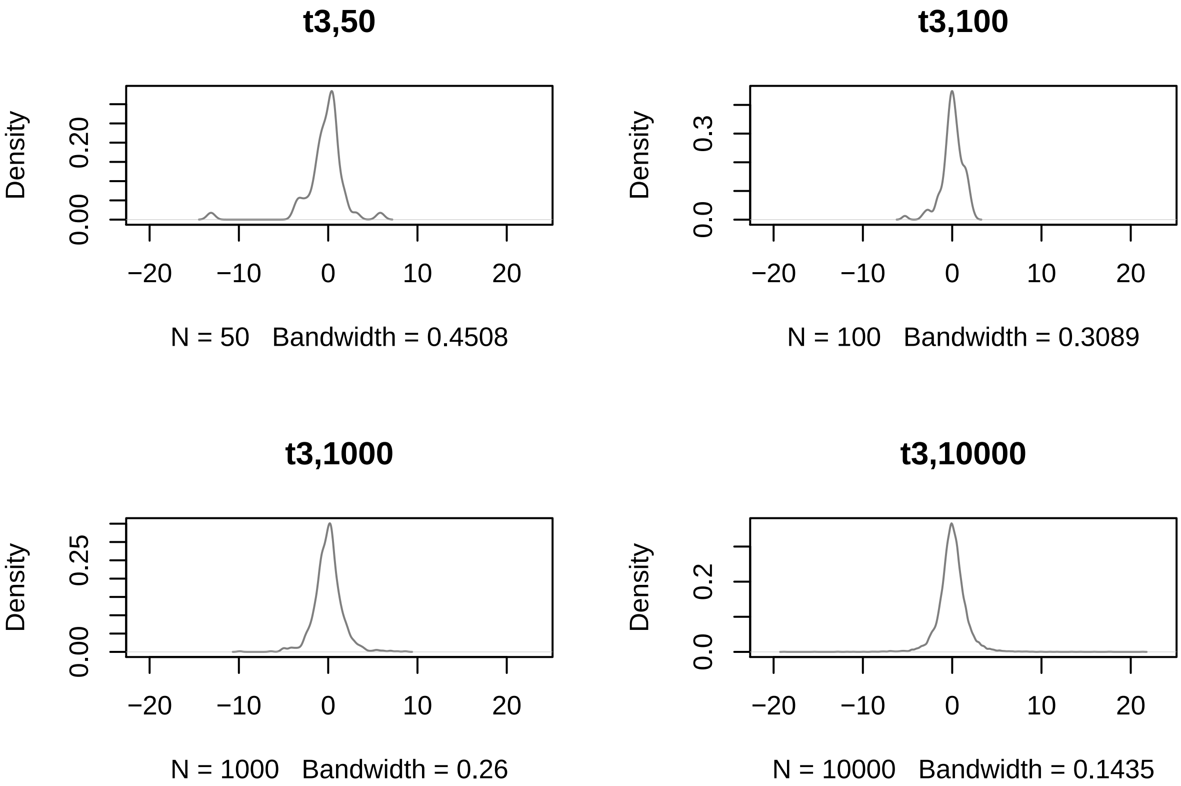
iv. **(5 marks)** *density plots*. Produce the three arrays of changing n, one for each distribution ($N(0,1)$, t_3, and χ^2_3). Submit each arrangement of the four displayed plots and comment on how the quality of the display changes as n increases.

```
savePar <- par(mfrow=c(2,2))
plot(density(z50, bw='SJ'), col='grey50',
     xlim=zlims, main='N(0,1),50')
plot(density(z100, bw='SJ'), col='grey50',
     xlim=zlims, main='N(0,1),100')
plot(density(z1000, bw='SJ'), col='grey50',
     xlim=zlims, main='N(0,1),1000')
plot(density(z10000, bw='SJ'), col='grey50',
     xlim=zlims, main='N(0,1),10000')
```



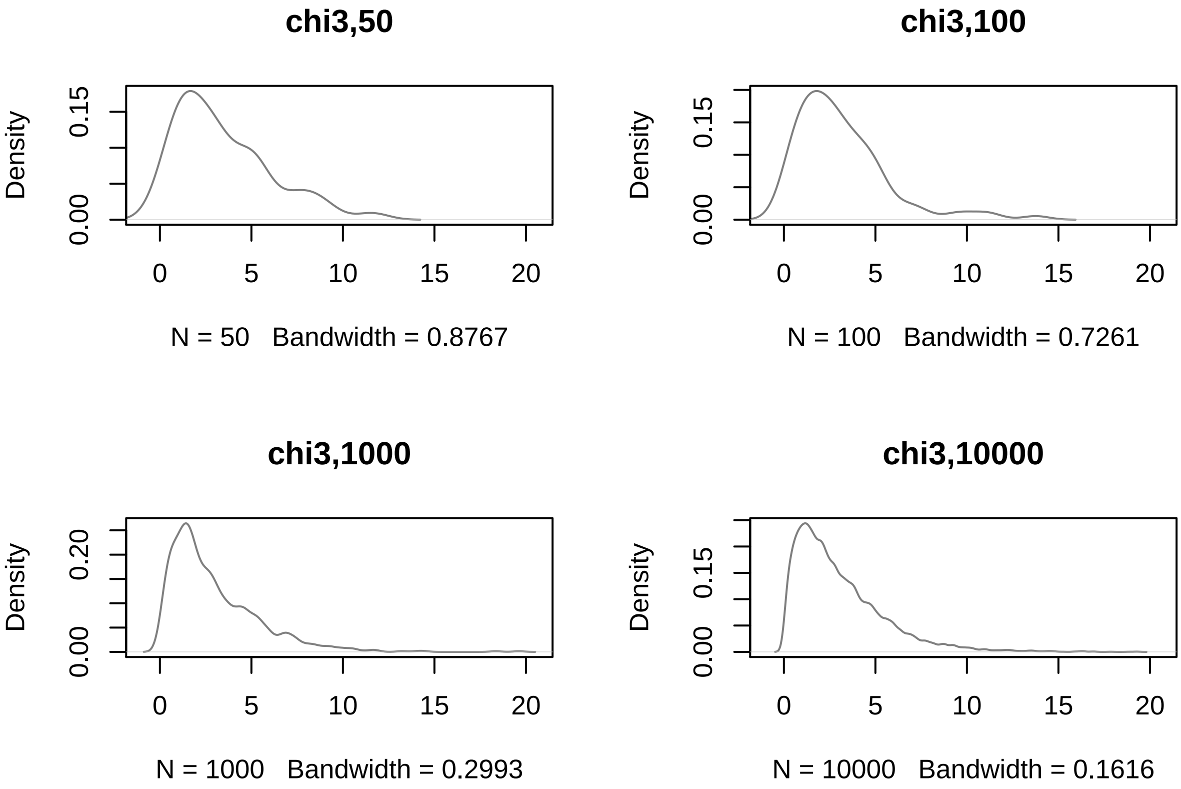
Comment: With sample size increases, we can clearly see that the distribution is symmetric and concentrated around 0.

```
savePar <- par(mfrow=c(2,2))
plot(density(t50, bw=’SJ’), col=’grey50’,
     xlim=tlims, main=’t3,50’)
plot(density(t100, bw=’SJ’), col=’grey50’,
     xlim=tlims, main=’t3,100’)
plot(density(t1000, bw=’SJ’), col=’grey50’,
     xlim=tlims, main=’t3,1000’)
plot(density(t10000, bw=’SJ’), col=’grey50’,
     xlim=tlims, main=’t3,10000’)
```



Comment: With sample size increases, the density plot becomes more symmetric and densities at tails get larger.

```
savePar <- par(mfrow=c(2,2))
plot(density(c50, bw='SJ'), col='grey50',
     xlim=clims, main='chi3,50')
plot(density(c100, bw='SJ'), col='grey50',
     xlim=clims, main='chi3,100')
plot(density(c1000, bw='SJ'), col='grey50',
     xlim=clims, main='chi3,1000')
plot(density(c10000, bw='SJ'), col='grey50',
     xlim=clims, main='chi3,10000')
```



Comment: With sample size increases, the data get more dense at left and the right tail gets heavier.