Assignment 1

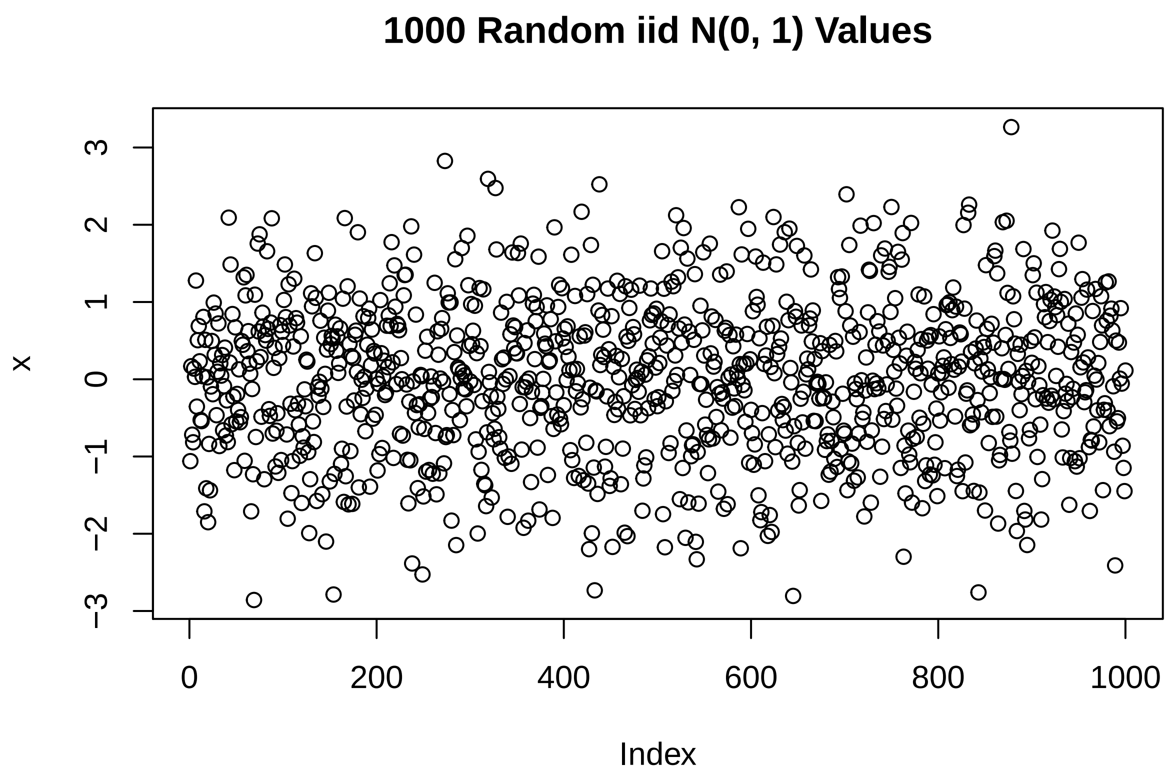
1.a

Family of Normal Functions

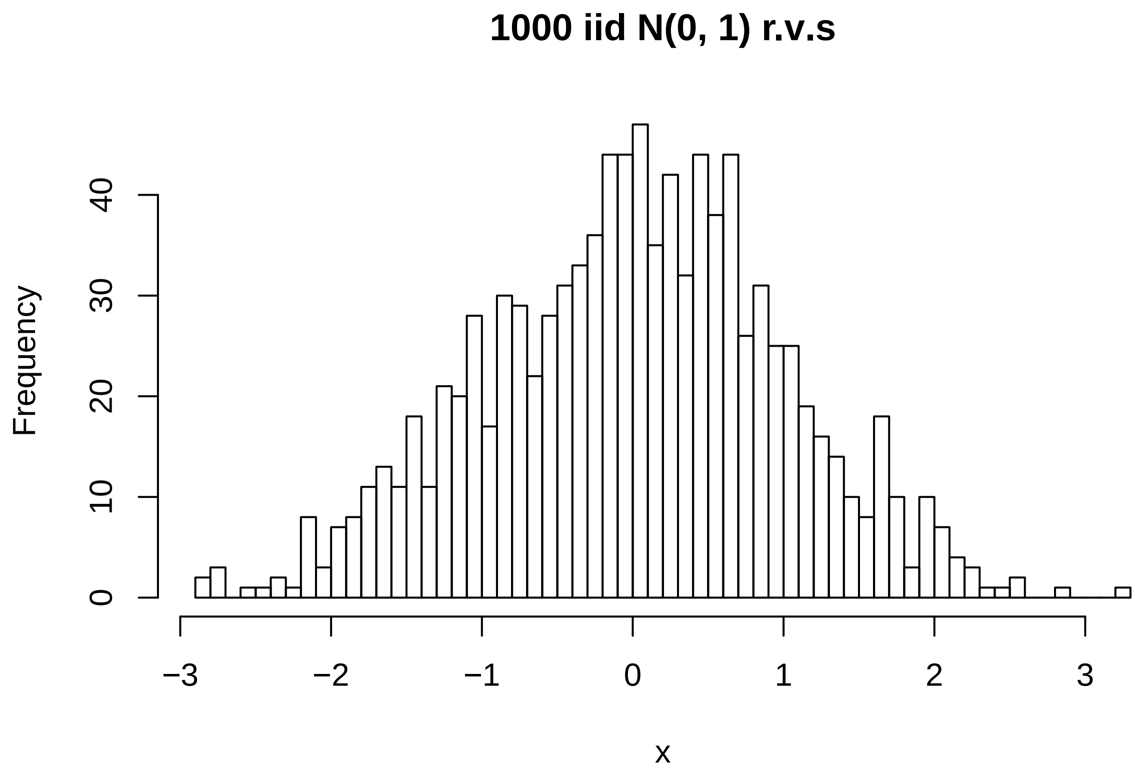
rnorm() is the random number generator

```r
# Generate a vector x of 1000 iid N(0, 1) values.
x = rnorm(n=1000, mean=0, sd=1)

# Create a plot of x to visualise the data
plot(x, main="1000 Random iid N(0, 1) Values")
```

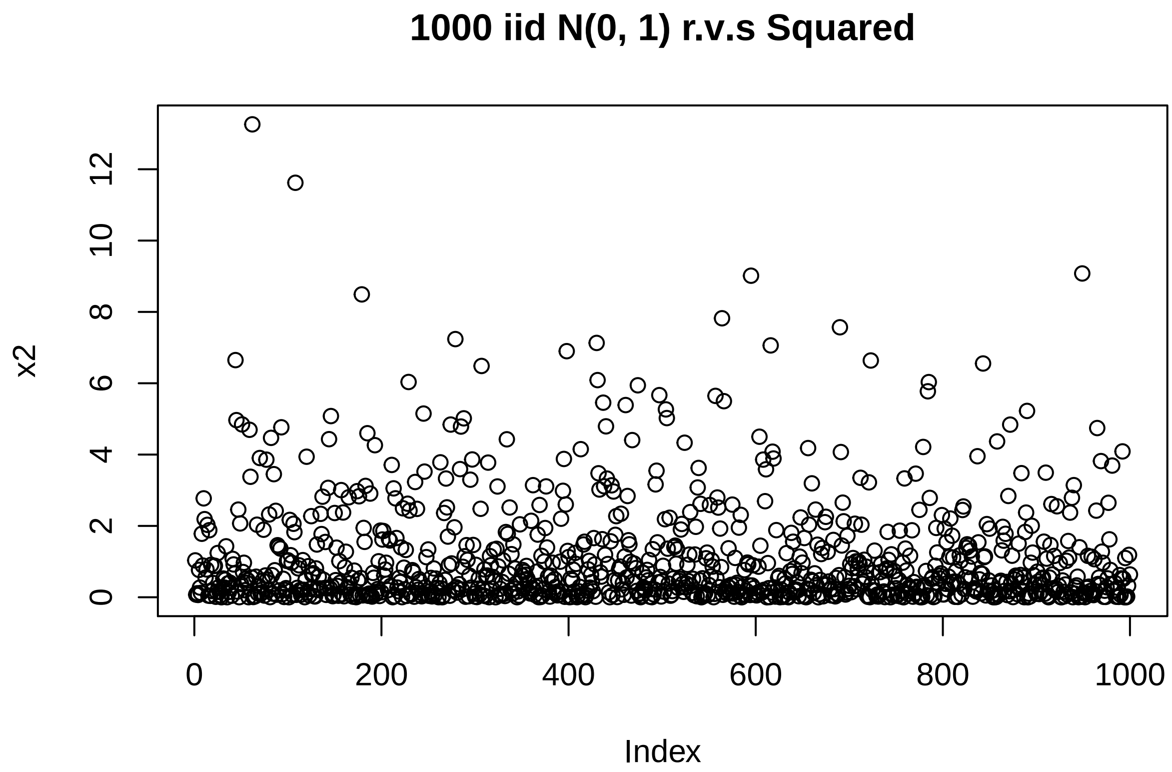



```r
# Create a histogram of x to visualise the data.
hist(x, main="1000 iid N(0, 1) r.v.s", nclass = 50)
```

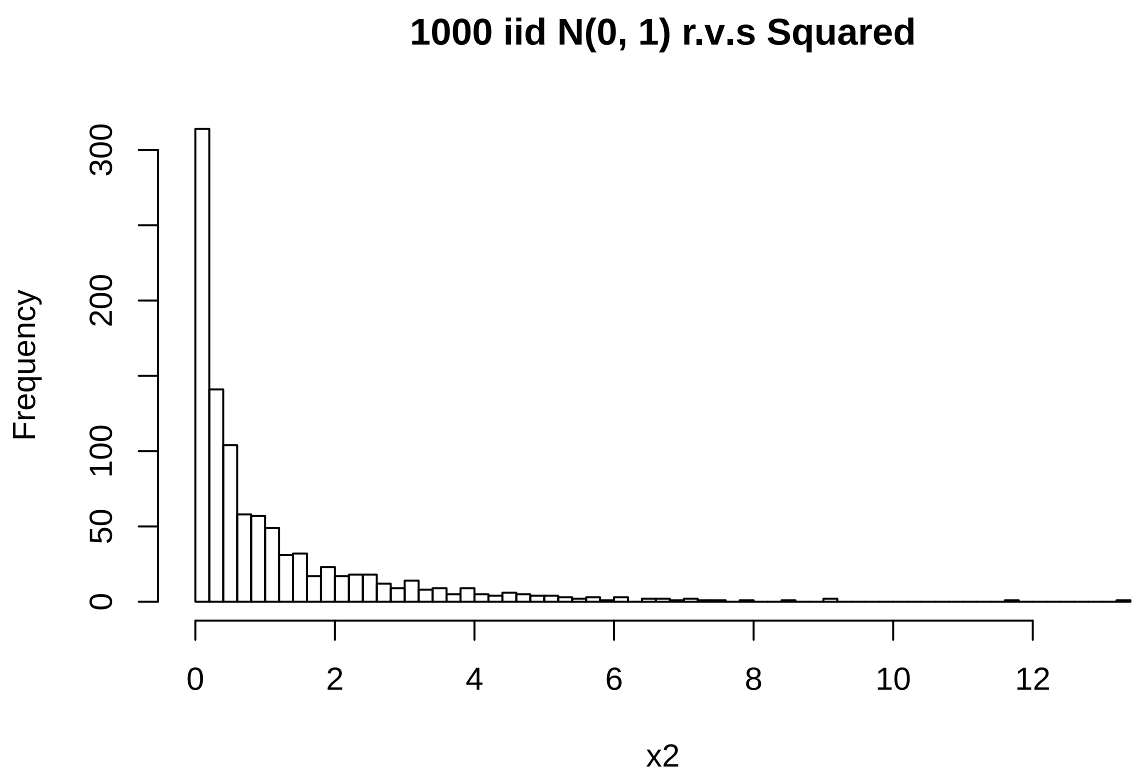



```r
# Firstly we generate a vector x of 1000 iid N(0, 1) values.
x2 = (rnorm(n=1000, mean=0, sd=1))^2

# Create a plot of x to visualise the data
plot(x2, main="1000 iid N(0, 1) r.v.s Squared")
```




```
# Create a histogram of x to visualise the data.
hist(x2, main="1000 iid N(0, 1) r.v.s Squared", nclass = 50)
```



dnorm() is the density function

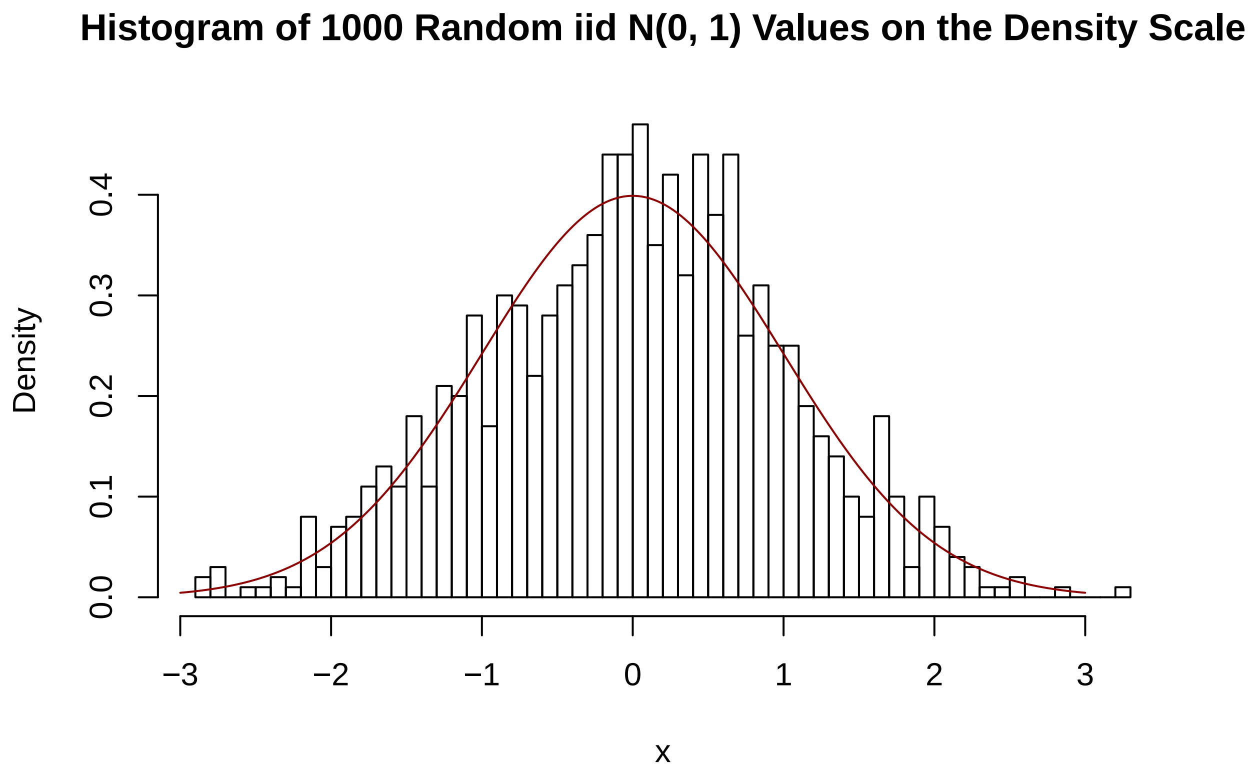
```
# Using dnorm() to plot the PDF of x. ***Please do not mark***.

# Plot the histogram of x on the density scale.
hist(x,
     probability = T,
     main="Histogram of 1000 Random iid N(0, 1) Values on the Density Scale",
     nclass = 50)

# Create a grid of x values.
x_grid=seq(-3,3,.01)

# Calculate the vector of the corresponding values of x.
density_grid=dnorm(x_grid, mean =0, sd=1)

# Use the lines function to superimpose a line onto the histogram.
lines(x_grid, density_grid, col='dark red', cex=3)
```

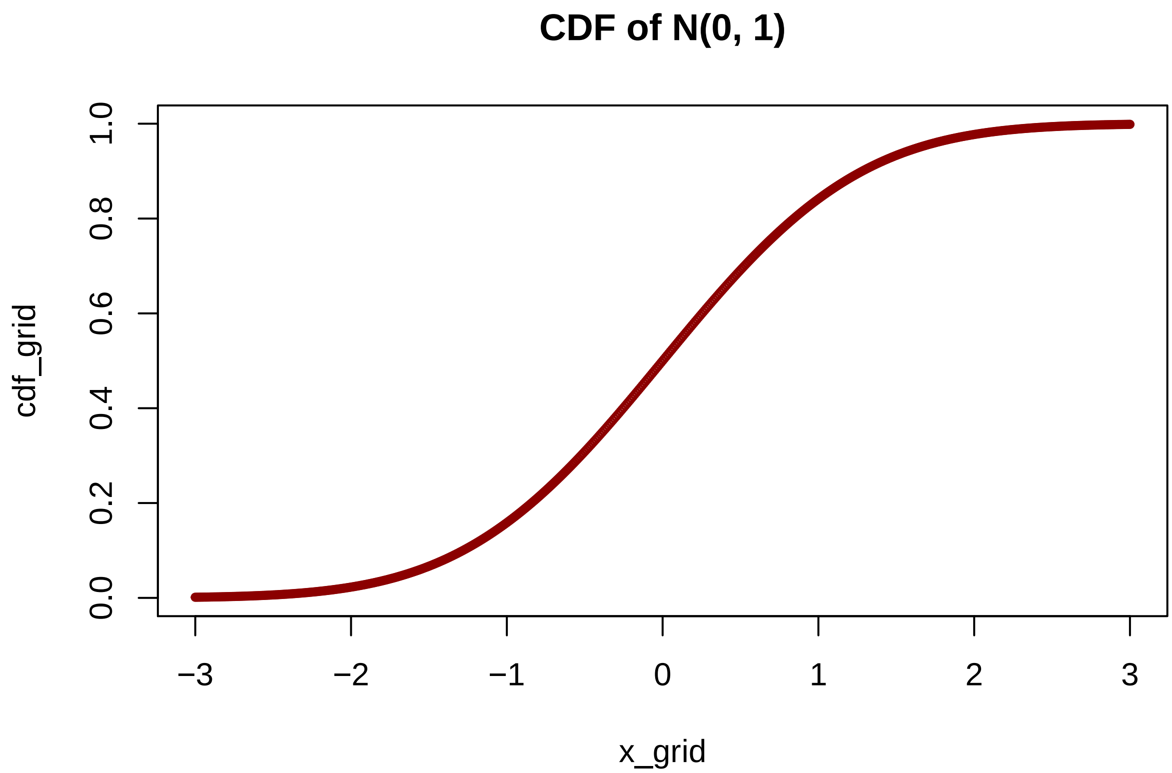


pnorm() is the cumulative distribution function

```
# Using pnorm() to plot the CDF of x. ***Please do not mark***.

# Calculate the cdf.
cdf_grid=pnorm(x_grid, mean=0, sd=1)

# Create a plot of the cdf.
plot(x_grid, cdf_grid, col='dark red', cex = .5,
     main="CDF of N(0, 1)")
```



qnorm() is the quantile function

```r
# Using qnorm() to find various quantiles. ***Please do not mark***.

qnorm(0.95)
```

```
## [1] 1.644854
```

```r
qnorm(0.975)
```

```
## [1] 1.959964
```

```r
qnorm(0.995)
```

```
## [1] 2.575829
```

```r
qnorm(0.9995)
```

```
## [1] 3.290527
```

```r
qnorm(1-0.95)
```

```
## [1] -1.644854
```

```r
qnorm(1-0.975)
```

```
## [1] -1.959964
```

```r
qnorm(1-0.995)
```

```
## [1] -2.575829
```

```r
qnorm(1-0.9995)
```

```
## [1] -3.290527
```

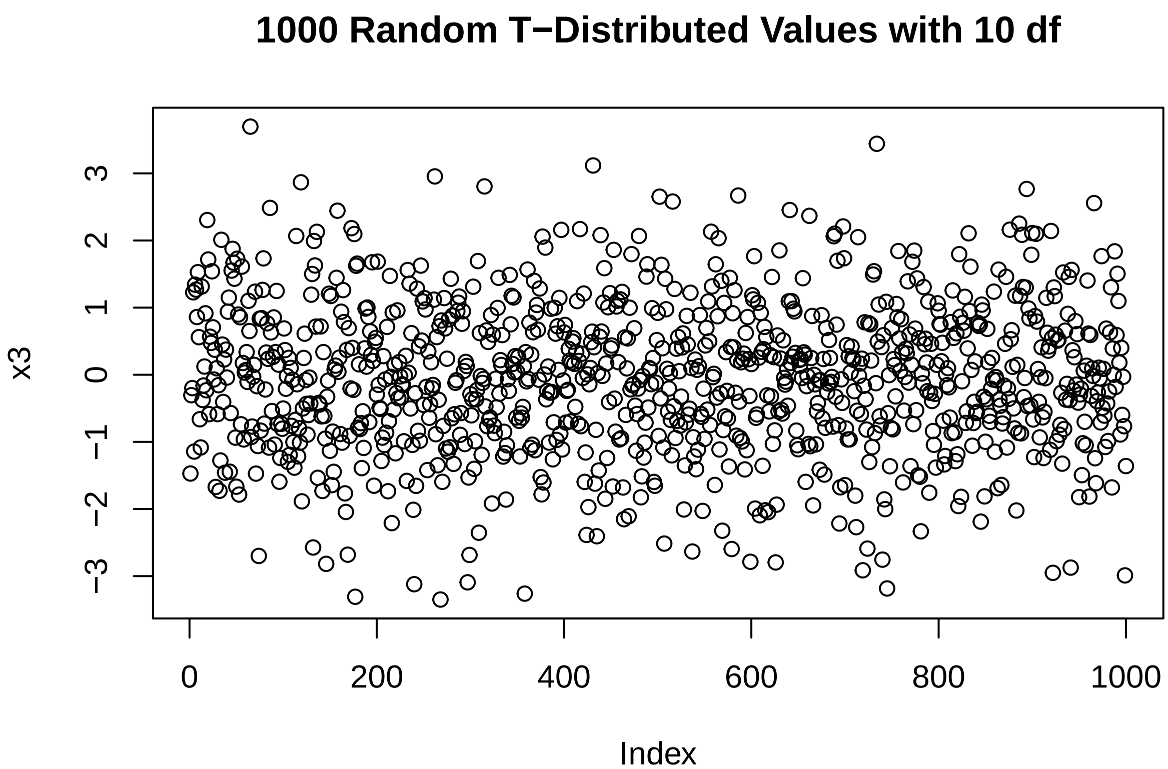
1.b

Family of T Functions

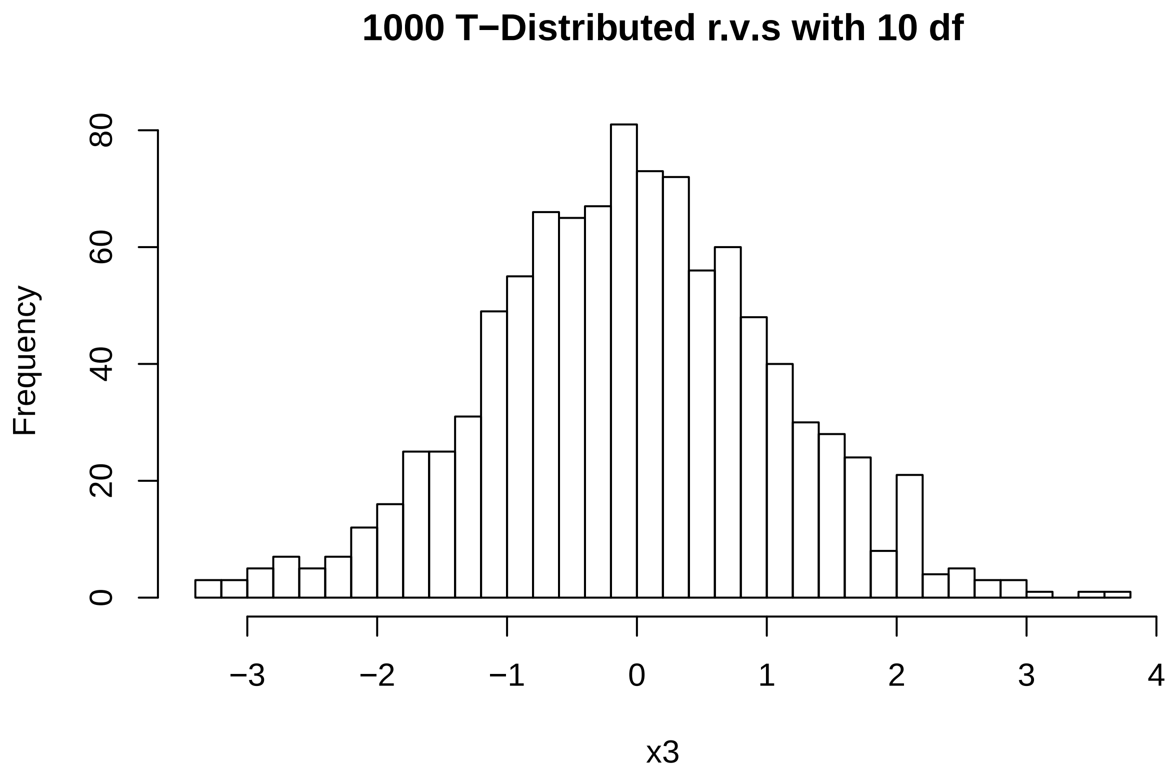
rt() is the random number generator

```r
# Generate a vector x of 1000 T-Distributed values.
x3 = rt(n=1000, df=10)

# Create a plot of x to visualise the data
plot(x3, main="1000 Random T-Distributed Values with 10 df")
```

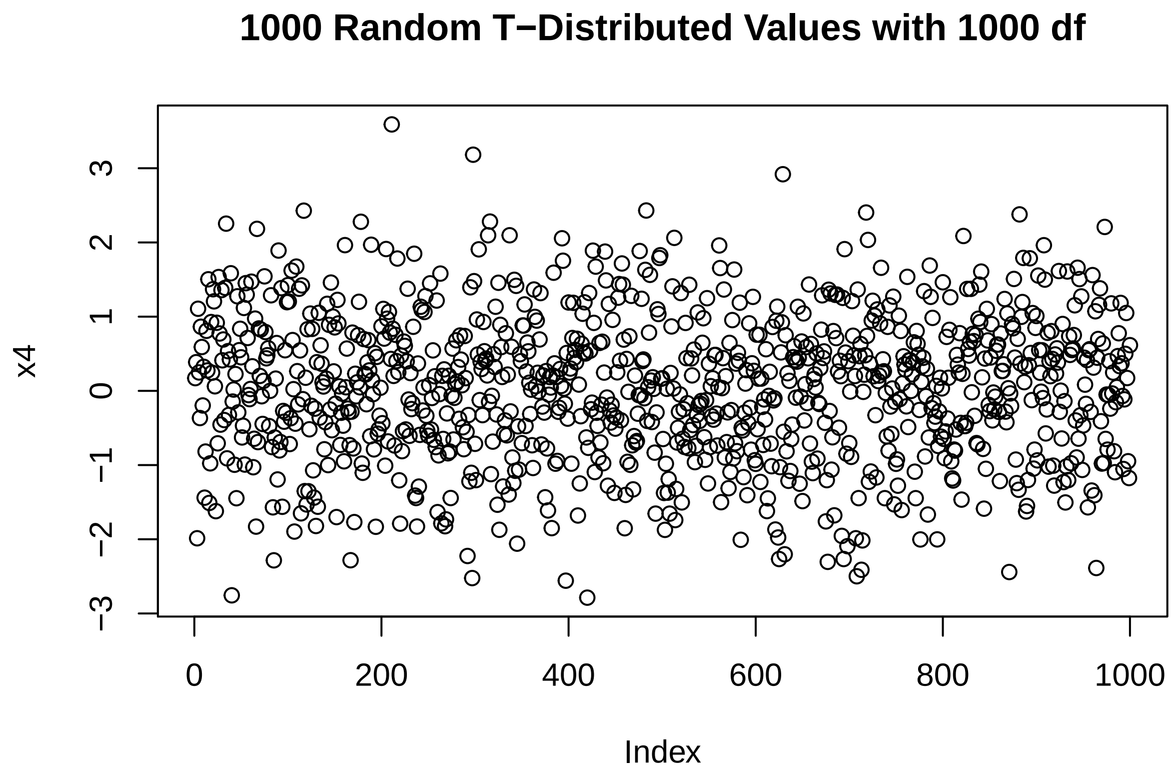



```r
# Create a histogram of x to visualise the data.
hist(x3, main="1000 T-Distributed r.v.s with 10 df", nclass = 50)
```

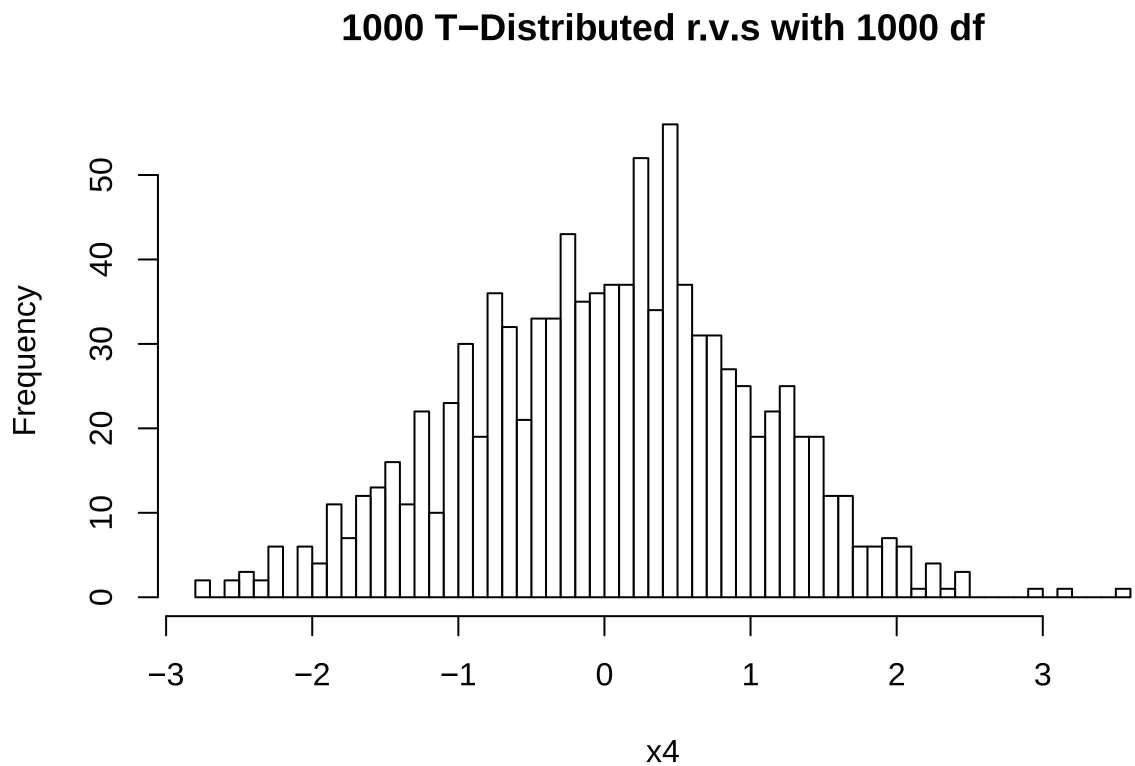



```r
# Generate a vector x of 1000 iid N(0, 1) values.
x4 = rt(n=1000, df=1000)

# Create a plot of x to visualise the data
plot(x4, main="1000 Random T-Distributed Values with 1000 df")
```




```r
# Create a histogram of x to visualise the data.
hist(x4, main="1000 T-Distributed r.v.s with 1000 df", nclass = 50)
```



Question: "How do these 2 histograms compare with the first histogram from part (a)?"

Answer: There is a convergence in distribution between a Student's T distribution with mean μ, n degrees of freedom (df) and scale σ^2 and a Normal distribution with mean μ and variance σ^2 if the degrees of freedom n becomes large (a convergence to infinity).

Therefore, the histogram with 1000 df is more normally distributed than the histogram with 10 df.

This can be seen in subsection dt() the density function.

Proof: Let X_n be a t-distributed random variable (r.v) and be given by:

$$X_n = \mu + \sigma \frac{Y}{\sqrt{\chi_n^2/n}}$$

Where Y in $N(0, 1)$, and χ_n^2 is a Chi-square r.v with n df and is independent of Y.

Furthermore, χ_n^2 can be stated as a sum of identically and independently distributed $N(0, 1)$ r.v.s $Z_1, \ldots, Z_n$:

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2$$

Dividing both sides by n:

$$\frac{\chi_n^2}{n} = \frac{1}{n} \sum_{i=1}^{n} Z_i^2$$

As n approaches infinity there is a convergence in the expected value of the probability of Z_i^2 to one via the law of large numbers:

$$\mathrm{E}\left(Z_i^2\right) = 1$$

Therefore, X_n converges in its distribution to:

$$X = \mu + \sigma Y$$

Which is $N(\mu, \sigma^2)$.[1]

dt() is the density function

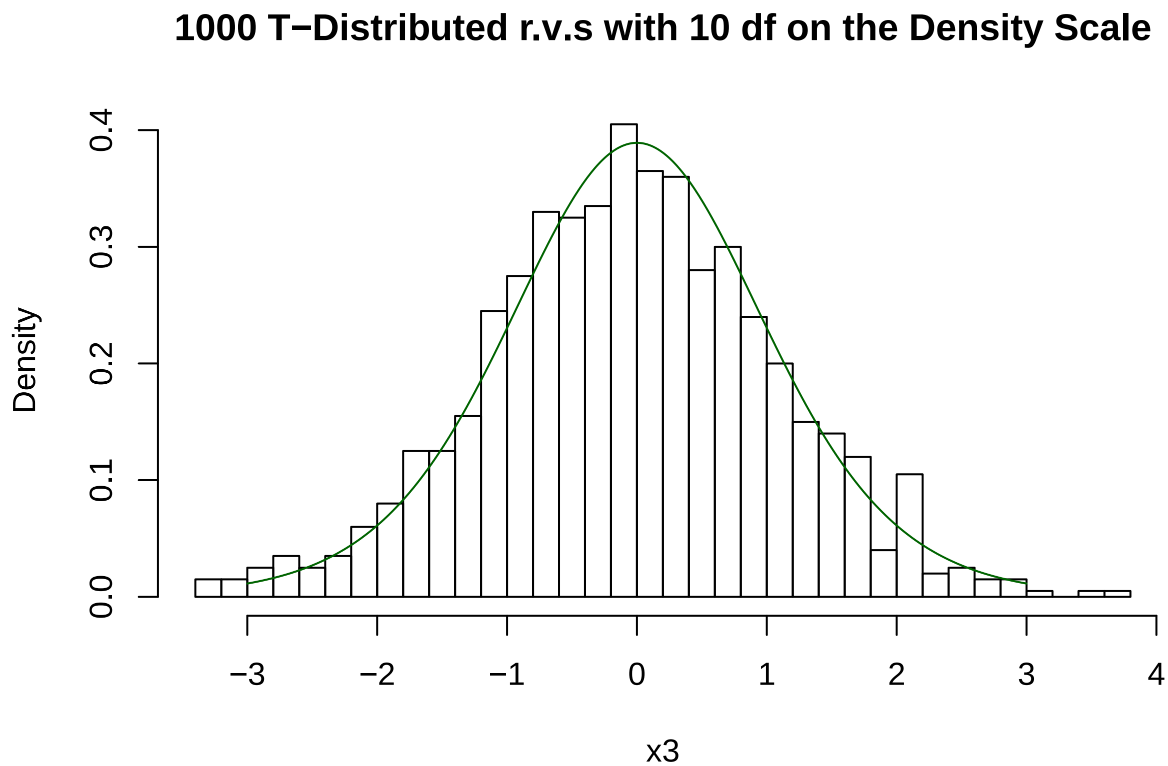
```
# Using dt() to plot the PDF of x. ***Please do not mark***.

# Plot the histogram of x on the density scale.
hist(x3,
     probability = "TRUE",
     main="1000 T-Distributed r.v.s with 10 df on the Density Scale",
     nclass = 50)

# Create a grid of x values.
x_grid=seq(-3,3,.01)

# Calculate the vector of the corresponding values of x.
density_dt_grid=dt(x_grid, df = 10)

# Use the lines function to superimpose a line onto the histogram.
lines(x_grid, density_dt_grid, col='dark green', cex=3)
```



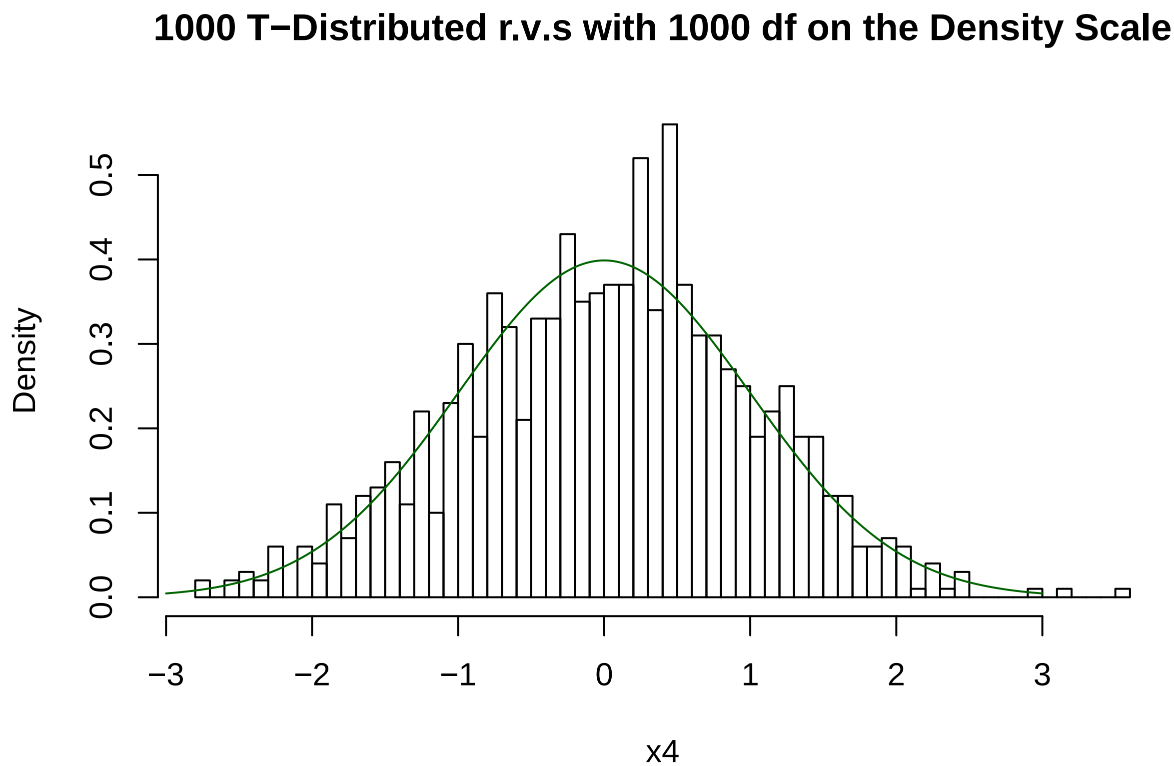

```
# Using dt() to plot the PDF of x. ***Please do not mark***.

# Plot the histogram of x on the density scale.
hist(x4,
     probability = "TRUE",
     main="1000 T-Distributed r.v.s with 1000 df on the Density Scale",
     nclass = 50)

# Create a grid of x values.
x_grid=seq(-3,3,.01)

# Calculate the vector of the corresponding values of x.
density_dt_grid=dt(x_grid, df = 1000)

# Use the lines function to superimpose a line onto the histogram.
lines(x_grid, density_dt_grid, col='dark green', cex=3)
```

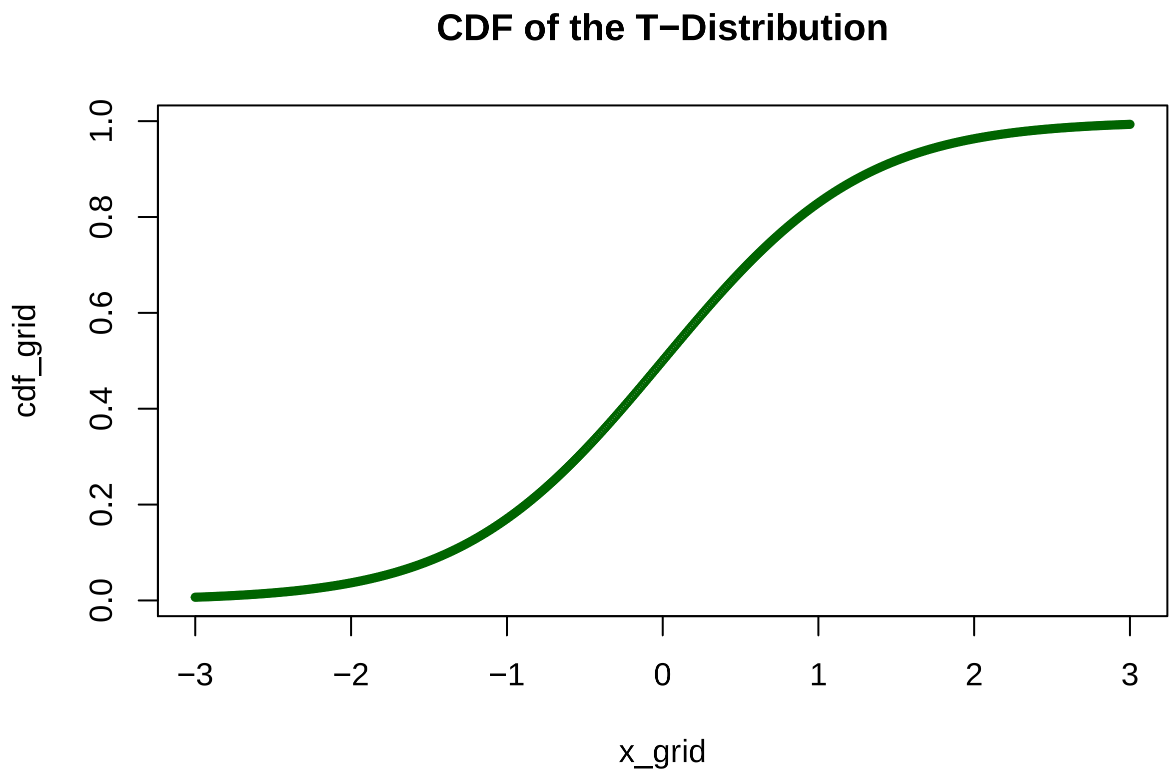


pt() is the cumulative distribution function

```
# Using pt() to plot the CDF of x. ***Please do not mark***.

# Calculate the cdf.
cdf_grid=pt(x_grid, df = 10)

# Create a plot of the cdf.
plot(x_grid, cdf_grid, col='dark green', cex = .5,
     main= "CDF of the T-Distribution")
```



qt() is the quantile function

```r
# Using qt() to find various quantiles. ***Please do not mark***.

qt(c(0.05, 0.95), df = 10)
```

```
## [1] -1.812461  1.812461
```

```r
qt(c(0.025, 0.975), df = 10)
```

```
## [1] -2.228139  2.228139
```

```r
qt(c(0.005, 0.995), df = 10)
```

```
## [1] -3.169273  3.169273
```

```r
qt(c(0.0005, 0.9995), df = 10)
```

```
## [1] -4.586894  4.586894
```

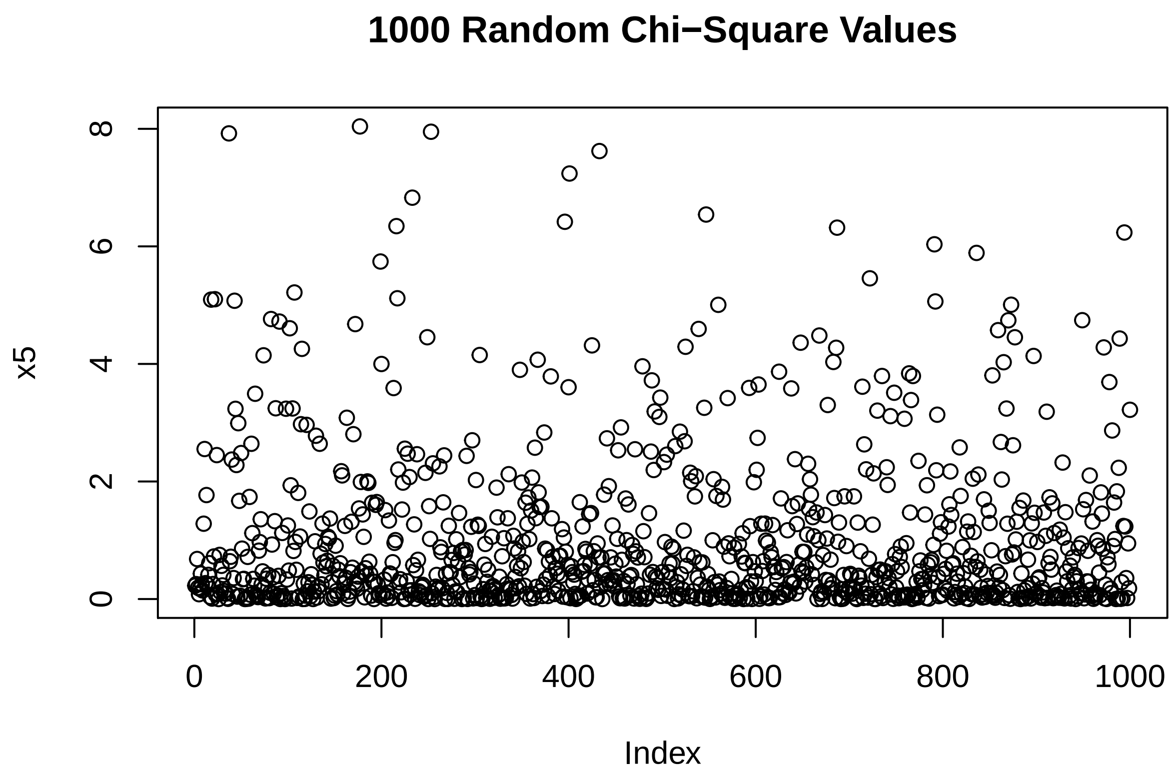
1.c

Family of Chisquare Functions

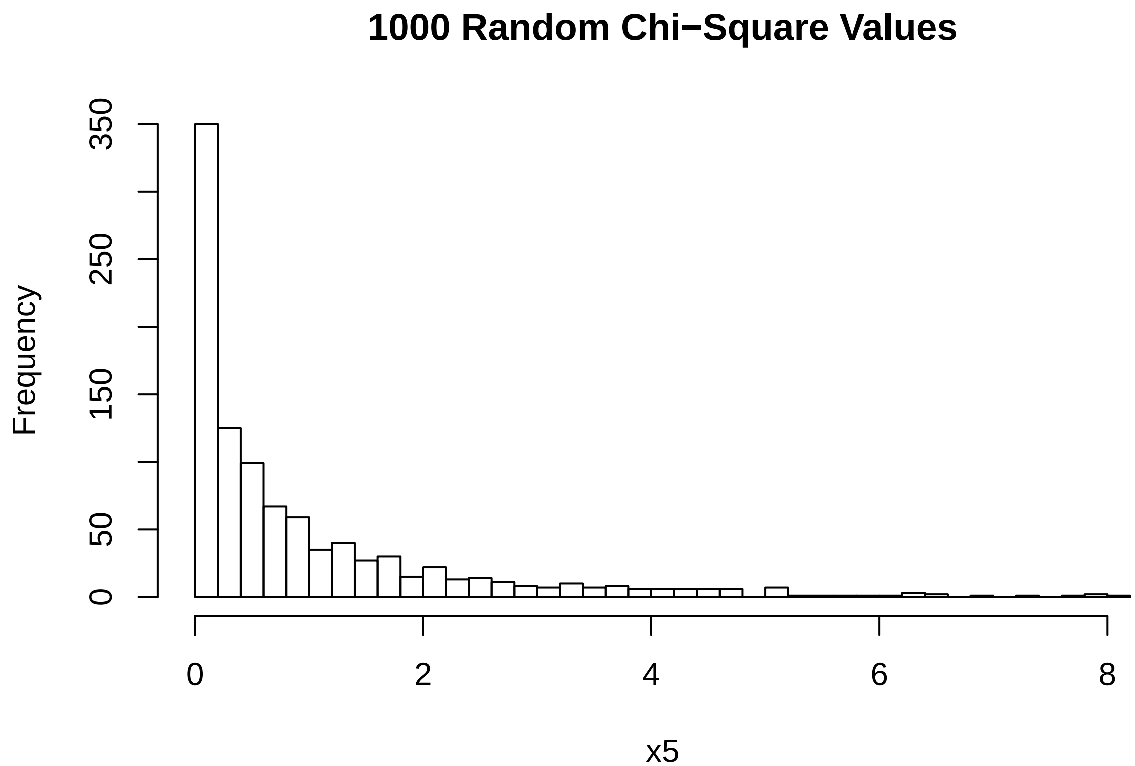
rchisq() is the random number generator

```
# Generate a vector x of 1000 chisquare r.v.s.
x5 = rchisq(n=1000, df=1)

# Create a plot of x to visualise the data
plot(x5, main="1000 Random Chi-Square Values")
```




```
# Create a histogram of x to visualise the data.
hist(x5, main="1000 Random Chi-Square Values", nclass=50)
```



Question: "How does this histogram compare with the second histogram from part (a)?"

Answer: The histograms will be similar in appearance because the sum of iid $N(0,1)$ r.v.s is χ^2_n. Letting Z be $N(0,1)$ with square X:

$$X = Z^2$$

It follows that X is a χ^2_n r.v with 1 df. [2]

dchisq() is the density function

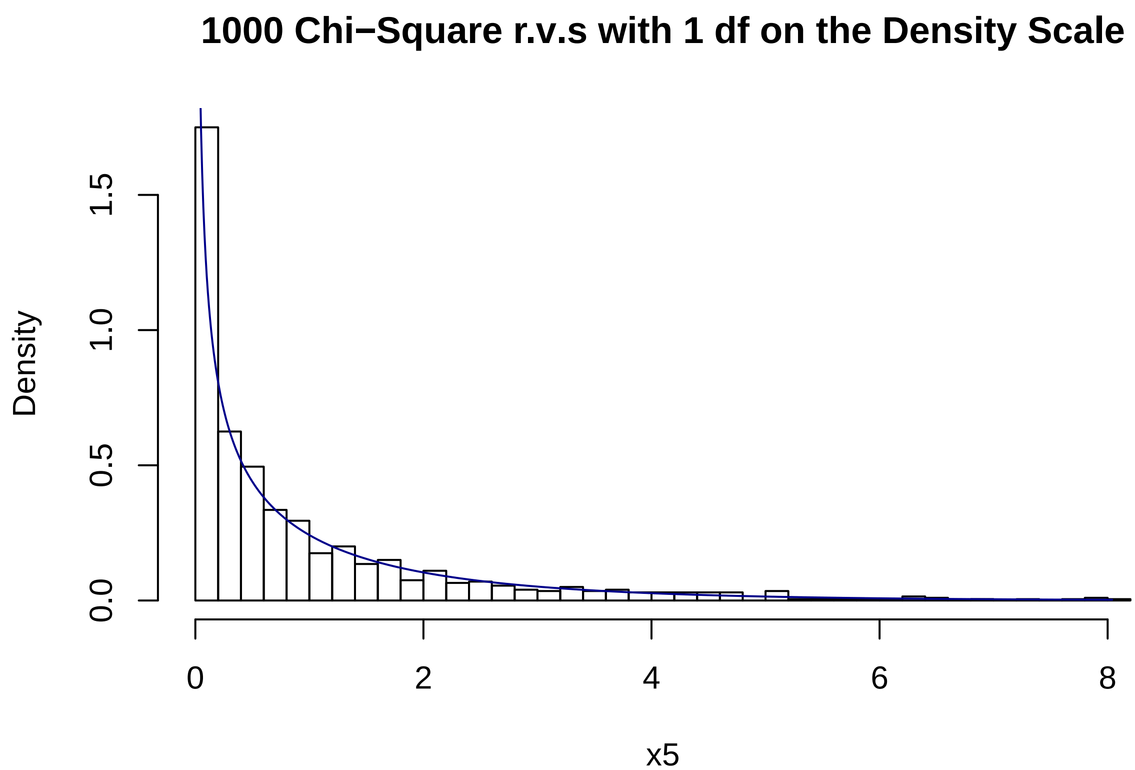
```
# Using dchisq() to plot the PDF of x. ***Please do not mark***.

# Plot the histogram of x on the density scale.
hist(x5,
     probability = "TRUE",
     main="1000 Chi-Square r.v.s with 1 df on the Density Scale",
     nclass=50)

# Create a grid of x values.
x_chisq_grid=seq(0, max(x5), 0.01)

# Calculate the vector of the corresponding values of x.
density_chisq_grid=dchisq(x_chisq_grid, df = 1)
```

```r
# Use the lines function to superimpose a line onto the histogram.
lines(x_chisq_grid, density_chisq_grid, col='dark blue', cex=3)
```

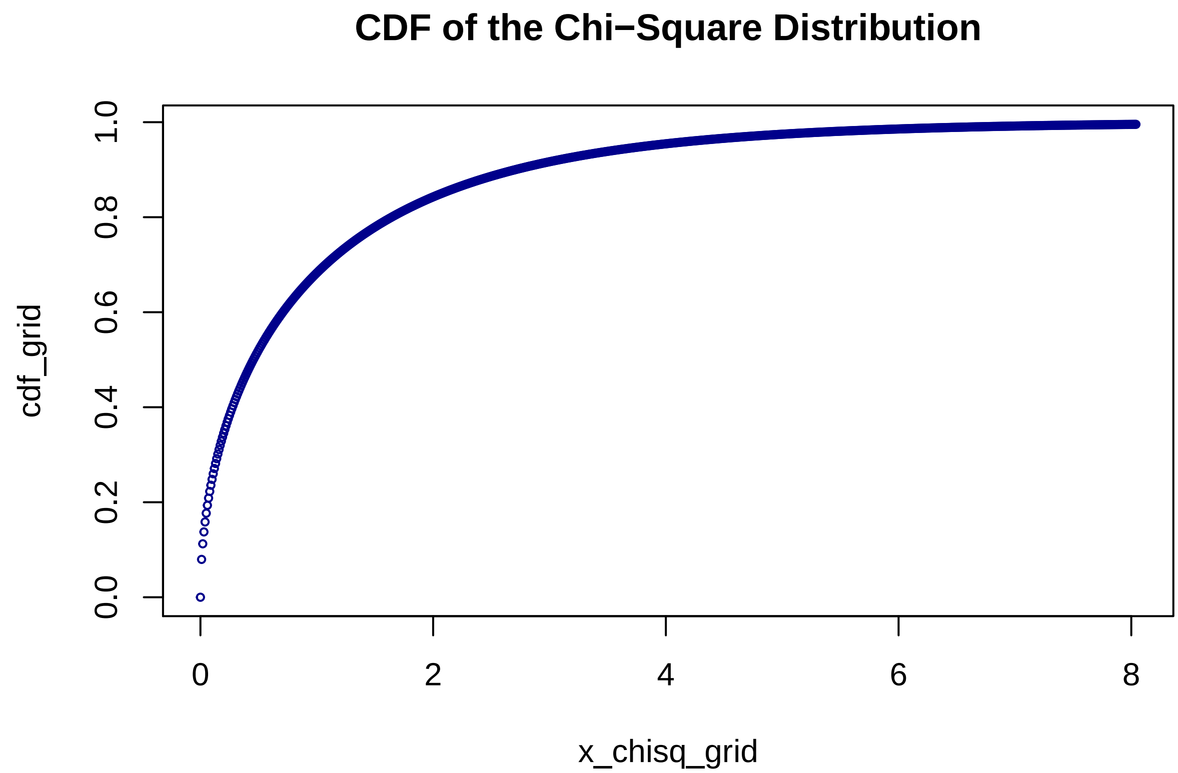


pchisq() is the cumulative distribution function

```r
# Using pchisq() to plot the CDF of x. ***Please do not mark***.

# Calculate the cdf.
cdf_grid=pchisq(x_chisq_grid, df = 1)

# Create a plot of the cdf.
plot(x_chisq_grid, cdf_grid, col='dark blue', cex = .5,
     main="CDF of the Chi-Square Distribution")
```



qchisq() is the quantile function

```r
# Using qt() to find various quantiles. ***Please do not mark***.

qchisq(0.0005, df = 1)
```

```
## [1] 3.926991e-07
```

```r
qchisq(0.005, df = 1)
```

```
## [1] 3.927042e-05
```

```r
qchisq(0.025, df = 1)
```

```
## [1] 0.0009820691
```

```r
qchisq(0.05, df = 1)
```

```
## [1] 0.00393214
```

```r
qchisq(0.05, df = 1)
```

```
## [1] 0.00393214
```

```r
qchisq(0.95, df = 1)
```

```
## [1] 3.841459
```

```r
qchisq(0.975, df = 1)
```

```
## [1] 5.023886
```

```r
qchisq(0.995, df = 1)
```

```
## [1] 7.879439
```

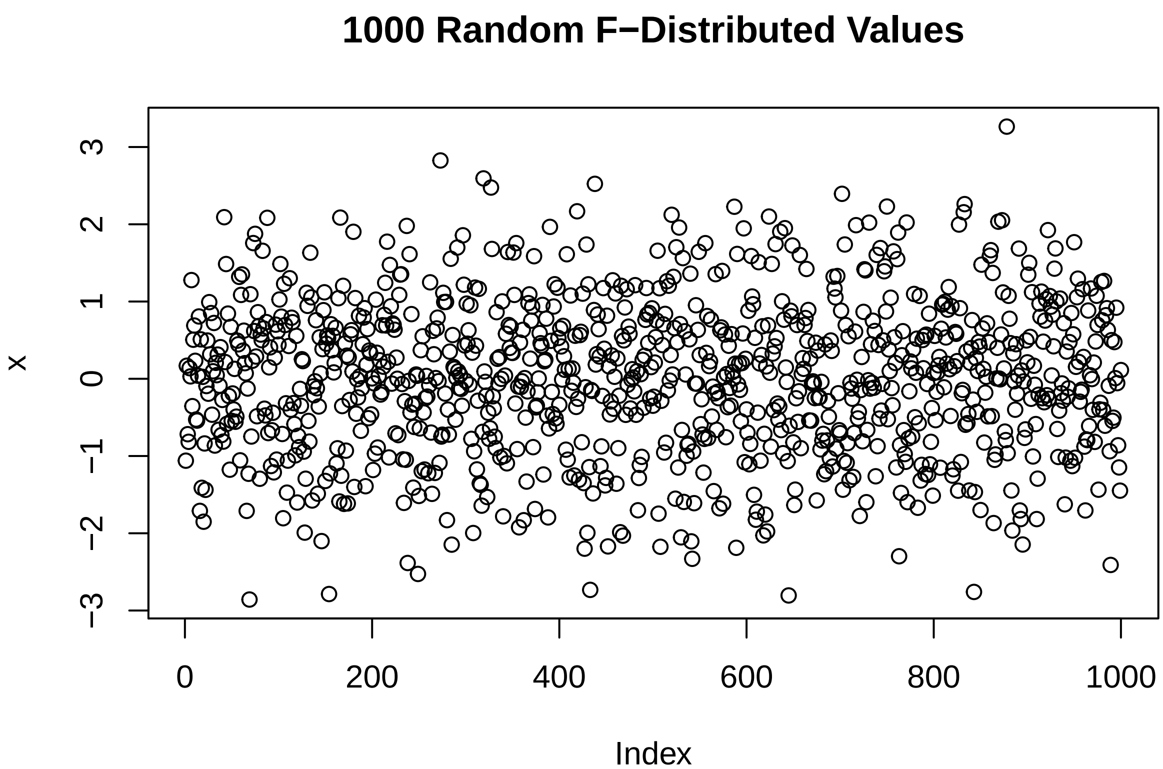
1.d

Family of F Functions

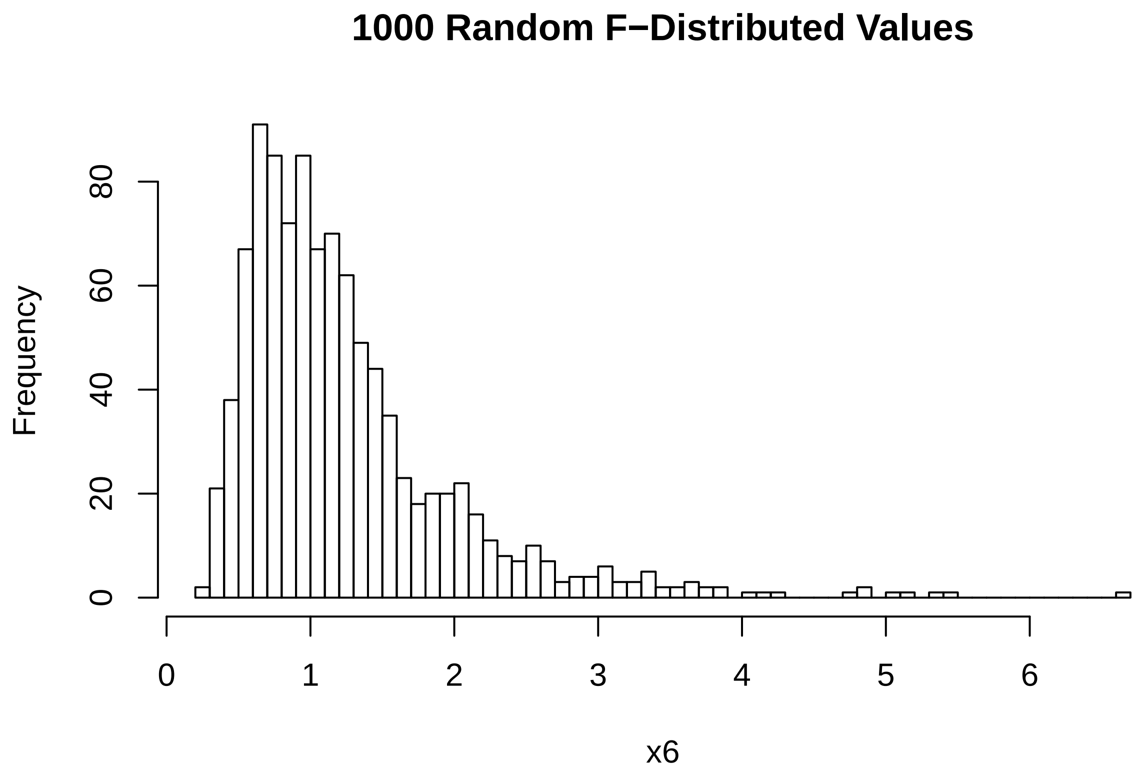
rf() is the random number generator

```r
# Generate a vector x of 1000 F-Distributed values.
x6 = rf(n=1000, df1 = 50, df2 = 10)

# Create a plot of x to visualise the data
plot(x, main="1000 Random F-Distributed Values")
```




```r
# Create a histogram of x to visualise the data.
hist(x6, main="1000 Random F-Distributed Values", nclass=50)
```



df() is the density function

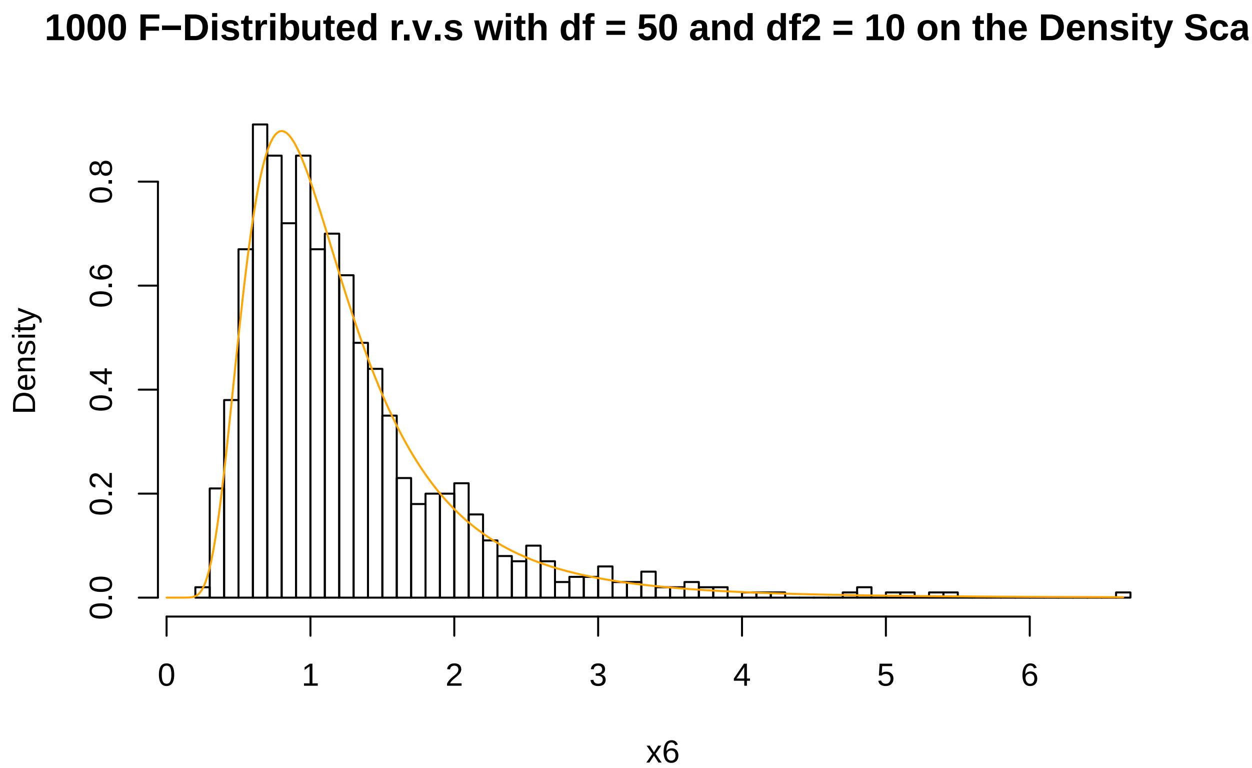
```r
# Using df() to plot the PDF of x. ***Please do not mark***.

# Plot the histogram of x on the density scale.
hist(x6,
     probability = "TRUE",
     main="1000 F-Distributed r.v.s with df = 50 and df2 = 10 on the Density Scale",
     nclass=50)

# Create a grid of x values.
x_f_grid=seq(0, max(x6), 0.01)

# Calculate the vector of the corresponding values of x.
density_f_grid=df(x_f_grid, df1 = 50, df2 = 10)

# Use the lines function to superimpose a line onto the histogram.
lines(x_f_grid, density_f_grid, col='orange', cex=3)
```

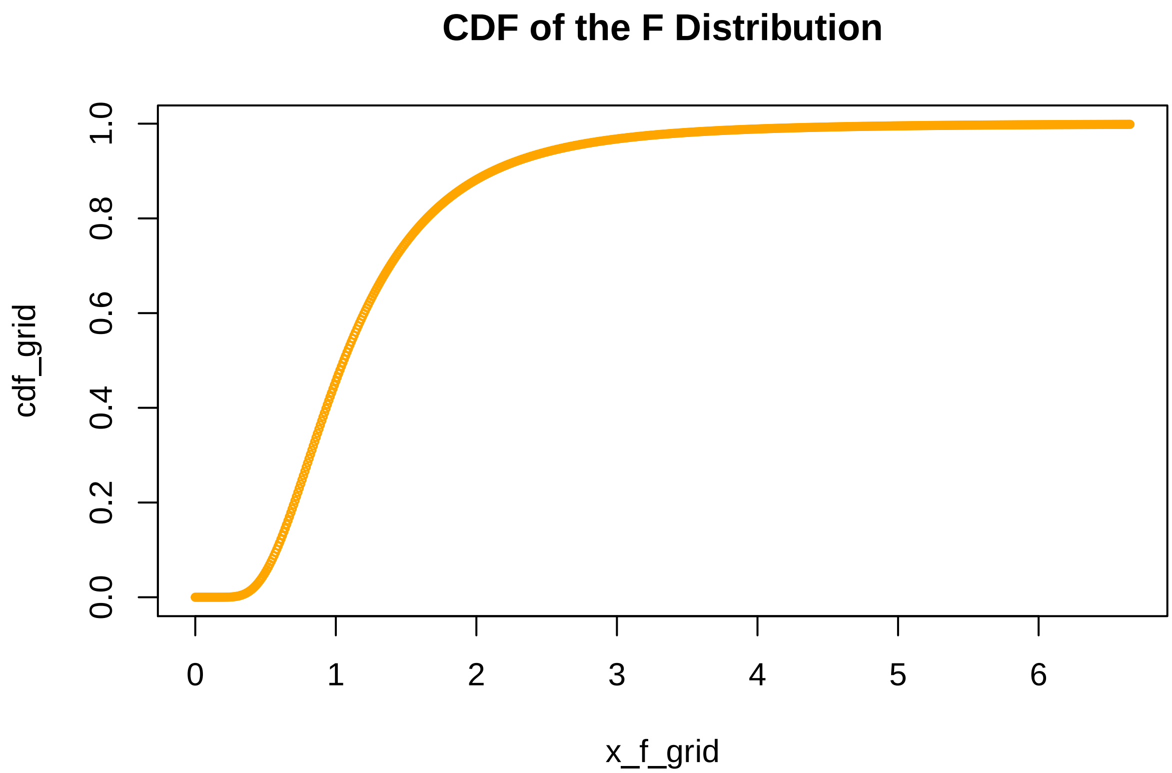


pf() is the cumulative distribution function

```
# Using pf() to plot the CDF of x. ***Please do not mark***.

# Calculate the cdf.
cdf_grid=pf(x_f_grid, df1 = 50, df2 = 10)

# Create a plot of the cdf.
plot(x_f_grid, cdf_grid, col='orange', cex = .5,
     main="CDF of the F Distribution")
```



qf() is the quantile function

```r
# Using qf() to find various quantiles. ***Please do not mark***.

qf(0.0005, df1 = 50, df2 = 10)
```

```
## [1] 0.2517149
```

```r
qf(0.005, df1 = 50, df2 = 10)
```

```
## [1] 0.3347259
```

```r
qf(0.025, df1 = 50, df2 = 10)
```

```
## [1] 0.4316309
```

```r
qf(0.05, df1 = 50, df2 = 10)
```

```
## [1] 0.4935486
```

```r
qf(0.05, df1 = 50, df2 = 10)
```

```
## [1] 0.4935486
```

```r
qf(0.95, df1 = 50, df2 = 10)
```

```
## [1] 2.637124
```

```r
qf(0.975, df1 = 50, df2 = 10)
```

```
## [1] 3.221372
```

```r
qf(0.995, df1 = 50, df2 = 10)
```

```
## [1] 4.902156
```

2.a

```r
data <- c(11.4, 15.1, 20.3,
          12.0, 17.2, 21.5,
          12.1, 14.8, 21.4,
          13.0, 16.7, 21.3,
          12.1, 13.8, 22.0,
          12.5, 14.2, 20.9,
          11.8, 15.7, 24.4,
          11.7, 16.1, 21.1,
          12.2, 13.2, 22.7,
          10.8, 15.8, 21.9)

m <- matrix(data = data, nrow = 10, ncol = 3, byrow = TRUE)

m.t <- t(m)

Tube <- m.t[1,]
Bus <- m.t[2,]
Bike <- m.t[3,]

data <- c(Tube, Bus, Bike)

t <- rep(c("Tube", "Bus", "Bike"), each=10)

transport <- factor(t)

type_of_transport <- data.frame(data, transport)
```

2.b

```r
type_of_transport_aov <- aov(data ~ transport, data = type_of_transport)
summary(type_of_transport_aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## transport    2  496.2   248.1     226 <2e-16 ***
## Residuals   27   29.6     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case the value of the F-statistic is 226 with p-value <2e-16, so we reject the null hypothesis at the .1% significance level (also 1% and 5% respectively).

We deduce that there is very strong evidence indeed that the type of transport has an effect on the length of commute.

2.c

Let $X_1, X_2, \ldots, X_{10} \sim N\left(\mu_X, \sigma_X^2\right)$, $Y_1, Y_2, \ldots, Y_{10} \sim N\left(\mu_Y, \sigma_Y^2\right)$ and $Z_1, Z_2, \ldots, Z_{10} \sim N\left(\mu_Z, \sigma_Z^2\right)$ be independent random samples for tube, bus and bike respectively.

```
sum(Tube)
```

```
## [1] 119.6
```

```
sum(Bus)
```

```
## [1] 152.6
```

```
sum(Bike)
```

```
## [1] 217.5
```

```
# User defined function to find the sum of squares for the different transport types.
sum_of_square <- function(transport_type){
        sum_ = sum(unlist(lapply(transport_type, function(x) x^2)))
        return(sum_)
}

sum_of_square(Tube)
```

```
## [1] 1433.64
```

```
sum_of_square(Bus)
```

```
## [1] 2343.44
```

```
sum_of_square(Bike)
```

```
## [1] 4742.27
```

The observed statistics are given by:

$$\sum_{i=1}^{10} x_i = 119.6, \sum_{i=1}^{10} x_i^2 = 1433.64, \sum_{j=1}^{10} y_j = 152.6, \sum_{j=1}^{8} y_j^2 = 2343.44, \sum_{k=1}^{10} z_k = 217.5, \sum_{k=1}^{10} z_k^2 = 4742.27$$

Where μ_i, σ_i^2 $i = X, Y, Z$ are unknown.

[1] Taboga, Marco (2017). "Student's t distribution", Lectures on probability theory and mathematical statistics, Third edition. Kindle Direct Publishing. Online appendix. https://www.statlect.com/probability-distributions/student-t-distribution.

[2] Taboga, Marco (2017). "Chi-square distribution", Lectures on probability theory and mathematical statistics, Third edition. Kindle Direct Publishing. Online appendix. https://www.statlect.com/probability-distributions/chi-square-distribution.