ST 502 HW3 Question 5

Halid Kopanski

2/2/2022

Data Creation and Set Up

```
library(ggplot2)

par(mfrow = c(2 , 2))

set.seed(1000)

N = 1000

df_data <- data.frame(cbind(rnorm(N, 0, 1),
                            rnorm(N, 0, 1),
                            rnorm(N, 0, 1))
                      )

x_names <- c("X1", "X2", "X3")

df_z <- data.frame(cbind(

  df_data[,1] + 2*df_data[,2] + 3*df_data[,3],

  df_data[,1]^2 + df_data[,2]^2 + df_data[,3]^2,

  0.5*(df_data[,1] - df_data[,3])^2,

  (2*df_data[,3]^2) / (df_data[,1]^2 + df_data[,2]^2)
                        )

                   )

z_names <- c("Za", "Zb", "Zc", "Zd")

colnames(df_z) <- z_names

chart_colors <- c("#003f5c", "#2f4b7c", "#665191", "#a05195",
                  "#d45087", "#f95d6a", "#ff7c43", "#ffa600")
```

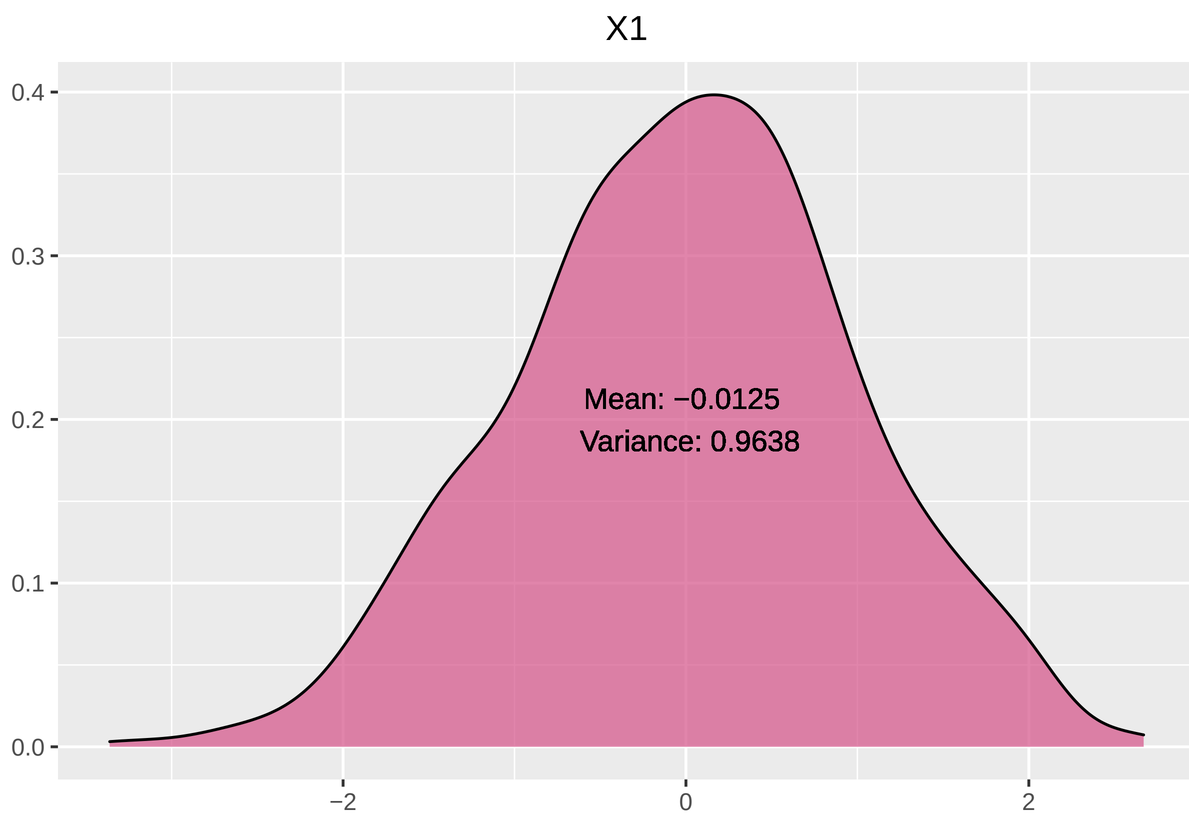
Plots of X random variables

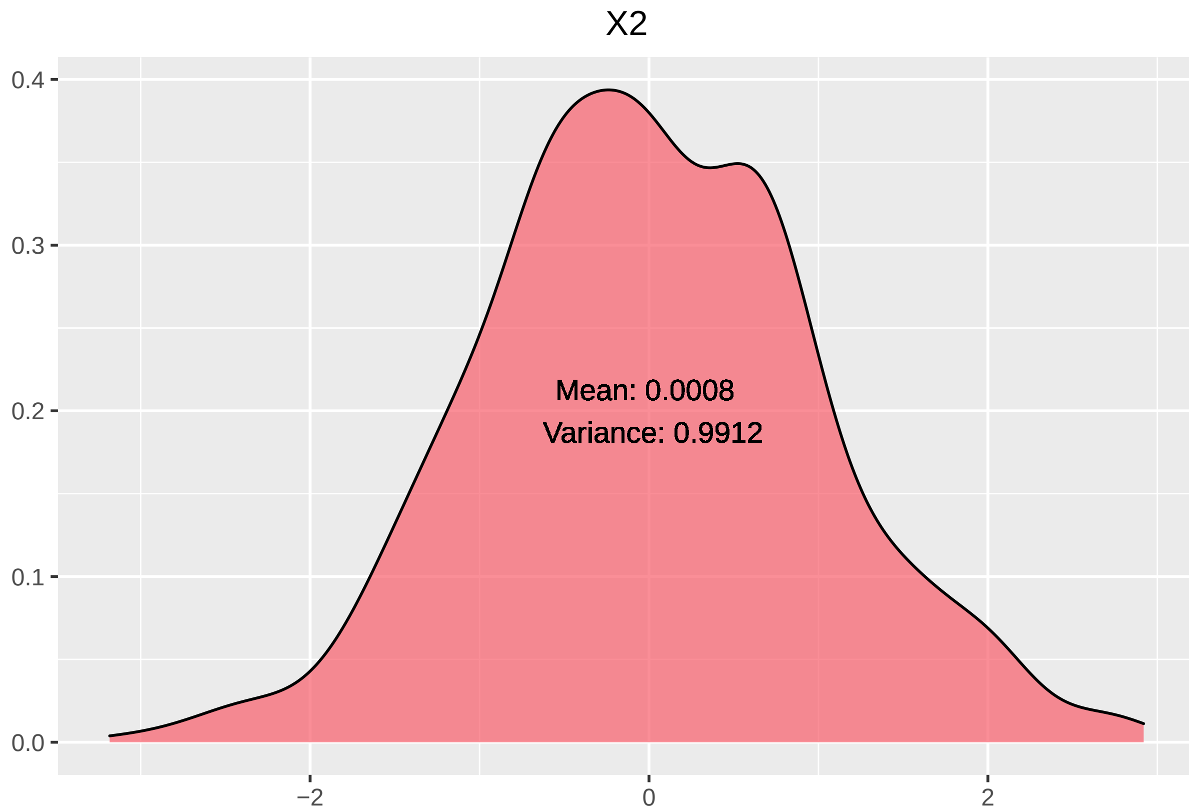
The mean and variance for each variable is quite close to 0 and 1, respectively. The reason for the values not be exact is due to noise in the generated data. Generally speaking, all three random variables follow a normal like distribution. Which was expected.

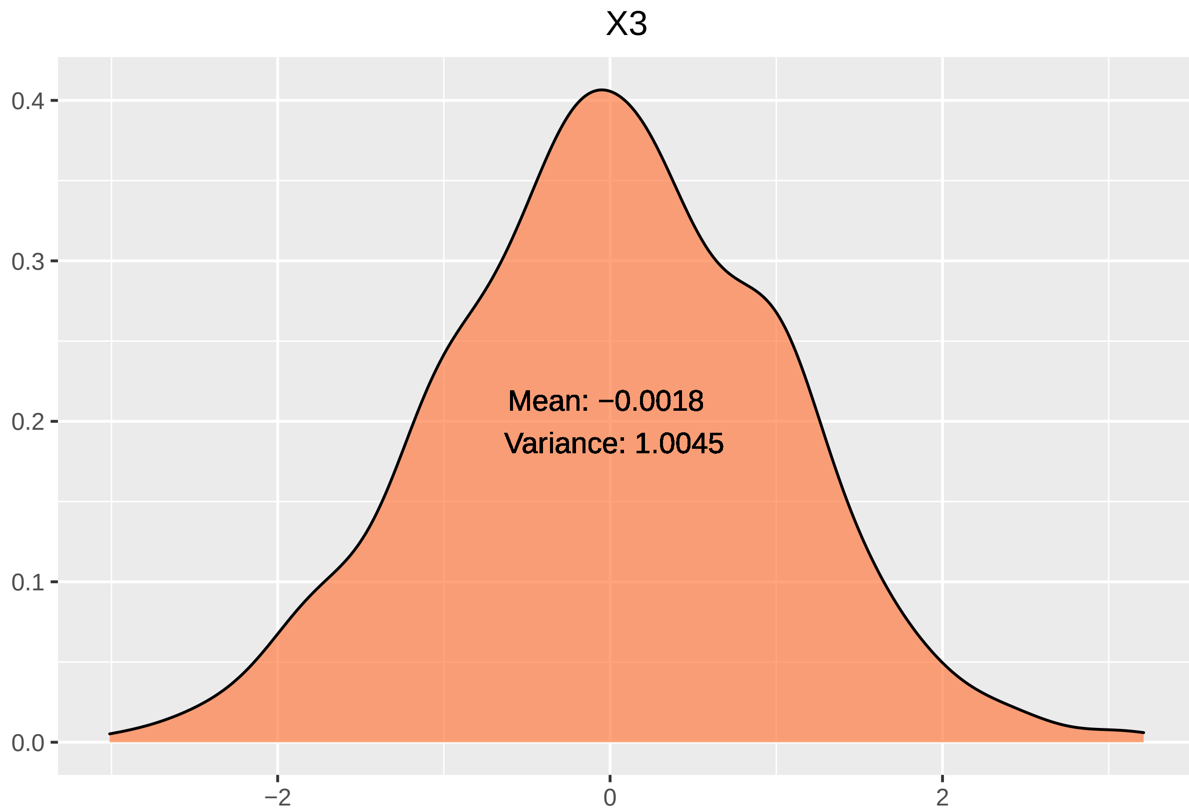
```
for (i in 1:3){
  stmnt <- sprintf("The mean of %s is %.4f and the variance is %.4f",
                   x_names[i],
                   mean(df_data[ ,i]),
                   var(df_data[ ,i])
                   )

  print(stmnt)
}
```

```
## [1] "The mean of X1 is -0.0125 and the variance is 0.9638"
## [1] "The mean of X2 is 0.0008 and the variance is 0.9912"
## [1] "The mean of X3 is -0.0018 and the variance is 1.0045"
```

```
for (i in 1:3){

  stmnt <- sprintf("Mean: %.4f \n Variance: %.4f",
                   mean(df_data[ ,i]),
                   var(df_data[ ,i])
                   )

  dplot <- ggplot(df_data, aes( x = df_data[, i])) +
          geom_density(fill = chart_colors[i + 4], alpha = 0.7) +
          ggtitle(paste0("X", i)) + xlab("") + ylab("") +
          theme(plot.title = element_text(hjust = 0.5)) +
          geom_text(x = 0, y = 0.2, label = stmnt) + facet_grid()

  print(dplot)
}
```


X1
Mean: −0.0125
Variance: 0.9638
0.4
0.3
0.2
0.1
0.0
−2
0
2


X2
0.4
0.3
0.2
0.1
0.0
Mean: 0.0008
Variance: 0.9912
−2
0
2


X3
Mean: −0.0018
Variance: 1.0045
0.4
0.3
0.2
0.1
0.0
−2
0
2

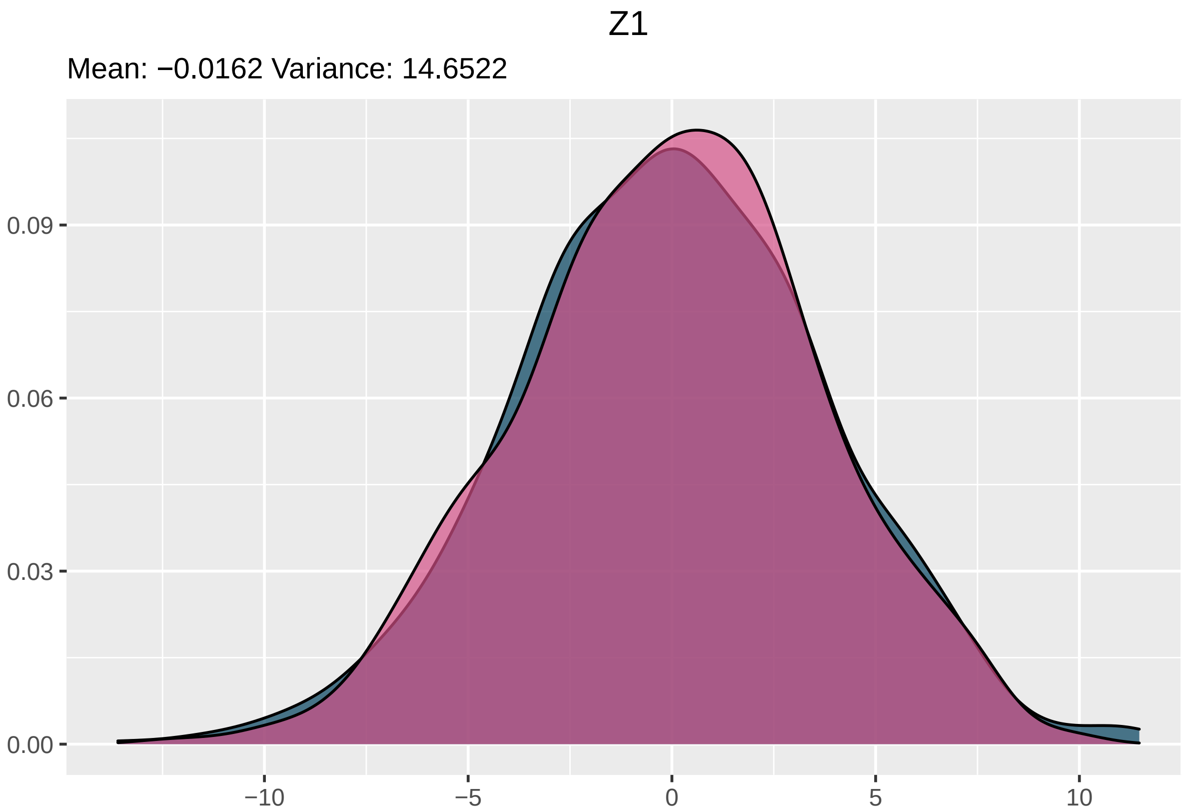
Plots of Z Variables

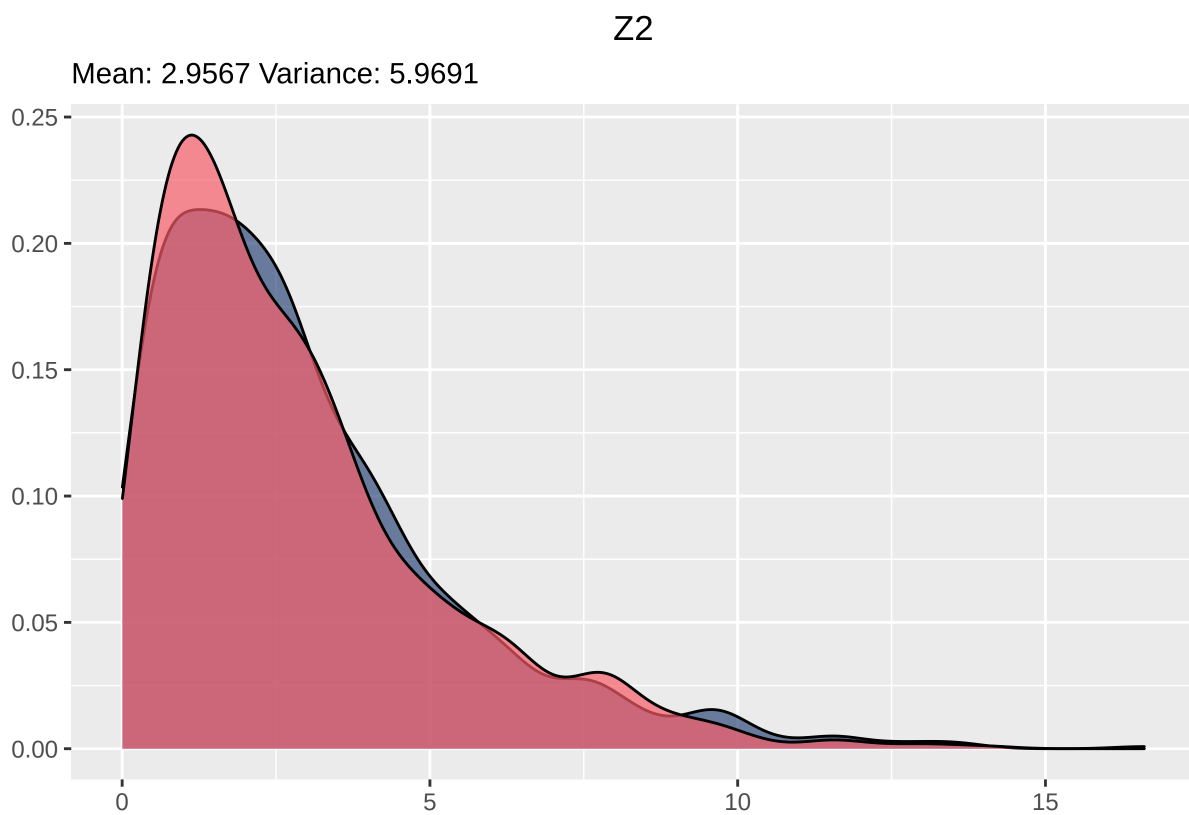
Plotting the Z variables, we can see that they form a normal, chi squared, gamma, and F distributions as expected from the equations used to construct each Z variable.

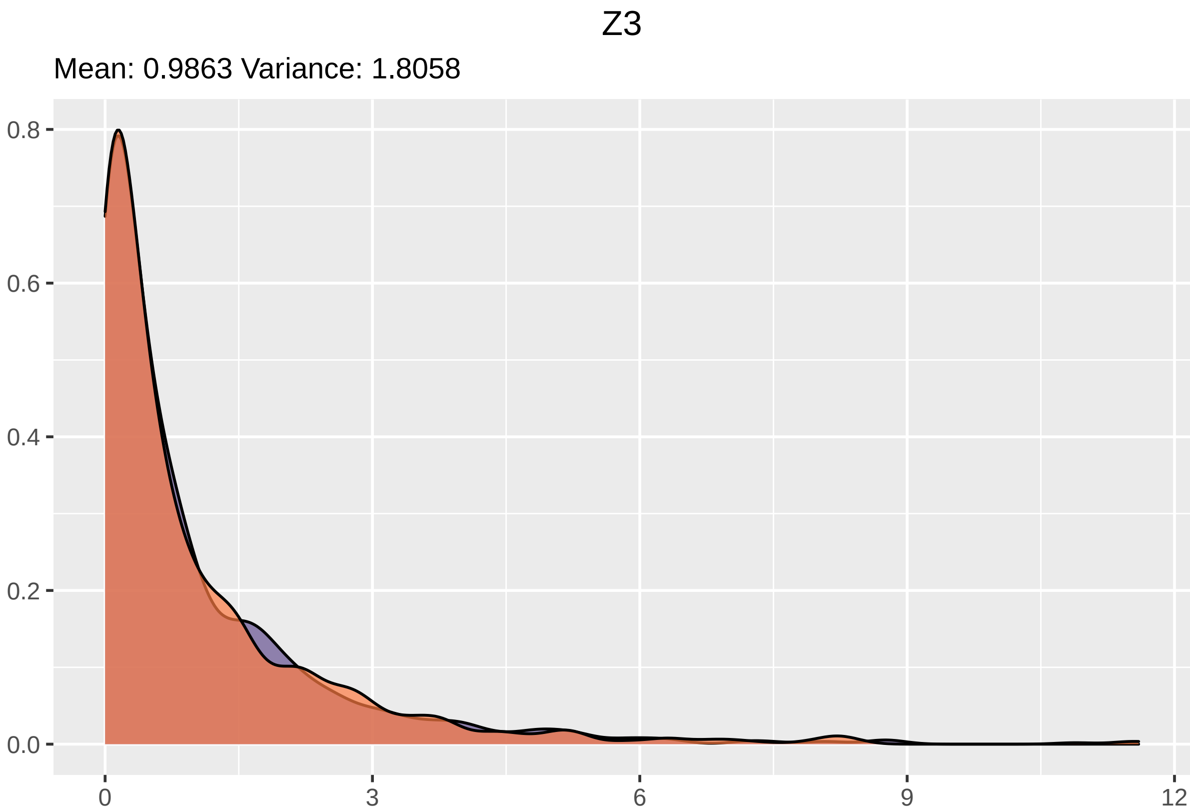
The mean and variance of each Z variable is quite close to the computed ones found in Problem 4. The differences are most likely due the noise observed in the X variable data. When plotting the Z1, Z2, and Z3 variables against data generated directly from the suspected distribution. This can be observed in the charts where each Z variable is overlaid on the equivalent distribution. They roughly matched the peak and spread.

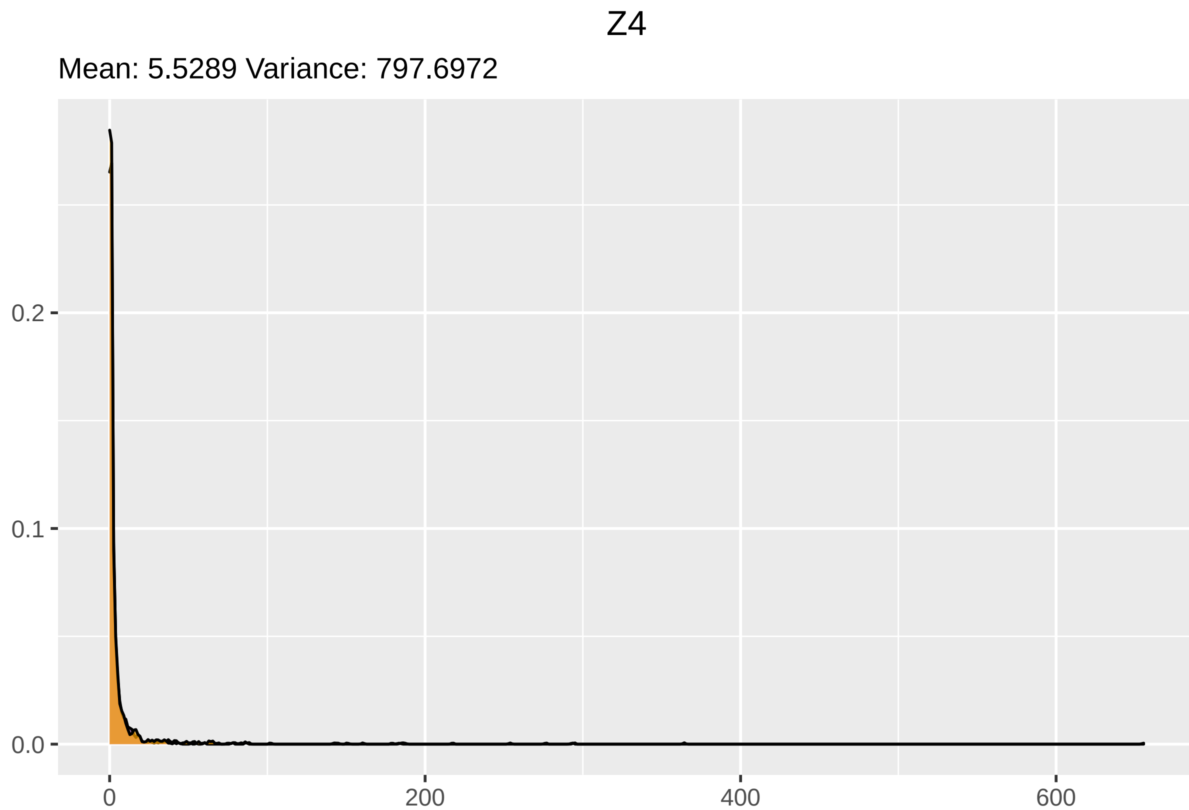
Z4 was difficult to ascertain since the Z variable and the distribution were both extreme. Zooming into the plot helped view the comparison better.

```r
set.seed(1000)

df_distros <- data.frame(cbind(rnorm(N, 0, 14^0.5),
                               rchisq(N, 3),
                               rchisq(N, 1),
                               rf(N, 1, 2)))

for (i in 1:4){
  stmnt <- sprintf("Mean: %.4f Variance: %.4f",
                   mean(df_z[ ,i]),
                   var(df_z[ ,i])
                   )
  dplot <- ggplot() +
           geom_density(aes( x = df_z[, i]), fill = chart_colors[i], alpha = 0.7) +
           geom_density(aes( x = df_distros[, i]), fill = chart_colors[i + 4], alpha = 0.7) +
           ggtitle(label = paste0("Z", i), subtitle = stmnt) + xlab("") + ylab("") +
           theme(plot.title = element_text(hjust = 0.5))

  print(dplot)
}
```


Z1
Mean: −0.0162 Variance: 14.6522
0.09
0.06
0.03
0.00
−10
−5
0
5
10

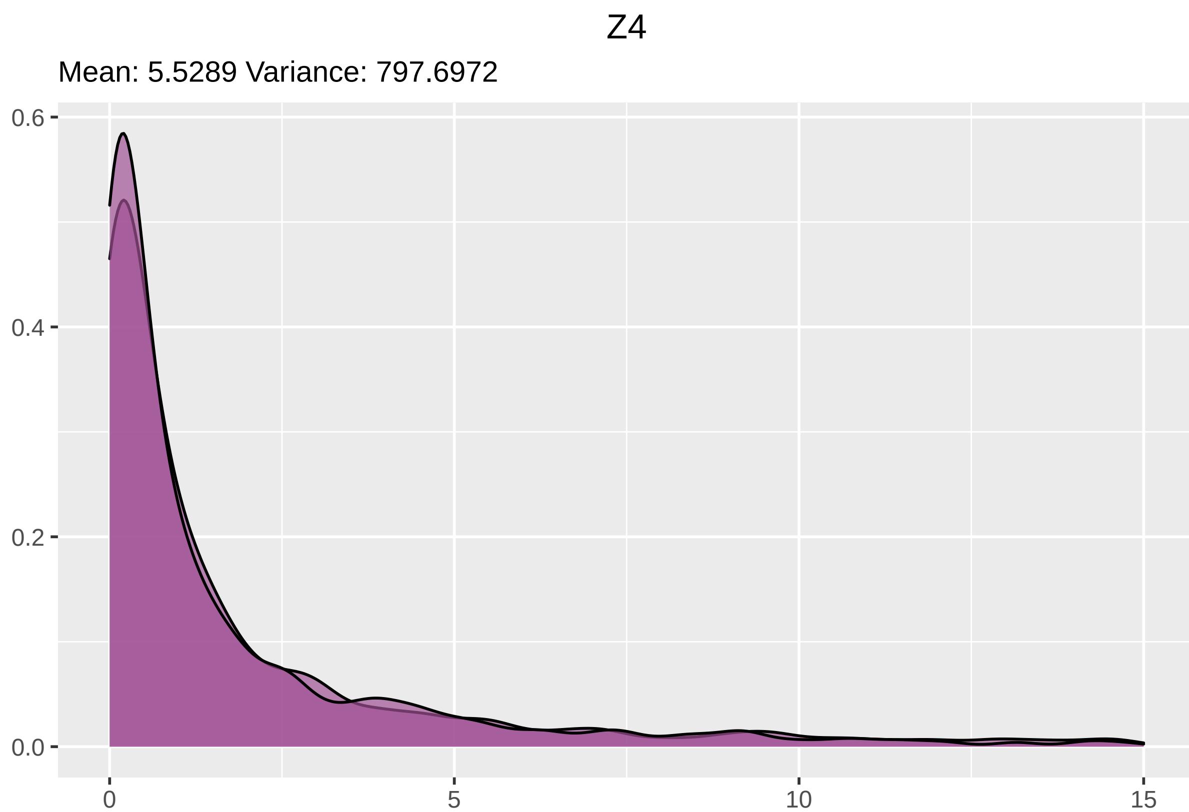

Z2
Mean: 2.9567 Variance: 5.9691
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15


Z3
Mean: 0.9863 Variance: 1.8058
0.8
0.6
0.4
0.2
0.0
0
3
6
9
12


Z4
Mean: 5.5289 Variance: 797.6972
0.2
0.1
0.0
0
200
400
600

The following chart zooms into the F distribution and reveals that the two variables match quite closely:

```
ggplot() +
           geom_density(aes( x = df_z[, 4]), fill = chart_colors[4], alpha = 0.7) +
           geom_density(aes( x = df_distros[, 4]), fill = chart_colors[4], alpha = 0.7) +
           ggtitle(label = paste0("Z", i), subtitle = stmnt) + xlab("") + ylab("") +
           theme(plot.title = element_text(hjust = 0.5)) + xlim(0, 15)
```




```
mean_sd_info <- rep(0, 4)

for (i in 1:4){
    x_info <- c(mean(df_z[ ,i]), var(df_z[ ,i]), mean(df_distros[ ,i]), var(df_distros[ ,i]))
    mean_sd_info <- rbind(mean_sd_info, x_info)
}

mean_sd_info <- mean_sd_info[2:5, ]

colnames(mean_sd_info) <- c("Z Mean", "Z Var", "Distro Mean", "Distro Var")
rownames(mean_sd_info) <- c("Z1", "Z2", "Z3", "Z4")

knitr::kable(mean_sd_info)
```

	Z Mean	Z Var	Distro Mean	Distro Var
Z1	-0.0162438	14.652224	-0.0466806	13.493580
Z2	2.9567271	5.969068	2.8589621	5.440282
Z3	0.9863173	1.805752	1.0299216	2.208856
Z4	5.5289443	797.697230	5.3966160	499.198341

As can be seen in the above table, the Z variables and their respective distribution mean and variance were quite close but not exactly the same. This is due to how the data was generated. All generated data sets used in this document displayed noise. This noise caused noticeable changes in the density plots. Overall, the Z variable and the respective distribution matched quite closely, verifying the results in the question 4.