Question 1

The probability density function f for the χ^2_ν distribution is

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2},$$

where, as usual, $\Gamma(z)$ is the gamma function:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x.$$

(a) Show that the gamma function has the property $\Gamma(z+1) = z\Gamma(z)$ (Hint: use integration by parts).

(b) Show that if $Y \sim \chi^2_\nu$ then $\mathrm{E}(Y) = \nu$ (Hint: try to get the integral into a form that is a constant times 'something' that integrates to 1).

(c) Show that if $Y \sim \chi^2_\nu$ then $\mathrm{E}\left(Y^2\right) = \nu(\nu+2)$.

Solution to Question 1

Part (a)

When the argument is $z+1$, the gamma function is:

$$\Gamma(z+1) = \int_0^\infty x^z e^{-x} \mathrm{d}x.$$

Let's use integration by parts, and we choose

$$\begin{aligned} u &= x^z & &\Rightarrow \mathrm{d}u = z x^{z-1} \mathrm{d}x \\ dv &= e^{-x}\mathrm{d}x & &\Rightarrow v = -e^{-x} \end{aligned}$$

Then

$$\begin{aligned} \int_a^b u\mathrm{d}v &= [uv]_a^b - \int_a^b v\mathrm{d}u \\ \Rightarrow \Gamma(z+1) = \int_0^\infty x^z e^{-x}\mathrm{d}x &= \left[-x^z e^{-x}\right]_0^\infty - \int_0^\infty \left(-e^{-x}\right) z x^{z-1}\mathrm{d}x \\ &= (0-0) + z\int_0^\infty x^{z-1}e^{-x}\mathrm{d}x \\ &= z\Gamma(z) \end{aligned}$$

Part (b)

$$\mathrm{E}(Y) = \int_0^\infty y f(y)\mathrm{d}y = \int_0^\infty y \cdot \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}\mathrm{d}y = \int_0^\infty \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{(\nu/2+1-1)} e^{-y/2}\mathrm{d}y$$

Now, we note three things:

1. $\int_0^\infty \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}\mathrm{d}y = 1$, for any value ν.
2. $2^{\nu/2+1} = 2^{\nu/2} \cdot 2 \Rightarrow 2^{\nu/2} = \frac{1}{2} \cdot 2^{\nu/2+1}$
3. $\Gamma(\nu/2+1) = (\nu/2)\Gamma(\nu/2) \Rightarrow \Gamma(\nu/2) = \frac{2}{\nu} \cdot \Gamma(\nu/2+1)$

Then, using these three things, we manipulate the integral above:

$$\begin{aligned}
\mathrm{E}(Y) &= \int_0^\infty \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{(\nu/2+1-1)} e^{-y/2}\mathrm{d}y \\
&= \int_0^\infty \frac{1}{\left(\frac{1}{2} \cdot 2^{\nu/2+1}\right) \frac{2}{\nu} \cdot \Gamma(\nu/2+1)} y^{(\nu/2+1-1)} e^{-y/2}\mathrm{d}y \\
&= \nu \int_0^\infty \frac{1}{2^{\nu/2+1}\Gamma(\nu/2+1)} y^{(\nu/2+1-1)} e^{-y/2}\mathrm{d}y \\
&= \nu \cdot 1 \\
\Rightarrow \mathrm{E}(Y) &= 1
\end{aligned}$$

where the integral is 1 because the parameter is $\nu+2$, since $(\nu+2)/2 = \nu/2+1$.

Part (c)

The same idea as for Part (b), but we first note $2^{\nu/2} = 2^{\nu/2+2} \cdot \frac{1}{4}$, and

$$\Gamma(\nu/2+2) = \left(\frac{\nu}{2}+1\right)\Gamma(\nu/2+1) = \left(\frac{\nu}{2}+1\right)\left(\frac{\nu}{2}\right)\Gamma(\nu/2) = \left(\frac{\nu(\nu+2)}{4}\right)\Gamma(\nu/2)$$

Then,

$$\begin{aligned}
\mathrm{E}\left(Y^2\right) &= \int_0^\infty y^2 f(y)\mathrm{d}y = \int_0^\infty y^2 \cdot \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}\mathrm{d}y = \int_0^\infty \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2+2-1} e^{-y/2}\mathrm{d}y \\
&= \int_0^\infty \frac{1}{\left(\frac{1}{4} \cdot 2^{\nu/2+2-1}\right) \frac{4}{\nu(\nu+2)}\Gamma(\nu/2+2)} y^{\nu/2+2} e^{-y/2}\mathrm{d}y \\
&= \nu(\nu+2) \int_0^\infty \frac{1}{2^{\nu/2+2}\Gamma(\nu/2+2)} y^{\nu/2+2-1} e^{-y/2}\mathrm{d}y \\
&= \nu(\nu+2) \cdot 1 \\
&= \nu(\nu+2)
\end{aligned}$$

Question 2

Prove Boole's inequality: for a set of events A_i, $i = 1, 2, \ldots n$,

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathrm{P}\left(A_i\right).$$

Hint: use induction.

Solution to Question 2

We shall use induction to prove the inequality. First prove it true for $n = 2$:

$$\mathrm{P}\left(A_1 \cup A_2\right) = \mathrm{P}\left(A_1\right) + \mathrm{P}\left(A_2\right) - \mathrm{P}\left(A_1 \cap A_2\right) \leq \mathrm{P}\left(A_1\right) + \mathrm{P}\left(A_2\right),$$

since $\mathrm{P}\left(A_1 \cap A_2\right) \geq 0$.

Assume the inequality is true for $n = k$, i.e. for events $A_1, A_2, \ldots, A_k$,

$$\mathrm{P}\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathrm{P}\left(A_i\right).$$

We now prove the inequality is true for $n = k + 1$. For events for events $A_1, A_2, \ldots, A_k$ and A_{k+1}, if we set $B = \bigcup_{i=1}^{k} A_i$, then

$$\begin{aligned}
\mathrm{P}\left(\bigcup_{i=1}^{k+1} A_i\right) &= \mathrm{P}\left(\left(\bigcup_{i=1}^{k} A_i\right) \bigcup A_{k+1}\right) \\
&= \mathrm{P}\left(B \bigcup A_{k+1}\right) \\
&\leq \mathrm{P}\left(B\right) + \mathrm{P}\left(A_{k+1}\right) \qquad \text{(By case } n = 2\text{)} \\
&= \mathrm{P}\left(\bigcup_{i=1}^{k} A_i\right) + \mathrm{P}\left(A_{k+1}\right) \\
&\leq \sum_{i=1}^{k} \mathrm{P}\left(A_i\right) + \mathrm{P}\left(A_{k+1}\right) \qquad \text{(By case } n = k\text{)} \\
&= \sum_{i=1}^{k+1} \mathrm{P}\left(A_i\right).
\end{aligned}$$

This proves the result.

Question 3

Download the dataset `data_ps9.csv` (link on Blackboard below problem sheet). This dataset contains 200 observations for each of the random variables X, Y and Z, where the ith row shows the simultaneous measurement of the three variables at time i. Using R, perform an exploratory data analysis to:

(a) Investigate whether or not there is any relationship between any of the variables.

(b) Guess the distributions of X, Y and Z.

Solution to Question 3

Part (a):

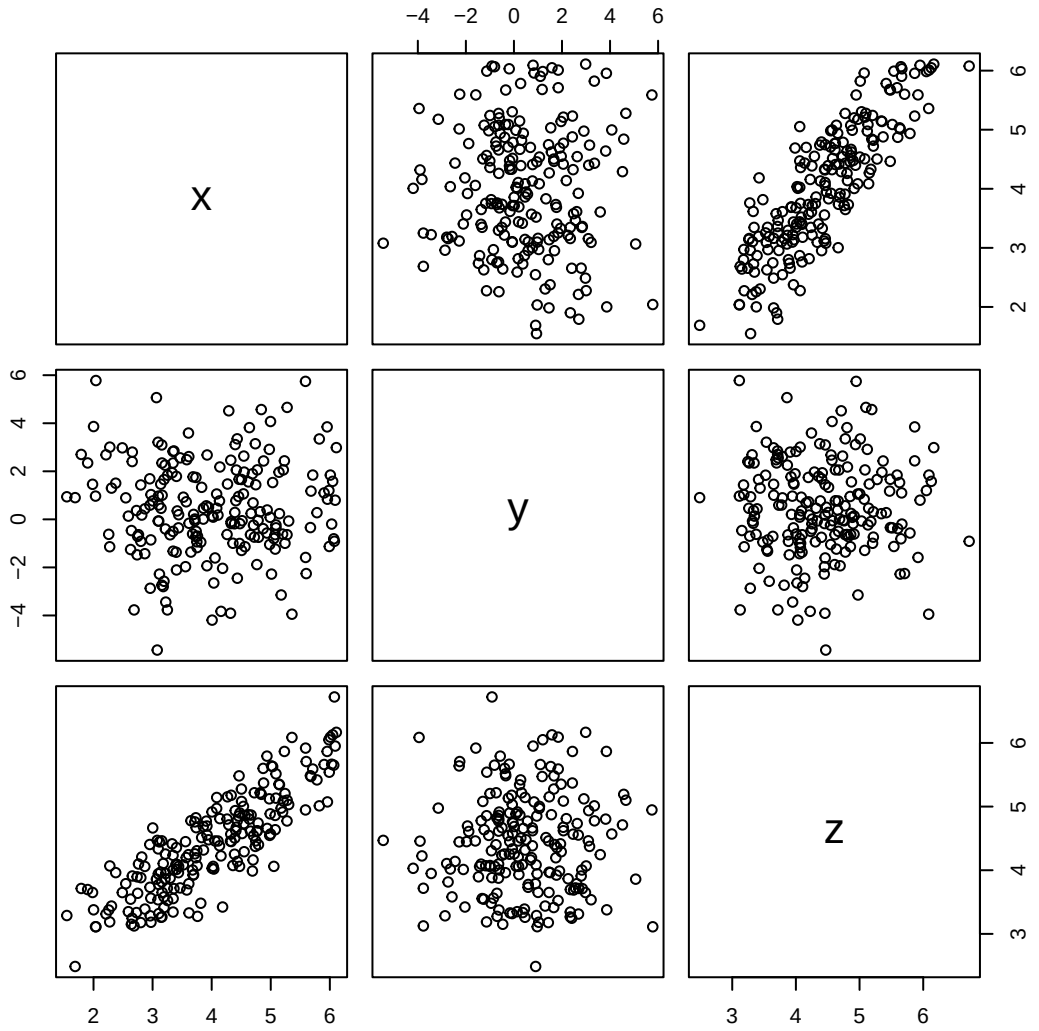
First, we need to read in the data:

```
# read the data into a data frame (see Problem sheet 8)
df <- read.table("data_ps9.csv", sep=",", header=T)

#let's see how many values are in the data.frame
print(dim(df))
#> [1] 200   3

#let's see the first five rows
print(head(df, n=5))
#>       x       y     z
#> 1 3.103  0.5960 3.776
#> 2 4.185 -2.0391 3.422
#> 3 5.588  5.7418 4.947
#> 4 2.870  0.4374 3.390
#> 5 3.920 -1.9331 4.695
```

So there are 200 rows and 3 columns in the data frame, and the columns are labelled `x`, `y` and `z`.

Now, to investigate any relationships between the random variables `x`, `y` and `z`, since there are exactly 200 observations each, we could plot the random variables against each other, e.g. `x` vs `y`, etc. R has a function called `pairs` which does this in one line and produces the plots in a nice layout:

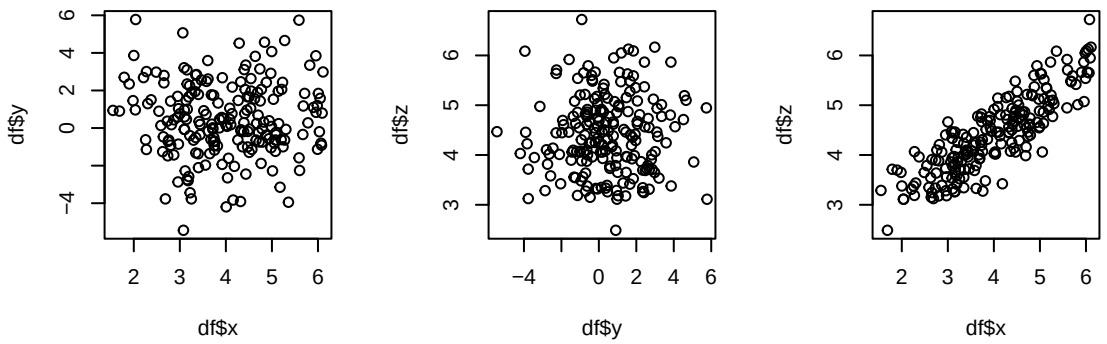
```
# use the 'pairs' function to plot variables against each other
pairs(df)
```



One reads the figures as row vs column; i.e. in the 3×3 matrix of figures above, the (sub)plot in the second row and first column is `y` vs `x`, i.e. `y` on the y-axis and `x` on the x-axis.

Alternatively, one could use the following code to do this 'manually':

```
# create a plot with 1 row and 3 columns
par(mfrow=c(1,3))
plot(x=df$x, y=df$y)
plot(x=df$y, y=df$z)
plot(x=df$x, y=df$z)
```

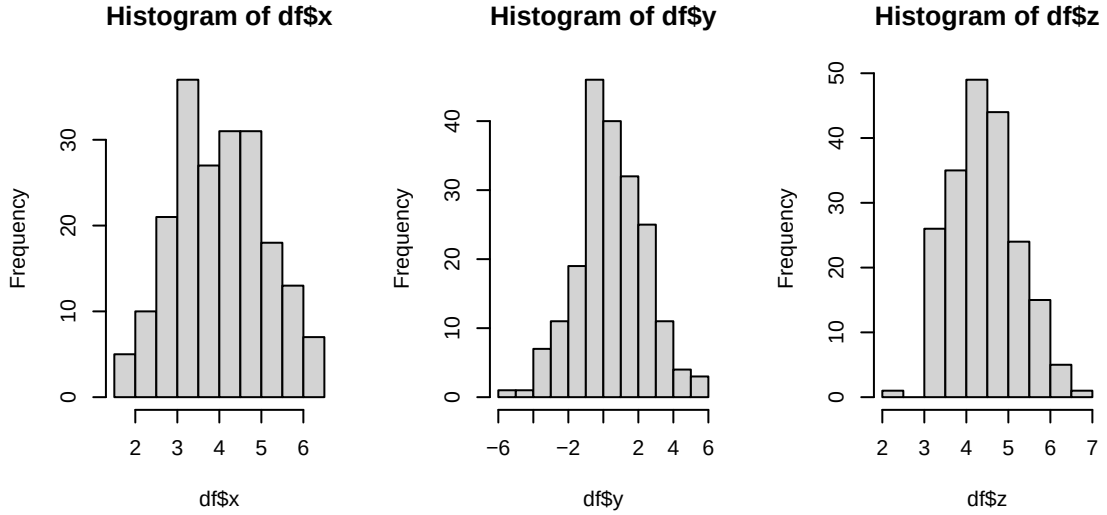


From the figures, there does not seem to be any obvious relationship between `x` and `y`, or between `y` and `z`, since the scatterplots show a somewhat random cloud of points. However, looking at `z` vs `x`, it seems that when a small `x` value was measured at time i, a small `z` value was measured at time i. This suggests there is some relationship between `x` and `z`.

Part (b):

To get an overview of the distribution of a random variable, a histogram is a useful plot. The following code plots histograms of the `x`, `y` and `z`.

```
# create a plot with 1 row and 3 columns
par(mfrow=c(1,3))
hist(df$x)
hist(df$y)
hist(df$z)
```



It is difficult to draw conclusions from these histograms, but they do suggest that the random variables follow normal distributions; perhaps the histogram for `y` most clearly suggests this.

We can compute the mean and standard deviation of these datasets to obtain:

```
# this creates the pairs of means/sd's
x_param <- c(mean(df$x), sd(df$x))
y_param <- c(mean(df$y), sd(df$y))
z_param <- c(mean(df$z), sd(df$z))

#create the data frame
df_param <- data.frame(x=x_param, y=y_param, z=z_param)

# set the row names
row.names(df_param) <- c("mean", "sd")

#print the data framw
print(df_param)
#>           x     y      z
#> mean 3.999 0.457 4.4206
#> sd   1.074 1.972 0.7801
```

For example, from these values, one might say that Y follows a normal distribution with mean approximately 0.46 and standard deviation approximately 2. However, we cannot be certain.

It might be interesting to see how this data was generated. Here is the code used to generate this data:

```
    set.seed(2)
    filename <- "data_ps9.csv"
    n <- 200
    rho <- 0.6
    x <- rnorm(n, mean=4, sd=1)
    y <- rnorm(n, mean=0, sd=2)
    z <- rnorm(n, mean=5, sd=1)
    z <- (1-rho) * z + rho * x
    df <- data.frame(x, y, z)
    write.csv(df, file=filename, quote=F, row.names=F)
```

So, in fact X and Y are normal, but Z is a mixture between X and another normal distribution; this explains the relationship between X and Z that we observed in Part (a).