Question 1

Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables that follow a $\mathrm{N}\left(\mu, \sigma^2\right)$ distribution, and define $\overline{X} = \frac{1}{n}\sum_{i=1} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}\left(\overline{X} - X_i\right)^2$, as usual. Show that the random variable T, where

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

can be written in the form

$$T = \frac{U}{\sqrt{V/p}},$$

where

- $U \sim \mathrm{N}(0,1)$,
- p is some function of n,
- $V \sim \chi^2_p$, the chi-squared distribution with p degrees of freedom,
- U and V are independent random variables.

Solution to Question 1

Recall Corollary 3.1.3 in the notes which states that, given the assumptions above, $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Recall also Theorem 3.2.2 in the notes that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. These results will inform our approach to rewrite T in the desired form.

We perform several manipulations on the quantity T to obtain

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \times \frac{\sqrt{n}}{\sqrt{n}} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S}$$

$$= \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} \times \frac{\left(\frac{1}{\sigma}\right)}{\left(\frac{1}{\sigma}\right)} = \frac{\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}{\frac{S}{\sigma}} = \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{S}{\sigma}} = \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}}$$

$$= \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}} \times \frac{1}{\left(\sqrt{\frac{n-1}{n-1}}\right)} = \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\left[\frac{(n-1)S^2}{\sigma^2}\right]/(n-1)}}$$

Therefore, by setting

$$U = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}, \qquad V = \frac{(n-1)S^2}{\sigma^2}$$

we have shown that we can write

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{U}{\sqrt{V/(n-1)}}.$$

and we can take $p = n - 1$. We still need to show that U and V follow the appropriate distributions.

Since $\overline{X}$ is a linear transformation of the independent normal $X_1, X_2, \ldots, X_n$, the sample mean $\overline{X}$ also follows a normal distribution (Corollary 1.6.2 in the notes). Furthermore, since U is a linear transformation of $\overline{X}$, U also follows a normal distribution. Now, from Proposition 1.2.6, since the $X_1, X_2, \ldots, X_n$ are independent with mean μ and variance σ^2 (since they are i.i.d. $\mathrm{N}\left(\mu, \sigma^2\right)$), we have that $\mathrm{E}\left(\overline{X}\right) = \mu$ and $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$. Therefore, we can compute the mean and variance of U:

$$\mathrm{E}\left(U\right) = \mathrm{E}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{1}{\frac{\sigma}{\sqrt{n}}}\mathrm{E}\left(\overline{X} - \mu\right) = \frac{1}{\frac{\sigma}{\sqrt{n}}}\left[\mathrm{E}\left(\overline{X}\right) - \mu\right] = 0$$

$$\mathrm{Var}\left(U\right) = \mathrm{Var}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{1}{\left(\frac{\sigma}{\sqrt{n}}\right)^2}\mathrm{Var}\left(\overline{X} - \mu\right) = \frac{n}{\sigma^2}\mathrm{Var}\left(\overline{X}\right) = \frac{n}{\sigma^2}\left(\frac{\sigma^2}{n}\right) = 1.$$

Therefore, U is random variable following a $\mathrm{N}\left(0, 1\right)$ distribution.

For V, Theorem 3.1.3 in the notes gives us that $V = \frac{(n-1)S^2}{\sigma^2}$ follows a χ^2_{n-1} distribution.

Finally, regarding independence, Theorem 3.1.3 also gives us that $V = \frac{(n-1)S^2}{\sigma^2}$ and $\overline{X}$ are independent. Since U is simply a linear transformation of $\overline{X}$, we therefore have that U and V are independent.

Therefore, we have shown that $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ can be written as $\frac{U}{\sqrt{V/(n-1)}}$, where $U \sim \mathrm{N}\left(0, 1\right)$ and $V \sim \chi^2_{n-1}$ and U and V are independent.

Question 2

Suppose the following 11 values are the transaction amounts (in £) of online purchases for a particular credit card customer in a given month.

$$45,\ 81,\ 52,\ 23,\ 147,\ 92,\ 76,\ 124,\ 287,\ 103,\ 65$$

Tukey's criterion states that, given the lower quartile $q_{0.25}$, the upper quartile $q_{0.75}$ and the interquartile range IQR, if a value x is either $x < q_{0.25} - k\text{IQR}$ or $x > q_{0.75} + k\text{IQR}$, for $k = 1.5$, then x is considered to be an outlier.

(a) Compute the lower and upper quartiles, and the interquartile range for this dataset.

(b) According to Tukey's criterion, are any of these transaction amounts outliers?

(c) If any of the transactions is an outlier, would you take any action? What could be the consequences of (i) inaction (doing nothing) or (ii) taking action (preventing the transaction from going through)?

(d) If you were designing your own fraud detector for this customer (not using Tukey's criterion) for the next month, how high would a value need to be for you to decide that a value is anomalous and potentially fraudulent? In other words, at what value would you set the threshold?

Solution to Question 2

Part (a):

Sorting the data,

$$23,\ 45,\ 52,\ 65,\ 76,\ 81,\ 92,\ 103,\ 124,\ 147,\ 287$$

One finds the median as the 6th value (since there are 11 values), and therefore the lower quartile is at index $(1+6)/2 = 3.5$. This means that $q_{0.25}$ is the average of the 3rd and 4th order statistics (ordered values), i.e. $q_{0.25} = (52 + 65)/2 = 117/2 = 58.5$.

The upper quartile is computed similarly; it is 3.5 units away from the largest values, i.e. the average of 103 and 124. Therefore $q_{0.75} = 227/2 = 113.5$.

The IQR is therefore $113.5 - 58.5 = 55$.

Part (b):

Using Tukey's criterion for outliers, the lower limit is $58.5 - 1.5(55) = -24$ and the upper limit is $113.5 + 1.5(55) = 196$. Therefore, according to Tukey's criterion, the value 287 is an outlier.

Part (c):

Whether or not you take action in this case is a personal choice. The transaction with value 287 is an outlier, so if one were strictly following the criterion one woudl take action. On the other hand, 287 does not seem to be very different to the others. If one does not take action, it is risking that a fraudulent transaction goes through and the customer (or the company) loses money. If action is taken, and the transaction is blocked or delayed, this could have consequences for the customer if this is not a fradulent transaction. The point is - it is a trade-off, and finding the optimal decision rule while balancing these aspects is not easy.

Part (d):

There is no right answer to this question, and it is meant to make you think of your own possible outlier detection algorithm. For example, one approach would be to consider a threshold based on standard deviations from the mean. Denoting the sample mean of this data set by $\overline{x}$, one can compute $\overline{x} \approx 100$. The sample standard deviation is $s \approx 72$. One option would be to set the upper limit as $\overline{x} + 10s \approx 820$.

Question 3 (R question)

It is suggested that the following question is done in an R Markdown document.

(a) Use `dnorm` to plot the probability density function of the standard normal random distribution on the interval $[-4, 4]$.

Hint: Use the `seq` function to generate 1000 evenly spaced points on the interval $[-4, 4]$.

(b) Use `dgamma` to plot the probability density function of a $\Gamma(2, 0.5)$ random variable on the interval $[0, 20]$. Note that we are using the shape/rate parametrisation here, i.e. $\alpha = 2$ is the shape and $\beta = 0.5$ is the rate.

(c) Now do the following:

(i) For $X_1, X_2, \ldots, X_n \sim \Gamma(\alpha, \beta)$, use R to sample observations $x_1, x_2, \ldots, x_n$, where $n = 1000$ and $\alpha = 2$ and $\beta = 0.5$.

(ii) From these $x_1, x_2, \ldots, x_n$ values, compute the the standardised $z_1, z_2, \ldots, z_n$, where

$$z_i = \frac{x_i - \mathrm{E}[X_i]}{\sqrt{\mathrm{Var}[X_i]}}.$$

Hint: For $X \sim \Gamma(\alpha, \beta)$, $\mathrm{E}[X] = \frac{\alpha}{\beta}$ and $\mathrm{Var}[X] = \frac{\alpha}{\beta^2}$.

(iii) Compute the weighted sum

$$S = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i.$$

(Note the square root in the fraction $1/\sqrt{n}$; this is **not** the sample mean.)

(iv) Repeat steps (a) to (c) t times (using a loop), and save the resulting sums $S_1, S_2, \ldots, S_t$ to a vector $\mathbf{S}$. It is suggested that t is set to $t = 10,000$.

(v) Plot a histogram of the values $S_1, S_2, \ldots, S_t$. In the `hist` function, set the parameters `freq=FALSE` and `breaks=30`.

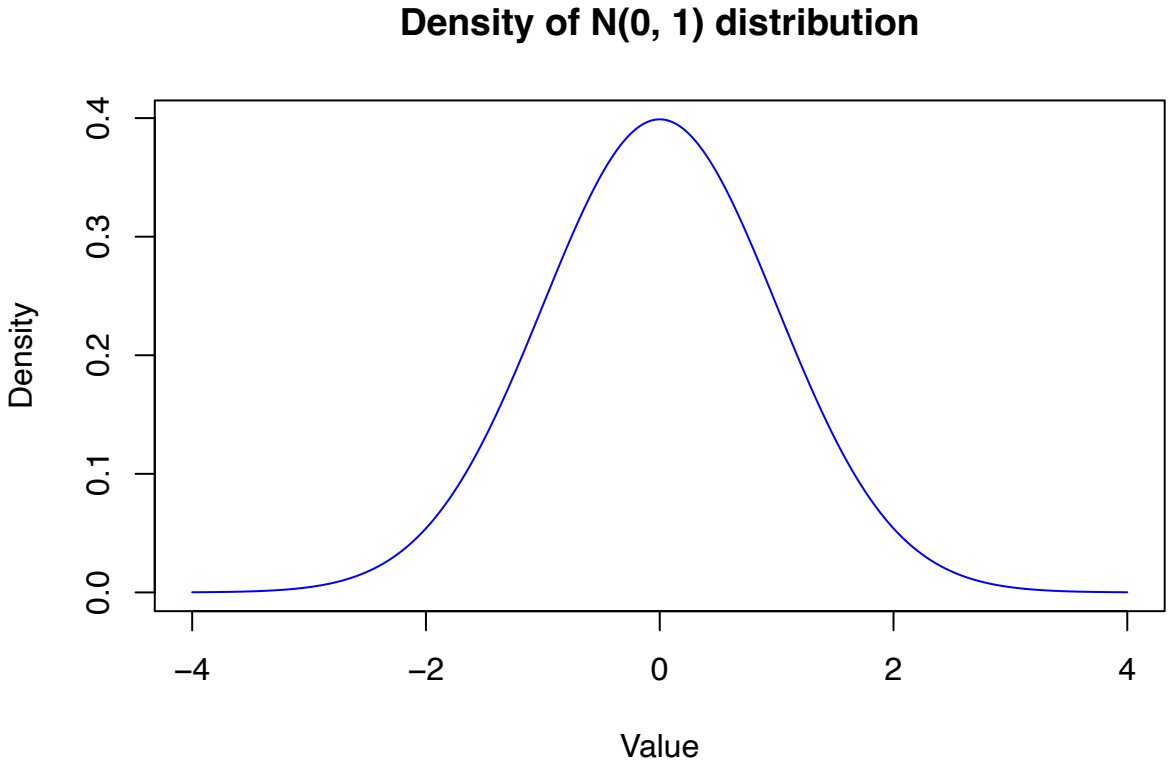
(vi) Does this histogram look familiar? Use the `lines` function in R to plot the probability density function of an appropriate distribution over the histogram.

Problem sheet 11 Question 3

Part (a)

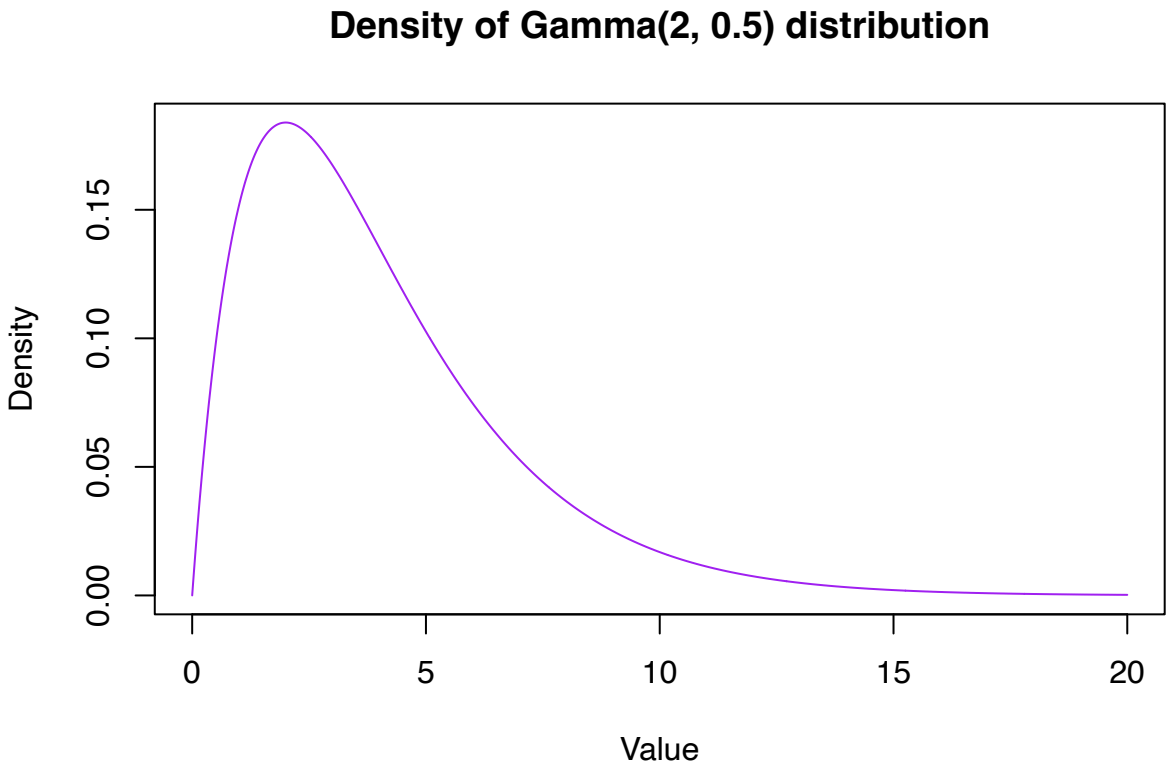
Plotting a standard normal density, i.e. $f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$ for $Z \sim N(0,1)$.

```
mu <- 0
sigma <- 1
x <- seq(-4, 4, length=1000)
y <- dnorm(x, mean=mu, sd=sigma)
xlab <- "Value"
ylab <- "Density"
main <- paste0("Density of N(", mu, ", ", sigma, ") distribution")
plot(x, y, type='l', xlab=xlab, ylab=ylab, main=main, col="blue")
```



Part (b)

```
alpha <- 2
beta <- 0.5
x <- seq(0, 20, length=1000)
y <- dgamma(x, shape=alpha, rate=beta)
xlab <- "Value"
ylab <- "Density"
main <- paste0("Density of Gamma(", alpha, ", ", beta, ") distribution")
plot(x, y, type='l', xlab=xlab, ylab=ylab, main=main, col="purple")
```



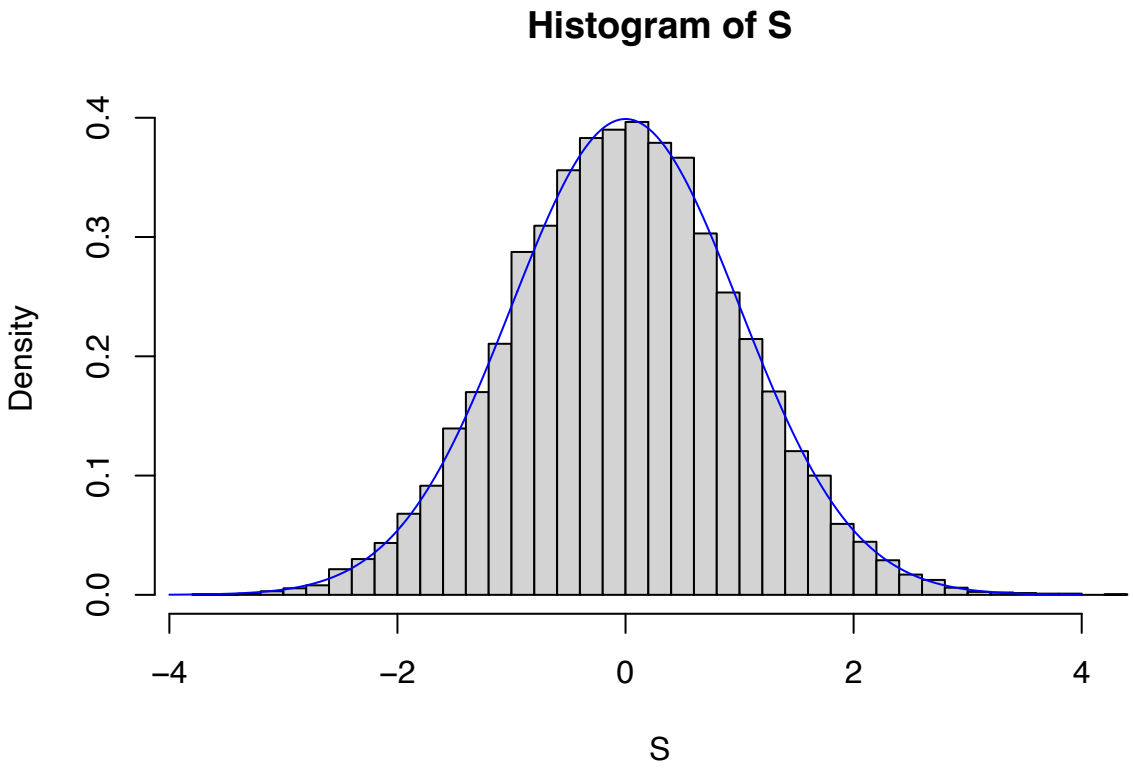
Part (c)

```
n <- 1e3
numtrials <- 1e4
set.seed(1)
#shape
alpha <- 2
#rate
beta <- 0.5

#computing mean and variance
mu <- alpha/beta
sigma_sq <- alpha/(beta^2)
sigma <- sqrt(sigma_sq)

# initialise the sums, and run the trials
S <- rep(0, numtrials)
for (i in seq_len(numtrials)){
    x <- rgamma(n, shape=alpha, rate=beta)
    #standardise
    x <- (x - mu)/sigma
    S[i] <- sum(x) / sqrt(n)
}

# plot the histogram
hist(S, freq=F, breaks=30)
z <- seq(-4, 4, length=1000)
# add the normal density plot
d <- dnorm(z, mean=0, sd=1)
lines(z, d, col="blue")
```



Note how this histogram appears to show a standard normal distribution. This is not an accident; for interest look up the **central limit theorem** in a standard textbook. You will learn more about this in Year 2.