Problems:

1. Poisson Distribution

(a) Generate 200 random samples of size n = 10 from a Poisson distribution with mean λ = 12.

The simulation is performed in R using the function rpois() to generate 200 independent random samples of size n = 10 from a Poisson distribution (λ = 12). The R code is shown:

```
N=200 # random samples
n=10 # size of random samples
lambda = 12 # Poisson distribution parameter

# Generating random samples
samples = matrix(rpois(N * n, lambda), nrow = N)
```

i. Calculate sample means for each sample. Report the first 10 sample means.

The mean for each sample is calculated using the function mean() as shown in the following R code. Also, the first 10 sample means are reported as shown below.

```
# Calculating sample mean for each sample
mean_samples = matrix(0, nrow = N) # Initialize
for(i in 1:N){
  mean_samples[i] = mean(samples[i,1:n])
}
```

```
# Reporting first 10 sample means
mean_10 = mean_samples[1:10]
print(mean_10)
```

Output:

```
> mean_10
 [1] 11.9 10.6 10.2 12.9 13.0 11.0 10.9 10.2 11.5 10.0
```

ii. Draw a histogram of the sample means (where the y-axis is the density) and fit a density estimate (default density estimator is ok).

Histograms can provide insights on skewness, behavior in the tails, presence of multi-modal behavior, and data outliers. The histogram of the sample means can be obtained using hist() statement and is shown in Figure 1. Notice that in this plot the y-axis is probability densities instead of frequencies. Also, a density estimate can be fit using density() and plot() to display the distribution graphic as shown in Figure 2. The implementation using R code is also shown below.

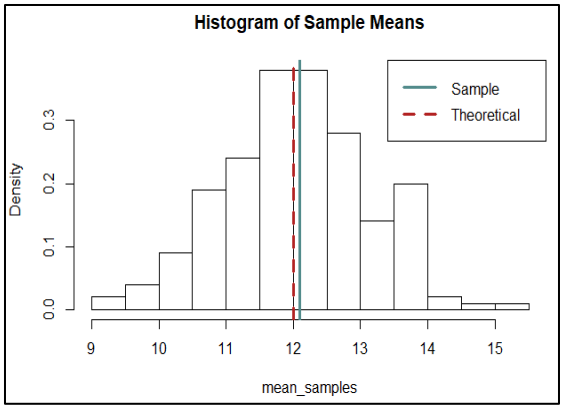

Figure 1

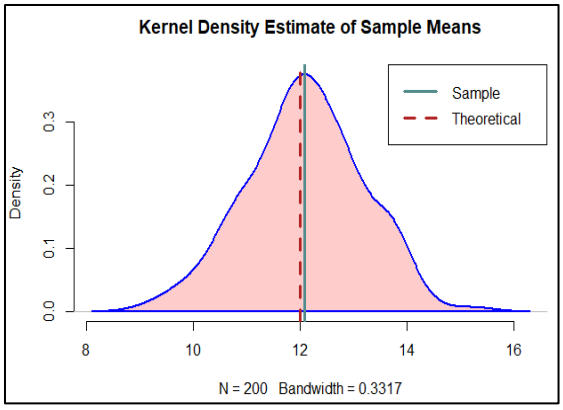

Figure 2

```
# Plotting histogram of sample means (y-axis: density)
hist(mean_samples, freq = FALSE, main = 'Histogram of Sample Means')
# The option freq=FALSE plots probability densities instead of frequencies
# Adding a dashed line for the mean of sample means
abline(v=mean(mean_samples), lwd=3, col='darkslategray4')
# Theoretical mean and standard deviation
theoretical_mean = lambda
theoretical_sd = sqrt(lambda/n)
# Adding a dashed line for the theoretical mean
abline(v=theoretical_mean, lty=2, lwd=3, col='firebrick')
legend(c("Sample", "Theoretical"),x='topright', lty=c(1,2),
       lwd=c(3,3), col=c('darkslategray4', 'firebrick'))

# Fitting a density estimate (default estimator)
d = density(mean_samples)
plot(d, main = "Kernel Density Estimate of Sample Means", bty = 'n')
polygon(d, col = "#FFCCCC", border = 'blue', lwd = 2)
# Adding a dashed line for the sample mean
abline(v=mean(mean_samples), lwd=3, col='darkslategray4')
# Adding a dashed line for the theoretical mean
abline(v=theoretical_mean, lty=2, lwd=3, col='firebrick')
legend(c("Sample", "Theoretical"),x='topright', lty=c(1,2),
       lwd=c(3,3), col=c('darkslategray4', 'firebrick'))
```

iii. What is your finding about the sampling distribution of the sample mean, based on your histogram. Be sure to give the distribution name along with its parameter estimates.

In Figure 1 and Figure 2 above, looking at the histogram and density plots of the distribution of sample means, even though the sample size is small (n = 10), the distribution still appears approximately Gaussian (as it shows to follow a bell curve), centered almost on its true theoretical value (μ = mean of original Poisson distribution = λ = 12). In these plots, the position of the mean of simulated sample means is indicated by a vertical solid line and the theoretical mean of distribution is shown by vertical dashed line. The mean of means is 12.014 and standard deviation is 1.02805. To get the variance we can square the value of standard deviation to get 1.05689.

```
> mean(mean_samples)
[1] 12.014
> sd(mean_samples)
[1] 1.028051
```

(b) Generate 200 random samples of size *n* = 50 from a Poisson distribution with mean λ = 12.

The same set of R codes shown above can be used to generate 200 independent random samples of size $n = 50$ from a Poisson distribution ($\lambda = 12$). The distribution of sample means is displayed using histogram and density plots as shown in Figure 3 and Figure 4, respectively.

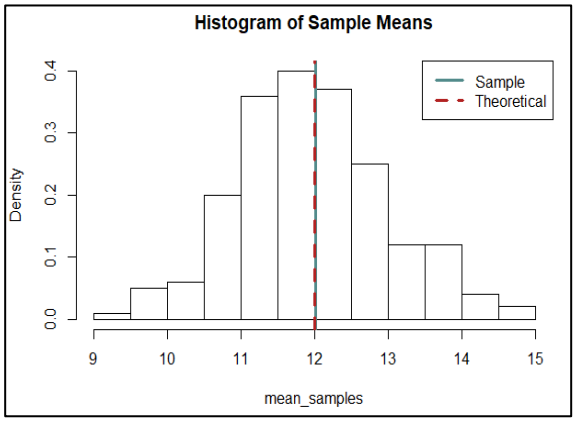

Figure 3

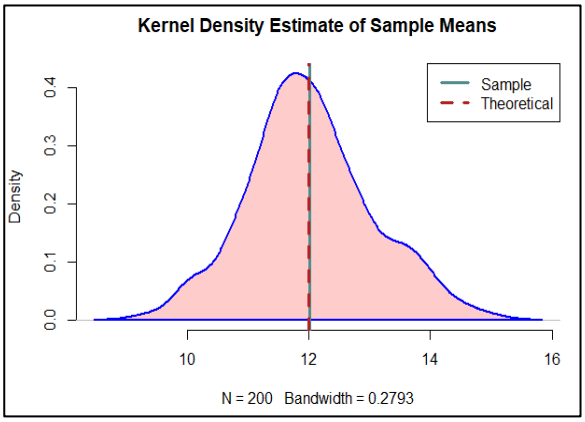

Figure 4

i. What distribution is the new set of sample means? State the exact distribution expected from theory (including parameter values), and state the distribution obtained from your samples (including estimated parameter values).

The theoretical mean for this distribution of means is $\lambda = 12$ and the theoretical standard deviation is $\sqrt{\lambda / n} = 0.48989$ (i.e. variance = 0.24). Thus ideally, we would expect the distribution of means to be normal with a mean of 12 and a variance of 0.24.

From Figure 3 and Figure 4 above, looking at the histogram and density plots of the distribution of sample means, the distribution does appear to be approximately normal, centered almost on its true theoretical value. In these plots, the position of the mean of simulated sample means is indicated by a vertical solid line and the theoretical mean of distribution is shown by vertical dashed line. The mean of means is 12.0053 and standard deviation is 0.48273. To get the variance we can square the value of standard deviation to get 0.23303.

```
> mean(mean_samples)
[1] 12.0053
> sd(mean_samples)
[1] 0.4827287
```

We can also look at the Quantile-Quantile plot in Figure 5 to examine how closely this distribution of sample means approximates a normal distribution. The closer the values lie to the line $y = x$, the better the fit. Here we see that the plot suggests a high degree of normality. From the Central Limit Theorem, this is what we expect to see. The R code is also shown below.

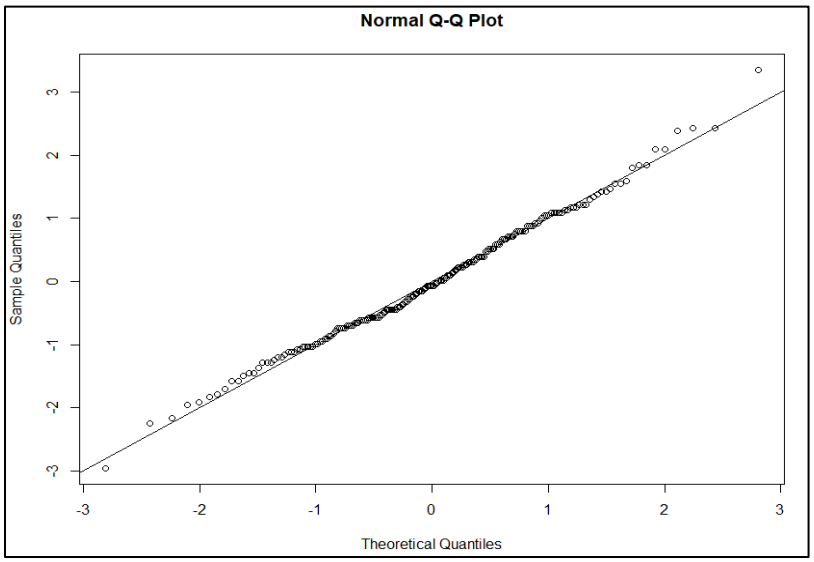

Figure 5

```
# Standardizing the sample means to draw QQ-plot
mean_samples_norm<-(mean_samples-mean(mean_samples))/sd(mean_samples)
qqnorm(mean_samples_norm) # drawing the QQplot
abline(0,1) # drawing a 45-degree reference line
```

ii. Based on this distribution, construct a 95% confidence interval for the mean λ. Assume σ is known.

Based on the conclusion above, since the sampling distribution of mean roughly follows a normal distribution, a 95% confidence interval for the mean λ should lie within $z_{0.025}$ standard errors of the mean (based on Central Limit Theorem). Previously we calculated the mean and standard deviation of the sampling distribution to be 12.0053 and 0.48273, respectively. Thus, the 95% confidence interval is:

$$12.0053 \pm 1.96 * 0.48273/\sqrt{50}$$

$$(11.8715, 12.1391)$$

Thus, we are 95% confident that the true mean λ lies within the interval between 11.8715 and 12.1391 assuming that the sampling distribution is normally distributed. The R code is shown below.

```
> error = qnorm(0.975)*sd(mean_samples)/sqrt(n)
> left = mean(mean_samples)-error
> right = mean(mean_samples)+error
> left
[1] 11.8715
> right
[1] 12.1391
```

2. Chi-Square Distribution

(a) Generate 200 random samples of size *n* = 20 from a Chi-Square distribution with degrees of freedom *df* = 6.

The simulation is performed in R using the function rchisq() to generate 200 independent random samples of size *n* = 20 from a Chi-Square distribution (*df* = 6). The R code is shown:

```
N=200 # random samples
n=20 # size of random samples
df=6 # degrees of freedom

# Generating random samples of chi-sq dist
samples = matrix(rchisq(N * n, df), nrow = N)
```

i. Find the first quantile Q1 from each sample (note: qchisq() gives the quantile from the exact distribution, NOT the sample). Report the first 10 quantiles.

There are several quantiles of an observation variable. The first quantile, or lower quantile, is the value that cuts off the first 25% of the data when it is sorted in ascending order. The second quantile, or median, is the value that cuts off the first 50%. The third quantile, or upper quartile, is the value that cuts off the first 75%. Thus, the first quantile for each sample is calculated using the function quantile() as shown in the following R code. Also, the first quantile of first 10 samples are reported below.

```
# Calculating the first quantile for each sample
quantile_samples = matrix(0, nrow = N) # Initialize
for(i in 1:N){
  quantile_samples[i] = quantile(samples[i,1:n], probs = 0.25)
}
```

```
# Reporting first 10 sample quantiles
quantile_10 = quantile_samples[1:10]
print(quantile_10)
```

Output:

```
> quantile_10
 [1] 3.617057 5.092080 3.099026 3.340888 4.329461 3.590075
 [7] 3.975380 3.446277 3.920886 3.834909
```

ii. Draw a histogram of the sample Q1's (where the *y*-axis is the density) and fit a density estimate (default density estimator is ok).

The histogram of the sample Q1's can be obtained using hist() statement and is shown in Figure 6. Notice that in this plot the y-axis is probability densities instead of frequencies. Also, a density estimate can be fit using density() and plot() to display the distribution graphic as shown in Figure 7. In these plots, the position of the mean of simulated sample Q1's is indicated by a vertical solid line and has a value of 3.57929. The implementation using R code is also shown below.

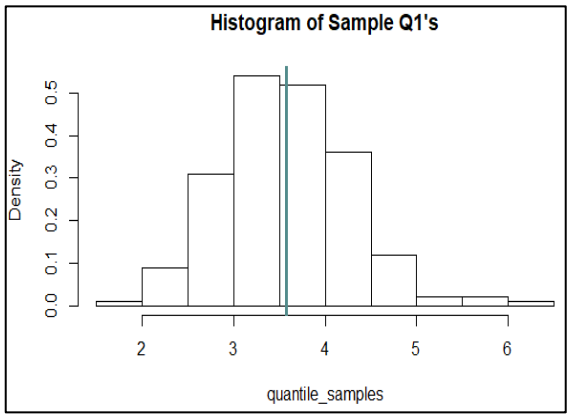

Figure 6

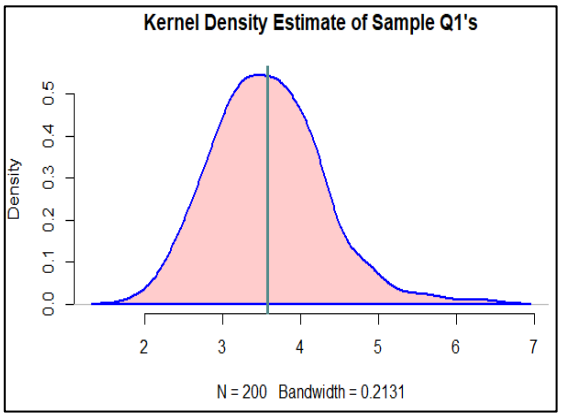

Figure 7

```
# Plotting histogram of sample Q1's (y-axis: density)
hist(quantile_samples, freq = FALSE, main = "Histogram of Sample Q1's")
# The option freq=FALSE plots probability densities instead of frequencies
# Adding a dashed line for the sample mean
abline(v=mean(quantile_samples), lwd=3, col='darkslategray4')

# Fitting a density estimate for sample Q1's (default estimator)
d = density(quantile_samples)
plot(d, main = "Kernel Density Estimate of Sample Q1's", bty = 'n')
polygon(d, col = "#FFCCCC", border = 'blue', lwd = 2)
# Adding a dashed line for the sample mean
abline(v=mean(quantile_samples), lwd=3, col='darkslategray4')
```

(b) Generate 200 random samples of size *n* = 100 from a Chi-Square distribution with degrees of freedom *df* = 6.

The same set of R codes shown above can be used to generate 200 independent random samples of size $n = 100$ from a Chi-Square distribution ($df = 6$).

i. Draw a histogram of the sample first quantiles (where the *y*-axis is the density) and fit a density estimate (default density estimator is ok).

The distribution of sample Q1's is displayed using histogram and density plots as shown in Figure 8 and Figure 9, respectively. In these plots, the position of the mean of simulated sample Q1's is indicated by a vertical solid line and has a value of 3.51580.

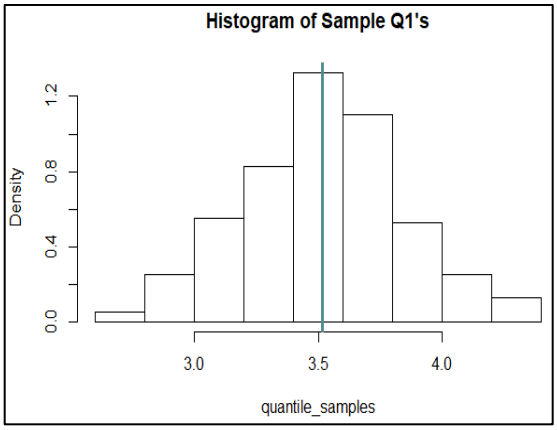

Figure 8

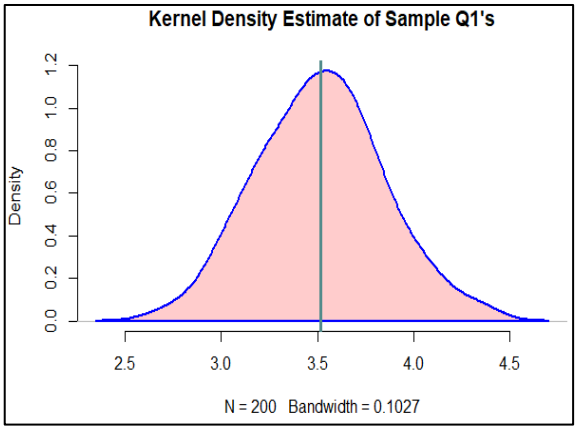

Figure 9

ii. Are there any differences between the distribution from Problem 2(a)ii and 2(b)i? Describe any differences you may see.

The simulated mean and variance of sample Q1's for distribution in Problem 2(a)ii are 3.57929 and 0.49768, respectively. For distribution in Problem 2(b)i these values are 3.51580 and 0.10846, respectively. We can see that increasing the size of sample does not significantly change the estimated mean of sample Q1's. However, the variance becomes smaller with increasing sample size.

From Figure 6 and Figure 7 we see that for Problem 2(a)ii, although the distribution seems to have a bell curve, it has some positive skewness and looks to approximate the distribution of original function i.e. Chi-square. This seems reasonable because the sample size is only 20. Inversely, from Figure 8 and Figure 9, the distribution of sample Q1's for Problem 2(b)i seems to follow increasingly normal distribution because of larger sample size.

To further support these findings, the Shapiro-Wilk Normality Test is performed on the sample Q1's for both these problems. The R code and results of the test are shown below.

```
## Shapiro-Wilk Normality Test
## H_0: population is normal
## H_1: population is not normal
shapiro.test(quantile_samples_norm)
```

Result of test for Problem 2(a)ii:

```
        Shapiro-Wilk normality test

data:  quantile_samples_norm
W = 0.98137, p-value = 0.009348
```

Result of test for Problem 2(b)i:

```
        Shapiro-Wilk normality test

data:  quantile_samples_norm
W = 0.99666, p-value = 0.9458
```

From above results, we see that the p-value is very small for Problem 2(a)ii and thus we reject the null hypothesis indicating that the distribution of sample Q1's is not normal. However, for Problem 2(b)i, the p-value is very large and thus we do not reject the null hypothesis indicating that the distribution of sample Q1's possibly follows a normal distribution.

Thus, we can conclude that with larger sample size the distribution of sample Q1's will have an increasingly smaller variance and an increasingly normal distribution.