1.

```r
set.seed(01052001)
p <- 0.05
n_samp <- 10000
samp_sizes <- c(15, 30, 50)

samp <- function(n, p, n_samp) {
  replicate(n_samp, rbinom(1, n, p))
}

samp_15 <- samp(15, p, n_samp)
samp_30 <- samp(30, p, n_samp)
samp_50 <- samp(50, p, n_samp)
```

2. (a)

```r
results <- sapply(samp_sizes, function(n) {
  samp <- samp(n, p, n_samp)  # Use the samp() function defined earlier
  c(Mean = mean(samp), Standard_Error = sd(samp)) / sqrt(n)
})


means_and_se <- data.frame(
  Sample_Size = samp_sizes,
  Mean = results["Mean", ],
  Standard_Error = results["Standard_Error", ]
)
print(means_and_se)
```

```
##   Sample_Size      Mean Standard_Error
## 1          15 0.1958697      0.2184487
## 2          30 0.2733866      0.2133643
## 3          50 0.3536807      0.2183409
```

2. (b)

```r
set.seed(01052001)
p <- 0.05
n_samp <- 10000
samp_sizes <- c(15, 30, 50)

get_stats <- function(n) {
  samp <- replicate(n_samp, mean(rbinom(n, size = 1, prob = p)))
  c(Mean = mean(samp), SD = sd(samp))
}

stats <- sapply(samp_sizes, get_stats)

mean_diff <- diff(stats["Mean", ])
sd_diff <- diff(stats["SD", ])

# Print results
cat("Mean Differences:", mean_diff, "\n")
```

```
## Mean Differences: 0.00021 6.266667e-05
```

```r
cat("Standard Deviation Differences:", sd_diff, "\n")
```

```
## Standard Deviation Differences: -0.01597098 -0.009108748
```

2. (c)

```r
theo_stats <- data.frame(
  "Sample_Size" = samp_sizes,
  "Theoretical_Mean" = samp_sizes * p,
  "Theoretical_SE" = sqrt(samp_sizes * p * (1 - p)) / sqrt(samp_sizes)
)
theo_stats
```

```
##   Sample_Size Theoretical_Mean Theoretical_SE
## 1          15             0.75      0.2179449
## 2          30             1.50      0.2179449
## 3          50             2.50      0.2179449
```

The empirical and theoretical means should closely match for p=0.05, as both are based on the same probability. Similarly, the empirical and theoretical standard errors should align, with minor differences due to random variation. As the sample size increases, empirical values converge to theoretical ones, consistent with the law of large numbers.

3. (a)

```r
quartiles_15 <- quantile(samp_15, prob = c(0.25, 0.75))
quartiles_30 <- quantile(samp_30, prob = c(0.25, 0.75))
quartiles_50 <- quantile(samp_50, prob = c(0.25, 0.75))



quartiles_15
```

```
## 25% 75%
##   0   1
```

```r
quartiles_30
```

```
## 25% 75%
##   1   2
```

```r
quartiles_50
```

```
## 25% 75%
##   1   3
```

3. (b)

```r
true_quartiles_15 <- qnorm(c(0.25, 0.75), 15*p, sd= sqrt(p* (1-p)*15))
true_quartiles_30 <- qnorm(c(0.25, 0.75), 30*p, sd= sqrt(p* (1-p)*30))
true_quartiles_50 <- qnorm(c(0.25, 0.75), 50*p, sd= sqrt(p* (1-p)*50))

true_quartiles_15
```

```
## [1] 0.1806651 1.3193349
```

```r
true_quartiles_30
```

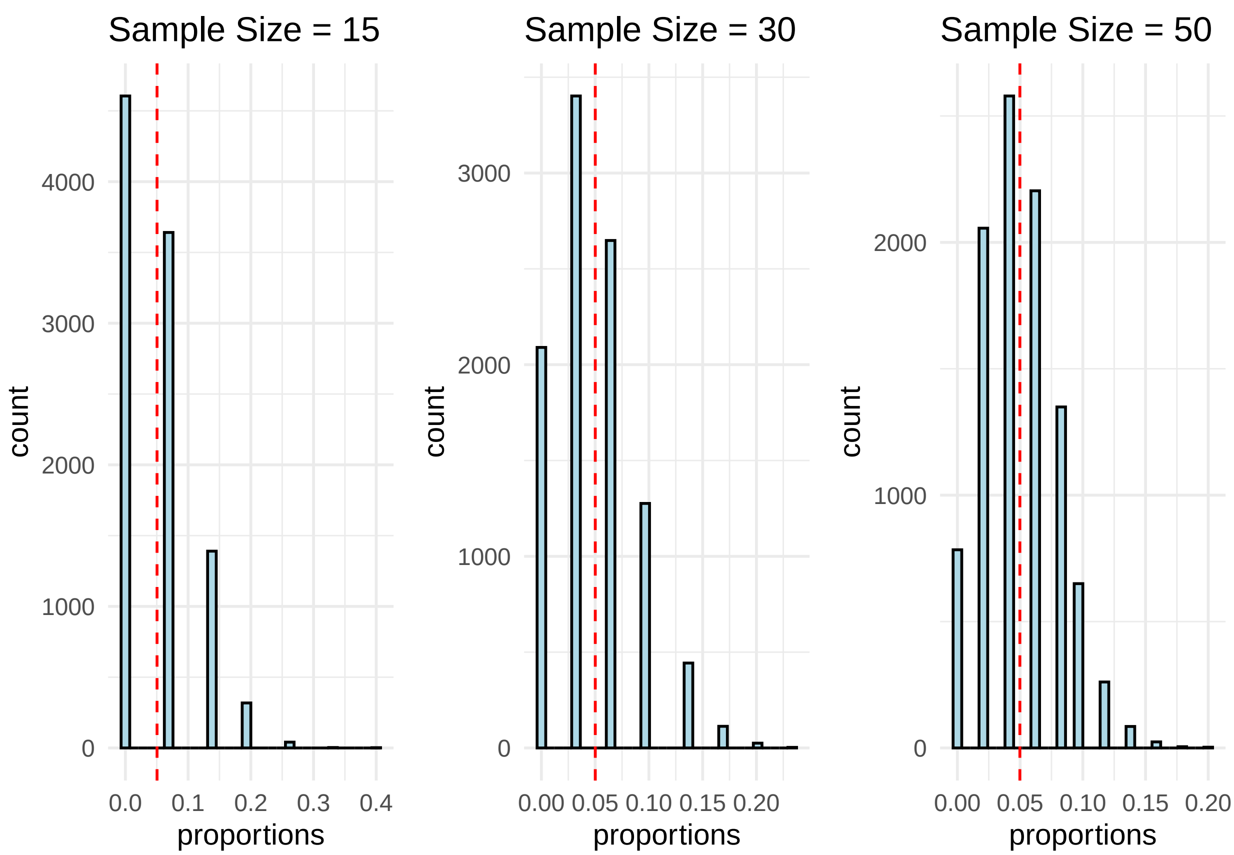
```
## [1] 0.6948389 2.3051611
```

```r
true_quartiles_50
```

```
## [1] 1.460541 3.539459
```

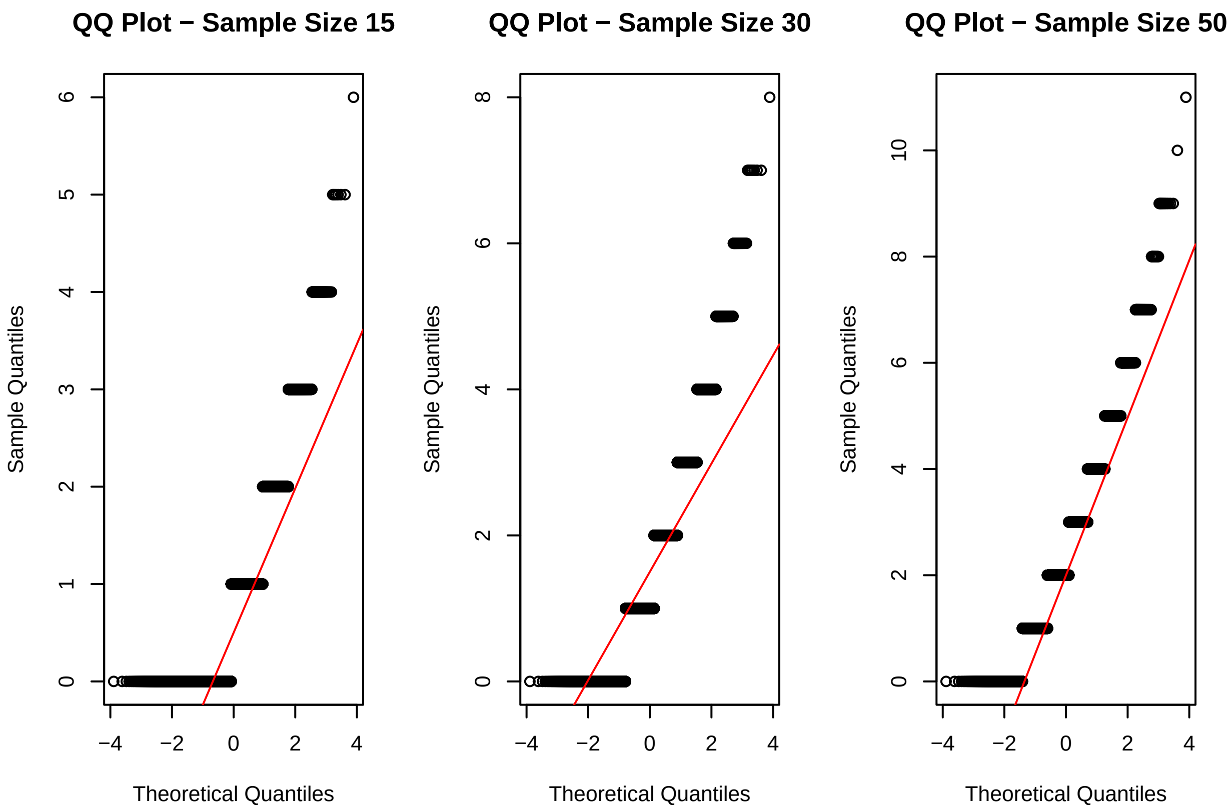
3. (c) True quartiles are based on the normal approximation of the binomial distribution, while empirical quartiles come from simulated data. As sample size increases, empirical quartiles should better match true quartiles. Smaller samples may show larger deviations because of random variation.
4.

```r
library(ggplot2)
library(gridExtra)

generate_proportions <- function(n) {
  rbinom(n_samp, n, p) / n
}

proportions_15 <- generate_proportions(15)
proportions_30 <- generate_proportions(30)
proportions_50 <- generate_proportions(50)

plot_histogram <- function(proportions, n) {

  ggplot(data.frame(proportions), aes(x = proportions)) +
    geom_histogram(bins = 30, fill = "lightblue", color = "black") +
    geom_vline(xintercept = mean(proportions), color = "red", linetype = "dashed") +
    ggtitle(paste("Sample Size =", n)) +
    theme_minimal()
}

hist_15 <- plot_histogram(proportions_15, 15)
hist_30 <- plot_histogram(proportions_30, 30)
hist_50 <- plot_histogram(proportions_50, 50)

grid.arrange(hist_15, hist_30, hist_50, ncol = 3)
```



5.

```
generate_qqplot <- function(samp, n) {
  qqnorm(samp, main = paste("QQ Plot - Sample Size", n))
  qqline(samp, col = "red")
}

par(mfrow = c(1, 3))
generate_qqplot(samp(15, p, n_samp), 15)
generate_qqplot(samp(30, p, n_samp), 30)
generate_qqplot(samp(50, p, n_samp), 50)
```



6. The Central Limit Theorem (CLT) states that for large sample sizes, the sampling distribution of the sample proportion approximates normality. When p=0.05, the CLT's accuracy depends on sample size. Small samples like n=15 deviate from normality, while larger samples such as n=30 and n=50 show more normal distributions. This demonstrates the CLT's reliability with larger samples.

```
# Load the data
download.file("http://www.openintro.org/stat/data/atheism.RData", destfile =
"atheism.RData")
load("atheism.RData")
```

7.

```
selected_nationality <- "Canada"

subset_data <- subset(atheism, nationality == selected_nationality & year == 2012)

head(subset_data)
```

```
##       nationality    response year
## 10107      Canada non-atheist 2012
## 10108      Canada non-atheist 2012
## 10109      Canada non-atheist 2012
## 10110      Canada non-atheist 2012
## 10111      Canada non-atheist 2012
## 10112      Canada non-atheist 2012
```

8.

```
samp_size <- nrow(subset_data)

atheist_count <- sum(subset_data$response == "atheist")
```

```r
samp_proportion <- atheist_count / samp_size

cat("Sample Size:", samp_size, "\n")
```

```
## Sample Size: 1002
```

```r
cat("Sample Proportion of Atheists:", samp_proportion, "\n")
```

```
## Sample Proportion of Atheists: 0.08982036
```

9. (a) It is reasonable to assume independence if participants were randomly selected and the sample size represents a small fraction of the total population
10. (b)

```r
p_hat <- mean(subset_data$response == "atheist")

n <- nrow(subset_data)

successes <- n * p_hat
failures <- n * (1 - p_hat)

cat("Successes:", successes, "\n")
```

```
## Successes: 90
```

```r
cat("Failures:", failures, "\n")
```

```
## Failures: 912
```

10.

```r
p_hat <- mean(subset_data$response == "atheist")
n <- nrow(subset_data)
se <- sqrt(p_hat * (1 - p_hat) / n)
z <- 1.96  # for 95% confidence

ci <- p_hat + c(-1, 1) * z * se
ci
```

```
## [1] 0.07211629 0.10752443
```

My interpretation: The interval suggests that we are 95% confident that the true proportion of atheists in the chosen nationality in 2012 lies within this range.