Keck_Brandon_STAT620_HW3

Brandon Keck

2024-09-29

(1)

```
set.seed(11101987) # Scorpio

# Generate 10,000 samples of size n = 15
prop_15 <- rep(NA, 10000)
for(i in 1:10000){
  success15 <- rbinom(1, 15, prob = 0.05) # n = 15, p = 0.05
  prop_15[i] <- success15 / 15 # proportion of successes
}

# Generate 10,000 samples of size n = 30
prop_30 <- rep(NA, 10000)
for(i in 1:10000){
  success30 <- rbinom(1, 30, prob = 0.05) # n = 30, p = 0.05
  prop_30[i] <- success30 / 30 # proportion of successes
}

# Generate 10,000 samples of size n = 50
prop_50 <- rep(NA, 10000)
for(i in 1:10000){
  success50 <- rbinom(1, 50, prob = 0.05) # n = 50, p = 0.05
  prop_50[i] <- success50 / 50 # proportion of successes
}
```

(2) Mean and standard error:

(a)

Sample Size	Mean	SD
n = 15	0.0495	0.0565
n = 30	0.05	0.0396
n = 50	0.0502	0.0309

(2)

(b)

The means of all sample sizes are relatively close approximately equal to 0.05, with small variations between each. The standard deviations however have some noticeable differences. For example when n = 15, SD = 0.0565, while n = 50 it decreases to 0.0309. This indicates that, although the sample means are similar, the variability of the sample proportions decreases as the sample size increases.

(2)

(c) Theoretical mean and standard deviation

```
p <- 0.05
sqrt(p*(1-p)/15) # n = 15
```

```
## [1] 0.05627314
```

```
sqrt(p*(1-p)/30) # n = 30
```

```
## [1] 0.03979112
```

```
sqrt(p*(1-p)/50) # n = 50
```

```
## [1] 0.03082207
```

The theoretical mean of the sampling distribution is set at 0.05, which is aligned with the observed means from the sampling distributions. The standard errors calculated from the theoretical mean-approximately 0.0563, 0.0398, and 0.0308 for sample sizes = 15, n = 30 and n = 50 respectively show a similar patter to those derived from the empirical sampling distributions.

(3)

(a)

```
# calculate the percentiles for sample of size n = 15
quantile(prop_15, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##   0.0   0.2
```

```
# calculate the percentile for sample of size n = 30
round(quantile(prop_30, c(0.025, 0.975)), 2)
```

```
##  2.5% 97.5%
##  0.00  0.13
```

```r
# calculate the percentile for sample size n = 50
quantile(prop_50, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##  0.00  0.12
```

(3)

(b)

```r
# Calculate the first (25th) and third (75th) quartiles
c(qnorm(0.025, mean(prop_15), sd(prop_15)), qnorm(0.975, mean(prop_15), sd(prop_15))) # when n = 15
```

```
## [1] -0.06120784  0.16020784
```

```r
c(qnorm(0.025, mean(prop_30), sd(prop_30)), qnorm(0.975, mean(prop_30), sd(prop_30))) # when n = 30
```

```
## [1] -0.02755412  0.12752078
```

```r
c(qnorm(0.025, mean(prop_50), sd(prop_50)), qnorm(0.975, mean(prop_50), sd(prop_50))) # when n = 50
```

```
## [1] -0.01029824  0.11065824
```

(3)

(c)

The intervals between the true quartiles and the empirical quartiles are relatively close, meaning that the empirical distributinos align well with the theoretical. This suggests that as sample sizes increase, the sample proportions approach the true population proportion. However, it is worth noting that the true quaritles contain negative values, while the empirical quartiles are all positive.

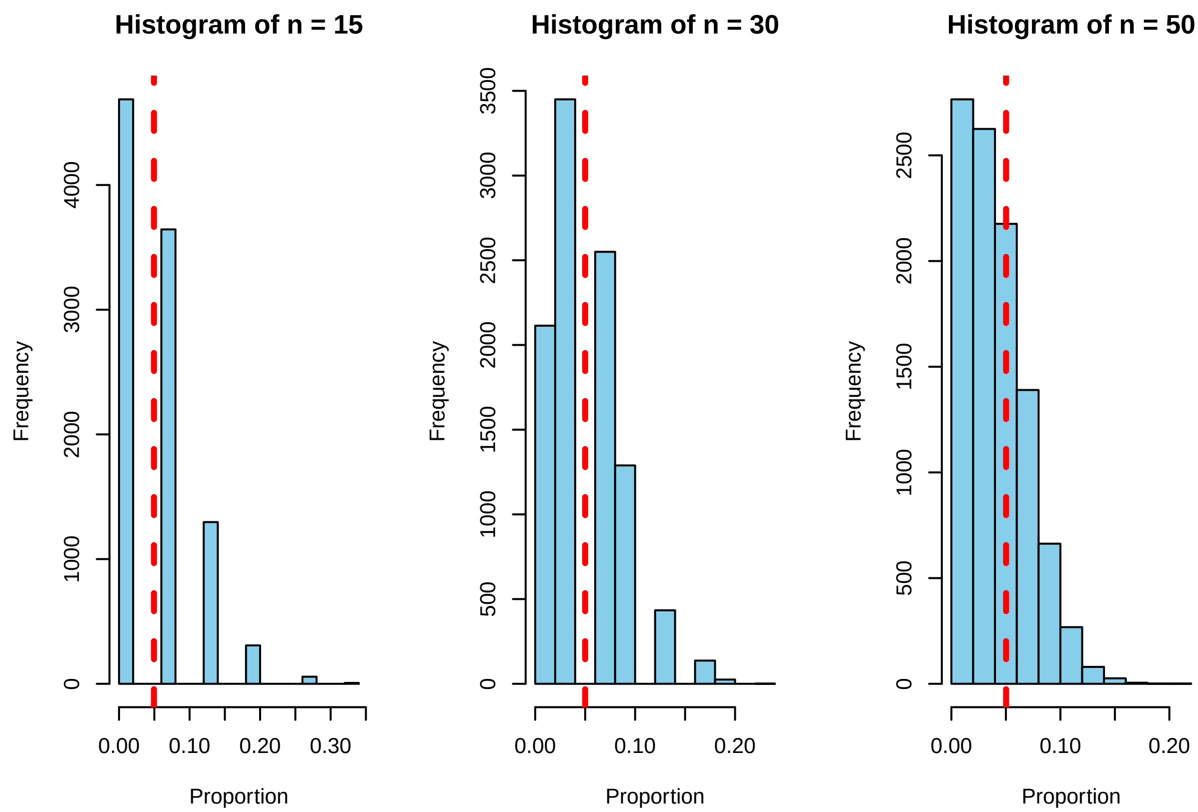
(4)

```r
par(mfrow = c(1, 3)) # displaying three plots on the same page

# Histogram of n = 15
hist(prop_15, main = "Histogram of n = 15", xlab = "Proportion", ylab = "Frequency", col = "skyblue")
abline(v = mean(prop_15), col = "red", lwd = 3, lty = "dashed") # abline to add a vertical line and lty

# Histogram of n = 30
hist(prop_30, main = "Histogram of n = 30", xlab = "Proportion", ylab = "Frequency", col = "skyblue")
abline(v = mean(prop_30), col = "red", lwd = 3, lty = "dashed") # abline to add a vertical line and lty

# Histogram of n = 50
hist(prop_50, main = "Histogram of n = 50", xlab = "Proportion", ylab = "Frequency", col = "skyblue")
abline(v = mean(prop_50), col = "red", lwd = 3, lty = "dashed") # abline to add a vertical line and lty
```



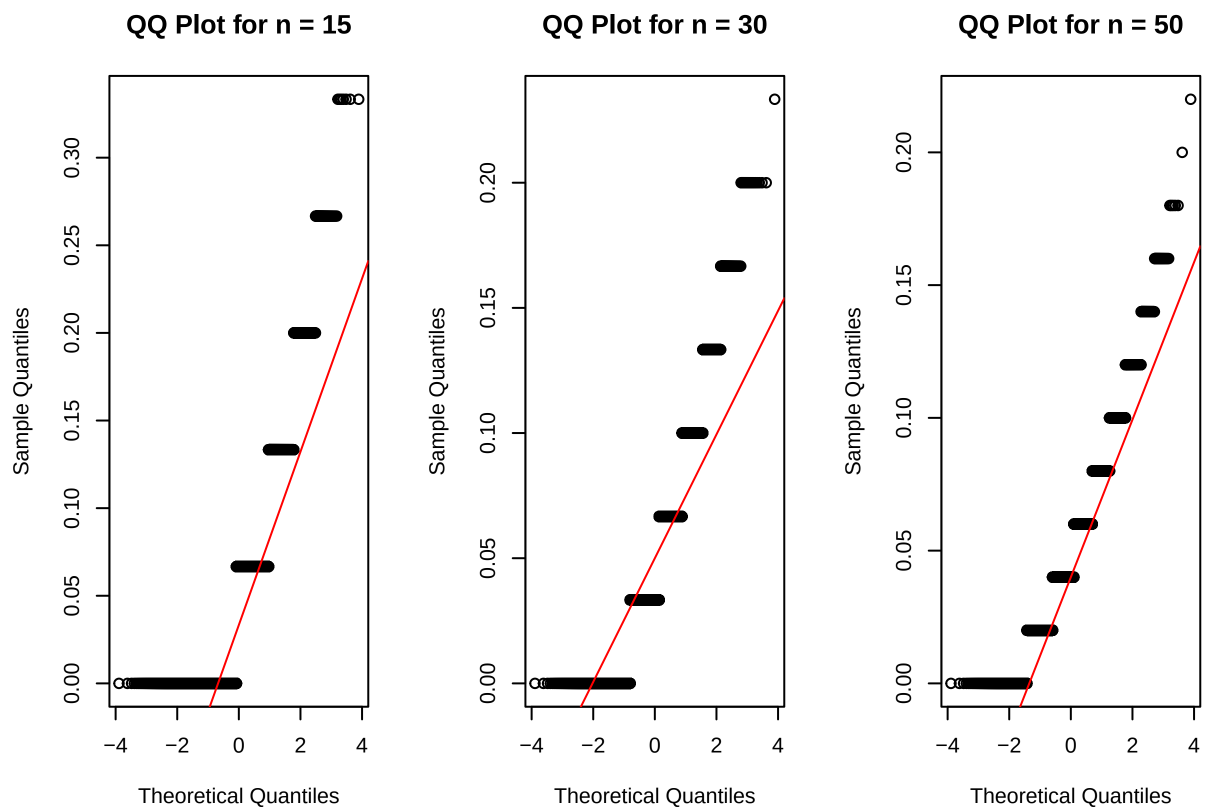
(5)

```
par(mfrow = c(1, 3)) # displaying three plots on the same page

# QQ plot for prop_15
qqnorm(prop_15, main = "QQ Plot for n = 15") # Creates QQ Plot n = 15
qqline(prop_15, col = "red") # adds reference line

# QQ plot for prop_30
qqnorm(prop_30, main = "QQ Plot for n = 30") # Creates QQ Plot n = 30
qqline(prop_30, col = "red") # adds reference line

# QQ plot for prop_50
qqnorm(prop_50, main = "QQ Plot for n = 50") # Creates QQ Plot n = 50
qqline(prop_50, col = "red") # adds reference line
```



Evaluating normality the data has a discrete nature. Since the points do not follow a straight line it implies that the data doesn't follow a normal distribution. The sampling distributions of the sample proportion indicate deviation from normality.

(6)

The Central Limit Theorem states that as the sample size increases, the distribution of the sample proportion p will approach normality. However, because the population proportion is quite small and the sample sizes are limited the validty of the CLT for a small p is compromised. The above QQ plots illustrate significant deviation from the reference line. This means the sampling distributions for smaller sample sizes do not approximate normality.

(7)

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

```
# install.packages("dplyr")
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
glimpse(atheism)
```

```
## Rows: 88,032
## Columns: 3
## $ nationality <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan"~
## $ response    <fct> non-atheist, non-atheist, non-atheist, non-atheist, non-at~
## $ year        <int> 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012~
```

```
# Nationality to filter by
chosen_nationality <- "Ireland"
subset_data <- atheism %>%
  filter(nationality == chosen_nationality, year == 2012)
# Use the filter function to filter 'atheism' to match the chosen nationality and the year 2012
```

(8)

```
sample_size <- nrow(subset_data)
sample_proportion_atheists <- subset_data %>%
  filter(response == "atheist") %>%
  summarise(proportion = n() / sample_size)

cat("Sample size:", sample_size, "\n")
```

```
## Sample size: 1010
```

```
cat("Sample Proportion of Atheists:",
sample_proportion_atheists$proportion, "\n")
```

```
## Sample Proportion of Atheists: 0.0990099
```

From a sample of size 1010 and a sample proportion of Atheists of 0.0990099 this tells us that in 2012 the proportion of Atheists in Ireland was about 9.9%.

(9)

(a)

It is reasonable to assume that each response was independent of one another. Each respondent's response is unlikely to affect the other one.

(b)

```r
n <- 1010 # Sample size of Ireland
p <- 0.0990099 # Proportion of Atheist in Ireland

# Calculate success and failures
success_condition <- n * p
failures_condition <- n * (1-p)

success_condition
```

```
## [1] 100
```

```r
failures_condition
```

```
## [1] 910
```

(10)

```r
phat <- sample_proportion_atheists # Proportion of Atheist
n <- sample_size # Sample size

ci_low <- phat - qnorm(0.975) * sqrt((phat*(1-phat))/n) # upper bound
ci_high <- phat + qnorm(0.975) * sqrt((phat*(1-phat))/n) # lower bound


print(c(ci_low, ci_high))
```

```
## $proportion
## [1] 0.08059002
##
## $proportion
## [1] 0.1174298
```

We are 95% confident that the true proportion of atheists among the population in Ireland from which this sample was drawn is between 8.06% and 11.74%. This means that if we were to take many random samples and compute a confidence interval for each sample, approximately 95% of those intervals would capture the true proportion of atheists in the population.

(11)

I honestly found this whole assignment to be very difficult. Normally I'm able to complete a homework by myself. However, this time I really needed the input of fellow classmates in order to fully understand my mistakes. It was nice being able to work together on certain parts and share our knowledge with one another and what worked for them and what did not.

(12)

M- Meeting Expectations