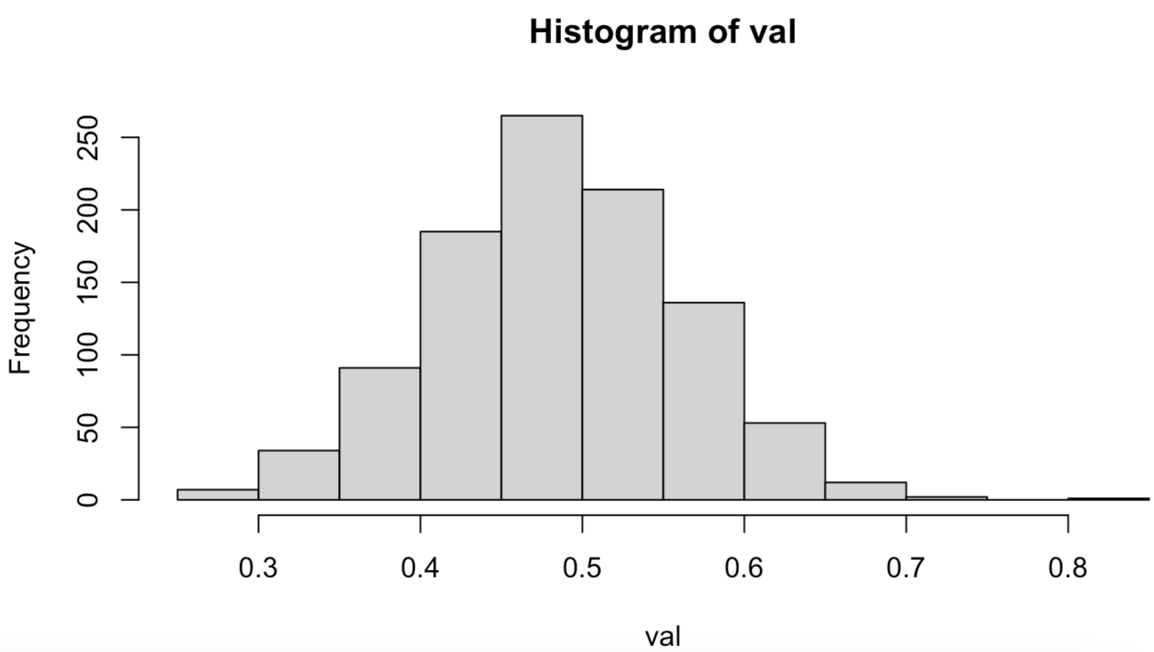
DATA 557 Homework 1

Amrit Bhat

Q 1.1

```
# 1.1
# Q: Should we plot the distribution?
val <- rbinom(n=1000,size = 40,prob = 0.5)/40
hist(val)
```

The distribution of the number of heads in the set of coin flips is



Q 1.2

```
#1.2
val <- rbinom(n=1000,size = 40,prob = 0.5)/40
mean <- mean(val)
SD <- sd(val)
print(paste("Mean: ",mean,"Standard Deviation: ",SD))
```

```
> print(paste("Mean: ",mean,"Standard Deviation: ",SD))
[1] "Mean:  0.5036 Standard Deviation:  0.0778819979559179"
```

Amrit Bhat

Q 1.3

The Z-statistic for conducting a test of the null hypothesis that the coin is fair is defined as below

```
#1.3
Z-Statistic | p = 0.5 = X-20/3.162277
```

Q 1.4

```
#1.4
z-statistic | no. of heads = 15
z_score = (15-20)/3.162277
p_value = 2*(pnorm(z_score))
print(paste("z_score: ",z_score, "P-value of experiment: ",p_value,
            "\n Therefore H0 is Rejected:",p_value<0.05))
```

```
[1] "z_score:  -1.58113916016845 P-value of experiment:  0.113846222550158 \n Therefore H0 is Rejected: FALSE"
```

Q 1.5

```
#1.5
type_1_error = 0.1
print(paste("Is H0 rejected with type_1_error = 0.1? ", p_value<type_1_error))
```

```
[1] "Is H0 rejected with type_1_error = 0.1?  FALSE"
```

Q 1.6

```
#1.6
p_value <- sum(dbinom(c(0:15,25:40),size=40,p=0.5))
print(paste("p_value using normal approximation: ",p_value))
```

```
[1] "p_value using normal approximation:  0.153859944162832"
```

Q 1.7

```
#1.7
# 95% confidence level corresponds to 1.96 times Standard Error as confirmed below
SD_multiplier_95 <- 1.96
p_hat <- 15/40
SE <- sqrt(p_hat*(1-p_hat)/40)
print(paste("Confidence level percentage for SD_multiplier_95: ", (1 - 2*(1-pnorm(SD_multiplier_95)))*100))
lower_bound_95 <- (p_hat-SD_multiplier_95*SE)
upper_bound_95 <- (p_hat+SD_multiplier_95*SE)
print(paste("95% Confidence level bounds (lower,upper): (",lower_bound_95,",",upper_bound_95,")"))
print(paste("Does the confidence interval include the value 0.5?",(lower_bound_95<0.5) & (upper_bound_95>0.5)))
```

```
[1] "95% Confidence level bounds (lower,upper): ( 0.22496875325453 , 0.52503124674547 )"
> print(paste("Does the confidence interval include the value 0.5?",(lower_bound_95<0.5) & (upper_bound_95>0.5)))
[1] "Does the confidence interval include the value 0.5? TRUE"
```

Q 1.8

```
#1.8
# 90% confidence level corresponds to 1.645 times Standard Error as confirmed below
SD_multiplier_90 <- 1.645
print(paste("Confidence level percentage for SD_multiplier_90: ", (1 - 2*(1-pnorm(SD_multiplier_90)))*100))
lower_bound_90 <- (p_hat-SD_multiplier_90*SE)
upper_bound_90 <- (p_hat+SD_multiplier_90*SE)
print(paste("90% Confidence level (lower,upper): (",lower_bound_90,",",upper_bound_90,")"))
```

```
[1] "Confidence level percentage for SD_multiplier_90:  90.0030188921757"
> lower_bound_90 <- (p_hat-SD_multiplier_90*SE)
> upper_bound_90 <- (p_hat+SD_multiplier_90*SE)
> print(paste("90% Confidence level (lower,upper): (",lower_bound_90,",",upper_bound_90,")"))
[1] "90% Confidence level (lower,upper): ( 0.249080917910052 , 0.500919082089948 )"
```

Final comment:
The 90% confidence interval range is narrower compared to 95% confidence interval, therefore there's more chance for the null hypothesis to be rejected

Q 2.1

```
#2.1
p_hat <- 305/400
n <- 400
SE <- sqrt(p_hat*(1-p_hat)/n)
print(paste("Estimated standard error of the proportion of drivers wearing seatbelts after the
intervention: ", SE))
```

```
[1] "Estimated standard error of the proportion of drivers wearing seatbelts after the \nintervention:  0.0212775556631865"
```

Amrit Bhat

Q 2.2

```
#2.2
p_hat <- 305/400
n <- 400
SE <- sqrt(p_hat*(1-p_hat)/n)
SD_multiplier_95 <- 1.96
lower_bound_95 <- (p_hat-SD_multiplier_95*SE)
upper_bound_95 <- (p_hat+SD_multiplier_95*SE)
print(paste("95% Confidence level bounds (lower,upper): (",lower_bound_95,",",upper_bound_95,")"))
```

```
[1] "95% Confidence level bounds (lower,upper): ( 0.720795990900154 , 0.804204009099845 )"
```

Final comment:
Based on above range of 95% confidence interval, the p-value would be less than 0.05, therefore the researcher can reject the null hypothesis that the proportion of drivers wearing their seatbelt after the intervention is equal to 0.7 (or unchanged from before)

Q 2.3

```
z_score = (305-280)/sqrt(280*0.3)
p_value_normal_approx = 2*(1-pnorm(z_score))
print(paste("p_value using normal approximation is: ", p_value_normal_approx))
```

```
[1] "p_value using normal approximation is:  0.00637730142168835"
```

Final comment:
As p_value is lower than type 1 error probability 0.05, the researcher can reject the null hypothesis that proportion of drivers wearing seatbelt after intervention is same as before intervention. This conclusion is same as the conclusion from the confidence interval found earlier.

Amrit Bhat

Q 2.4

```
n_dri_w_seatbelts <- 0.7*400
print(n_dri_w_seatbelts)   #   <- null hypothesis value
# As null hypothesis value (280) is lower than new hypothesis sample (305)
# We find p_value between 0:280-25 and 305:400
p_value_binomial <- sum(dbinom(c(0:255,305:400),size=400,p=0.7))
print(paste("p_value using binomial distribution: ",p_value_binomial))
```

```
[1] "p_value using binomial distribution:  0.0074436166718936"
```

As p-value is lower than type 1 error probability 0.05, the researcher can reject the null hypothesis that proportion of drivers wearing seatbelt after intervention is same as before intervention.

```
print(paste("The p-value using normal approximation and binomial distribution are
            different by: ",p_value_normal_approx-p_value_binomial))
```

```
[1] "The p-value using normal approximation and binomial distribution are \n            different by:  -0.00106631525020524"
```

Final comment: The above p-values are not very different.

Q 2.5

```
#2.5
lower <- qbinom(0.025,size = 400, prob = 0.7)
upper <- 400*0.7 + (400*0.7-lower)
power <- sum(dbinom(c(0:lower,upper:400),size=400,p=0.8))
print(paste("P(Reject H0|p=0.8):",power))
```

```
[1] "P(Reject H0|p=0.8): 0.996898117505343"
```

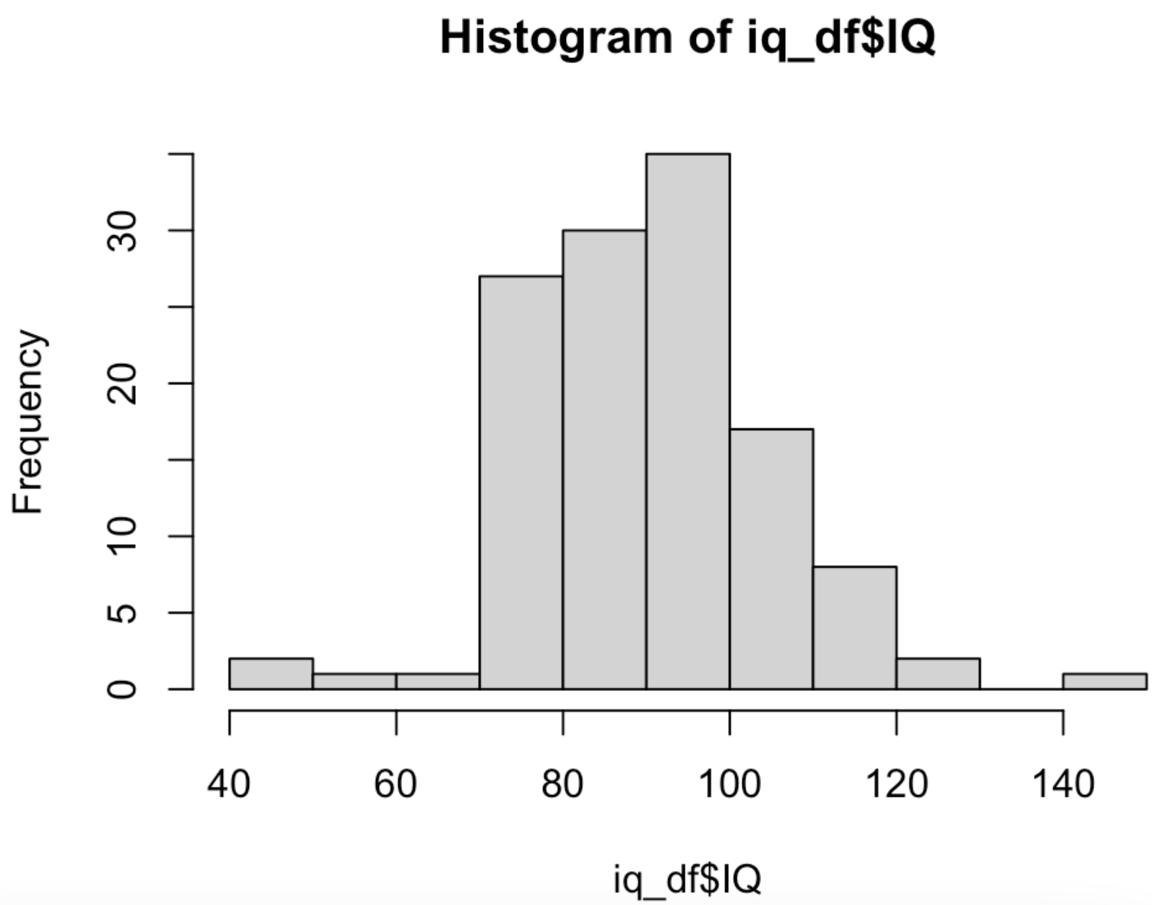
Amrit Bhat

Q 3.1

```
# 3.1
iq_df <- read.csv("/Users/amrit/Documents/Courses/Applied Statistics & Experimental Design/Assignments/A1/iq.csv")
hist(iq_df$IQ)
print(paste("Mean of IQ data : ",mean(iq_df$IQ)))
print(paste("Standard Deviation of IQ data: ", sd(iq_df$IQ)))

normal_values <- rnorm(length(iq_df$IQ), mean=100,sd=15)
hist(normal_values)``
print(paste("Mean of Normal distribution with same parameters: ",mean(normal_values)))
print(paste("Standard Deviation Normal distribution with same parameters: ", sd(normal_values)))
```

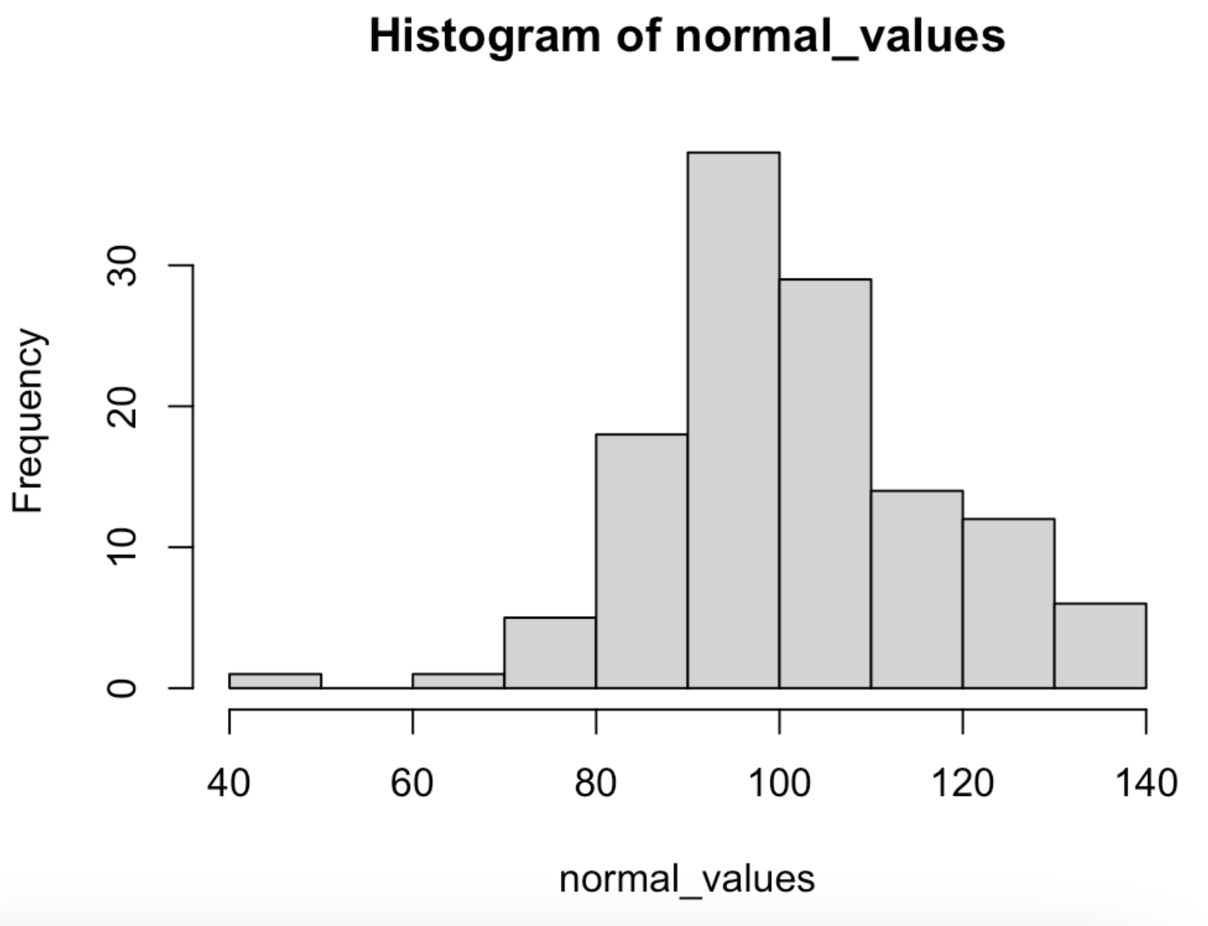
```
> print(paste("Mean of IQ data : ",mean(iq_df$IQ)))
[1] "Mean of IQ data :  91.0806451612903"
> print(paste("Standard Deviation of IQ data: ", sd(iq_df$IQ)))
[1] "Standard Deviation of IQ data:  14.403927182528"
```



DATA 557 Homework 1

Amrit Bhat

```
> print(paste("Mean of Normal distribution with same parameters: ",mean(normal_values)))
[1] "Mean of Normal distribution with same parameters:  101.462794871654"
> print(paste("Standard Deviation Normal distribution with same parameters: ", sd(normal_values)))
[1] "Standard Deviation Normal distribution with same parameters:  15.2777948106384"
```



Final comment:
As seen above, the distribution of the IQ variable approximately resembles a normal distribution as their graphic representations along with their mean and standard deviation is similar.

Q 3.2

```
#3.2
print(paste("Mean of IQ data : ",mean(iq_df$IQ)))
print(paste("Standard Deviation of IQ data: ", sd(iq_df$IQ)))
lower <- qnorm(0.025,mean=100, sd=15)
print(paste("The lower bound for the 95% confidence interval is: ",lower))
```

```
> print(paste("Mean of IQ data : ",mean(iq_df$IQ)))
[1] "Mean of IQ data :  91.0806451612903"
> print(paste("Standard Deviation of IQ data: ", sd(iq_df$IQ)))
[1] "Standard Deviation of IQ data:  14.403927182528"
> lower <- qnorm(0.025,mean=100, sd=15)
> print(paste("The lower bound for the 95% confidence interval is: ",lower))
[1] "The lower bound for the 95% confidence interval is:  70.6005402318992"
```

Final comment:
The mean is very well inside the lower bound of the 95% confidence interval.

Q 3.3

```
# 3.3
SE <- sqrt(var(iq_df$IQ)/length(iq_df$IQ))
t_statistic <- (mean(iq_df$IQ)-100)/SE
critical.value = qt(0.975,df=length(iq_df$IQ)-1)
print(paste("t_statistic: ",t_statistic, " critical.value: ",critical.value))
print(paste("Reject null hypothesis that population mean is 100?: ",abs(t_statistic)>critical.value))
```

```
[1] "t_statistic:  -6.89546196410128  critical.value:  1.97943868509329"
> print(paste("Reject null hypothesis that population mean is 100?: ",abs(t_statistic)>critical.value))
[1] "Reject null hypothesis that population mean is 100?:  TRUE"
```

Q 3.4

```
# 3.4
p_value <- 2*(1-pt(abs(t_statistic),df=length(iq_df$IQ)-1))
print(paste("The p_value is: ",p_value))
```

```
[1] "The p_value is:  2.48647324951889e-10"
```

Final comment: As the p-value is less than 0.05, the probability of the population mean being as or more extreme is higher therefore null hypothesis that the population mean is 100 is rejected

Q 3.5

```
# 3.5
population_mean <- 100
x_bar <- mean(iq_df$IQ)
lower <- x_bar - 1.96*SE
upper <- x_bar + 1.96*SE
print(paste("95% Confidence level bounds (lower,upper): (",lower,",",upper,")"))
print(paste("Is null hypothesis rejected according to confidence interval method?",(lower<population_mean) & (upper>population_mean)))
```

```
> print(paste("95% Confidence level bounds (lower,upper): (",lower,",",upper,")"))
[1] "95% Confidence level bounds (lower,upper): ( 88.5453639031405 , 93.6159264194401 )"
> print(paste("Is null hypothesis rejected according to confidence interval method?",(lower<population_mean) & (upper>population_mean)))
[1] "Is null hypothesis rejected according to confidence interval method? FALSE"
```

Final comment:

As the bounds of confidence interval (88.5453639031405 , 93.6159264194401) don't include population mean of 100, the null hypothesis that the population mean is 100 is rejected. This conclusion is the same as hypothesis test method.

Q 3.6

```
# 3.6
## Hypothesis test for significance level 0.01
SE <- sqrt(var(iq_df$IQ)/length(iq_df$IQ))
t_statistic <- (mean(iq_df$IQ)-100)/SE
critical.value = qt(0.995,df=length(iq_df$IQ)-1)
print(paste("t_statistic: ",t_statistic, " critical.value: ",critical.value))
print(paste("Reject null hypothesis that population mean is 100?: ",abs(t_statistic)>critical.value))
```

```
[1] "t_statistic:  -6.89546196410128  critical.value:  2.61639177642797"
> print(paste("Reject null hypothesis that population mean is 100?: ",abs(t_statistic)>critical.value))
[1] "Reject null hypothesis that population mean is 100?:  TRUE"
```

```
# Finding bounds for 99% confidence interval
population_mean <- 100
x_bar <- mean(iq_df$IQ)
SD_multiplier_99 <- 2.575
print(paste("Confidence level percentage for SD_multiplier_99: ", (1 - 2*(1-pnorm(SD_multiplier_99)))*100))
lower <- x_bar - SD_multiplier_99*SE
upper <- x_bar + SD_multiplier_99*SE
print(paste("99% Confidence level bounds (lower,upper): (",lower,",",upper,")"))
print(paste("Is null hypothesis rejected according to confidence interval method?",(lower>population_mean) | (upper<population_mean)))
```

```
> print(paste("99% Confidence level bounds (lower,upper): (",lower,",",upper,")"))
[1] "99% Confidence level bounds (lower,upper): ( 87.7498547328537 , 94.4114355897269 )"
> print(paste("Is null hypothesis rejected according to confidence interval method?",(lower>population_mean) | (upper<population_mean)))
[1] "Is null hypothesis rejected according to confidence interval method? TRUE"
```

Final comment:
As the bounds of confidence interval (87.7498547328537 , 94.4114355897269) don't include population mean of 100, the null hypothesis that the population mean is 100 is rejected. This conclusion is the same as hypothesis test method with 0.01 significance level.