DATA 557: Homework 1

Student: Corbin Charpentier

Question 1

1.1. What is the distribution of the number of heads assuming the coin is fair?

Binomial(0.5, 40)

1.2. The sample proportion of heads has an approximately normal distribution. What are the mean and standard deviation of this distribution assuming the coin is fair?

$$\mu_{\hat{p}} = P_{fair} = 0.5$$
$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5 * 0.5}{40}} \approx 0.079$$

1.3. Define the Z-statistic for conducting a test of the null hypothesis that the coin is fair (i.e., has probability of a head equal to 0.5).

$$Z = \frac{\hat{p} - 0.5}{\sigma_{\hat{p}}}$$

1.4. Suppose the experiment results in 15 heads and 25 tails. Conduct a test of the null hypothesis with type I error probability 0.05 using the normal approximation. State the Z-statistic, the p-value, and the conclusion of the test (do you reject the null hypothesis or not).

$$H_0 : \mu_{\hat{p}} = 0.5$$

```
samp_prop <- 15/40
se_est    <- sqrt((samp_prop*(1 - samp_prop))/40)
z_stat    <- (samp_prop - 0.5)/se_est

print(paste(c("Z-statistic:", z_stat), collapse=" "))
```

```
## [1] "Z-statistic: -1.63299316185545"
```

```
p_value <- 0.0516 * 2 # Looked it up in a table
print(paste(c("p-value:", p_value), collapse=" "))
```

```
## [1] "p-value: 0.1032"
```

Since |Z-statistic| is less than 1.96, we do not reject the null hypothesis.

1.5. If you had decided to use a type I error probability of 0.1 instead of 0.05 would your conclusion be different? Explain.

I would also not reject the null hypothesis since the the absolute value of our critical value (1.63) is less than 1.645, the Z score threshold for rejecting the null hypothesis.

1.6. Calculate the p-value using the binomial distribution. Do you reach the same conclusion with the binomial distribution as with the normal approximation?

```
cumu_probs  <- pbinom(seq.int(0, 40, by=1), size=40, prob=0.5)
p_value_bin <- cumu_probs[16] + (1 - cumu_probs[25])
print(paste(c("p-value:", p_value_bin), collapse=" "))
```

```
## [1] "p-value: 0.153859944162832"
```

I did reach the same conclusion: do not reject the null hypothesis.

1.7. Calculate a 95% confidence interval for the probability of a head using the normal approximation. Does the confidence interval include the value 0.5?

```
interval_bound = 1.96*se_est
conf_int <- c(samp_prop - interval_bound, samp_prop + interval_bound)
print(paste(c("95% confidence interval: (", conf_int[1], ",", conf_int[2], ")"), collapse=""))
```

```
## [1] "95% confidence interval: (0.22496875325453,0.52503124674547)"
```

Yes, it does include 0.5.

1.8. Calculate a 90% confidence interval for the probability of a head using the normal approximation. How does it compare to the 95% confidence interval?

```
interval_bound = 1.645*se_est
conf_int <- c(samp_prop - interval_bound, samp_prop + interval_bound)
print(paste(c("95% confidence interval: (", conf_int[1], ",", conf_int[2], ")"), collapse=""))
```

```
## [1] "95% confidence interval: (0.249080917910052,0.500919082089948)"
```

This confidence interval is narrower than the 95% confidence interval.

Question 2

A study is done to determine if enhanced seatbelt enforcement has an effect on the proportion of drivers wearing seatbelts. Prior to the intervention (enhanced enforcement) the proportion of drivers wearing their seatbelt was 0.7. The researcher wishes to test the null hypothesis that the proportion of drivers wearing their seatbelt after the intervention is equal to 0.7 (i.e., unchanged from before). The alternative hypothesis is that the proportion of drivers wearing their seatbelt is not equal to 0.7 (either < 0.7 or > 0.7). After the intervention, a random sample of 400 drivers was selected and the number of drivers wearing their seatbelt was found to be 305.

2.1. Calculate the estimated standard error of the proportion of drivers wearing seatbelts after the intervention.

```r
samp_prop <- 305/400
se_est    <- sqrt((samp_prop*(1 - samp_prop))/400)
print(paste(c("Standard Error of sample proportion", se_est), collapse=" "))
```

```
## [1] "Standard Error of sample proportion 0.0212775556631865"
```

2.2. Calculate a 95% confidence interval for the proportion of drivers wearing seatbelts after the intervention. What conclusion would you draw based on the confidence interval?

```r
interval_bound = 1.96*se_est
conf_int <- c(samp_prop - interval_bound, samp_prop + interval_bound)
print(paste(c("95% confidence interval: (", conf_int[1], ",", conf_int[2], ")"), collapse=""))
```

```
## [1] "95% confidence interval: (0.720795990900154,0.804204009099845)"
```

Conclusion: Since the null hypothesis (0.7) lies outside of the 95% confidence interval, we reject the null hypothesis.

2.3. Conduct a test of the null hypothesis with type I error probability 0.05 using the normal approximation. Should the null hypothesis be rejected? How does your conclusion compare to the conclusion from the confidence interval?

```r
z_stat <- (samp_prop - 0.7)/se_est
print(paste(c("Z-statistic:", z_stat), collapse=" "))
```

```
## [1] "Z-statistic: 2.93736747723023"
```

Since the z-statistic is more than 1.96 standard deviations away from the mean, we again reject the null hypothesis.

2.4. Calculate the approximate p-value using the normal approximation and the exact p-value using the binomial distribution. Are the two p-values very different?

```r
p_value_normal <- 0.0016 * 2 # Looked it up in a table using z stat
print(paste(c("p-value via normal approximation:", p_value_normal), collapse=" "))
```

```
## [1] "p-value via normal approximation: 0.0032"
```

```r
cum_probs   <- pbinom(seq.int(0, 400), size=400, prob=0.7)
p_value_bin <- cum_probs[255] + (1 - cum_probs[305])
print(paste(c("p-value via binomial:", p_value_bin/2), collapse=" "))
```

```
## [1] "p-value via binomial: 0.00315016220743965"
```

2.5. Calculate the power of the test to detect the alternative hypothesis that the proportion of drivers wearing their seatbelt after the intervention is equal to 0.8.

```r
power <- 1 -  pnorm(0.1-se_est*1.96) + (1 - pnorm(0.1+se_est*1.96))
print(paste(c("Power:", power), collapse=" "))
```

```
## [1] "Power: 0.920413334490716"
```

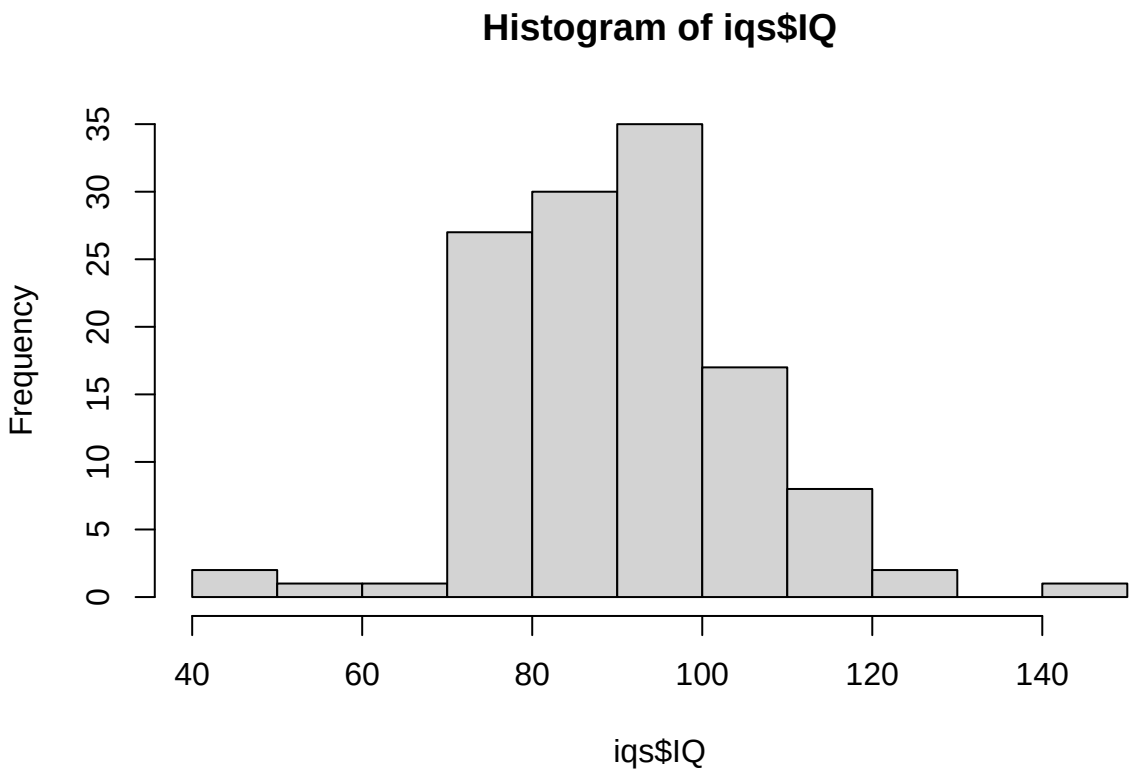
Question 3

Data set: 'iq.csv' (data set posted on canvas)

The data come from a study of lead exposure and IQ in children. IQ scores were measured on a sample of children living in a community near a source of lead. The IQ scores were age-standardized using established normal values for the US population. Such age-standardized scores have a mean of 100 and a standard deviation of 15 in the US population.

3.1. Create a histogram of the IQ variable. Is the distribution approximately normal?

```
iqs <- read.csv(file = 'data/iq.csv')
hist(iqs$IQ)
```



3.2. Calculate the sample mean and sample SD of IQ. How do they compare numerically to the US population values?

```
l<- length(iqs$IQ)
se <- sd(iqs$IQ)/sqrt(l)
smean <- mean(iqs$IQ)
print(paste(c("Standard error: ", se, ". Sample mean: ", smean), collapse=""))
```

```
## [1] "Standard error: 1.29351084599481. Sample mean: 91.0806451612903"
```

The mean is about one standard deviation lower for this sample than for the population. Standard error seems to be a close estimate of the population standard deviation.

3.3. Test the null hypothesis that the mean IQ score in the community is equal to 100 using the 2-sided 1-sample t-test with a significance level of 0.05. State the value of the test statistic and whether or not you reject the null hypothesis at significance level 0.05.

```r
t_stat <- (smean - 100)/se
print(paste(c("t-statistic:", t_stat), collapse=" "))
```

```
## [1] "t-statistic: -6.89546196410128"
```

```r
print("Threshold:")
```

```
## [1] "Threshold:"
```

```r
qt(0.05, l, lower.tail=TRUE)
```

```
## [1] -1.657235
```

Since abs(t-statistic), our critical value, is much greater than 1.657, the 0.05 significance threshold for 120 degress of freedom (using table), we reject the null hypothesis.

3.4. Give the p-value for the test in the previous question. State the interpretation of the p-value.

```r
p_value_t <- pt(q=t_stat, df=l, lower.tail=TRUE)*2
print(paste(c("p-value: ", p_value_t), collapse=" "))
```

```
## [1] "p-value:  2.42515019757264e-10"
```

```r
# ~97
```

Intrepretation: assuming the null hypothesis is true, it is extemely unlikely we would get this result by chance.

3.5. Compute a 95% confidence interval for the mean IQ. Do the confidence interval and hypothesis test give results that agree or conflict with each other? Explain.

```r
t.test(iqs$IQ, mu=100, alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  iqs$IQ
## t = -6.8955, df = 123, p-value = 2.486e-10
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
##  88.52022 93.64107
## sample estimates:
## mean of x
##  91.08065
```

Indeed, the confidence interval and hypothesis test yield consistent results.

3.6. Repeat the hypothesis test and confidence interval using a significance level of 0.01 and a 99% confidence interval.

```r
t.test(iqs$IQ, mu=100, conf.level=0.99, alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  iqs$IQ
```

```
## t = -6.8955, df = 123, p-value = 2.486e-10
## alternative hypothesis: true mean is not equal to 100
## 99 percent confidence interval:
##  87.69631 94.46498
## sample estimates:
## mean of x
##  91.08065
```

```
print("Threshold:")
```

```
## [1] "Threshold:"
```

```
qt(0.01, l, lower.tail=TRUE)
```

```
## [1] -2.356797
```

We still reject the null hypothesis.