DATA 557: Homework Assignment 1

Hriday Baghar

January 20, 2022

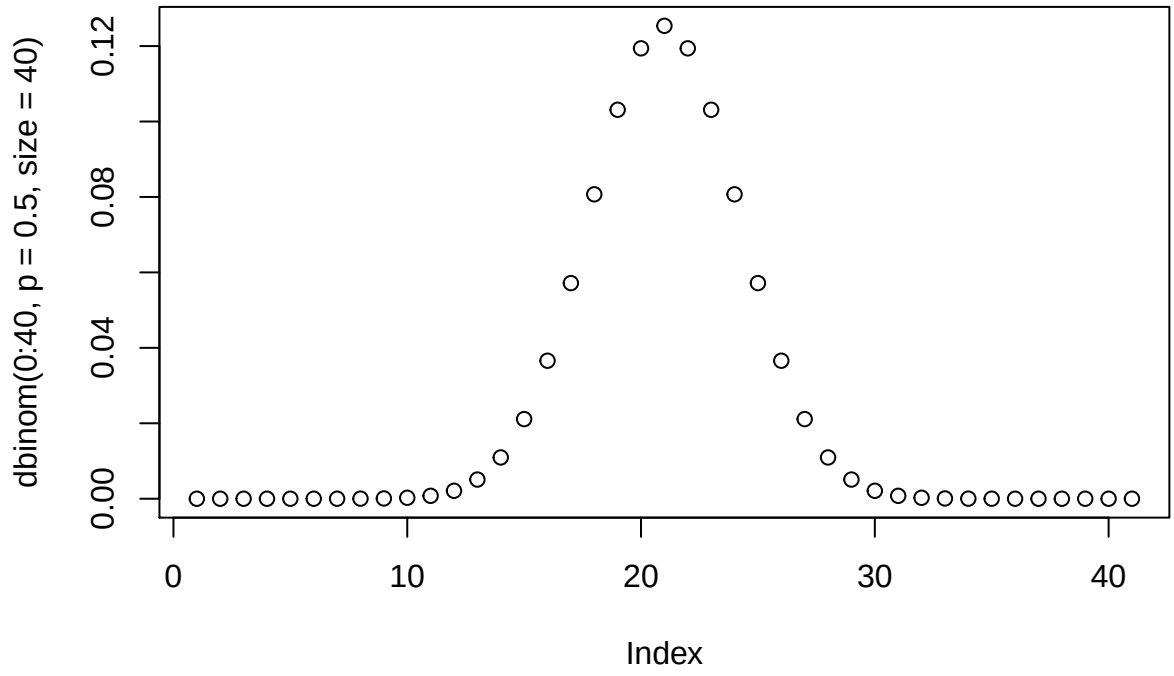
Question 1

Suppose that you flip a coin 40 times and count the number of heads.

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

The number of heads follows a binomial distribution, that approximates a normal distribution assuming the coin is fair.

```
plot(dbinom(0:40, p=0.5, size=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?

Since we assume that the coin is fair we have, $\mu = p = 0.5$.
We calculate standard deviation using the formula,

$$\sigma = \sqrt{\frac{p(1-p)}{n}}$$

```
N <- 40
p <- 0.5
cat("SD = ", round(sqrt(p*(1-p)/N),3))
```

```
## SD =  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).

We have n=40 and p=0.5, which gives us

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{X - 20}{\sqrt{10}}$$

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).

Let us assume X is the number of heads obtained We are given $\alpha = 0.05$, $X = 15$ and $Z = \frac{X-20}{\sqrt{10}}$

For the given α the rejection region would be $|Z| > 1.96$
We calculate the Z-statistic for the given value of X,

$$|Z| = |\frac{X - 20}{\sqrt{10}}| = |\frac{15 - 20}{\sqrt{10}}| = |-1.58|$$

We see that $|Z| = |-1.58|$ which is not in the rejection region $|Z| > 1.96$ for $\alpha = 0.05$
The p-value is given by,

```
2*pnorm(-1.58)
```

```
## [1] 0.1141069
```

The p-value 0.114 > 0.05, therefore **we do not reject the null hypothesis** - we have no evidence that the coin is unfair.

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.

If we used $\alpha = 0.1$ instead our **conclusion would still be the same (cannot reject H_0)** since our p-value of 0.114 is still greater than the significance level of 0.1

We can also see the new rejection region is

```
qnorm(0.95)
```

```
## [1] 1.644854
```

Since the new rejection region is $|Z| > 1.64$ and our test statistic still does not fall in that region, **our conclusion does not change**.

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?

Using the binomial distribution the p-value is,

```
sum(dbinom(c(0:15, 25:40), size = 40, p=0.5))
```

```
## [1] 0.1538599
```

We reach the same conclusion using the binomial distribution even though there is a slight difference in the p-value.

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?

We use the following formula to construct a 95% confidence interval:

$$(\hat{p} - 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

where $\hat{p} = 15/40 = 0.375$

```r
phat <- 15/40
se <- sqrt(phat*(1-phat)/40)
ci.lower <- round(phat - 1.96*se,3)
ci.upper <- round(phat + 1.96*se,3)
cat("95% confidence interval = (",ci.lower,",", ci.upper,")")
```

```
## 95% confidence interval = ( 0.225 , 0.525 )
```

The confidence interval contains the value 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?

We use 1.64 as the new cutoff value for the Z score, and use that to calculate the confidence interval

```r
ci.lower <- round(phat - 1.64*se,3)
ci.upper <- round(phat + 1.64*se,3)
cat("90% confidence interval = (",ci.lower,",", ci.upper,")")
```

```
## 90% confidence interval = ( 0.249 , 0.501 )
```

The 90% confidence interval is narrower than the 95% confidence interval but it still contains 0.5.

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.

We use the formula

```r
n <- 400
phat <- 305/n
se <- sqrt(phat*(1-phat)/n)
cat("Estimated SE = ", round(se,3))
```

```
## Estimated SE =  0.021
```

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
ci.lower <- round(phat - 1.96*se,3)
ci.upper <- round(phat + 1.96*se,3)

cat("95% confidence interval = (",ci.lower,",", ci.upper,")")
```

```
## 95% confidence interval = ( 0.721 , 0.804 )
```

Based on the confidence interval we can conclude that the true proportion of drivers wearing seatbelts after the intervention falls in the range (0.721, 0.804) 95% of the time. It appears that the proportion after the intervention has increased, but this should be back by a hypothesis test.

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?

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{0.7625 - 0.7}{0.0229}$$

```
p <- 0.7
z <- (phat-p)/sqrt(p*(1-p)/n)

cat('Z statistic = ',round(z,3))
```

```
## Z statistic =  2.728
```

Since $|Z| > 1.96$ we **reject the null hypothesis**, i.e. proportion of drivers wearing their seatbelt is not equal to 0.7. This conclusion is consistent with our confidence interval which does not contain the value 0.7.

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?

```
p.val.norm <- 2*(1-pnorm(z))
cat('p-value (using normal approximation) = ', round(p.val.norm,3))
```

```
## p-value (using normal approximation) =  0.006
```

To obtain the result using binomial distribution, we want to check the probability of obtaining a result as or more extreme than 305, given mean $np = 400 * 0.7 = 280$. We want values that are $305 - 280 = 25$ further from the mean. Lower limit for rejection region would then be $280 - 25 = 255$

```
p.val.bin <- sum(dbinom(c(0:255,305:400), size = 400, p=0.7))
cat('\np-value (using binomial distribution) = ', round(p.val.bin,3))
```

```
##
## p-value (using binomial distribution) =  0.007
```

The p-values are similar.

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.

We must define the rejection region under the Null Hypothesis $p = 0.7$. We find that by applying qbinom(). We then consider the rate at which the Null Hypothesis would be rejected given that p is equivalent to an alternate value (0.8 in this case)

```
region <- qbinom(0.025, p = 0.7, size = 400)
l <- 0.7*400 - region; u <- 0.7*400 + region
cat("Lower limit = ",l,"\nUpper Limit = ",u)
```

```
## Lower limit =  268
## Upper Limit =  292
```

```
power <- sum(dbinom(c(0:l,u:400), size = 400, p=0.8))
cat("\nPower = ",power)
```

```
##
## Power =  0.999707
```

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.

```
iq.csv <- read.csv('iq.csv')
str(iq.csv)
```

```
## 'data.frame':    124 obs. of  2 variables:
##  $ ID: int  101 102 103 104 105 106 107 108 109 110 ...
##  $ IQ: int  70 85 86 76 84 96 94 56 115 97 ...
```

```
summary(iq.csv)
```

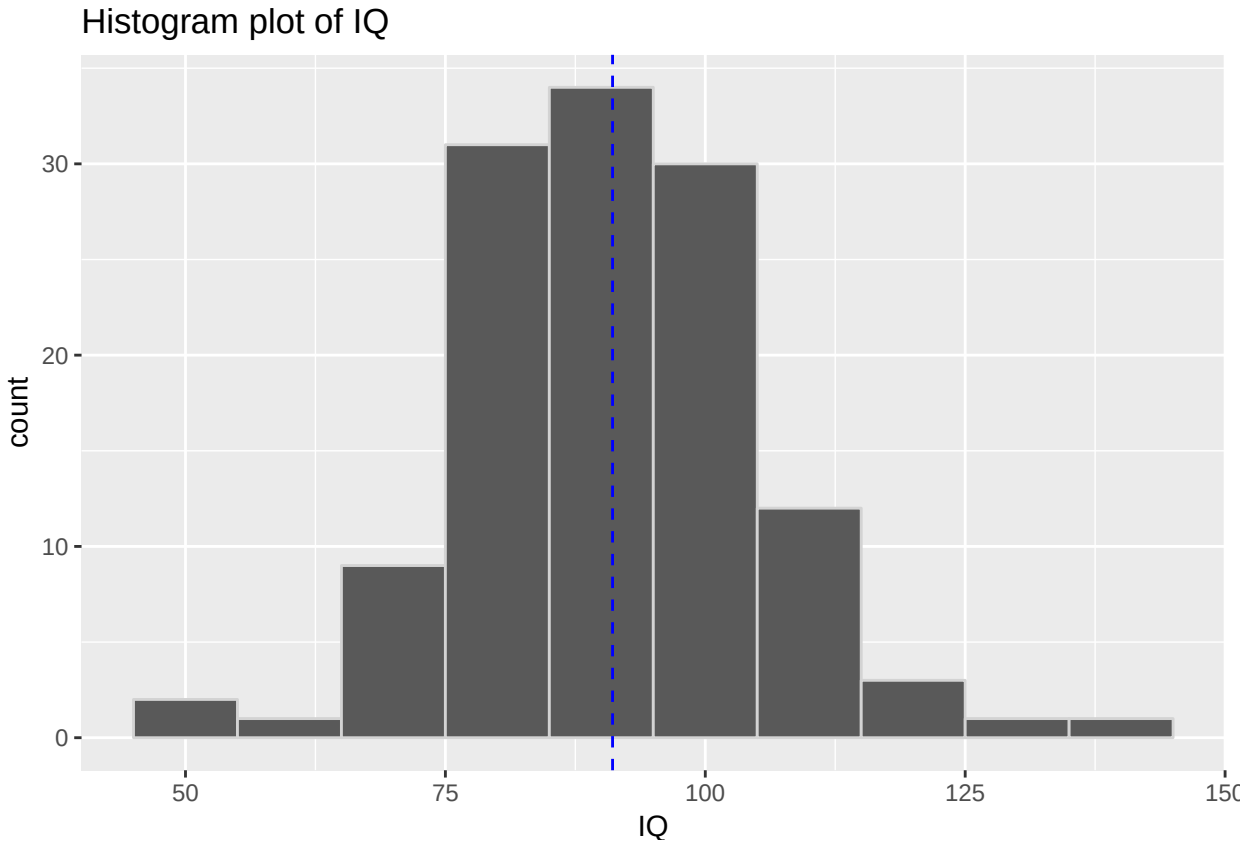
```
##        ID              IQ
##  Min.   :101.0   Min.   : 46.00
##  1st Qu.:131.8   1st Qu.: 81.50
##  Median :162.5   Median : 91.00
##  Mean   :240.2   Mean   : 91.08
##  3rd Qu.:311.2   3rd Qu.: 99.00
##  Max.   :607.0   Max.   :141.00
```

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

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
iq.hist <- ggplot(data = iq.csv, aes(IQ)) +
  geom_histogram(binwidth = 10, color = "lightgray") +
  ggtitle("Histogram plot of IQ") +
  geom_vline(xintercept = mean(iq.csv$IQ), color = "blue", linetype = 2)
iq.hist
```



The distribution is approximately normal. The blue line marks the mean.

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

```
mu <- mean(iq.csv$IQ)
s <- sd(iq.csv$IQ)
cat("Mean = ", mu, "\nSD = ", s)
```

```
## Mean =  91.08065
## SD =  14.40393
```

A mean of 91 and standard deviation of 15 are lesser than the US population values.

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.

```
se <- s/sqrt(nrow(iq.csv))
z <- (mu - 100)/(se)
thresh <- qt(0.975, df = nrow(iq.csv) - 1)

cat("Test statistic (|Z|) = ", round(abs(z),3))
```

```
## Test statistic (|Z|) =  6.895
```

```
cat("\nThreshold value (for alpha = 0.05) = ", round(thresh,3))
```

```
##
## Threshold value (for alpha = 0.05) =  1.979
```

Based on the results above, we see that $|Z| > 1.979$ which falls within the rejection region. Therefore we **reject the null hypothesis** that the mean IQ score in the community is equal to 100.

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

```
2*pt(z, df = nrow(iq.csv)-1)
```

```
## [1] 2.486475e-10
```

The p-value is $2.48 * 10^{-10}$ which is extremely low. It would hence be extremely rare to observe a mean IQ of ~91 given the population average is 100. We **reject the null hypothesis**.

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.

```
lower.ci <- round(mu - thresh*se,3)
upper.ci <- round(mu + thresh*se,3)

cat("95% confidence interval = (",lower.ci,",", upper.ci,")")
```

```
## 95% confidence interval = ( 88.52 , 93.641 )
```

The confidence interval and hypothesis test are consistent with each other. Since 100 does not fall within the confidence interval and the p-value is less than the significance level, we know that the null hypothesis does not stand.

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

```
thresh <- qt(0.995, df = nrow(iq.csv) - 1)
cat("Test statistic (|Z|) = ", round(abs(z),3))
```

```
## Test statistic (|Z|) =  6.895
```

```
cat("\nThreshold value (for alpha = 0.01) = ", round(thresh,3))
```

```
##
## Threshold value (for alpha = 0.01) =  2.616
```

```
lower.ci <- round(mu - thresh*se,3)
upper.ci <- round(mu + thresh*se,3)

cat("\n99% confidence interval = (",lower.ci,",", upper.ci,")")
```

```
##
## 99% confidence interval = ( 87.696 , 94.465 )
```

The conclusions of the hypothesis test and confidence interval do not change after adjusting the significance level as we can see $|Z| > 2.616$