DATA 557
Winter 2022
Homework Assignment 1

Instructions

Submit your solutions **in pdf format** to the dropbox on the canvas page by **12:00PM, Friday January 21**. You may use any program to generate your pdf file. (RStudio is recommended but not required.)

For each question you will be given 1 point for complete credit, ½ point for partial credit, and 0 points for no credit. Assignment of credit will be based on the correctness of your answers as well as your reasoning (when requested as part of the question). You do not need to submit R code for this assignment.

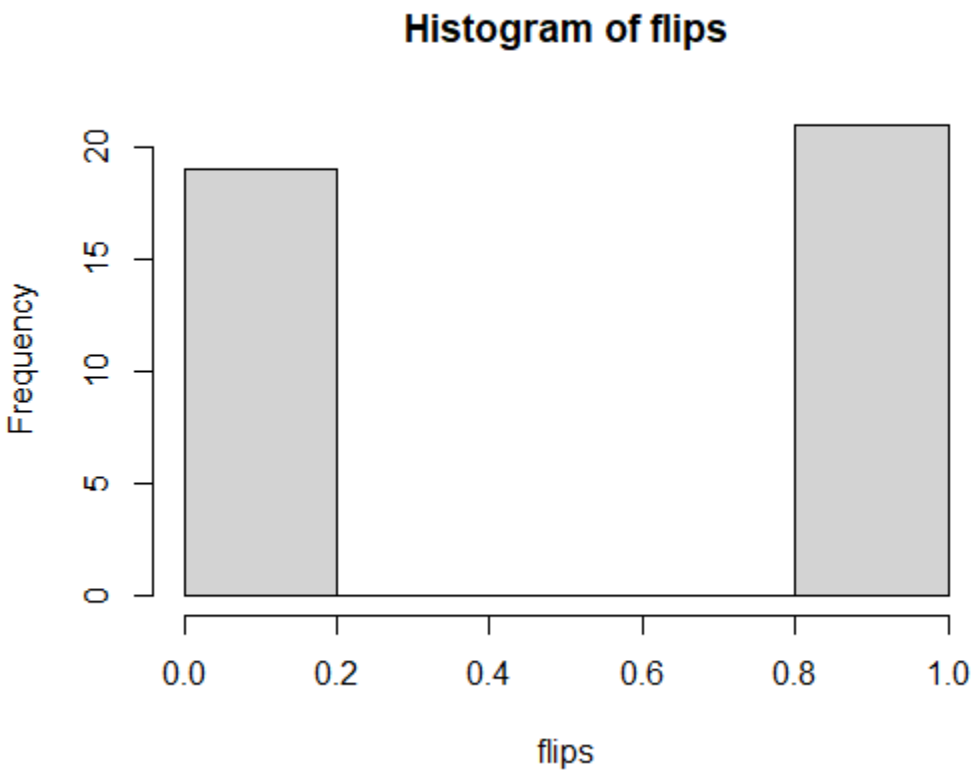
You may work together to help each other solve problems, but you should create your own solutions and hand in your own work without copying others' work.

Question 1

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

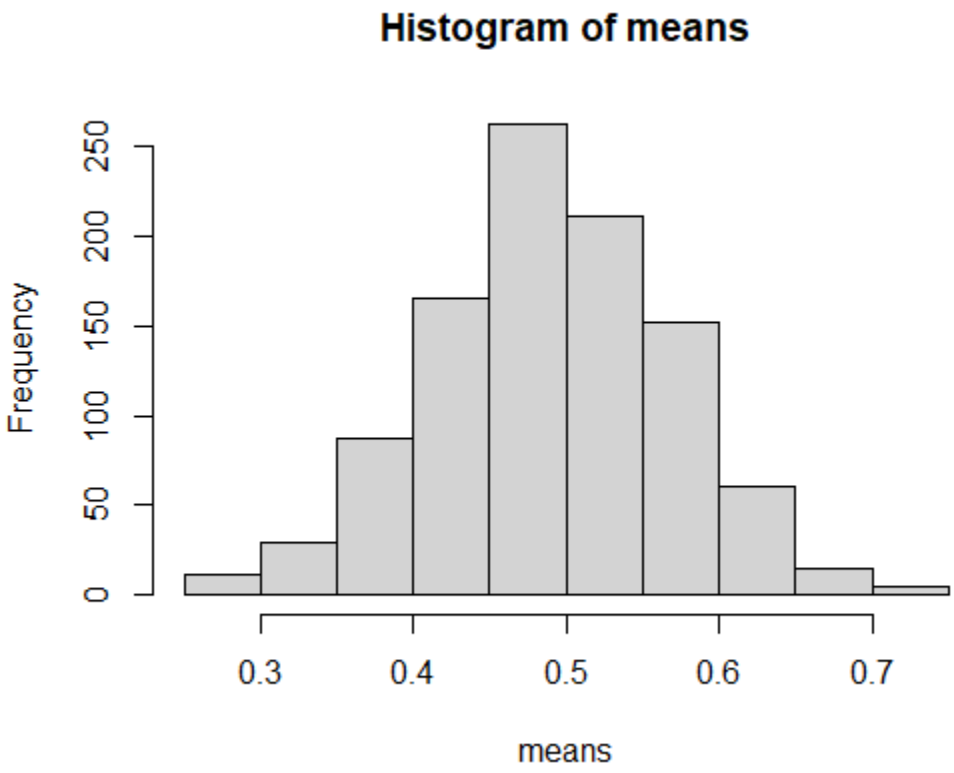
1.1. What is the distribution of the number of heads assuming the coin is fair?

Ans. **Binomial** distribution



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?

Ans. Mean = **0.5**, Standard Deviation = **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).
H0 = 0.5 (coin is fair)

Ans. $Z = \frac{X - Mean}{Std.Dev.} = \frac{\boldsymbol{p - 0.5}}{\boldsymbol{0.079}}$

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).

Ans. From the problem we know that,

n = 40
p = 0.5
X = 15

$Z = \frac{X - np}{sqrt(n * p * (1-p))}$ = **-1.58** (approx. -1.6)

P-value as the tail probability in the normal distribution (approximate p-values):

p-value = pnorm(-1.58)+1-pnorm(1.58) = **0.114**

There is about a 11% chance of getting a result as or more extreme than 15 in 40 tosses, if the coin is fair.
We know that alpha = 0.05.
p< alpha if an only if X lies in the rejection region with significance level.
Since p is not smaller than alpha and p is not small enough, **we do not have enough evidence to question/reject a 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.

Ans. Alpha = 0.1
P value = 0.114
There would be **no** difference in the conclusion since p value > alpha value still.
We cannot conclude that a significant difference exists and hence do not have enough evidence to reject 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?

Ans. p-value using binomial distribution:
sum(dbinom(c(0:15,25:40),size = 40, p = 0.5)) = **0.15**
Yes, the **same conclusion** is reached with the binomial distribution as well as the normal distribution: do not have enough evidence to 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?

Ans. We know that p_cap = 15/40 = 0.375
Standard Error = sqrt(p_cap * (1-p_cap)/n) = 0.0765
To calculate the confidence interval :
Low = p_cap -1.96 * standard error = 0.224
High = p_cap +1.96 * standard error = 0.525
Hence, the 95% confidence interval is **(0.224,0.525)**.
Yes, it includes 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?

Ans. We know that p_cap = 15/40 = 0.375
Standard Error = sqrt(p_cap * (1-p_cap)/n) = 0.0765
To calculate the confidence interval :
Low = p_cap -1.645 * standard error = 0.249
High = p_cap +1.645 * standard error = 0.501
Hence, the 90% confidence interval is **(0.249,0.501)**.
The 90% confidence interval is smaller than the 95% confidence interval and **Yes**, it includes the value 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.

Ans. Given in the question that 305 out of 400 drivers wear seatbelts.
Therefore, sample proportion p_cap = 305/400 = 0.7625
To calculate the Standard Error:

$$\frac{\sqrt{p\_cap} * (1 - p_{cap})}{\sqrt{n}} = \mathbf{0.0212}$$

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?

Ans. The null hypothesis : $H_0 = 0.7$
The alternative hypothesis: $H_1 \neq 0.7$
To calculate the 95% confidence interval:
p_cap ± 1.96 * standard error = **(0.720,0.804)**
Since the value 0.7 does not lie within the confidence interval range, we are **rejecting** 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?

Ans. $Z = \frac{X - np}{sqrt(n*p*(1-p))} = \frac{305 - 400*0.7}{9.165} = 2.729$
The Z score for significance level 0.05 is 1.96.
Since the Z statistic value > Z score value, the null hypothesis should be rejected.
The **conclusion is the same** as the one from the confidence intervals.

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?

Ans. p-value using normal approximation:
pnorm(-2.72)+1-pnorm(2.72) = **0.0065**
p-value using binomial distribution:
sum(dbinom(c(0:255,305:400),size=400,p=0.7)) = **0.0074**
The two p-values are approximately the same.

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.

Ans. Standard deviation = sqrt(n*p*(1-p))
The confidence interval range is:
Low = n * p – 1.96 * standard deviation = 262.04 (~262)
High = n * p + 1.96 * standard deviation = 297.95 (~298)

The power of the test where the probability is 0.8:
Sum (dbinom(c (0:262,298:400),size = 400, p =0.8)) = **0.9968**

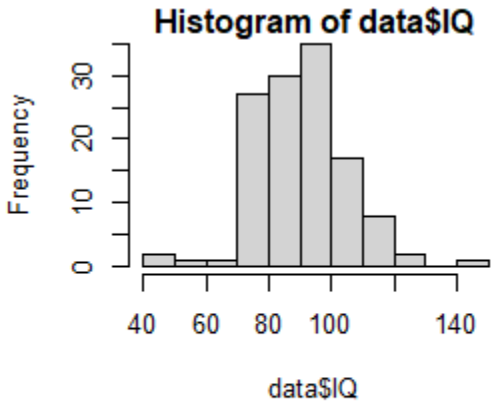
Question 3

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?

Ans. No, the distribution is not approximately normal. Given in the question that actual population size is normally distributed. The sample size of 124 is small enough that the distribution does not look like an approximate normal distribution. The tails of the distribution (as shown in the figure) are too thick and there are not equal number of points on each side of the curve.



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

Ans. Mean = **91.08**, Standard Deviation = **14.40**
The US population mean and SD are given to be 100 and 15 in the question which means that due to a small sample size the sample mean and SD are both smaller than the actual 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.

Ans. The null hypothesis:

$H_0 = 100$

The rejection rule is

Reject H_0 if $|Z| = \frac{\dot{X} - \mu}{s/\sqrt{n}} > t_{\alpha,n-1}$

Calculating the test statistic :

$T = \frac{100-91.08}{14.40/\sqrt{124}} =$ **6.8986**

The test statistic value is much greater than the t-critical value (2.87) and hence 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.

Ans. Calculating the p-value:

p-value = 2 * (1 – pnorm(6.8986)) = **5.274003e-12**

The p-value is much smaller than 0.001 and hence, 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.

Ans. Calculating the 95% confidence interval:

91.08 ± 1.96 * 1.293 = **(88.545,93.614)**

[The Standard error is calculated by std dev/$\sqrt{n}$ = 14.40/$\sqrt{124}$ = 1.293]

The null hypothesis does not fall in the given confidence interval range and hence, we reject the null hypothesis.

The results of the confidence interval and the hypothesis test **agree** with each other.

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

Ans. For the hypothesis test, the test statistic = 6.8986 remains the same. The t-critical value at significance level 0.01is 2.63. (derived from the t-table)

Since the t statistic vale is much greater than the t-critical value, we reject the null hypothesis.

Calculating the 99% confidence interval:

91.08 ± 2.576 * 1.293 = **(87.749, 94.410)**

The null hypothesis does not fall under the given confidence interval and hence, we **reject** the null hypothesis.