Homework 6

Han Nguyen - TXN200004

11/08/2025

Question 1

A cola beverage company claims that their cans are filled with 16.00 ounces of cola. The company also states that the fill of the cola cans follows a Gamma distribution with a shape parameter of 256000 and a rate parameter of 16000. Mark doesn't want to pay for 16.01 ounces of soda unless he is getting at least that much soda. Thus, he samples 34 cola cans from this beverage company to test their claim.

What is the probability that the average fill of the sampled soda cans is greater than 16.01 ounces?

```
# Given parameters
shape_param <- 256000
rate_param <- 16000
n <- 34  # sample size
target_value <- 16.01

# Calculate population parameters from Gamma distribution
# For Gamma distribution: mean = shape/rate, variance = shape/rate^2
pop_mean <- shape_param / rate_param
pop_variance <- shape_param / (rate_param^2)
pop_sd <- sqrt(pop_variance)

cat("Population mean:", pop_mean, "ounces\n")
```

```
## Population mean: 16 ounces
```

```
cat("Population variance:", pop_variance, "\n")
```

```
## Population variance: 0.001
```

```
cat("Population standard deviation:", pop_sd, "ounces\n\n")
```

```
## Population standard deviation: 0.03162278 ounces
```

```
# Using Central Limit Theorem with Normal approximation
# The sample mean follows approximately Normal(mean = pop_mean, sd = pop_sd/sqrt(n))
sampling_mean <- pop_mean
sampling_sd <- pop_sd / sqrt(n)

cat("Sampling distribution of the mean:\n")
```

```
## Sampling distribution of the mean:
```

```
cat("Mean of sample mean:", sampling_mean, "ounces\n")
```

```
## Mean of sample mean: 16 ounces
```

```r
cat("Standard deviation of sample mean:", sampling_sd, "ounces\n\n")
```

```
## Standard deviation of sample mean: 0.005423261 ounces
```

```r
# Find P(sample mean > 16.01)
prob <- 1 - pnorm(target_value, mean = sampling_mean, sd = sampling_sd)
cat("P(sample mean > 16.01) =", prob, "\n")
```

```
## P(sample mean > 16.01) = 0.03259821
```

Answer: The probability that the average fill of the sampled soda cans is greater than 16.01 ounces is **0.032598**.

Conclusion: Since the probability is approximately 0.032598 (or about 3.26%), this is a very small probability. This means it would be extremely unlikely to observe a sample mean greater than 16.01 ounces if the cans are truly filled according to the company's claimed distribution with a mean of 16.00 ounces.

Question 2

The number of minutes for app engagement by a tablet user follows a normal distribution with $\mu = 8.2$ minutes, and $\sigma = 1$ minute. Suppose, we take a sample of 60 tablet users.

```r
# Given parameters
mu <- 8.2  # population mean
sigma <- 1  # population standard deviation
n <- 60  # sample size
```

a)

What are the mean and standard deviation of the sampling distribution of the sample mean?

```r
# Sampling distribution of the sample mean
# Mean of sample mean = mu
# Standard deviation of sample mean = sigma / sqrt(n)
mean_xbar <- mu
sd_xbar <- sigma / sqrt(n)

cat("Mean of sampling distribution:", mean_xbar, "minutes\n")
```

```
## Mean of sampling distribution: 8.2 minutes
```

```r
cat("Standard deviation of sampling distribution:", sd_xbar, "minutes\n")
```

```
## Standard deviation of sampling distribution: 0.1290994 minutes
```

Answer: The mean of the sampling distribution is **8.2** minutes and the standard deviation of the sampling distribution is **0.1291** minutes.

b)

Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.

```r
# 90th percentile
percentile_90 <- qnorm(0.90, mean = mean_xbar, sd = sd_xbar)
cat("90th percentile:", percentile_90, "minutes\n")
```

```
## 90th percentile: 8.365448 minutes
```

Answer: The 90th percentile for the sample mean time is **8.3654** minutes. This means that 90% of samples of 60 tablet users will have an average app engagement time less than 8.3654 minutes, and only 10% of samples will have an average greater than this value.

c)

Find the probabilities that the sample mean is between ±1 standard deviation, ±2 standard deviations, and ±3 standard deviations.

```
# Within 1 standard deviation: P(mu - 1*sd < X_bar < mu + 1*sd)
lower_1sd <- mean_xbar - 1 * sd_xbar
upper_1sd <- mean_xbar + 1 * sd_xbar
prob_1sd <- pnorm(upper_1sd, mean = mean_xbar, sd = sd_xbar) -
            pnorm(lower_1sd, mean = mean_xbar, sd = sd_xbar)

# Within 2 standard deviations
lower_2sd <- mean_xbar - 2 * sd_xbar
upper_2sd <- mean_xbar + 2 * sd_xbar
prob_2sd <- pnorm(upper_2sd, mean = mean_xbar, sd = sd_xbar) -
            pnorm(lower_2sd, mean = mean_xbar, sd = sd_xbar)

# Within 3 standard deviations
lower_3sd <- mean_xbar - 3 * sd_xbar
upper_3sd <- mean_xbar + 3 * sd_xbar
prob_3sd <- pnorm(upper_3sd, mean = mean_xbar, sd = sd_xbar) -
            pnorm(lower_3sd, mean = mean_xbar, sd = sd_xbar)

cat("P(within 1 SD) =", prob_1sd, "\n")
```

```
## P(within 1 SD) = 0.6826895
```

```
cat("P(within 2 SD) =", prob_2sd, "\n")
```

```
## P(within 2 SD) = 0.9544997
```

```
cat("P(within 3 SD) =", prob_3sd, "\n")
```

```
## P(within 3 SD) = 0.9973002
```

Answer:

- Within ±1 standard deviation: **0.682689** or about 68.27%
- Within ±2 standard deviations: **0.9545** or about 95.45%
- Within ±3 standard deviations: **0.9973** or about 99.73%

d)

Is there a different way to do part (c) that involves not using R, and not using any form of calculations?

Answer: Yes, there is a different way using the **Empirical Rule** (68-95-99.7 Rule) for normal distributions. This rule states that for any normal distribution:

- Approximately 68% of values fall within ±1 standard deviation of the mean
- Approximately 95% of values fall within ±2 standard deviations of the mean
- Approximately 99.7% of values fall within ±3 standard deviations of the mean

Since the sampling distribution of the sample mean is normally distributed, we can directly apply this rule without any calculations. The values we calculated in part (c) should be very close to 68%, 95%, and 99.7%

respectively.

Question 3

A study involving stress is done on a college campus among the students. The stress scores, ranging from 0 to 5, follow a binomial distribution with N = 5 and p = 0.5. Using a sample of 75 students, find:

```
# Given parameters
N <- 5  # number of trials for binomial
p <- 0.5  # probability of success
n <- 75  # sample size (number of students)

# Population parameters for Binomial(N, p)
pop_mean <- N * p
pop_variance <- N * p * (1 - p)
pop_sd <- sqrt(pop_variance)

cat("Population parameters (for individual student):\n")
```

```
## Population parameters (for individual student):
```

```
cat("Mean:", pop_mean, "\n")
```

```
## Mean: 2.5
```

```
cat("Variance:", pop_variance, "\n")
```

```
## Variance: 1.25
```

```
cat("Standard deviation:", pop_sd, "\n\n")
```

```
## Standard deviation: 1.118034
```

a)

The probability that the average stress score for the 75 students is less than 2.25.

```
# By Central Limit Theorem, sample mean is approximately Normal
# Mean of sample mean = pop_mean
# SD of sample mean = pop_sd / sqrt(n)
mean_xbar <- pop_mean
sd_xbar <- pop_sd / sqrt(n)

cat("Sampling distribution of sample mean:\n")
```

```
## Sampling distribution of sample mean:
```

```
cat("Mean:", mean_xbar, "\n")
```

```
## Mean: 2.5
```

```
cat("Standard deviation:", sd_xbar, "\n\n")
```

```
## Standard deviation: 0.1290994
```

```
# Find P(sample mean < 2.25)
prob_a <- pnorm(2.25, mean = mean_xbar, sd = sd_xbar)
cat("P(sample mean < 2.25) =", prob_a, "\n")
```

```
## P(sample mean < 2.25) = 0.02640376
```

Answer: The probability that the average stress score for the 75 students is less than 2.25 is **0.026404** or about 2.64%.

b)

The 90th percentile for the average stress score for the 75 students.

```
# Find the 90th percentile
percentile_90_mean <- qnorm(0.90, mean = mean_xbar, sd = sd_xbar)
cat("90th percentile for sample mean:", percentile_90_mean, "\n")
```

```
## 90th percentile for sample mean: 2.665448
```

Answer: The 90th percentile for the average stress score for the 75 students is **2.6654**. This means 90% of samples of 75 students will have an average stress score below this value.

c)

The probability that the total of the 75 stress scores is less than 200.

```
# Total (sum) of 75 scores
# By CLT, sum is approximately Normal with:
# Mean of sum = n * pop_mean
# SD of sum = sqrt(n) * pop_sd
mean_sum <- n * pop_mean
sd_sum <- sqrt(n) * pop_sd

cat("Sampling distribution of total (sum):\n")
```

```
## Sampling distribution of total (sum):
```

```
cat("Mean:", mean_sum, "\n")
```

```
## Mean: 187.5
```

```
cat("Standard deviation:", sd_sum, "\n\n")
```

```
## Standard deviation: 9.682458
```

```
# Find P(sum < 200)
prob_c <- pnorm(200, mean = mean_sum, sd = sd_sum)
cat("P(total < 200) =", prob_c, "\n")
```

```
## P(total < 200) = 0.9016472
```

Answer: The probability that the total of the 75 stress scores is less than 200 is **0.901647** or about 90.16%.

d)

The 90th percentile for the total stress score for the 75 students.

```
# Find the 90th percentile for the sum
percentile_90_sum <- qnorm(0.90, mean = mean_sum, sd = sd_sum)
cat("90th percentile for total:", percentile_90_sum, "\n")
```

```
## 90th percentile for total: 199.9086
```

Answer: The 90th percentile for the total stress score for the 75 students is **199.9086**. This means 90% of samples will have a total stress score below this value.

Question 4

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the data allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the data included in their basic contract, the excess data used follows an exponential distribution with a mean of 2 Gigabytes (Gb). Consider a random sample of 80 customers who exceed the data allowance included in their basic cell phone contract.

```
# Given parameters
pop_mean <- 2  # mean of exponential distribution in Gb
n <- 80  # sample size
target <- 2.5  # target value in Gb

# For exponential distribution with mean = 2
# rate parameter = 1/mean
rate <- 1 / pop_mean

# Population variance for exponential = mean^2
pop_variance <- pop_mean^2
pop_sd <- pop_mean  # for exponential, sd = mean

cat("Population parameters (Exponential distribution):\n")
```

```
## Population parameters (Exponential distribution):
```

```
cat("Mean:", pop_mean, "Gb\n")
```

```
## Mean: 2 Gb
```

```
cat("Rate:", rate, "\n")
```

```
## Rate: 0.5
```

```
cat("Variance:", pop_variance, "\n")
```

```
## Variance: 4
```

```
cat("Standard deviation:", pop_sd, "Gb\n\n")
```

```
## Standard deviation: 2 Gb
```

a)

Suppose that one customer who exceeds the data limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess data use is larger than 2.5 Gb.

```
# For a single customer, use exponential distribution
# P(X > 2.5) = 1 - P(X <= 2.5)
prob_single <- 1 - pexp(target, rate = rate)
cat("P(single customer > 2.5 Gb) =", prob_single, "\n")
```

```
## P(single customer > 2.5 Gb) = 0.2865048
```

Answer: The probability that a single randomly selected customer's excess data use is larger than 2.5 Gb is **0.286505** or about 28.65%.

b)

Find the probability that the average excess data used by the 80 customers in the sample is larger than 2.5 Gb.

```
# By Central Limit Theorem, sample mean is approximately Normal
# Mean of sample mean = pop_mean
# SD of sample mean = pop_sd / sqrt(n)
mean_xbar <- pop_mean
sd_xbar <- pop_sd / sqrt(n)

cat("Sampling distribution of sample mean:\n")
```

```
## Sampling distribution of sample mean:
```

```
cat("Mean:", mean_xbar, "Gb\n")
```

```
## Mean: 2 Gb
```

```
cat("Standard deviation:", sd_xbar, "Gb\n\n")
```

```
## Standard deviation: 0.2236068 Gb
```

```
# P(sample mean > 2.5)
prob_sample_mean <- 1 - pnorm(target, mean = mean_xbar, sd = sd_xbar)
cat("P(sample mean > 2.5 Gb) =", prob_sample_mean, "\n")
```

```
## P(sample mean > 2.5 Gb) = 0.01267366
```

Answer: The probability that the average excess data used by the 80 customers in the sample is larger than 2.5 Gb is **0.012674** or about 1.27%.

c)

Explain why the probabilities in (a) and (b) are different.

Answer: The probabilities are different because:

1. **Part (a)** examines a **single individual customer** whose data usage follows an exponential distribution with mean 2 Gb and high variability (standard deviation = 2 Gb). There is substantial probability (28.65%) that one individual could exceed 2.5 Gb.
2. **Part (b)** examines the **average of 80 customers**. By the Central Limit Theorem, the sample mean has the same mean (2 Gb) but much lower variability (standard deviation = 0.2236 Gb). The averaging process reduces variability, making it much less likely (1.27%) for the average to deviate far from the population mean of 2 Gb.

In summary, individual observations have high variability, but averages of many observations have low variability due to the law of large numbers. This makes extreme values much less likely for sample means compared to individual observations.

Question 5

There were 70 enrolled students in STAT 3355 during the year 2020. The population of adults, 18 years or older, in the United States was 258.3 million in 2020. A student surveyed 30 of her classmates in 2020 and found that 22 students liked to play video games.

If this student computed a 95% confidence interval, would it have contained the value of 65%, which was known to be the proportion of adults that liked to play video games in the United States in 2020?

```r
# Given data
n <- 30  # sample size
x <- 22  # number who like video games
p_hat <- x / n  # sample proportion
p_us <- 0.65  # US population proportion
confidence_level <- 0.95
alpha <- 1 - confidence_level
z_critical <- qnorm(1 - alpha/2)

cat("Sample proportion:", p_hat, "\n")
```

```
## Sample proportion: 0.7333333
```

```r
cat("US population proportion:", p_us, "\n\n")
```

```
## US population proportion: 0.65
```

```r
# Calculate standard error
se <- sqrt(p_hat * (1 - p_hat) / n)
cat("Standard error:", se, "\n")
```

```
## Standard error: 0.08073734
```

```r
# Calculate 95% confidence interval
margin_error <- z_critical * se
lower_bound <- p_hat - margin_error
upper_bound <- p_hat + margin_error

cat("\n95% Confidence Interval:\n")
```

```
##
## 95% Confidence Interval:
```

```r
cat("Lower bound:", lower_bound, "\n")
```

```
## Lower bound: 0.575091
```

```r
cat("Upper bound:", upper_bound, "\n")
```

```
## Upper bound: 0.8915756
```

```r
cat("Interval: [", lower_bound, ",", upper_bound, "]\n\n")
```

```
## Interval: [ 0.575091 , 0.8915756 ]
```

```r
# Check if 0.65 is in the interval
contains_065 <- (p_us >= lower_bound) & (p_us <= upper_bound)
cat("Does the interval contain 0.65?", contains_065, "\n")
```

```
## Does the interval contain 0.65? TRUE
```

Answer: The sample proportion is $\hat{p} = 0.7333$ (or 73.33%). The 95% confidence interval is [0.5751, 0.8916] or [57.51%, 89.16%].

Conclusion: The confidence interval **DOES** contain the value of 65%. This suggests that the sample comes from a class of 70 students, which is a very different and much smaller population than the entire U.S. adult population of 258.3 million. The class sample is not representative of the U.S. population, so we would not necessarily expect the class proportion to match the national proportion.

Question 6

An elevator can safely hold 3,500 lbs. A sign in the elevator limits the passenger count to 15. If the adult population has a mean weight of 180 lbs with a 25 lbs standard deviation, how unusual would it be, if the central limit theorem applied, that an elevator holding 15 people would be carrying more than 3,500 pounds?

```
# Given parameters
mu_weight <- 180  # mean weight in lbs
sigma_weight <- 25  # standard deviation in lbs
n_people <- 15  # number of people
max_weight <- 3500  # maximum safe weight

# By CLT, total weight of 15 people is approximately Normal
# Mean of total = n * mu
# SD of total = sqrt(n) * sigma
mean_total <- n_people * mu_weight
sd_total <- sqrt(n_people) * sigma_weight

cat("Distribution of total weight for 15 people:\n")
```

```
## Distribution of total weight for 15 people:
```

```
cat("Mean:", mean_total, "lbs\n")
```

```
## Mean: 2700 lbs
```

```
cat("Standard deviation:", sd_total, "lbs\n\n")
```

```
## Standard deviation: 96.82458 lbs
```

```
# Find P(total weight > 3500)
prob_exceed <- 1 - pnorm(max_weight, mean = mean_total, sd = sd_total)
cat("P(total weight > 3500 lbs) =", prob_exceed, "\n")
```

```
## P(total weight > 3500 lbs) = 1.110223e-16
```

```
# Calculate z-score to show how unusual this is
z_score <- (max_weight - mean_total) / sd_total
cat("Z-score:", z_score, "\n")
```

```
## Z-score: 8.262364
```

Answer: The probability that an elevator holding 15 people would be carrying more than 3,500 pounds is 1.110223×10^{-16} or basically 0%.

Conclusion: This is extremely unusual. The z-score is 8.26, which means 3,500 lbs is 8.26 standard deviations above the mean. The expected total weight is only 2700 lbs, which is well below the 3,500 lb limit, making it very unlikely that 15 randomly selected adults would exceed the weight limit.

Question 7

A restaurant sells an average of 25 bottles of wine per night, with a variance of 25. Assuming the central limit theorem applies, what is the probability that the restaurant will sell more than 600 bottles in the next 30 days?

```
# Given parameters (Poisson distribution)
lambda <- 25  # average per night
variance_per_night <- 25
n_days <- 30
```

```r
target_total <- 600

# For Poisson distribution: mean = variance = lambda
# Total sales over 30 days follows Poisson(30 * lambda)
# Mean of total = n * lambda
# Variance of total = n * variance (for Poisson, this equals n * lambda)
# SD of total = sqrt(n * variance)

mean_total <- n_days * lambda
variance_total <- n_days * variance_per_night
sd_total <- sqrt(variance_total)

cat("Distribution of total sales over 30 days:\n")
```

```
## Distribution of total sales over 30 days:
```

```r
cat("Mean:", mean_total, "bottles\n")
```

```
## Mean: 750 bottles
```

```r
cat("Variance:", variance_total, "\n")
```

```
## Variance: 750
```

```r
cat("Standard deviation:", sd_total, "bottles\n\n")
```

```
## Standard deviation: 27.38613 bottles
```

```r
# By CLT, use normal approximation
# P(total > 600)
prob_exceed <- 1 - pnorm(target_total, mean = mean_total, sd = sd_total)
cat("P(total sales > 600 bottles) =", prob_exceed, "\n")
```

```
## P(total sales > 600 bottles) = 1
```

Answer: The probability that the restaurant will sell more than 600 bottles in the next 30 days is **1** or about 100%.

Conclusion: Since the expected total sales over 30 days is 750 bottles, selling more than 600 bottles would be very likely. The value of 600 is below the expected value of 750 bottles.

Question 8

Currently, there are 54 enrolled students in STAT 3355. It is known that 13.1% of the population in U.S. are left-handed. A student wishes to find the proportion of left-handed people in this class. She surveys 30 students and finds that only 2 are left-handed.

If she computes a 95% confidence interval, would it contain the value of 13.1%?

```r
# Given data
n <- 30  # sample size
x <- 2  # number of left-handed students
p_hat <- x / n  # sample proportion
p_us <- 0.131  # US population proportion
confidence_level <- 0.95
alpha <- 1 - confidence_level
z_critical <- qnorm(1 - alpha/2)
```

```r
cat("Sample proportion:", p_hat, "\n")
```

```
## Sample proportion: 0.06666667
```

```r
cat("US population proportion:", p_us, "\n\n")
```

```
## US population proportion: 0.131
```

```r
# Calculate standard error
se <- sqrt(p_hat * (1 - p_hat) / n)
cat("Standard error:", se, "\n")
```

```
## Standard error: 0.045542
```

```r
# Calculate 95% confidence interval
margin_error <- z_critical * se
lower_bound <- p_hat - margin_error
upper_bound <- p_hat + margin_error

cat("\n95% Confidence Interval:\n")
```

```
##
## 95% Confidence Interval:
```

```r
cat("Lower bound:", lower_bound, "\n")
```

```
## Lower bound: -0.02259402
```

```r
cat("Upper bound:", upper_bound, "\n")
```

```
## Upper bound: 0.1559274
```

```r
cat("Interval: [", lower_bound, ",", upper_bound, "]\n\n")
```

```
## Interval: [ -0.02259402 , 0.1559274 ]
```

```r
# Check if 0.131 is in the interval
contains_0131 <- (p_us >= lower_bound) & (p_us <= upper_bound)
cat("Does the interval contain 0.131?", contains_0131, "\n")
```

```
## Does the interval contain 0.131? TRUE
```

Answer: The sample proportion is $\hat{p} = 0.0667$ (or 6.67%). The 95% confidence interval is [-0.0226, 0.1559] or [-2.26%, 15.59%].

Conclusion: The confidence interval **DOES** contain the value of 13.1%. This suggests that the class proportion is consistent with the national proportion.

Question 9

For the `babies` dataset in the package `UsingR`, the variable `age` contains the mother's age and the variable `dage` contains the father's age.

Find a 95% confidence interval for the difference in mean age. Does it contain 0? What do you assume about the data?

```r
# Load the required package and dataset
library(UsingR)
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
## Loading required package: Hmisc
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
data(babies)

# Extract ages (remove NAs)
mother_age <- babies$age[!is.na(babies$age) & !is.na(babies$dage)]
father_age <- babies$dage[!is.na(babies$age) & !is.na(babies$dage)]

# Calculate differences (paired data)
age_diff <- father_age - mother_age

# Summary statistics
n <- length(age_diff)
mean_diff <- mean(age_diff)
sd_diff <- sd(age_diff)
se_diff <- sd_diff / sqrt(n)

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

```
## Sample size: 1236
```

```r
cat("Mean difference (father - mother):", mean_diff, "years\n")
```

```
## Mean difference (father - mother): 3.365696 years
```

```r
cat("Standard deviation of differences:", sd_diff, "years\n")
```

```
## Standard deviation of differences: 6.803471 years
```

```r
cat("Standard error:", se_diff, "years\n\n")
```

```
## Standard error: 0.193518 years
```

```r
# 95% confidence interval using t-distribution
alpha <- 0.05
t_critical <- qt(1 - alpha/2, df = n - 1)

margin_error <- t_critical * se_diff
lower_bound <- mean_diff - margin_error
upper_bound <- mean_diff + margin_error

cat("95% Confidence Interval for difference:\n")
```

```
## 95% Confidence Interval for difference:
```

```r
cat("Lower bound:", lower_bound, "years\n")
```

```
## Lower bound: 2.986035 years
```

```r
cat("Upper bound:", upper_bound, "years\n")
```

```
## Upper bound: 3.745356 years
```

```r
cat("Interval: [", lower_bound, ",", upper_bound, "]\n\n")
```

```
## Interval: [ 2.986035 , 3.745356 ]
```

```r
# Check if 0 is in the interval
contains_zero <- (0 >= lower_bound) & (0 <= upper_bound)
cat("Does the interval contain 0?", contains_zero, "\n")
```

```
## Does the interval contain 0? FALSE
```

```r
# Summary statistics for context
cat("\nSummary statistics:\n")
```

```
##
## Summary statistics:
```

```r
cat("Mean mother age:", mean(mother_age), "years\n")
```

```
## Mean mother age: 27.37136 years
```

```r
cat("Mean father age:", mean(father_age), "years\n")
```

```
## Mean father age: 30.73706 years
```

Answer: The 95% confidence interval for the difference in mean age (father age - mother age) is [2.986, 3.7454] years.

Does it contain 0? No, the interval **DOES NOT** contain 0.

Since the interval does not contain 0 and is entirely positive, we can conclude with 95% confidence that there is a significant difference in mean ages - fathers are on average 3.37 years older than mothers in this dataset.

Assumptions: We assume that:

1. The data represents **paired observations** (each baby has both a mother and a father), so we use paired t-test methodology by analyzing the differences.
2. The differences in ages are **approximately normally distributed** (or the sample size is large enough for the Central Limit Theorem to apply).
3. The sample is **representative** of the population we want to make inferences about.
4. Observations are **independent** of each other.