Kingdom of Cambodia

Nation Religion King

Institute of Technology of Cambodia

Department of Applied Mathematics and Statistics

Report of Final Project

Subject: Statistics

Name	ID	Score
1. KHUN Limchheang	e20230393	…….
2. KIM Putdararith	e20230638	…….
3. LAY Sokleab	e20231117	…….

Lecturer: Dr. PHAUK Sokkhey (Course)

Dr. HAS Sothea (TD)

Academic year 2025 - 2026

Contents

Statistical Investigation of Daily Music Listening Time Among Undergraduate Students in Cambodia

Group Members: KHUN Limchheang, KIM Putdararith, LAY Sokleab

January 2026

---

1 Introduction

In contemporary Cambodian society, digital media consumption—particularly music—has become an integral component of daily student life. Music serves multiple roles: as a cognitive aid during study, a source of emotional regulation, and a form of cultural engagement. Understanding the magnitude and variability of this behavior provides a window into student lifestyles, time allocation, and digital habits.

This report presents a **comprehensive statistical investigation** of daily music listening time (in minutes) among undergraduate students in Cambodia. The analysis is grounded in the theoretical framework of **Fundamental Statistics (Chapters 1–4)** and fulfills the course's objective of carrying out a complete, self-contained statistical inquiry—from data collection to inference—while emphasizing **statistical reasoning, contextual interpretation, and methodological rigor**.

The central research question is: *What is the average daily music listening time among Cambodian university students, and is it significantly greater than one hour?*

This project contributes to the broader goal of applying foundational statistical tools to real-world, self-collected data in a transparent and pedagogically meaningful way.

2 Dataset Description

2.1 Data Collection and Source

Data were collected during the period of December 26–28, 2025, via a self-administered online survey distributed through academic and social networks to undergraduate students across Cambodia. Participation was voluntary and anonymous. The survey asked respondents to report their **average daily music listening time in minutes** over a typical week.

2.2 Variable Definition

The primary variable of interest is:

$$X = \text{Daily music listening time (minutes)}$$

This variable is **quantitative, continuous**, and measured on a ratio scale—making it suitable for all required statistical operations (mean, variance, confidence intervals, hypothesis testing).

Additional variables gender, age, and preferred music genre were recorded for contextual enrichment but are **excluded from formal statistical analysis**, in strict adherence to the project requirement of focusing on **one numerical variable**.

2.3 Sample Size and Data Cleaning

The initial dataset comprised 38 observations. To comply with the project constraint of $15 \leq n \leq 30$, we used simple random sampling (without replacement) to select $n = 30$ observations. This ensures compliance while preserving the representativeness of the original responses.

Data cleaning addressed inconsistent entries:

- "2h" and "2 Hours" were converted to 120 minutes.
- "30minutes" was standardized to 30.
- Non-numeric or ambiguous entries (e.g., "a lot") were excluded prior to sampling.

The final dataset consists of 30 valid, numerical observations of X.

3 Methodology

This investigation follows a structured four-part analytical framework corresponding to Chapters 1–4 of the course:

1. **Descriptive Statistics (Chapter 1)**: Summarize the sample distribution using graphical and numerical summaries.
2. **Point Estimation (Chapter 2)**: Use the sample mean $\bar{X}$ to estimate the population mean μ.

3. **Confidence Interval (Chapter 3)**: Construct a 95% confidence interval for μ using the t-distribution.

4. **Hypothesis Testing (Chapter 4)**: Test the claim that the average listening time exceeds 60 minutes.

All numerical results are derived through **hand calculations**, with Python used solely for data visualization. Assumptions (e.g., independence, approximate normality for inference) are explicitly evaluated.

4 Results

4.1 Part A: Descriptive Statistics

A1. Graphical Summaries

Figures 1 and 2 display the distribution of X for the 30 respondents.

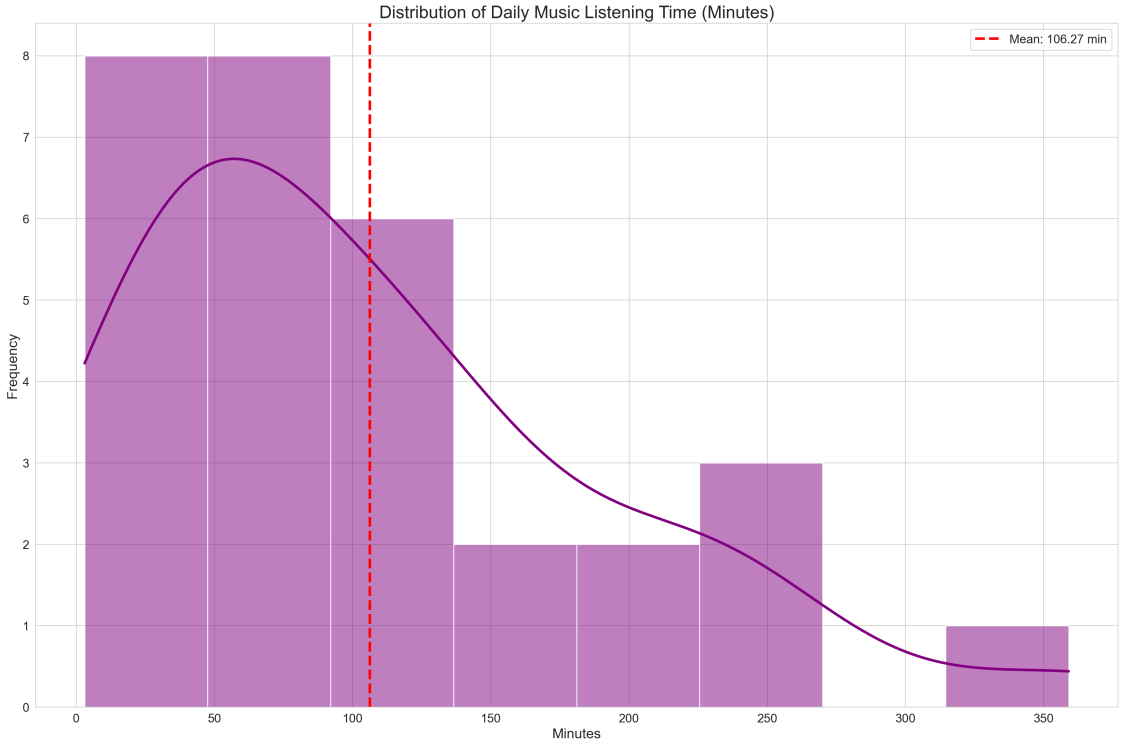

Figure 1: Histogram of daily music listening time (minutes). The distribution exhibits a pronounced right skew, suggesting most students listen less than 120 minutes daily, while a few report much higher usage (up to 360 minutes).

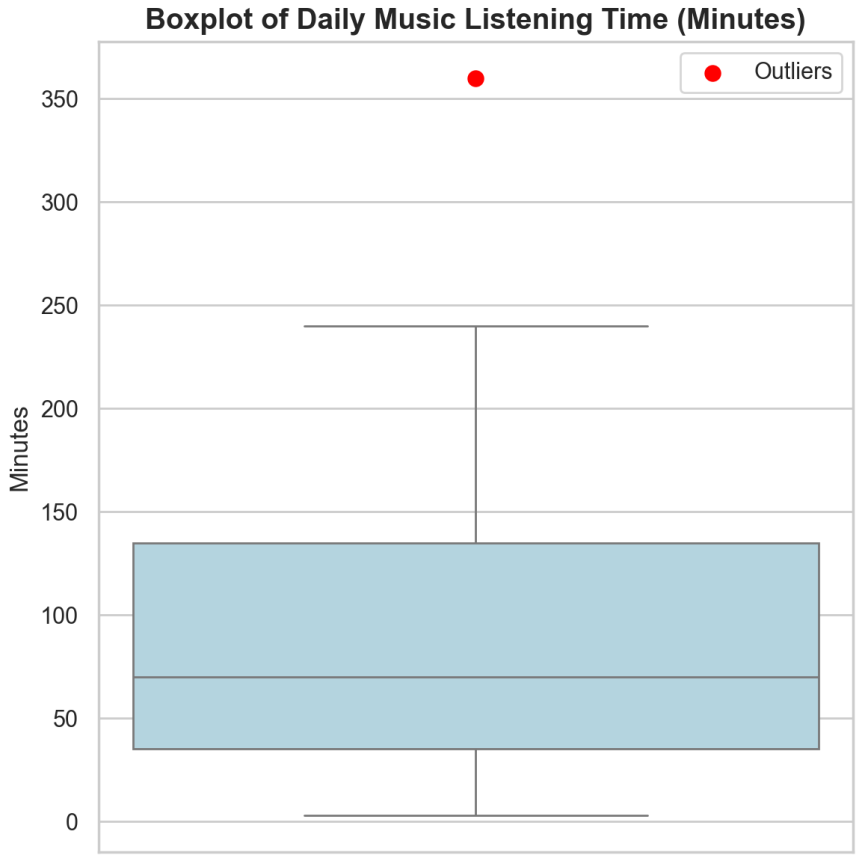

Figure 2: Boxplot of daily music listening time. The median (70 min) is substantially lower than the mean (106.27 min), reflecting skewness; a single high outlier at 360 minutes confirms the presence of extreme values.

A2. Numerical Descriptions (Hand Calculations)

Sorted data (in minutes):

3.0	25.0	30.0	30.0	30.0	30.0	30.0	30.0	50.0	60.0
60.0	60.0	60.0	60.0	60.0	80.0	120.0	120.0	120.0	120.0
120.0	120.0	140.0	150.0	200.0	200.0	240.0	240.0	240.0	360.0

- **Sample mean**:

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{3188}{30} = 106.2667 \text{ minutes}$$

- **Sample median**: The average of the 15th and 16th ordered values:

$$\tilde{x} = \frac{60 + 80}{2} = 70.0 \text{ minutes}$$

- **Mode**: 30, 60, and 120 minutes (appears 6 times)
- **Range**: $360 - 3 = 357$ minutes
- **Sample variance**:

$$s^2 = \frac{1}{29}\sum_{i=1}^{30}(x_i - \bar{x})^2 = \frac{208655.758}{29} \approx 7195.02615$$

- **Sample standard deviation**: $s = \sqrt{7195.02615} \approx 84.8235$ minutes
- **Quartiles and IQR**:
  - Q_1 (8th value) = 35.0
  - Q_3 (23rd value) = 135.0
  - IQR $= 135 - 35 = 100$ minutes

A3. Outlier Detection

Using the IQR rule:

$$\text{Lower fence} = Q_1 - 1.5 \times \text{IQR} = 35 - 150 = -115$$

$$\text{Upper fence} = Q_3 + 1.5 \times \text{IQR} = 135 + 150 = 285$$

The observation **360 minutes** exceeds the upper fence and is classified as a **high outlier**.

A4. Interpretation of Distribution

The histogram and boxplot together reveal a strongly right-skewed distribution of daily music listening time. Most students report listening for less than two hours per day, with common values clustered around 30, 60, and 120 minutes — suggesting rounding behavior typical in self-reported data. The median (70 minutes) is substantially lower than the mean (106.27 minutes), confirming the influence of a few high outliers. One observation (360 minutes) is identified as an outlier via the IQR rule and may reflect exceptional usage patterns.

Given the skew and presence of an outlier, the **median is the preferred measure of central tendency**, as it is robust to extreme values. The large standard deviation (84.82) and IQR (100) indicate **substantial heterogeneity** in listening habits. The outlier (360

minutes) may reflect a student who uses music continuously throughout the day (e.g., as ambient background during study or work), though self-reporting bias cannot be ruled out.

4.2 Part B: Point Estimation

B1. Population Parameters and Estimators

Assume the data constitute a random sample from a population with:

- True mean listening time: μ
- True variance: σ^2

The sample provides the following estimators:

- $\bar{X}$ estimates μ
- S^2 estimates σ^2

B2. Unbiasedness and Consistency

- **Unbiasedness**: The sample mean $\bar{X}$ is unbiased because its expected value equals the population mean: $\mathbb{E}[\bar{X}] = \mu$. This implies that, over repeated sampling, $\bar{X}$ would center exactly on μ.
- **Consistency**: As the sample size n increases, the variance of $\bar{X}$ decreases ($\text{Var}(\bar{X}) = \sigma^2/n \to 0$), causing $\bar{X}$ to converge in probability to μ. Thus, larger samples yield more precise estimates.

B3. Standard Error of the Mean

$$SE(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{84.8235}{\sqrt{30}} \approx \frac{84.8235}{5.477} \approx 15.487 \text{ minutes}$$

This value quantifies the **standard deviation of the sampling distribution of** $\bar{X}$. It indicates that, if we were to repeatedly draw samples of size 30 from this population, the sample means would typically vary by about **15 minutes** from the true mean μ.

4.3 Part C: Confidence Interval

C1. 95% Confidence Interval for μ

Because σ^2 is unknown and the sample size is modest ($n = 30$), we use the t**-distribution** with $df = n - 1 = 29$. The critical value is $t_{0.025,29} = 2.045$.

$$\bar{x} \pm t_{\alpha/2,df} \cdot \frac{s}{\sqrt{n}} = 106.267 \pm 2.045 \cdot 15.487 = 106.267 \pm 31.671$$

Thus, the 95% confidence interval is:

$$(74.596,\ 137.938) \text{ minutes}$$

C2. Interpretation

We are 95% confident that the **true average daily music listening time** for the population of Cambodian undergraduate students lies between **74.596 and 137.938 minutes**. This interval does not include 60 minutes, providing preliminary evidence that the average exceeds one hour.

C3. Required Sample Size for $E = 0.5$

To achieve a margin of error of $E = 0.5$ minutes at 95% confidence, and using $s = 84.8235$ as an estimate of σ, we compute:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{1.96 \cdot 84.8235}{0.5}\right)^2 \approx (332.49012)^2 \approx 110{,}550.26 = 110{,}550$$

This enormous required sample size around 100,000 observations—highlights the **impracticality of achieving sub-minute precision** in estimating average listening time without vast resources. Our current sample of $n = 30$ is appropriate for **exploratory inference** but insufficient for high-precision estimation.

4.4 Part D: Hypothesis Testing

D1. Hypotheses

We test the claim that students listen to **more than one hour (60 minutes)** of music per day on average:

$$H_0 : \mu = 60 \quad \text{vs.} \quad H_a : \mu > 60 \quad \text{(right-tailed test)}$$

D2. Test Statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{106.267 - 60}{84.8235/\sqrt{30}} = \frac{46.267}{15.487} \approx 2.988$$

D3. p-value

With $df = 29$, the p-value is the area to the right of $t = 2.988$ under the t_{29} curve. Using statistical software or interpolation:

$$p \approx 0.002833$$

D4. Decision

At $\alpha = 0.05$, since $p = 0.002833 < 0.05$, we **reject the null hypothesis**.

D5. Conclusion in Context

At the 5% significance level, there is **statistically significant evidence** to conclude that the average daily music listening time among Cambodian undergraduate students **exceeds 60 minutes**. This supports the hypothesis that music is a substantial component of daily student life.

5 Discussion and Conclusion

This investigation successfully applied the core statistical methods of Chapters 1–4 to a real-world dataset, fulfilling all project objectives. We found compelling evidence that Cambodian students listen to music for an average of approximately **106 minutes per day**, with considerable individual variation.

Alignment with Learning Objectives:

- We selected a **simple, real-world numerical variable** suitable for hand calculation.
- We summarized data using **appropriate graphical and numerical tools**.
- We constructed and interpreted a **confidence interval** for μ.
- We formulated and tested a **meaningful hypothesis** using a t-test.
- We communicated findings in a structured written report and will do so in an oral presentation.

Methodological Considerations:

- The assumption of **independence** is reasonable given anonymous, individual responses.

- The **right-skewed distribution** and small n violate the normality assumption for the t-test. However, the t-procedure is robust to mild skewness, and our sample size ($n = 30$) offers moderate protection via the Central Limit Theorem.

- **Self-reported data** may suffer from recall bias or over/underestimation.

Limitations:

- **Convenience sampling** limits generalizability beyond our peer group.

- **Cross-sectional design** captures only average daily habits, not variability over time.

Future Research: A larger, stratified random sample could improve external validity. Repeated measures over multiple days per student would allow for analysis of within-person variability and temporal patterns.

6 Appendix

6.1 Raw Data (n = 30)

30	240	140	30	60	120	80	120	120	30
360	60	50	30	60	120	30minutes	3	30	60
60	2h	25	240	200	60	150	200	2 Hours	240

6.2 Python Code for Visualization

```
# Import libraries
import numpy as np
import pandas as pd
import statistics as stat
import matplotlib.pyplot as plt
import seaborn as sns

# Set a clean style
sns.set_style("whitegrid")
plt.rcParams.update({'font.size': 10})

# Load data
df = pd.read_csv('Music_Listening_Habits_Survey_(Responses)_Form_Responses_1.csv')
df.sample(5)
```

```
# Randomly choose rows to drop
rows_to_drop = df.sample(n=8, random_state=42).index
# Drop these rows
df_cleaned = df.drop(rows_to_drop).reset_index(drop=True)

# Drop time column
df_cleaned = df_cleaned.drop(columns='Timestamp')

# Rename columns name
df_cleaned = df_cleaned.rename(columns={
    "What is your gender?": "gender",
    "How old are you? (Please enter a whole number only, e.g., 19)": "age",
    "On average, how many minutes per day do you listen to music?
    (Please answer with a number only, e.g., 120)": "minutes",
    "What type of music do you listen to most often?": "music_type"
})

# Data cleaning
def clean_minutes(value):
    value = str(value).lower().strip()

    if 'h' in value:
        number = float(value.replace("hours", "")
                            .replace("hour", "")
                            .replace("h", "")
                            .strip())
        return number * 60
    else:
        return float(value.replace("minutes", "")
                          .replace("minute", "")
                          .replace("mn", "")
                          .strip())

# Apply cleaning
df_cleaned['minutes'] = df_cleaned['minutes'].apply(clean_minutes)
```

```
# Histogram
fig, ax = plt.subplots(figsize=(18, 12))

sns.histplot(data=df_cleaned,
             x='minutes',
             kde=True,
             color='purple',
             bins=8,
             line_kws={'linewidth': 3},
             ax=ax)

ax.set_xlabel('Minutes', fontsize=16)
ax.set_ylabel('Frequency', fontsize=16)
ax.set_title('Distribution of Daily Music Listening Time (Minutes)', fontsize=20)
ax.tick_params(labelsize=14)
ax.axvline(df_cleaned['minutes'].mean(), color='red', linestyle='--',
           linewidth=3, label=f'Mean: {df_cleaned["minutes"].mean():.2f}')
ax.legend(fontsize=14)
plt.show()

# Boxplot
fig, ax = plt.subplots(figsize=(6, 6))
sns.boxplot(y=df_cleaned['minutes'],
            ax=ax,
            color='lightblue',
            width=0.4)
from matplotlib.cbook import boxplot_stats
stats = boxplot_stats(df_cleaned['minutes'])[0]
outliers = stats['fliers']
ax.scatter([0] * len(outliers),
           outliers,
           color='red',
           s=50,
           zorder=5,
           label='Outliers')
ax.set_title('Boxplot of Daily Music Listening Time (Minutes)', fontsize=14)
```

```
ax.set_ylabel('Minutes', fontsize=12)
ax.set_xticks([])
if len(outliers) > 0:
    ax.legend()
plt.tight_layout()
plt.show()
```