Introduction to Data Science **Sample Assignment**

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh** TAs: محمد جواد رنجبر

Problem 1: Central Limit Theorem and Sampling Distributions (15 pts)

🎵 The **popularity score** of a song is a measure between 0 and 100, representing its relative popularity on Spotify. Suppose the popularity scores in our dataset are drawn from an unknown distribution with mean μ\muμ and standard deviation σ\sigmaσ.

(a) **Demonstrate the Central Limit Theorem (CLT).**

- Randomly draw **1000 samples** of size **30** from the **'popularity'** column.
- Compute the **mean** of each sample and plot the **sampling distribution** of the sample means.
- Repeat this process for sample sizes **50, 100, and 500** and compare the distributions.

(b) **How does the shape of the sampling distribution change as sample size increases?**

- Compute the **mean and standard deviation** of the sampling distribution for each sample size.
- Compare them with the population mean and standard deviation.
- Explain why this supports the Central Limit Theorem.

Problem 2: Confidence Intervals for Song Duration (15 pts)

🎵 The **duration** of a song (in milliseconds) varies across different artists and albums. We assume song durations are normally distributed with unknown mean μ\muμ but known standard deviation σ=50,000\sigma = 50,000σ=50,000 ms.

(a) Construct **95% and 99% confidence intervals** for the true mean **duration** of Persian songs using a random sample of **50 songs**.

(b) Suppose we increase the sample size to **200 songs**.

- Compute the **new 95% confidence interval**.

- Explain why the confidence interval changes.

(c) What assumptions about the sample are necessary for these confidence intervals to be valid?

Problem 3: Hypothesis Testing – Explicit vs. Non-Explicit Songs (20 pts)

🎵 Explicit songs contain **mature language** and may be treated differently by streaming algorithms. We want to test whether **explicit songs are more popular than non-explicit songs**.

(a) **Formulate the hypotheses:**

- H_0 (Null Hypothesis): The mean popularity of explicit and non-explicit songs is **equal**.
- H_1 (Alternative Hypothesis): Explicit songs have **higher mean popularity**.

(b) **Perform a two-sample t-test:**

- Compute the mean and standard deviation of popularity for **explicit** and **non-explicit** songs.
- Conduct an **independent two-sample t-test** (assume unequal variances).
- Report the **p-value** and conclude whether we can reject H0H_0H0 at the **5% significance level**.

(c) **Effect size and practical significance:**

- Calculate **Cohen's d** to measure the practical significance of the difference.
- Explain whether the difference is meaningful in a real-world context.

Problem 4: Power of a Test – Danceability Before and After 2010 (18 pts)

🎵 **Danceability** is a feature measuring how suitable a song is for dancing (scale: 0 to 1). We hypothesize that modern Persian songs (released in **2010 and later**) are **more danceable** than older ones (before 2010).

(a) **State the hypotheses:**

- H_0: The mean **danceability** of songs released before 2010 is the **same** as those released in 2010 or later.
- H_1: The mean **danceability** of modern songs is **higher** than older ones.

(b) **Determine the sample size for a powerful test:**

- Assume a **true difference** of **0.05** in danceability scores.
- Use a power analysis to find the **minimum sample size** required to detect this difference at **α=0.05 and power = 0.8**.
- Repeat for power levels **0.9 and 0.95** and compare results.

(c) **Interpret the findings:**

- If the required sample size is large, what does this say about the strength of the effect?
- How does the **effect size** influence the power of the test?