# Assignment 4.1 Exercise

In this assignment, you will extend the concepts demonstrated in the lab portion to analyze income levels within a specific ethnic group: Native Americans. The goal is to understand how applying differential privacy impacts our analysis when focusing on a smaller, often underrepresented population group.

Specifically, you will explore how the Laplace mechanism affects the **income level counts** for Native Americans. Education and other government funding often depend on accurate data about income levels, and differential privacy can introduce noise that may distort these figures. Your task will be to determine how varying the privacy budget (epsilon) changes the accuracy of these counts and to reflect on the broader implications of these changes, particularly when allocating resources to underrepresented groups.

## Step 1: Calculate Original Counts for Income Levels

Display the original income counts for Native Americans:

```
Income Level Counts for Native Americans (Original Data):
Low: 45
Middle: 39
High: 27
```

## Step 2: Apply Laplace Mechanism to Add Noise

Apply the Laplace mechanism to add noise with at least two epsilon values.

**Answer:**

**The Native American Income Level after applying noise through Laplace mechanism.**

**Epsilon values considered are [ 0.1, 0.5. 1.0]**

**Epsilon - 0.1 Output**

```
Change in Representation for Native American Income Levels with Epsilon = 0.1:
Income Level: Low
 - Original Count: 45.00
 - Noisy Count: 46.20
 - Change in Representation (%): 2.66%

Income Level: Middle
 - Original Count: 39.00
 - Noisy Count: 32.89
 - Change in Representation (%): -15.68%

Income Level: High
 - Original Count: 27.00
 - Noisy Count: 33.87
 - Change in Representation (%): 25.46%
```

**Epsilon  0.5 Output**

```
Change in Representation for Native American Income Levels with Epsilon = 0.5:
Income Level: Low
 - Original Count: 45.00
 - Noisy Count: 48.00
 - Change in Representation (%): 6.67%

Income Level: Middle
 - Original Count: 39.00
 - Noisy Count: 38.00
 - Change in Representation (%): -2.57%

Income Level: High
 - Original Count: 27.00
 - Noisy Count: 31.45
 - Change in Representation (%): 16.47%
```

**Epsilon  1.0 Output**

```
Change in Representation for Native American Income Levels with Epsilon = 1.0:
Income Level: Low
 - Original Count: 45.00
 - Noisy Count: 44.68
 - Change in Representation (%): -0.72%

Income Level: Middle
 - Original Count: 39.00
 - Noisy Count: 38.87
 - Change in Representation (%): -0.33%

Income Level: High
 - Original Count: 27.00
 - Noisy Count: 27.75
 - Change in Representation (%): 2.76%
```

## Step 3: Generate Relative Error Visualization

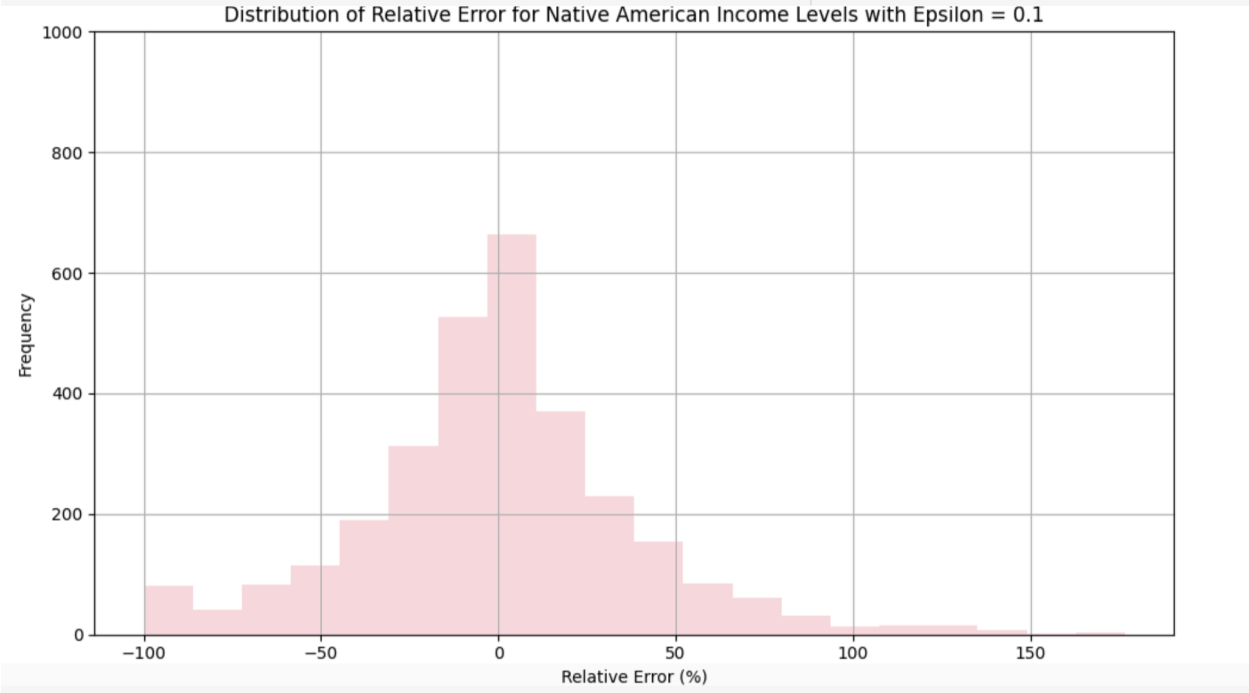
Screenshot the generated histograms for relative errors using the Laplace mechanism and paste your output for both below:

**Visualization shown for 3 Epsilon values : [ 0.1, 0.5, 1.0]**

**The graphs shown  below are  for the following:**

- **Distribution of Relative Error for Native American Level with Epsilon = 0.1**
- **Distribution of Relative Error for Native American Level with Epsilon = 0.5**
- **Distribution of Relative Error for Native American Level with Epsilon = 1**

- **Distribution of Relative Error for Native American Level with Epsilon = 0.1**

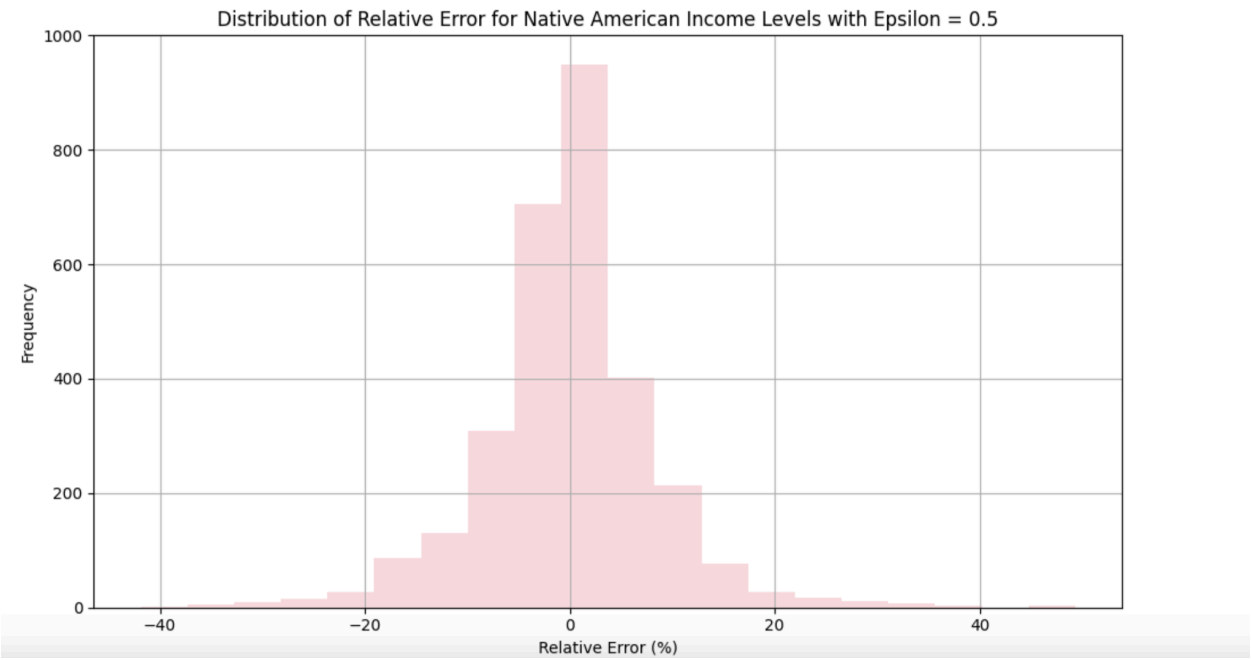



**Result Interpretation for Epsilon = 0.1**

The histogram for Epsilon = 0.1 shows a **wide and skewed distribution of relative errors, ranging from −100% to +150%.** Although **most errors are clustered near 0%**, the frequent and large outliers highlight **how strong privacy protections can severely distort the accuracy of income-level data for smaller groups like Native Americans**. This illustrates the trade-off between privacy and utility and suggests that Epsilon = 0.1 may be too low to support accurate policy decisions.

==========

- **Distribution of Relative Error for Native American Level with Epsilon = 0.5**

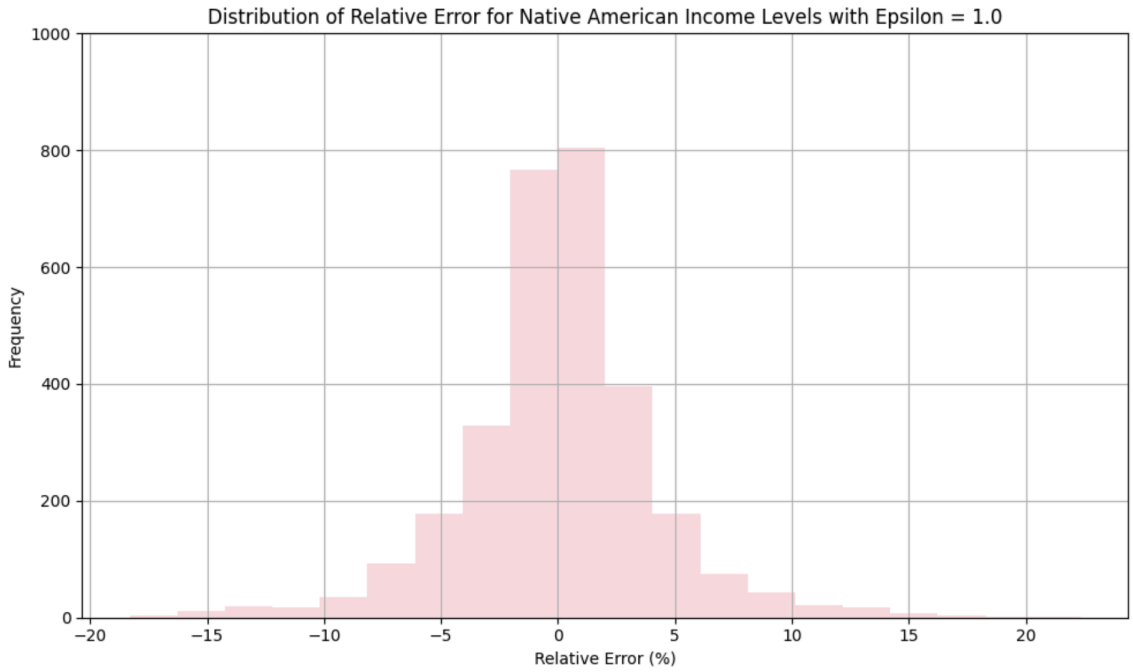



**Interpretation of the result for Epsilon 0.5**

The histogram for Epsilon = 0.5 shows a tighter, **more centered distribution of relative error, mostly within ±10%.** This indicates that the **Laplace noise introduced moderate variation, but preserved overall accuracy.** The graph suggests Epsilon = 0.5 provides a strong balance between privacy protection and data utility, especially important for small populations like Native Americans where accurate representation is critical for equitable resource allocation.

==========

- **Distribution of Relative Error for Native American Level with Epsilon = 1**




**Interpretation of the result for Epsilon = 1**

The histogram for Epsilon = 1.0 displays a very tight distribution of relative error, **mostly within ±5%, with a strong concentration near 0%.** This indicates that the **added noise had minimal impact on the accuracy of income level counts for Native Americans**. While this provides high utility for decision-making and analysis, it comes at the cost of reduced privacy. Therefore, Epsilon = 1.0 may be appropriate when high accuracy is essential, but it offers weaker privacy guarantees compared to lower Epsilon values.

==========

## Step 4: Observation and Reflection

1. Briefly describe your findings. How does changing the value of epsilon affect the relative error?

**Answer:**

E**psilon increases**, the **noise added to the data decreases**, resulting in **lower relative error** and **more accurate results**:

- At **Epsilon = 0.1**, relative errors ranged from about **−100% to +150%**. This caused significant distortion, making the data **less reliable**, especially for small groups like Native Americans.

- At **Epsilon = 0.5**, relative errors were mostly within **−40% to +50%**. This provided a **better balance** between accuracy and privacy, with more usable data and

moderate protection.

- At **Epsilon = 1.0**, relative errors were typically within **−10% to +10%**, showing **high accuracy** with only slight deviations from the original data — but with **weaker privacy protection**

2. Is there an epsilon value that produces useful results while still providing privacy protection?

Explain your reasoning.

**Answer**

Yes — based on the analysis, Epsilon = 0.5 **appears to provide a good balance between usefulness of the results and privacy protection.**

Reasoning:

1. Data Utility With Epsilon = 0.5, the relative error across all ethnic groups remained very small, mostly within ±1%. **This level of accuracy ensures that statistical trends and distributions remain reliable, making the data useful for analysis, reporting, or policymaking.**
2. Privacy Protection with Epsilon = 0.5 offers less privacy than Epsilon = 0.1, it still provides a meaningful layer of protection: **The noise is enough to mask individual contributions, especially in large datasets like this (10,000 rows).** It's suitable for aggregate reporting where individual identification risk is relatively low. Epsilon = 0.1 — **Too Noisy Causes high distortion**, especially in small groups (e.g., +10% or −8% errors). May misrepresent underrepresented communities and lead to inaccurate conclusions.

   **Epsilon = 1.0 — Too Weak for Strong Privacy Needs Offers high utility but weaker protection** — not ideal if sensitive personal data is involved or if adversaries have auxiliary information.

**Conclusion**

**Epsilon = 0.5 is a practical middle ground:** It preserves the structure and usefulness of the data while still maintaining a reasonable level of privacy.

This value is often recommended in practice (e.g., in research papers and DP libraries) as a starting point, with adjustments based on the sensitivity of the data and use case context.

3. Discuss how the trade-offs between privacy and data utility in differential privacy might impact smaller communities in a dataset (i.e. deflated or over-inflated representation). How might the added noise impact decision-making, especially in the context of census data?

**Answer :**

Differential privacy adds noise to data to protect individual identities, but this comes with a trade-off: as privacy increases, data accuracy decreases. **While this may be acceptable for large groups, it can significantly distort results for smaller, underrepresented communities like Native Americans.**

Because these groups have low population counts, even small amounts of noise can cause large percentage changes—either inflating or deflating their presence in the data.

This misrepresentation can lead to serious consequences in real-world decision-making. For example, **census data is used to allocate funding, plan services, and determine political representation. If noise causes a small group to appear smaller than it is, they may receive less support, funding, or recognition.**

Using too much noise (low epsilon) prioritizes privacy but risks erasing small communities from the data. On the other hand, too little noise (high epsilon) improves accuracy but reduces privacy. A balanced approach, such as a moderate epsilon (e.g., 0.5), offers a compromise between protecting individuals and preserving the visibility of smaller groups.

**In summary, differential privacy must be applied carefully to ensure it doesn't unintentionally harm the very communities it aims to protect. Maintaining this balance is key to ethical and equitable data use**