Statistics

|Symmetry and Skewness (التناظر والانحراف)

Prof. Dr. Mostafa Elhosseini

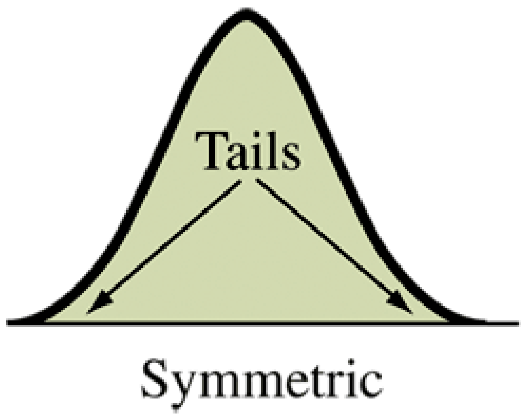
https://youtube.com/drmelhosseini

Agenda

- Symmetric Distribution (Normal - Bell Shaped – Gaussian)
- What Is Skewness?
- Positive Skew (Right Skew)
- Negative Skew (Left Skew)
- Why the Shape is Important?
- How to Quantify Skewness?
- Examples - Codes

Symmetric Distribution

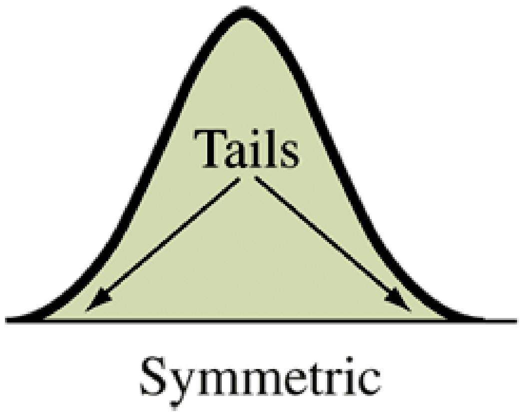
- The left and right sides **mirror** each other.
- Typically, the **mean = median = mode**.



- Example: A perfectly normal (bell-shaped) distribution of standardized **test scores** (under an **ideal scenario**) or measurement errors in a well-controlled experiment.

Symmetric Distribution

- The left and right sides mirror each other.
- Typically, the mean = median = mode.



- If you measure the **heights of adult males** in a homogeneous population, you might get a fairly symmetric (normal-like) distribution centered around a mean (e.g., 175 cm) and the **distribution tails decreasing at about the same rate in both directions**.

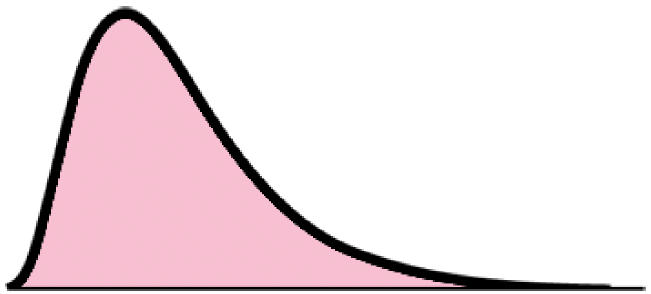
What Is Skewness?

- Skewness is a **measure of asymmetry** in a distribution.
  - When a distribution is **symmetric**, the left side of its histogram (or curve) mirrors the right side.
  - When data "**skew**" to one side, one tail of the distribution is longer or fatter than the other.

- Formally, skewness can be **quantified** (e.g., using **Pearson**'s or **Fisher**'s measure of skewness)

Types of Skewness

Positive Skew (Right Skew)



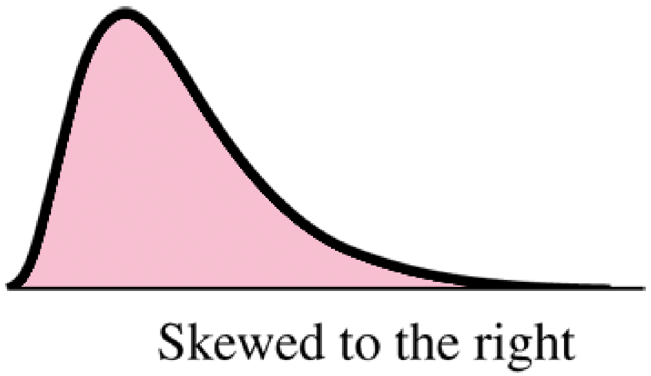
Skewed to the right

- The **right tail** of the distribution **is longer** (or fatter) than the left tail.
- Typically, the **mean > median > mode**.

- Example: Income distribution often has a long right tail—most people earn below or around an average range, but a <u>small number of very high incomes</u> stretch out the tail on the right side.

Types of Skewness

Positive Skew (Right Skew)

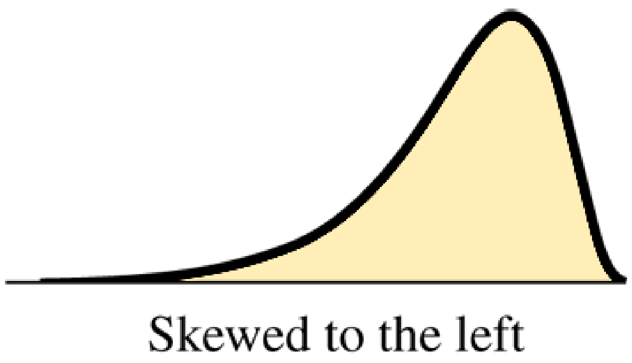


- The right tail of the distribution is longer (or fatter) than the left tail.
- Typically, the mean > median > mode.

- Consider the distribution of **time to complete a new task** (e.g., an assembly line step). Some workers might complete it quickly (left side of the distribution), but a **few may take a significantly longer time** (right tail). This pulls the mean to the right.

Types of Skewness

Negative Skew (Left Skew)

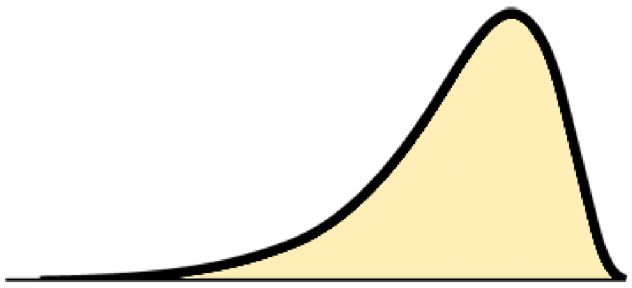


- The left tail of the distribution is longer (or fatter) than the right tail.
- Typically, the **mean < median < mode**.

- Example: **Exam scores in an easy test** scenario often have a longer left tail: most students score very high, and a few lower scorers stretch out the tail on the left side..

Types of Skewness

Negative Skew (Left Skew)

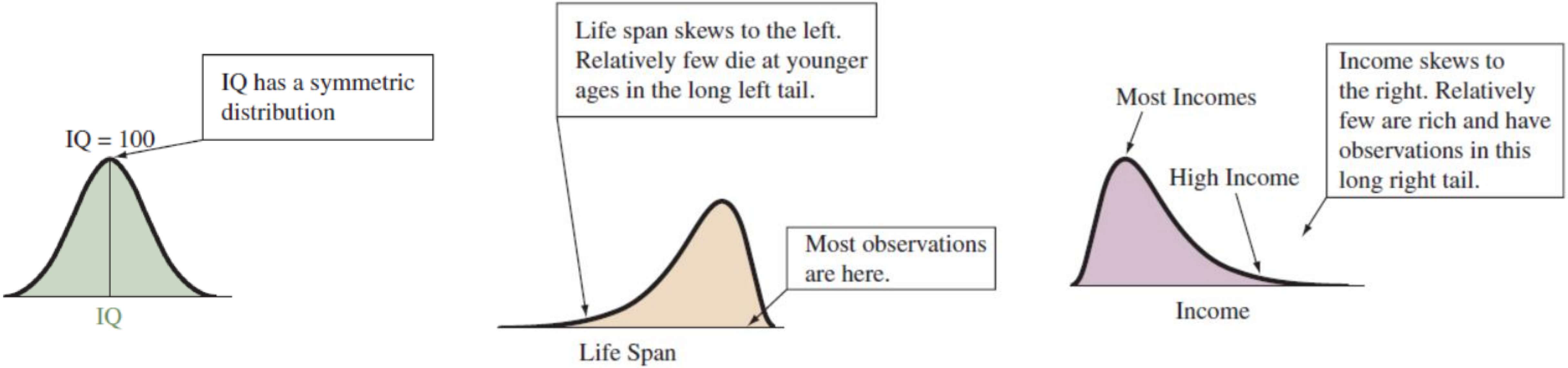


Skewed to the left

- The **left tail** of the distribution **is longer** (or fatter) than the right tail.
- Typically, the **mean < median < mode**.

- Consider a **reading test for native speakers** in their own language. Most participants might achieve high scores (right side of the distribution), while just a small group may achieve much lower scores (left tail).

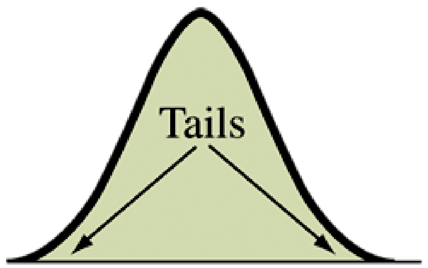
To conclude



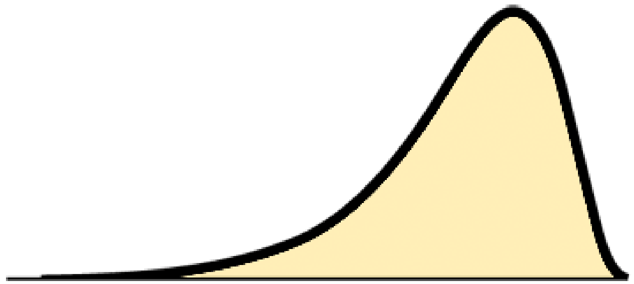
with small data sets, the shape of the distribution might not be so obvious

To conclude

Symmetric Distributions: if both left and right sides of the histogram are mirror images of each other.

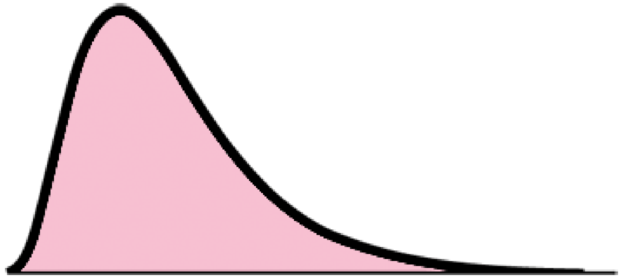


Symmetric



Skewed to the left

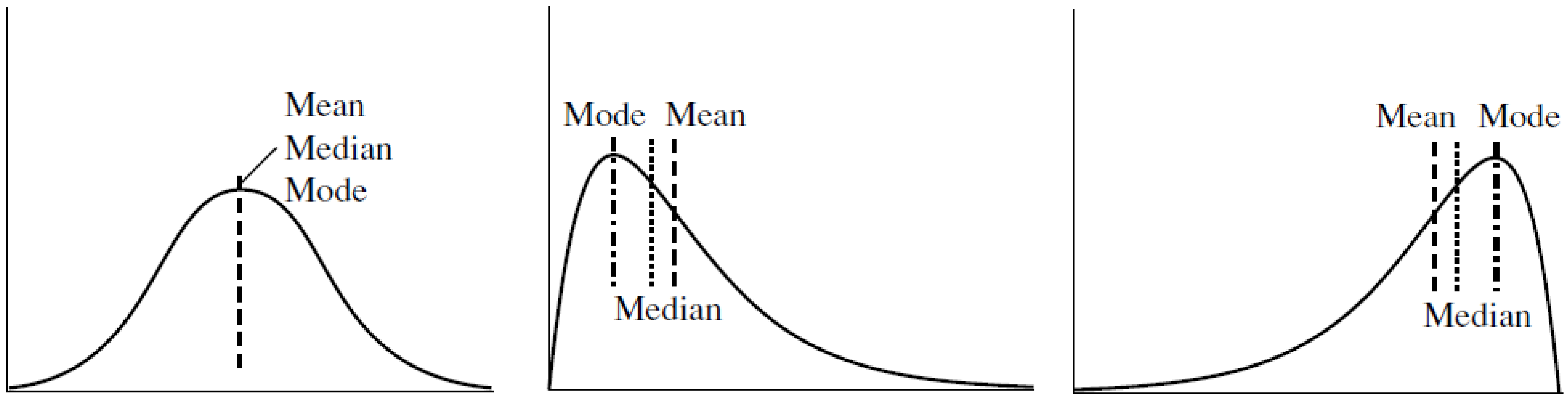
A distribution is **skewed to the left** if the left tail is longer than the right tail.



Skewed to the right

A distribution is **skewed to the right** if the right tail is longer than the left tail.

To conclude



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

Mean, median, and mode of symmetric versus positively and negatively skewed data.

Why the Shape is Important?

- In a **symmetric distribution**, the mean, median, and mode often **coincide** (or are very close). Therefore, using the **mean** is usually a **good representation** of the central tendency.

- In a skewed distribution, the mean, median, and mode can be far apart. Relying solely on the **mean might be misleading** because it can be dragged in the direction of the skew (towards the long tail). In such cases, the median may be a better indicator of the typical value.

Why the Shape is Important?

- Many **statistical tests** assume **normality** (a symmetric, bell-shaped distribution). If data are highly skewed, these assumptions may not hold, and results (p-values, confidence intervals) could be invalid or misleading.

- Understanding the shape helps in deciding whether **data transformation** (e.g., taking logarithms) or nonparametric methods would be more appropriate.

Why the Shape is Important?

- Skewness can hint at underlying phenomena—such as natural lower or upper bounds, outliers, or a long "tail" in one direction.

- For instance, in income distribution, it is common to have a few very high earners on the right-hand side (a right-skew), indicating that most people earn around a certain range, while a small number earn substantially more.

How to Quantify Skewness

- Statisticians use different formulas (or coefficients) to measure the degree and direction of asymmetry (skewness) in a distribution. Below are the most common methods and interpretations:
- Pearson’s Skewness Coefficients
  - Pearson’s First Coefficient of Skewness
  - Pearson’s Second Coefficient of Skewness
- Fisher-Pearson (Moment) Coefficient of Skewness

How to Quantify Skewness

- Pearson’s First Coefficient of Skewness

$$S_1 = \frac{\bar{X} - mode}{\sigma}$$

- $\bar{X}$: Mean
- σ: Standard Deviation

- Pearson’s First Coefficient is used when the mode can be reliably identified.

How to Quantify Skewness

- Pearson's Second Coefficient of Skewness

$$S_2 = \frac{3 \times (\bar{X} - Median)}{\sigma}$$

- $\bar{X}$: Mean
- σ: Standard Deviation

- Pearson's Second Coefficient is often preferred when the mode is unclear or difficult to determine.

How to Quantify Skewness

Interpreting Skewness Values

- There are no absolute cutoffs, but a common rule of thumb is:
- Skewness between -0.5 and 0.5: The distribution is roughly symmetric.
- Skewness between -1 and -0.5 or between 0.5 and 1: Moderate skew.
- Skewness <-1 or Skewness>1 : High (severe) skew.

How to Quantify Skewness

<u>Interpreting Skewness Values</u>

- If $(\bar{X} - mode)$ is positive $s_1 > 0$ → Right Skew
- If $(\bar{X} - mode)$ is Negative $s_1 < 0$ → Left Skew

- If $(\bar{X} - Median)$ is positive $s_1 > 0$ → Right Skew
- If $(\bar{X} - Median)$ is Negative $s_1 < 0$ → Left Skew

How to Quantify Skewness

Fisher-Pearson (Moment) Coefficient of Skewness

- This is the most commonly reported measure in statistical software (e.g., scipy.stats.skew in Python, or skewness() in R). It is based on the third central moment

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- x_i: is each data point
- $\bar{x}$: is the sample mean
- s: sample standard deviation
- n: is the sample size

How to Quantify Skewness

Fisher-Pearson (Moment) Coefficient of Skewness

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- $\frac{n}{(n-1)(n-2)}$ serves as a bias correction for small sample sizes
- $g_1 \cong 0 \rightarrow$ The distribution is (approximately) symmetric.
- $g_1 > 0 \rightarrow$ The distribution is positively (right) skewed (long right tail).
- $g_1 < 0 \rightarrow$ The distribution is negatively (left) skewed (long left tail).

Key Points When Measuring Skewness

- Extreme values (outliers) can strongly influence skewness coefficients.
- With small samples, skewness estimates may not reflect the true underlying distribution accurately. Bias corrections (like in the Fisher-Pearson method) help mitigate this.
- Use Pearson's First Coefficient when the mode is known and distinct.

Key Points When Measuring Skewness

- Use Pearson's First Coefficient when the mode is known and distinct.
- Use Pearson's Second Coefficient (median-based) or the Fisher-Pearson coefficient when the mode is unknown or less meaningful.
- In cases of severe skewness, transformations (e.g., log or square-root) can help normalize data before applying parametric statistical tests that assume (near) normality.

Simple Example

- 9 data points (already sorted for convenience): {1, 2, 2, 3, 4, 8, 9, 12, 15}
- We will compute:
  - Mean
  - Median
  - Mode
  - Standard Deviation (Sample)
  - Pearson's Skewness Coefficients (First & Second)
  - Fisher-Pearson (Moment) Coefficient (Optional, if you want to see a more advanced measure)

Simple Example

- Mean = $\bar{x} = \frac{1+2+2+3+4+8+9+12+15}{9} = \frac{56}{9} = 6.22$
- Median: The middle (5th) value is 4. Median = 4
- Mode = 2
- Standard Deviation (Sample) $s = 4.996$
- Pearson's First Coefficient $s_1 = \frac{\bar{X} - mode}{s} = \frac{6.22-2}{4.996} = 0.84$

Simple Example

- Mean = $\bar{x} = \frac{1+2+2+3+4+8+9+12+15}{9} = \frac{56}{9} = 6.22$
- Median: The middle (5th) value is 4. Median = 4
- Mode = 2
- Standard Deviation (Sample) $s = 4.996$

- Pearson's Skewness Coefficients (First & Second) $S_2 = \frac{3\times(\bar{X}-Median)}{s} = \frac{3\times(6.22-4)}{4.996} = 1.33$

Simple Example

- Mean = $\bar{x} = \frac{1+2+2+3+4+8+9+12+15}{9} = \frac{56}{9} = 6.22$
- Median: The middle (5th) value is 4. Median = 4
- Mode = 2
- Standard Deviation (Sample) $s = 4.996$
- Fisher-Pearson (Moment) Coefficient (Optional, if you want to see a more advanced measure) $g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3$
- $g_1 = +0.72$ (moderately positive skew)

Simple Example

- Interpretation:
- From these calculations:
  - Pearson’s First Skewness $S1 \approx 0.84$
  - Pearson’s Second Skewness $S2 \approx 1.33$
  - Fisher-Pearson (Moment) Coefficient $g1 \approx 0.72$
- All are positive values, confirming the data are right-skewed (positively skewed). This aligns with the pattern that a few larger values at the high end (9, 12, 15) pull the mean to the right, making it notably higher than the median and mode.

Heavily Left-Skewed Distribution Example

- We generate 100 data points from a Beta(20, 2) distribution, which is known to be heavily left (negatively) skewed.

- Statistical Measures
  - Mean
  - Median
  - Sample Standard Deviation
  - Fisher-Pearson Skewness (stats.skewstats.skew)

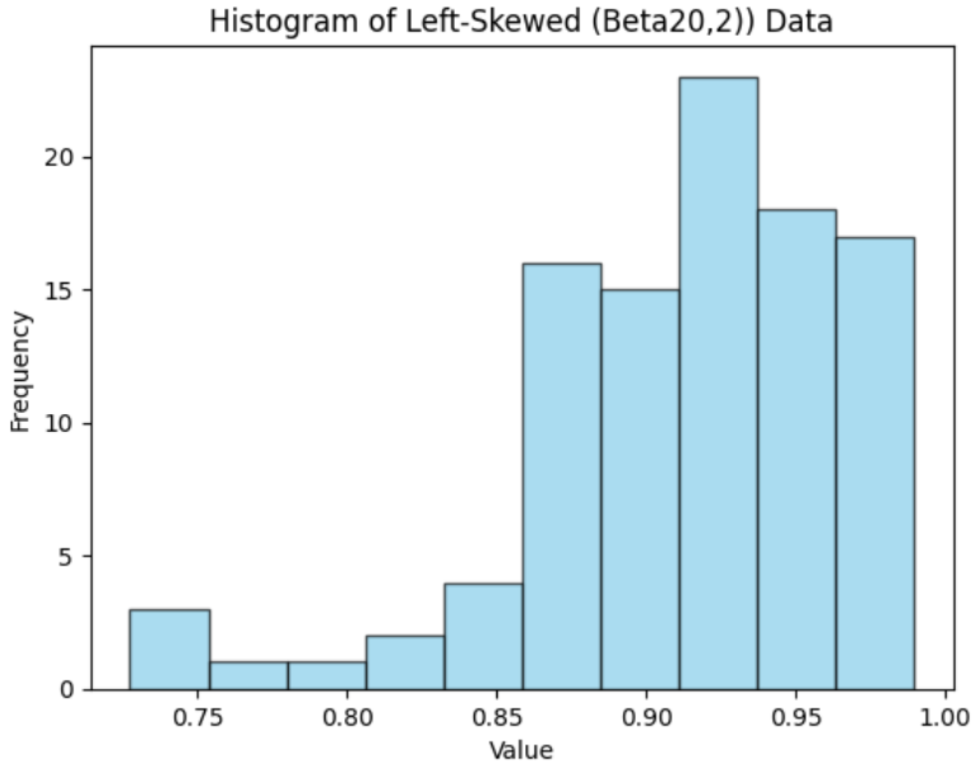
- Interpretation
  - If skewness < 0, the tail is to the left → negatively skewed.
  - The histogram visually confirms the shape: values cluster near 1 with a long tail toward 0.

Heavily Left-Skewed Distribution Example

- Most values near 1, with a long tail stretching toward 0
- The mean is typically slightly less than the median, reflecting the left-side tail pulling the mean downward.

```
mean_value = np.mean(data)
median_value = np.median(data)
std_value = np.std(data, ddof=1)
skew_value = stats.skew(data, bias=False)
```

```
------ Basic Statistics ------
Mean:   0.9116
Median: 0.9233
Std:    0.0550
Skew:   -1.2293
```



References

- Montgomery, D.C. and Runger, G.C., 2010. Applied statistics and probability for engineers. John wiley & sons.
- Devore, J.L., 2011. Probability and Statistics for Engineering and the Sciences. Cengage learning.
- Gubner, J.A., 2006. Probability and random processes for electrical and computer engineers. Cambridge University Press.
- Agresti, A. and Franklin, C., 2007. The art and science of learning from data. Upper Saddle River, New Jersey, 88.
- Navidi, W.C., 2008. Statistics for engineers and scientists (Vol. 2, No. 4). New York, NY, USA: McGraw-Hill Higher Education.