Understanding Histograms in Statistics

Statistics Course Notes

Introduction to Histograms

A **histogram** is a graphical representation of the distribution of numerical data. It provides a visual summary of:

- The **frequency** of data points
- The **spread** of the data
- Potential **patterns** or **outliers**

Histograms form the basis for deriving **probability density functions (PDFs)** using techniques like **kernel density estimation (KDE)**.

Key Components of a Histogram

- **Bins**: Equally spaced intervals that cover the range of data
- **Frequency**: Number of data points in each bin
- **Axes**:
  - Horizontal: Represents data ranges (bins)
  - Vertical: Represents frequency/count

Constructing a Histogram: Step-by-Step

Consider the age dataset: {23, 24, 25, 30, 34, 36, 40, 50, 60, 75, 80}

Step 1: Define Bins

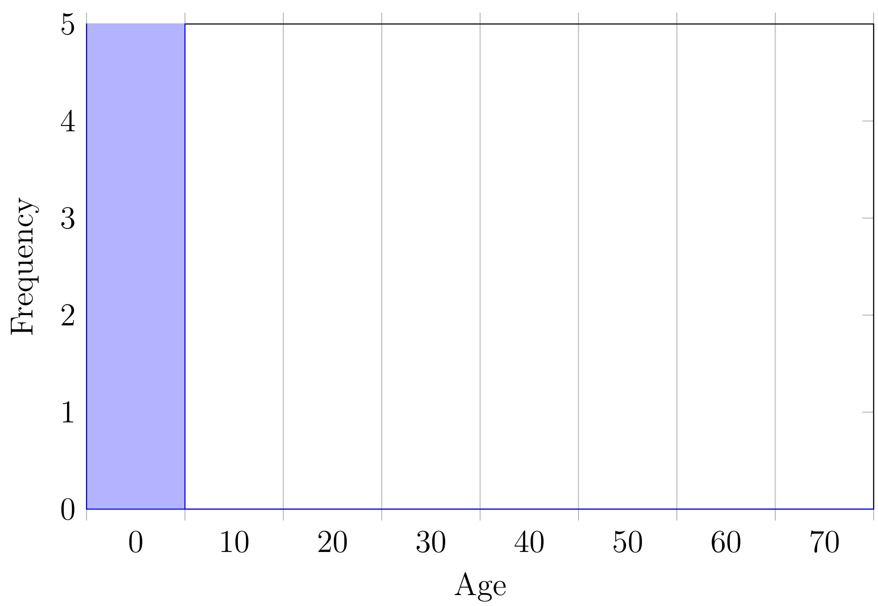
Choose bin size (e.g., 10-year intervals):

Bin Range	Description
0-10	
10-20	
20-30	Ages 23, 24, 25, 30
30-40	Ages 34, 36, 40
40-50	Age 50
50-60	Age 60
60-70	
70-80	Ages 75, 80

Step 2: Count Frequencies

Bin	Frequency
0-10	0
10-20	0
20-30	4
30-40	3
40-50	1
50-60	1
60-70	0
70-80	2

Step 3: Plot the Histogram



Important Concepts

1. Bin Size Selection

- Smaller bins → More detail but noisier
- Larger bins → Smoother but less detail
- Rule of thumb: Number of bins $\approx \sqrt{n}$ where n = data points

2. Boundary Conventions

- **Standard**: [lower, upper) - *includes* lower bound, *excludes* upper
- **Example**: [20, 30) contains 20 but not 30
- *Note*: Always specify your boundary convention!

3. Continuous vs. Discrete Data

- **Continuous**: Bars touch each other (shown above)
- **Discrete**: Gaps between bars (categorical data)

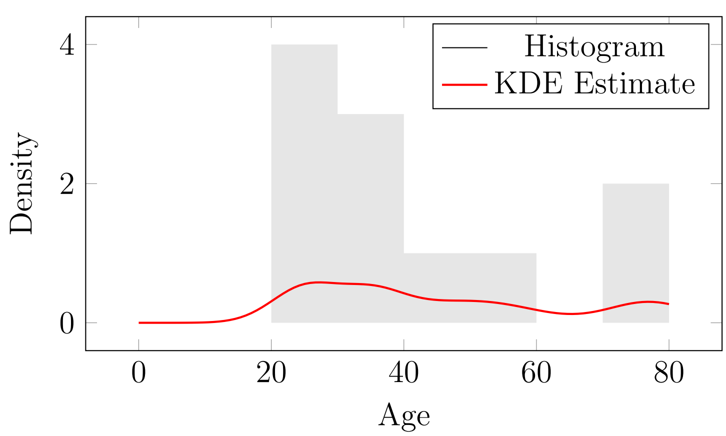
Histograms to Probability Density Functions (PDFs)

A histogram can be converted to a smooth PDF using **Kernel Density Estimation (KDE)**:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Where:

- K = kernel function (e.g., Gaussian)
- h = bandwidth (smoothing parameter)
- n = number of data points



Practical Considerations

1. **Outliers**: Can distort bin ranges - consider trimming
2. **Zero-frequency bins**: Indicate gaps in data distribution
3. **Software implementation**:

   ```
   # Python example
   import matplotlib.pyplot as plt
   ages = [23,24,25,30,34,36,40,50,60,75,80]
   plt.hist(ages, bins=8, edgecolor='black')
   plt.xlabel('Age')
   plt.ylabel('Frequency')
   plt.show()
   ```

Key Takeaways

- Histograms visualize **distribution** of continuous data
- Bin size critically impacts interpretation
- Histograms provide foundation for **probability density estimation**
- KDE creates smooth PDFs from histograms
- Always document bin boundaries and size