Writing Assignment 1: Distributions

Probability and Statistics

Author

Paul Beggs
BeggsPA@Hendrix.edu

Instructor

Prof. Lars Seme, M.S.

Due

September 17, 2025

1 Introduction

Probability and statistics rely on the concept of a distribution, which describes how the values of a random variable are spread across possible outcomes. Understanding a distribution's characteristics and properties are key to interpret data. Therein, the goal of this paper is to look into the relationship between relative frequency and the probability distribution function (pdf), the connection between cumulative frequency and the cumulative distribution function (cdf), and to explore five different distributions to investigate their properties.

2 Relative Frequency & the PDF

The relative frequency of a value in a dataset is the number of times that value occurs divided by the total number of observations. Similarly, the pdf describes the likelihood of each possible outcome of a random variable. Thus, when we plot the relative frequency against the values of x, we are essentially visualizing the pdf of the distribution.

3 Cumulative Frequency & the CDF

In a similar fashion, when we measure the cumulative frequency, we are looking at the total number of observations that fall below or at a certain value of x. So, when we take the cumulative frequency and plot it against x, we are visualizing the cdf of the distribution. We can utilize the cdf to look at an entire distribution at once, rather than just individual points.

4 Analysis of Distributions A through E

The following figures illustrate the relative frequency and cumulative frequency for each of the five distributions, A through E. Each distribution is represented by two graphs: one for the relative frequency (pdf) and one for the cumulative frequency (cdf). Each distribution contains 100 data points, with each point having an associated frequency. Because of this setup, we can find important statistics such as the mean, median, mode, variance, and intervals that contain the middle 50% and 90% frequency intervals of the data. These are presented in the Results section.

To find the statistics that we are after for each distribution, we heavily rely upon the relative and cumulative frequency data. That is, for the mean, we multiply each value of x by its relative frequency and sum the results. For the median, we look for the value of x where the cumulative frequency reaches 50%. The mode is determined by identifying the value of x with the highest relative frequency. The variance is calculated by finding the average of the squared differences from the mean, weighted by the relative frequencies. Finally, to find the intervals that contain the middle 50% and 90% of the data, we look for the values of x that correspond to cumulative frequencies of 25% and 75% for the 50% interval, and 5% and 95% for the 90% interval.

4.1 Comparing Distributions

Each distribution had different corresponding frequencies for each value of x. This led to different shapes and characteristics for each distribution. For example, Distribution A had consisted only of values 1 and 0, so the most frequently appearing value was 1, hence the mode was a range of x-values associated with a 1 in the distribution. On the other hand, Distribution E had a very tight clustering of values around 4, leading to a very small variance of 0.2. This distribution also had a normal shape, with the mean, median, and mode all being equal to 4.0. In contrast, Distribution C had a wider spread of values, evidenced with a high variance of 7.8. Additionally, this distribution had a tail that was similar to Distribution B, but included a spike of values around 9.0.

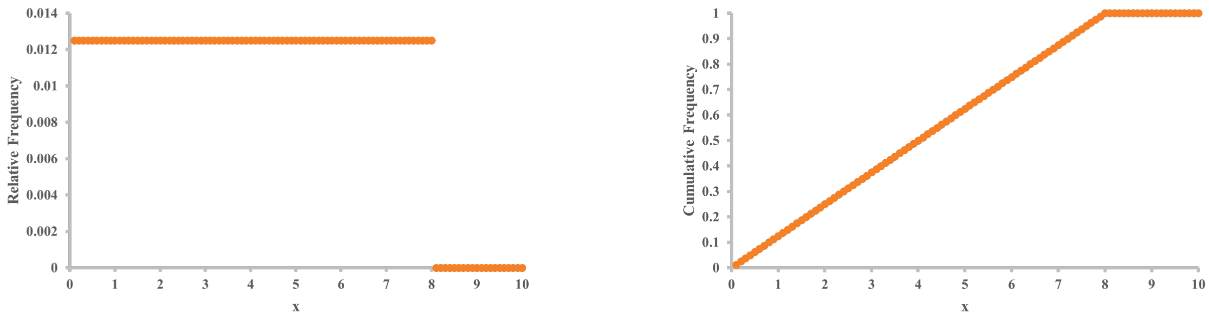
5 Results

- **Distribution A**
  - Mean: 4.1
  - Median: 3.9
  - Mode: [0.1, 8.0]
  - Variance: 5.3
  - 50% Interval: [2.0, 6.0]
  - 90% Interval: [0.4, 7.6]
- **Distribution B**
  - Mean: 4.0
  - Median: 3.5
  - Mode: 0.1
  - Variance: 7.7
  - 50% Interval: [1.6, 6.1]
  - 90% Interval: [0.3, 9.1]
- **Distribution C**
  - Mean: 4.0
  - Median: 3.4
  - Mode: 0.1
  - Variance: 7.8
  - 50% Interval: [1.5, 5.9]
  - 90% Interval: [0.3, 9.0]
- **Distribution D**
  - Mean: 4.0
  - Median: 3.9
  - Mode: 3.8
  - Variance: 4.1
  - 50% Interval: [2.4, 5.3]
  - 90% Interval: [0.8, 7.5]
- **Distribution E**
  - Mean: 4.0
  - Median: 4.0
  - Mode: 4.0
  - Variance: 0.2
  - 50% Interval: [3.7, 4.2]
  - 90% Interval: [3.3, 4.6]

6 Conclusion

Through this investigation of distributions and their properties, we have found that by calculating the relative frequency and cumulative frequency, and plot them against x, we can visualize the pdf and cdf of a distribution. This allows us to extract statistics such as the mean, median, mode, variance, and intervals that contain the middle 50% and 90% of the data. Each distribution had its own unique characteristics, which were reflected in these statistics. In some cases (such as Distribution C), the mean, median, and mode did not inform us about the spike around 9.0, which was only visible when looking at the graphs. This highlights the importance of visualizing data in addition to calculating summary statistics, as it can reveal patterns and features that may not be immediately apparent from the numbers alone.

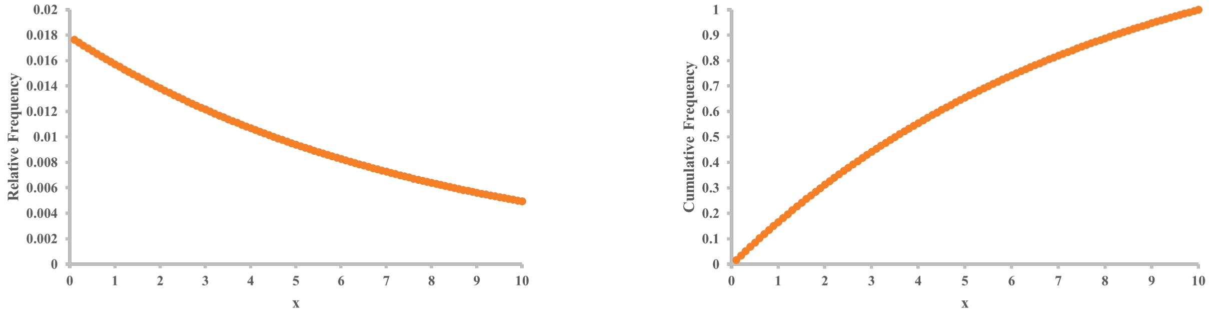
7 Graphs of Distributions A through E



(a) Relative Frequency vs. x

(b) Cumulative Frequency vs. x

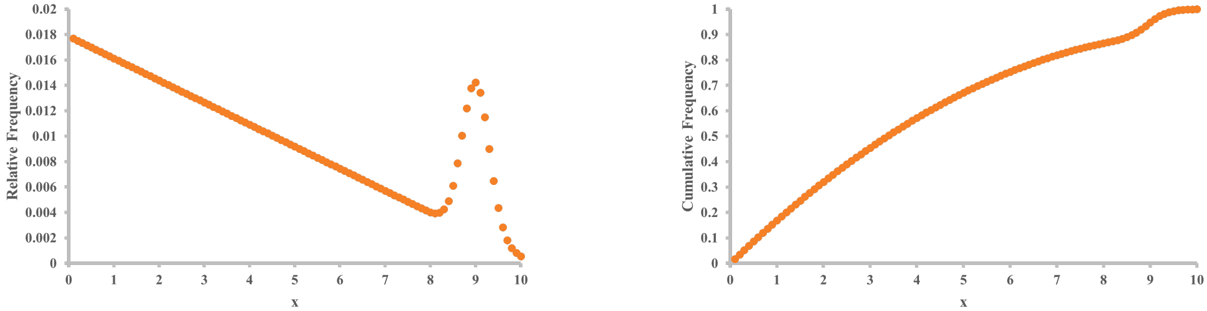
Figure 1: Distribution A



(a) Relative Frequency vs. x

(b) Cumulative Frequency vs. x

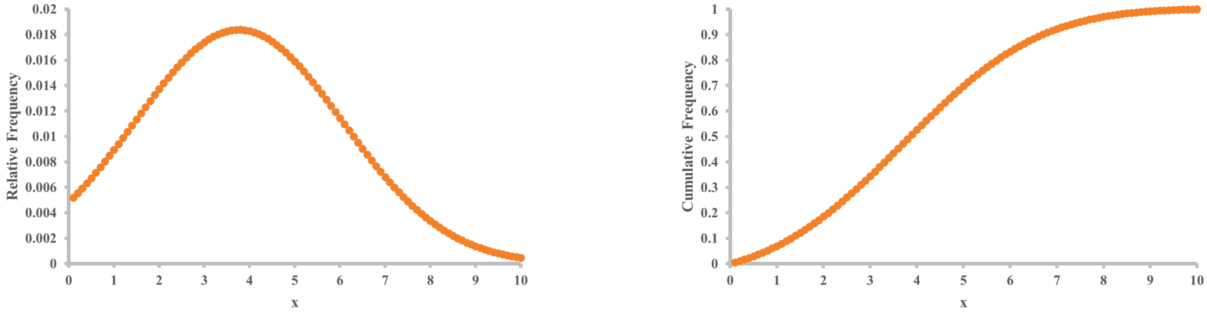
Figure 2: Distribution B



(a) Relative Frequency vs. x

(b) Cumulative Frequency vs. x

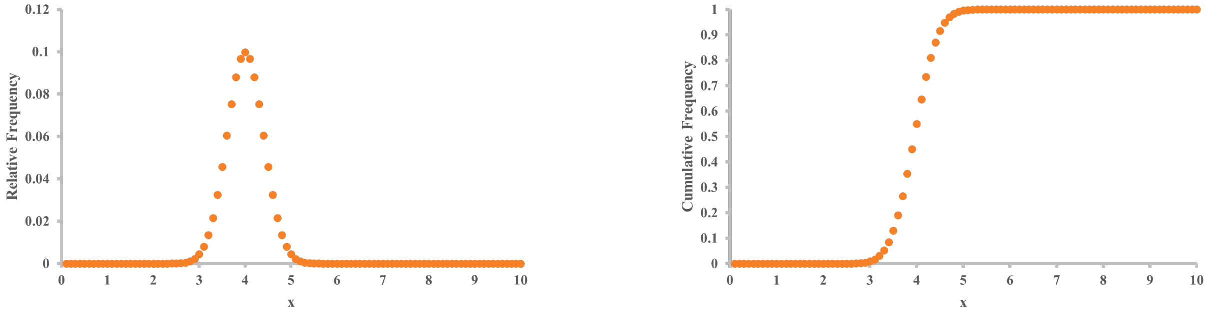
Figure 3: Distribution C



(a) Relative Frequency vs. x

(b) Cumulative Frequency vs. x

Figure 4: Distribution D



(a) Relative Frequency vs. x

(b) Cumulative Frequency vs. x

Figure 5: Distribution E