Random Variables, Densities, and CDFs

Gregory M. Shinault

Intro

Reminder

For a discrete RV X the most important object to understand the probabilities related to X is the probability mass function,

$$p_X(k) = \mathbb{P}(X = k)$$

for all $k \in \text{Ran}(X)$.

Goals for this Lecture

1. Define, use, and interpret the probability density function. This is the like the PMF, but used for continuous random variables.
2. Define and use the cumulative distribution function for any random variable.

This material corresponds to section 3.1 and 3.2 of the textbook.

Continuous Random Variables

Definition

The *probability density function* (PDF) of a continuous random variable is the function such that

$$\mathbb{P}(X \leq b) = \int_{-\infty}^{b} f_X(x)\, dx.$$

Facts

0. For $a \leq b$ and for integrable subsets $B \subset \mathbb{R}$

$$\mathbb{P}(a < X \leq b) = \int_a^b f_X(x)\, dx,$$

$$\mathbb{P}(X \in B) = \int_B f_X(x)\, dx.$$

1. A function f is a PDF if and only if $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
2. For a continuous RV X, $\mathbb{P}(X = c) = 0$ for all c.

Example

We say X is *uniformly distributed on* $[a, b]$ if it has the PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

We denote this by $X \sim \text{Unif}([a, b])$.

Let $X \sim \text{Unif}([-2, 4])$. Find $\mathbb{P}(X^2 \le 2)$.

Example

Let X be a continuous random variable with PDF

$$f_X(x) = \frac{c}{1 + x^2}$$

for all real x. Find the value of c.

Interpretation

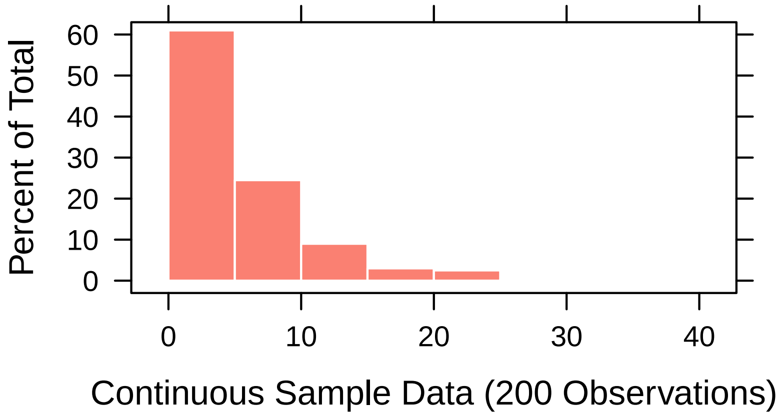
Naive Idea

We want to see how the PDF connects to actual data. To this end we look at a large sample of continuous data.

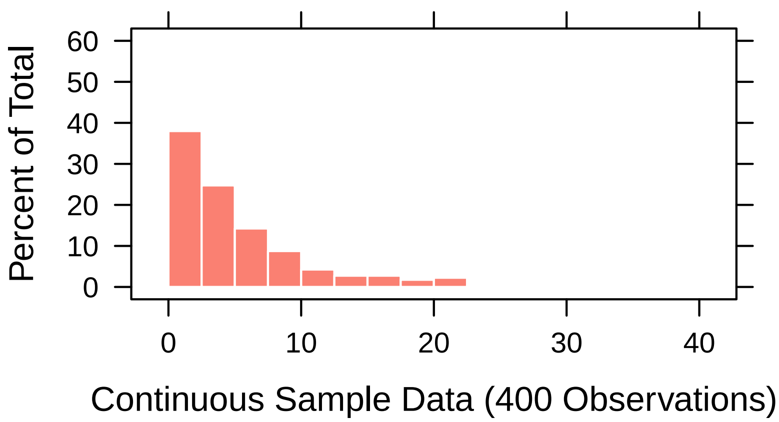
Reminder: The PMF of a distribution is analogous to the frequency bar chart for data generated from that distribution.

Expectation: The PDF of a distribution is analogous to the frequency histogram for data generated from that distribution. This turns out to be incorrect.

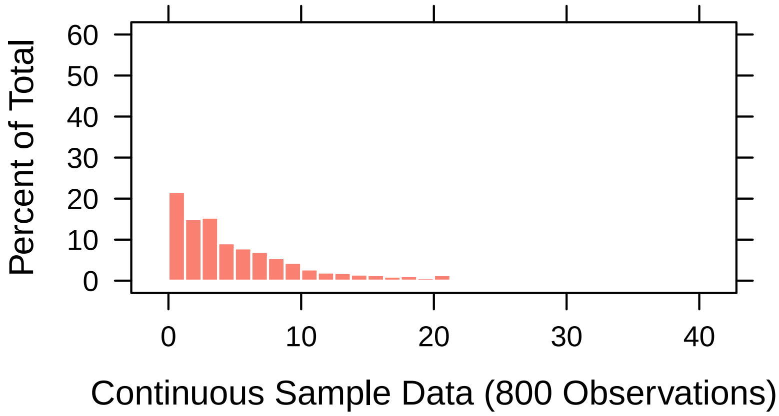
Probability Bins for 200 Samples



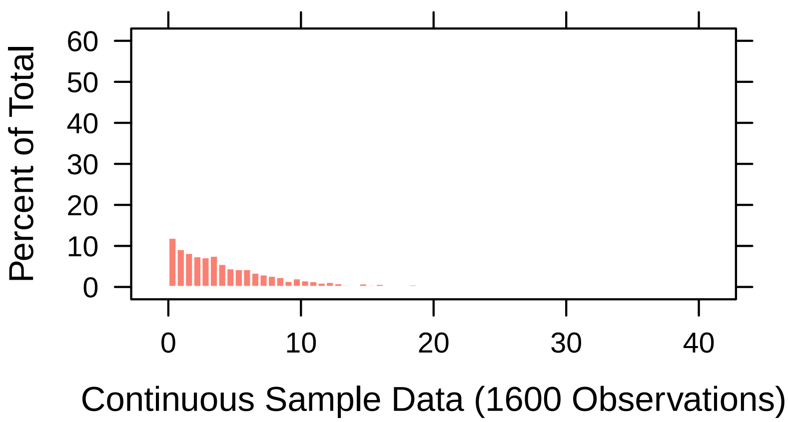
Probability Bins for 400 Samples



Probability Bins for 800 Samples



Probability goes to 0 as number of bins increases

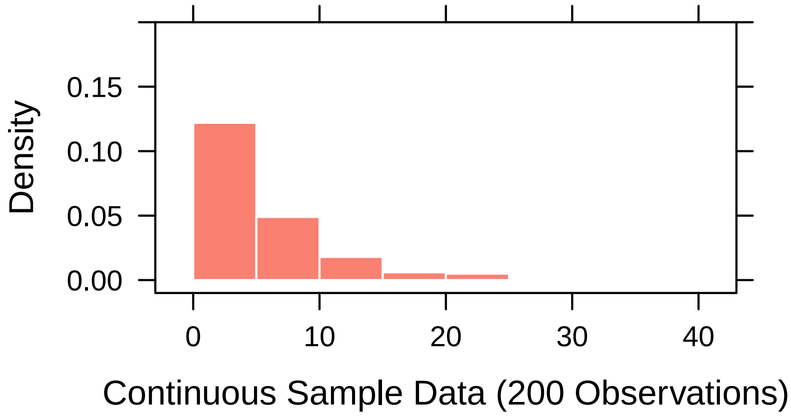


Why does this fail?

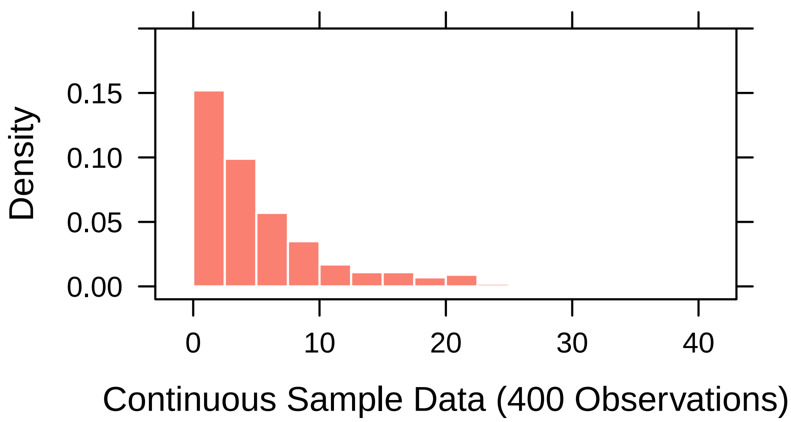
As the bins we are using to construct our frequency histogram get more narrow, the probability of a data point falling into that bin decreases to 0. So the frequency histogram cannot correspond to the PDF.

Fix: Plot the Frequency/Bin Width, rather than the frequency alone.

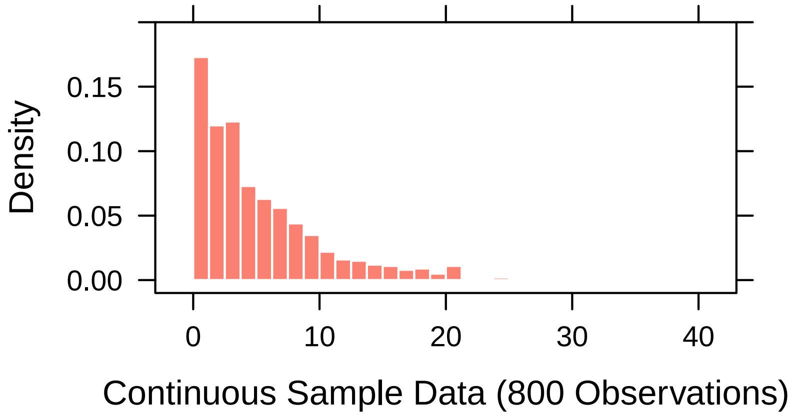
Fix: Plot (Frequency/Bin Width), Probability Density



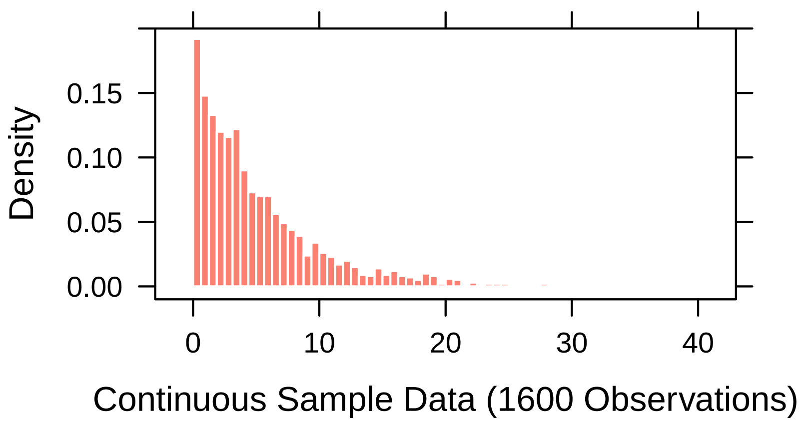
Fix: Plot (Frequency/Bin Width), Probability Density



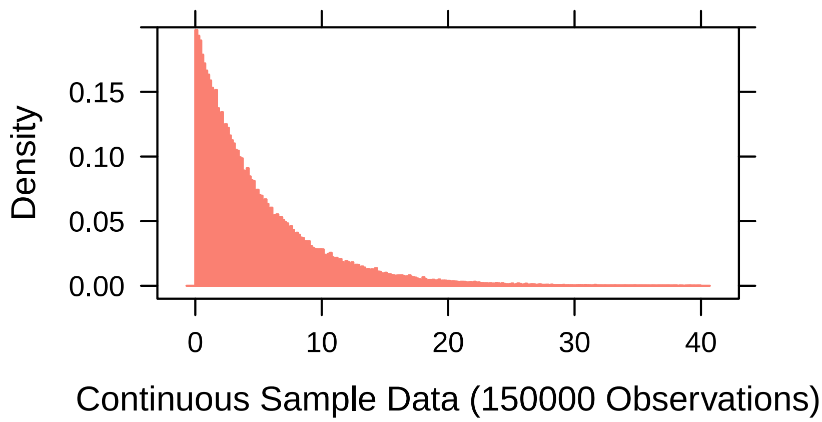
Fix: Plot (Frequency/Bin Width), Probability Density



Fix: Plot (Frequency/Bin Width), Probability Density



Fix: Plot (Frequency/Bin Width), Probability Density



Conclusion

The PDF corresponds to the frequency histogram, divided by bin width.

This is why it is a probability *density* function. The units are probability per unit of X.

Cumulative Distribution Function

Definition

The *cumulative distribution function* (CDF) of a random variable X is defined by

$$F_X(s) = \mathbb{P}(X \leq s).$$

Note the CDF is defined for all random variables: discrete, continuous, and mixed.

Relationship to PDF

If X is a continuous RV then

$$f_X(t) = \frac{d}{dt}\left[F_X(t)\right].$$

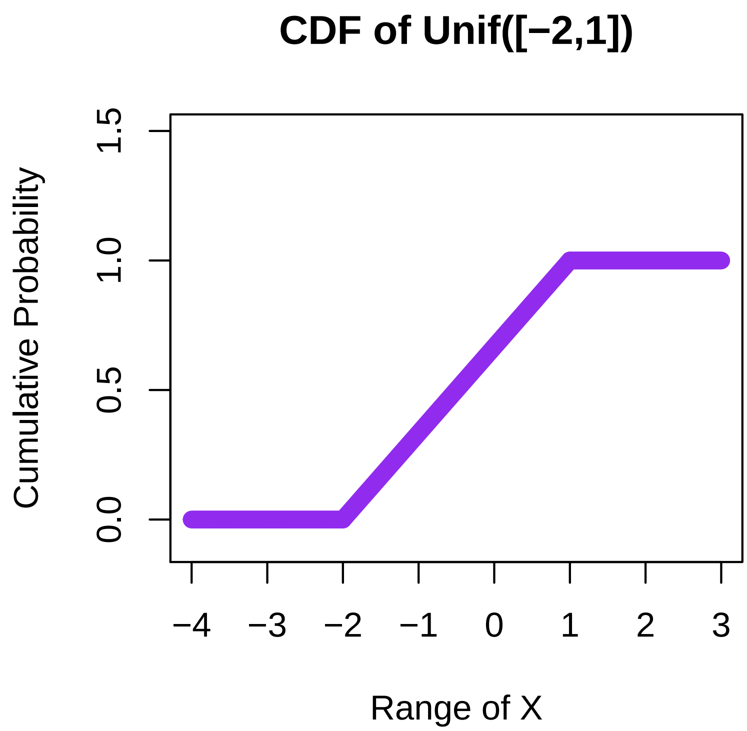
Exercise: Prove this fact.

What is the significance of this fact? If we want to find the PDF of a random variable, it is usually easiest to start by finding $P(X \leq x) = F_X(x)$. We then just differentiate to get the CDF.

Example

Find and sketch the CDF of $X \sim \text{Unif}(-2, 1)$.

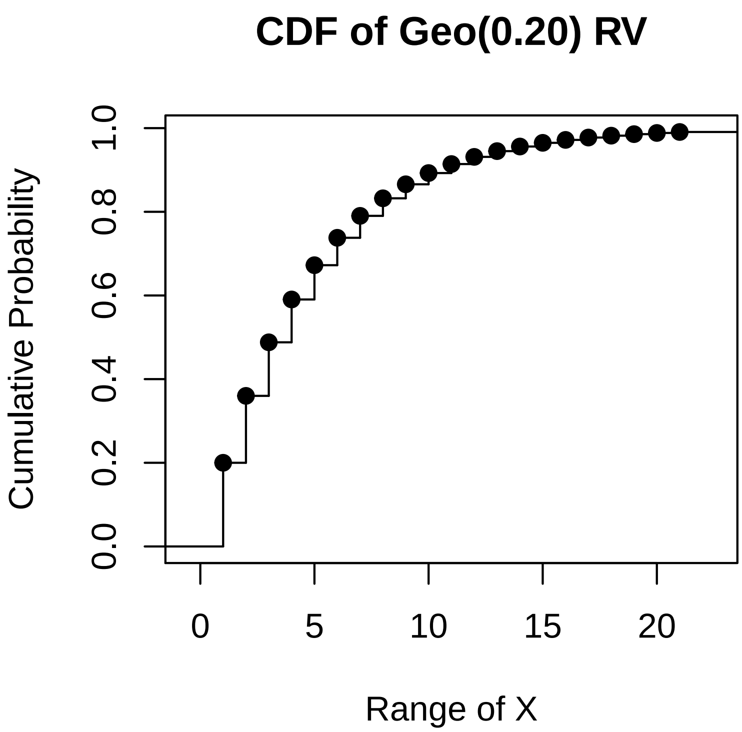
CDF Plot



Example

1. Find the CDF of $X \sim \text{Geo}(0.20)$.
2. Use the CDF to find $\mathbb{P}(4 \leq X \leq 7)$.

CDF Plot

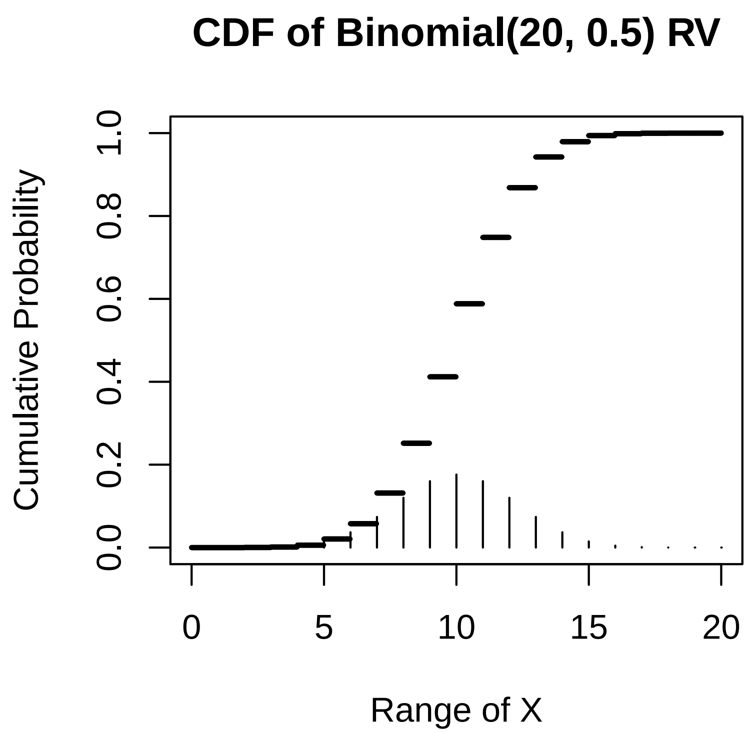


Example

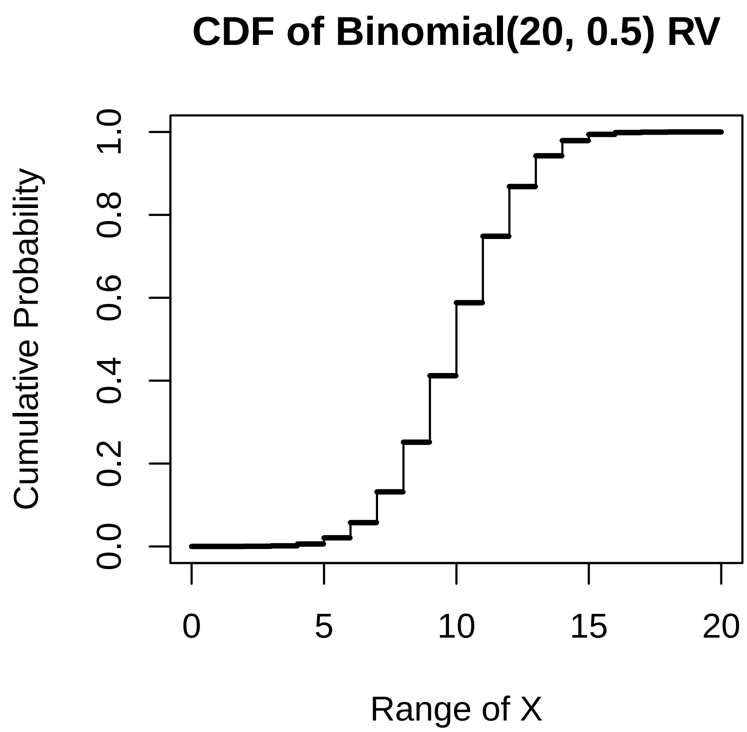
Throw a dart at a triangular board, whose corners we can place at $(0,0)$, $(0,3)$, and $(2,0)$. Let X be the x-coordinate of the point where the dart hits. Find the PDF of X.

More Graphs

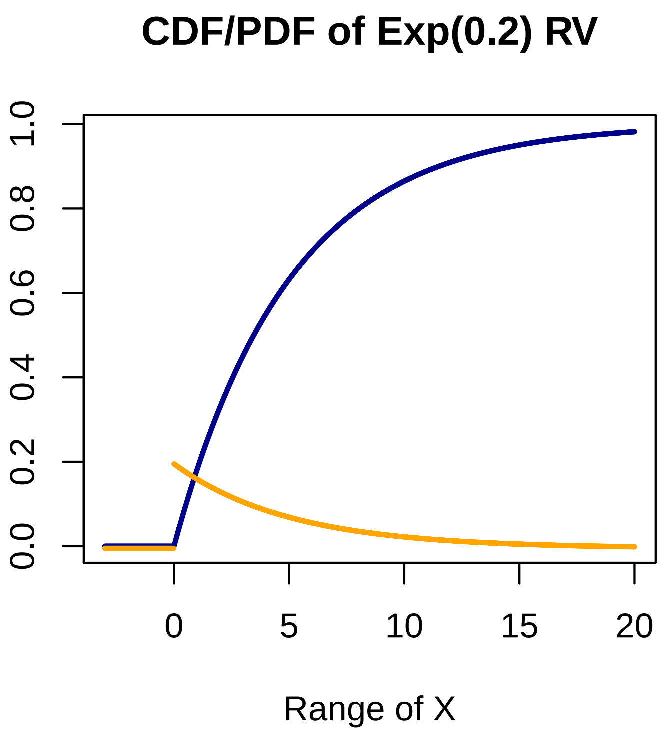
Binomial CDF and PMF



Binomial CDF connected by PMF



The CDF and PDF Plotted Together



A bit harder to interpret!

The Wrap Up

Summary

1. For continuous RVs we can use integrals of the PDF to compute probabilities.
2. The CDF is defined for all types of RVs and fully classifies its distribution.
3. For continuous RVs the CDF is often easier to use for probability computations.
4. To find the PDF of X, usually you should first find the CDF of X.

Next Step

The PMF and PDF give complete information about a random variable. The expectation and variance are quick summaries of the random variable.