

# Lab: Probability Distributions

Quantitative Methods - Winter 2022

## Introduction

In much of this course, we're working with **deterministic** models: Models for which, given a set of parameters and initial conditions, we obtain exactly the same result at each time step every time we evaluate the model.

In the real world, however, even if we manage to find good approximations for the demographic parameters describing the behavior of a natural population, it's quite unlikely that the population will behave exactly as we expect it to. The real world is more complex and "noisy" than our model - population trajectories depart from our expectations because of random variation in the number and timing of births and deaths (**demographic stochasticity**) and because random variation in the environment creates good years and bad years for growth, survival, and reproduction (**environmental stochasticity**).

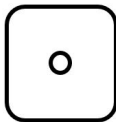
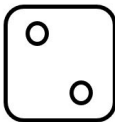
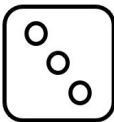
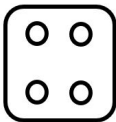
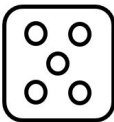
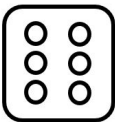
Additionally, it's often the case that we have data describing a natural biological phenomenon or population trajectory, and we'd like to use those data to find a good approximation for the parameters of a model that we think describes that system's behavior. Because, again, the real world is noisy, it's rarely the case that we can just "pick" parameters that produce a perfect fit of our model's expectations to the data.

Both of these scenarios necessitate the use of **probabilistic** (or "**stochastic**") models. To build stochastic population models, or to fit real data to deterministic models, we need to use what are called **probability distributions** - functions which map possible values of a **random variable** to probability (or probability "density").

## Discrete Random Variables

A **random variable** is any value (such as a demographic parameter or the response variable in a statistical model) which is taken to be the outcome of a random process.

Take, for example, the roll of a six-sided die:

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$
$X$						
<hr/>						
$P(X = x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

In this case, the random variable  $X$ , representing a roll of the die, has 6 potential outcomes  $x \in \{1, 2, 3, 4, 5, 6\}$ . Each outcome  $x$  corresponds to a probability  $P(X = x)$  that our die roll  $X$  equals that outcome, which (if we assume the die to be "fair") is  $P(X = x) = 1/6 \approx 0.167$  for  $x \in \{1, 2, 3, 4, 5, 6\}$ .

Because  $X$  has a countable number of potential outcomes, we call it a **discrete** random variable, as distinguished from **continuous** random variables that can take on any real number (or any real number

within an interval), described below.

## Probability Mass Functions

The probability distribution of a discrete random variable is described by a **probability mass function**  $f_X(x)$ - this is a function which maps an outcome  $x$  to a Probability value  $P(X = x)$ . The PMF for our six-sided die is very simple:

$$f_X(x) = P(X = x) = 1/6 ; x \in \{1, 2, \dots, 6\}$$

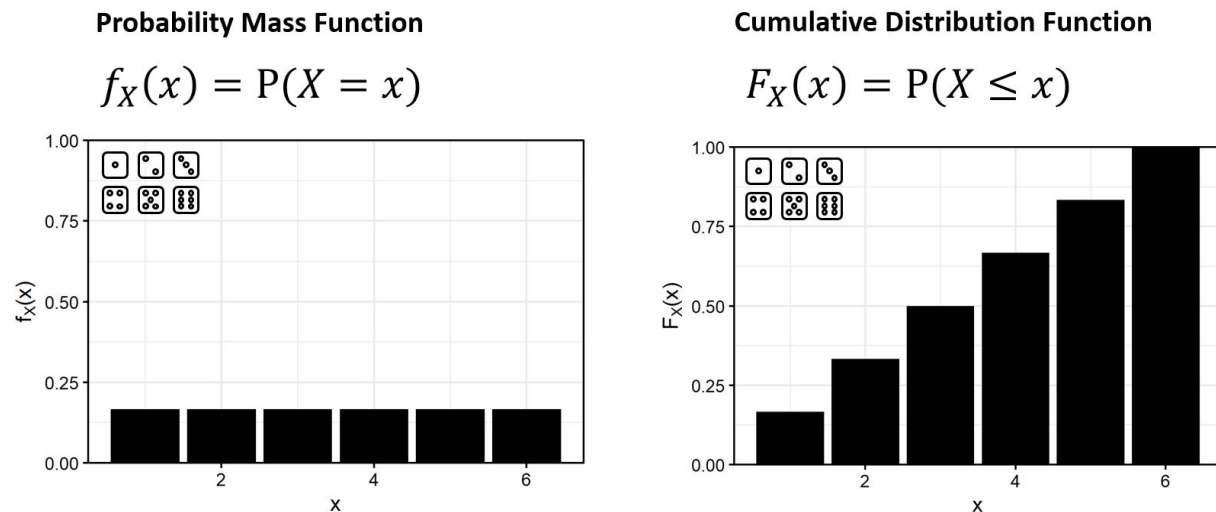
This distribution is also called a **discrete uniform** distribution (in this case, for the interval  $[1,6]$ , but other intervals are possible of course). Other common discrete probability distributions used in ecology are the **Bernoulli**, **Binomial**, **Poisson**, **Negative Binomial**, and **Multinomial** distributions.

## Cumulative Distribution Functions

The **cumulative distribution function**  $F_X(x)$  specifies the probability that the given outcome  $x$  of the random variable  $X$  will be *less than or equal to*  $x$ . For a six-sided die, the CDF is also very simple:

$$F_X(x) = P(X \leq x) = \sum_1^x 1/6 ; x \in \{1, 2, \dots, 6\}$$

Here is the PMF and CDF for our six-sided die plotted side by side:

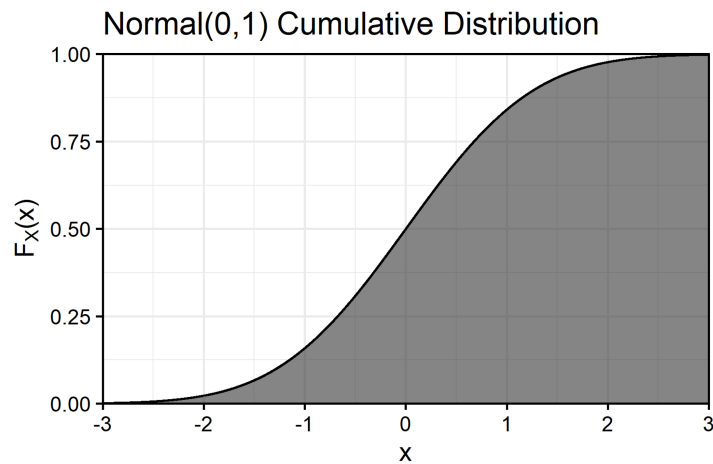


## Continuous Random Variables

**Continuous** random variables are conceptually a little harder than discrete random variables - they do not have a countable number of outcomes - meaning they have *infinitely many* outcome values, and that makes the probability of any one outcome  $P(X = x) \approx 0$ .

This means that the PMF for a continuous random variable, if we wrote one, would always yield 0, which isn't very useful. However, we can still represent continuous random variables using the CDF! For example, even though we might never register a temperature of exactly  $60.00^\circ$  on our thermometer, if the median daily temperature where I am is  $60^\circ$ , then the probability that the temperature on any given day is less than or equal to 60 is roughly 0.5.

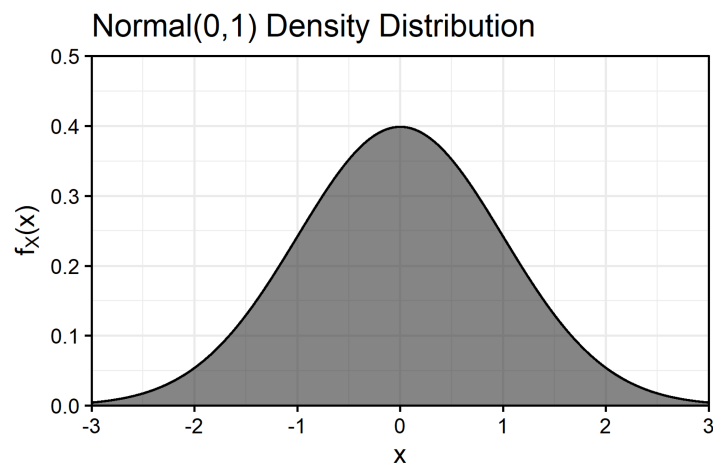
Thus, we usually define continuous probability distributions by first describing the CDF. Shown below, for example, is the CDF for the standard normal distribution (a normal distribution with a mean of 0 and a standard deviation of 1):



Thus, for the standard normal distribution, we can see that the probability we obtain a value  $x$  which is less than or equal to 0 is  $F_X(x) = P(X \leq 0) = 0.5$ , the probability we obtain a value less than 1 is  $F_X(x) = P(X \leq 1) \approx 0.84$ , and so on.

## Probability Density Functions

When most people picture the normal distribution, however, they picture a “bell curve” - this is, in fact, the **probability density function** (also denoted  $f_X(x)$ , as for the PMF) of the normal distribution, and it’s obtained by taking the derivative of the CDF with respect to  $x$ :



The PDF is **not** a PMF - it does not return the probability of an outcome  $x$ , it returns the rate of change of the CDF at  $x$  - but luckily, there are many applications for which we can use it as if it were a PMF and not have to worry about the distinction. And, we can read it as we’d expect to read a PMF - we can tell from the density function that values around 0 are more likely than values around 2, for example.

Common continuous probability distributions used in ecology are the **Normal** (or “Gaussian”), **Lognormal**, **Gamma**, **Beta**, **Exponential**, and **Uniform** distributions.

## R Functions

R has a number of probability distribution functions built-in (to see them, type `?Distributions`. You can use the help menu to get an overview of probability distribution functions in R, for example the Poisson (`?dpois`), or Normal (`?rnorm`). In brief:

`d_____`: probability (density/mass) function. This function returns the probability or probability density of a value, given the probability distribution function and your selected parameter values.

`p_____`: cumulative distribution function. This function returns the probability of observing a value, or anything smaller, given the probability distribution function. Useful to estimate the probability of observing a range of data.

`q_____`: quantile function. This function returns the value corresponding to a given quantile of a given distribution.

`r_____`: generates random samples. Used to generate data conditional on a probability distribution and selected parameter values.

## Lab Exercises

Open `distributions-lab.Rproj` to start a new RStudio session, and open `app.R` within that session. In the upper right-hand corner of the RStudio source viewer, click “Run App,” which should launch an interactive dashboard app in a new window. We’ll use this app for today’s lab.

### 1. Choosing Probability Distributions

Choose one or more appropriate distributions for the types of data shown below and justify your decision(s).

- The number of seals on a haul-out beach in the gulf of Alaska.
- Presence or absence of an invasive species in forest patches.
- The absolute distance that seastar larvae will settle from the location of spawning (assume it cannot be exactly 0).
- The number of abalone at time  $t$  surviving until time  $t + 1$ .
- The proportion of reef sharks on a reef captured by a camera trap.
- The number of prey (from an initial number  $n$ ) eaten by a predator during an experiment in aquaria.
- The body length of a cohort of adult whale sharks.

### 2. Exploring Distributions & Parameters

- You’re modeling a population of gophers at Hopkins Marine Station, and you want to incorporate predation by the local red-shouldered hawk. On average, the hawk eats 2 gophers per month, but it doesn’t catch the same number of gophers each month. Choose a distribution from which to simulate monthly hawk predation (justify your answer). What values should you choose for this distribution’s parameter(s), and why? What is the probability that your simulated hawk will eat 4 gophers in a given month (you can eyeball this from the plot)?
- At each of 15 sites, you’ve set up 10 enclosures where you’ve placed juvenile kelp, and after checking on these cages 3 months later you’d like to fit a model which relates the number of juvenile kelp surviving at each site to temperature at each site. Why is a binomial distribution with  $n = 10$  the best choice for this data? Vary the other parameter ( $p$ ), and observe what happens to the shape of the distribution. What is the most likely number of surviving kelp when we choose  $p = 0.1$ ? Choose a value of  $p$  for which the distribution’s shape is most symmetrical.

- c. You're simulating a spatial model for which a proportion of the available habitat will be suitable (and the other proportion unsuitable) for abalone, and you've chosen a Beta distribution to do this. Choose three pairs of values for  $\alpha$  and  $\beta$  that give you three Beta distributions which are all symmetrical around 0.5 (so that, on average, your simulations will have about 50% suitable habitat), each with the following properties:

- One in which simulations with a suitable habitat of around 0.5 will be the most likely.
- One in which some simulations will have lots of suitable habitat, and some will have very little suitable habitat, but few will have around 50% suitable habitat.
- One in which all possible percentages of suitable habitat are about equally likely, except for the extreme values.

Taking a look at the equation for the mean of the Beta distribution will probably make this pretty straightforward. What are the parameter combinations that you chose? Take a screenshot of these distributions (or draw them) and include them in your answers.

### 3. Using R's distribution functions

Answer these questions using the R console (if you're running the distribution dashboard locally, you'll have to close it to do so). Provide your code for all answers.

- Find the mean, variance, and 95% quantiles of 1000 random draws from a Poisson distribution with  $\lambda = 33$ .
- What is the probability  $P(X \leq 6)$  that a random draw from a Poisson distribution with  $\lambda = 4$  will be less than or equal to 6?
- What is the probability  $P(X = 3)$  of obtaining a value of 3 from a Binomial distribution with  $p = 0.3$  and  $n = 5$ ?
- What is the probability  $P(-1.5 \leq X \leq 1.5)$  that a value drawn from a standard normal distribution will be between -1.5 and 1.5 (it may help to approach this visually)?
- Find the value  $x$  that satisfies to  $P(X \leq x) = 0.8$ , if  $X$  is a Gamma random variable with  $k = 2$  and  $\theta = 1$ .

### 4. Samples and their means

- Read the intro text below the Normal distribution in the distribution dashboard. What is the difference between the distribution of the sample and the "sampling distribution?" How is the sample size distinguished from the number of samples?
- Drag the sliders for the sample size and the number of samples all the way to the right. Visit *all* of the distributions and spend a little time there, messing with their parameters. In general, what do you notice about the distribution of their sample means? Are there any distributions (/parameter values) for which this pattern does not hold? If so, what seems to be the reason?

### Attribution

Some questions used in this lab are adapted from one written by Tom Hobbs, Mary Collins, and Christian Che-Castaldo, as part of their Bayesian short course at SESYNC, and from material taught by Robin Elahi as part of his Bayesian course at Stanford's Hopkins Marine Station.