# Chapter 3: Common Families of Distributions (Sections 3.1-3.3)

## 3.1 Introduction

Statistical distributions are used to model populations. Therefore, we usually deal with families of distributions, where each family is indexed by one or more parameters that allow us to vary certain characteristics of the distribution, such as shape and/or spread, while staying with one functional form.

There are many common discrete and continuous distributions, and we’ll discuss a few of them, along with their interrelationships and common applications, in the following sections.

## 3.2 Discrete Distributions

A random variable  is said to have a discrete distribution if its support, , is countable. In most discrete distributions, the random variable has integer-valued outcomes. In this section, we’ll discuss the following discrete distributions: 1) Discrete Uniform; 2) Binomial; 3) Negative Binomial; 4) Geometric; 5) Hypergeometric; and 6) Poisson.

**Discrete Uniform Distribution**

|  |
| --- |
| **When Used?** This distribution puts equal mass on each of the outcomes  That is, each of the  outcomes has equal probability of being observed. |
| **PMF:** A random variable  has a **discrete uniform**  distribution if    where  is a specified integer. |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Example:** The German Tank Problem is a well-known example that uses a discrete uniform distribution to model the serial number of a tank captured during World War II in order to estimate  the total number of tanks produced. |
| **Fun Fact**: This distribution can be generalized so that the sample space is any range of integers,  with pmf |

**Bernoulli Distribution**

|  |
| --- |
| **When Used?** When an experiment has only two possible outcomes: “success” or “failure.” A Bernoulli random variable  is defined in the following manner:    where  and |
| **PMF:** A random variable  has a **Bernoulli** distribution if    Note: The notation  implies the value of the function is dependent on the value  A value such as  required for the calculation of any probability, is known as a **distributional parameter*.*** |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Examples:** Outcome of one coin flip; Whether a randomly selected student is infected with a disease; etc. |
| **Fun Facts**: A Bernoulli trial is named for James Bernoulli, one of the founding fathers of probability theory.  A Bernoulli distribution is a special case of the Binomial distribution where |
| **In R:** To find P(X = x) use:dbinom(x, n=1, p)  To find P(X ≤ x) use: pbinom(x, n=1, p)  To find smallest x\* such that P(X ≤ x\*) ≥ c, use: qbinom(c, n=1, p)  To simulate *m* random draws, use: rbinom(m, n=1, p) |

**Binomial Distribution**

|  |
| --- |
| **When Used?** When a random variable is a result of a sequence of independent Bernoulli trials, and we are interested in the number of successes observed during a fixed number of trials. This requires:   1. A fixed number of  identical trials. 2. Each trial results in one of two possible outcomes: “success” or “failure.” 3. The probability of success (denoted by  remains the same for each trial. The probability of failure is denoted as 4. The trials are independent. 5. The random variable,  is defined as the number of successes observed during the  trials. |
| **PMF:** A random variable  has a **binomial** distribution if    Note: The notation  implies the value of the function is dependent on the values of  and The values  and  are the **parameters** of the distribution. |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Fun Fact**: Recursive Relationship: |
| **Examples:** Number of heads observed in  coin flips; Number of  randomly selected voters who voted for a particular candidate; Number of  randomly selected students who are infected with a disease; etc. |
| **In R:** To find P(X = x) use: dbinom(x, n, p)  To find P(X ≤ x) use: pbinom(x, n, p)  To find smallest x\* such that P(X ≤ x\*) ≥ c, use: qbinom(c, n, p)  To simulate *m* random draws, use: rbinom(m, n, p) |

**Geometric Distribution**

|  |
| --- |
| **When Used?** When a random variable is a result of a sequence of independent Bernoulli trials, and we are interested in the trial on which the first success occurs. This requires:   1. Identical and independent trials. 2. Each trial results in one of two possible outcomes: “success” or “failure.” 3. The probability of success (denoted by  remains the same for each trial. The probability of failure is denoted as 4. The random variable,  is defined as the trial on which the first success occurs. |
| **PMF:** A random variable  has a **Geometric** distribution if    Note: The notation  implies the value of the function is dependent on the value  The value  is the **parameter** of the distribution. |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Examples:** Number of coin flips to observe first head flipped; Number of students randomly chosen until one infected with a disease is found; etc. |
| **Fun Facts**: The geometric distribution is a special case of the negative binomial distribution where  The geometric distribution has what’s known as the “memoryless property.” (see pg. 97)  The geometric distribution can also be re-parameterized to represent the number of failures before the  success  occurs. |
| **In R: Geom(p) – # Trials Geom\*(p) – # Failures**  To find P(X = x): dgeom(x-1, p)dgeom(x, p)  To find P(X ≤ x): pgeom(x-1, p) pgeom(x, p)  To find smallest x\* such that P(X ≤ x\*) ≥ c: qgeom(c, p) + 1qgeom(c, p)  To simulate *m* random draws: rgeom(m, p) + 1rgeom(m, p*)* |

**Example:** Suppose a baseball player has a batting average of .300. Assuming times at bat are independent, let the random variable ****represent the number of times at bat since the season started when he gets his first hit. If he has had no hits in his first 20 at bats, what is the probability it takes him more than 25 at bats to get his first hit of the season?

**Negative Binomial Distribution**

|  |
| --- |
| **When Used?** When a random variable is a result of a sequence of independent Bernoulli trials, and we are interested in the trial on which the  success occurs  This requires:   1. Identical and independent trials. 2. Each trial results in one of two possible outcomes: “success” or “failure.” 3. The probability of success (denoted by  remains the same for each trial. The probability of failure is denoted as 4. The random variable,  is defined as the trial on which the  success occurs |
| **PMF:** A random variable  has a **Negative Binomial** distribution if    Note: The notation  implies the value of the function is dependent on the values  and  The values  and  are the **parameters** of the distribution. |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Examples:** Number of coin flips to observe 7 heads; Number of students randomly chosen until 100 infected with a disease are found; etc. |
| **Fun Facts**: The negative binomial distribution can be re-parameterized to represent the number of failures before the  success occurs |
| **In R: NBin(r,p) – # Trials NBin\*(r,p) – # Failures**  To find P(X = x): dnbinom(x-r, r, p)dnbinom(x, r, p)  To find P(X ≤ x): pnbinom(x-r, r, p) pnbinom(x, r, p)  To find smallest x\* such that P(X ≤ x\*) ≥ c: qnbinom(c, r, p) + r qnbinom(c, r, p)  To simulate *m* random draws: rnbinom(m, r, p) + r rnbinom(m, r, p) |

What are the differences and similarities between Binomial and Negative Binomial random variables?

**Hypergeometric Distribution**

|  |
| --- |
| **When Used?** When a random variable represents the number of items with a certain characteristic observed in a sample of size  from a finite population where there are  total items with that characteristic and  total items without the characteristic The following conditions are also required:   1. Population contains a finite number of elements,  and the sample size,  is large relative to the population size. 2. Each item in the population possesses one of two characteristics. 3. The random variable,  is defined as the number of items with a certain characteristic observed in a sample of size |
| **PMF:** A random variable  has a **Hypergeometric** distribution if    Note: The notation  implies the value of the function is dependent on the values   and  The values   and  are the **parameters** of the distribution***.*** |
| **MGF:** Not useful |
| **Mean:** |
| **Variance:** |
| **Example:** Number of defective machine parts observed in 4 parts randomly sampled from a shipment of 25 total parts; etc. |
| **Fun Facts**: As  and  with  the Hypergeometric distribution approaches a Binomial distribution. |
| **In R:** To find P(X = x) use: dhyper(x, M, N-M, n)  To find P(X ≤ x) use: phyper(x, M, N-M, n)  To find smallest x\* such that P(X ≤ x\*) ≥ c, use: qhyper(c, M, N-M, n)  To simulate *m* random draws, use: rhyper(m, M, N-M, n) |

**Example:** A shipment of 50 refurbished smartphones were sent to a Bozeman distributor, and a family purchases six of them. Suppose 15 of the refurbished phones are still malfunctioning. What is the probability the family received at least one malfunctioning phone?

**Poisson Distribution**

|  |
| --- |
| **When Used?** When a random variable represents a number of occurrences over a certain amount of time or space. |
| **PMF:** A random variable  has a **Poisson** distribution if    Note: The notation  implies the value of the function is dependent on the value  The value is the **parameter** of the distribution***.*** |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Examples:** The number of accidents at an intersection in a week; the number of hits to a website each minute; the number of plants of a particular species found in a 1-*m*2 area; etc. |
| **Fun Facts**: Recursive Relationship:  As  and  with  the Binomial distribution approaches a Poisson distribution.  If  is a  random variable and  is a  random variable, where  and  are independent, then  is a  random variable |
| **In R:** To find P(X = x) use: dpois(x, *λ*)  To find P(X ≤ x) use: ppois(x, *λ*)  To find smallest x\* such that P(X ≤ x\*) ≥ c, use: qpois(c, *λ*)  To simulate *m* random draws, use: rpois(m, *λ*) |

**Example:** A certain type of tree has seedlings randomly dispersed in a large area, with the mean density of seedlings being approximately five per square yard. Let the random variable ****represent the number of such seedlings in 0.25 square yards. What is the probability there are at least four seedlings in 0.25 square yards?

## 3.3 Continuous Distributions

In this section, we’ll discuss some common families of continuous distributions: 1) Uniform; 2) Gamma; 3) Normal; 4) Beta; 5) Cauchy; 6) Lognormal; and 7) Double Exponential. This list of continuous distributions is by no means exhaustive.

**Uniform Distribution**

|  |
| --- |
| **When Used?** When a random variable takes on any value between two limits  and  with constant probability. |
| **PDF:** A random variable  has a **Uniform** distribution if    Note: The values  and  are the **parameters** of the distribution. The parameter  represents the minimum value the random variable can assume with non-zero probability, and  represents the maximum. |
| **MGF:** and |
| **Mean:** |
| **Variance:** |
| **Fun Fact:** The Uniform distribution is often used for noninformative priors in Bayesian Statistics. |
| **In R:** To get the density, use: dunif(x, a, b)  To find P(X ≤ x) use: punif(x, a, b)  To find x\* such that P(X ≤ x\*) = c, use: qunif(c, a, b)  To simulate *m* random draws, use: runif(m, a, b) |

**Gamma Distribution**

|  |
| --- |
| **When Used?** This is a large, flexible family of distributions with many uses, such as when a random variable describes the time between events or the time to an event occurring, such as an equipment failure (reliability analysis) or death (survival analysis). |
| **PDF:** A random variable  has a **Gamma** distribution if    where  For  Note: The values  and are the **parameters** of the distribution***.*** The  parameter is known as the shape parameter, because it most influences the peakedness of the distribution, and the  parameter is called the scale parameter, because most of its influence is on the spread of the distribution. For any integer   Also, |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Fun Facts:** If Gamma where  is an integer, then for any  where Poisson |
| **Special Cases:**  **Chi-Squared Distribution**  The chi-squared distribution is a special case of the gamma distribution where  and  where  is a positive integer (a.k.a. degrees of freedom). The chi-squared distribution plays an important role in statistical inference, especially when sampling from a normal distribution (see Chapter 5).  A random variable  has a  distribution if          **Exponential Distribution**  The exponential distribution is a special case of the gamma distribution where and it is often used to model lifetimes. The exponential distribution has the “memoryless” property.  A random variable  has an **Exponential** distribution if |
| **In R: Gamma** **Exp Chi-squared**  To get the density, usedgamma(x, *, 1/β*) dexp(x, *1/β*) dchisq(x, p)  To find P(X ≤ x) use pgamma(x, *, 1/β*) pexp(x, *1/β*) pchisq(x, p)  To find x\* such that P(X ≤ x\*) = c, use qgamma(c, *, 1/β*) qexp(c, *1/β*) qchisq(c,p)  To simulate *m* random draws, use rgamma(m,*, 1/β*) rexp(m, *1/β*) rchisq(m, p)  Note: R uses the rate parameterinstead of the scale parameter  for the gamma and exponential distribution functions. These merely make use of another parameterization of these distributions, denoted with the following notation:  Gamma\* or  Exp\* |
| **Derived Distributions:**  **Weibull Distribution**  If  Exponential then  has a **Weibull** distribution. The Weibull distribution is important in modeling failure time data, as well as fatigue and breaking strength of materials. It is also useful in survival analysis for modeling hazard functions, which give the probability that an object survives a little past time  given that the object survives to time  A random variable  has a **Weibull** distribution if    **Others** (see Exercise 3.24)   * If  Exponential then  has a **Rayleigh** distribution. * If  Gamma then  has an **Inverted Gamma** distribution. * If  Gamma  then  has a **Maxwell** distribution. * If  Exponential then  has a **Gumbel** distribution. |

**Normal Distribution**

|  |
| --- |
| **When Used?** Often! The normal distribution (also called the Gaussian distribution) is the most widely used continuous probability distribution, mainly because it is tractable analytically, it follows the familiar bell shape which fits with a lot of population models, and the central limit theorem says that, with a large enough sample, the normal distribution can be used to approximate a large variety of other distributions (e.g., Normal approximation to the Binomial). |
| **PDF:** A random variable  has a **Normal** distribution if    Note: The values  and are the **parameters** of the distribution***.*** |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Fun Facts:** The normal distribution was published by de Moivre in 1733!  The chi-squared, *t* and *F* distributions can be derived from the normal distribution (see Chapter 5).  The normal distribution follows the 68-95-99.7 Rule:    For large  and not extreme  the distribution of Binomial can be approximated by a Normal distribution. This approximation can be improved by using a “continuity correction” (see pg. 105). |
| **Special Cases:**  **Standard Normal Distribution**  The standard normal distribution is a special case of the normal distribution, where  and  If  then  A random variable  has a **Normal** distribution if          All normal probabilities may be calculated in terms of the standard normal. |
| **In R:** To get the density, use: dnorm(x, *µ, σ*)  To find P(X ≤ x) use: pnorm(x, *µ, σ*)  To find x\* such that P(X ≤ x\*) = c, use: qnorm(c, *µ, σ*)  To simulate *m* random draws, use: rnorm(m, *µ, σ*) |

**Beta Distribution**

|  |
| --- |
| **When Used?** When a random variable is defined over the interval  typically the beta distribution is used to model proportions, which naturally lie between 0 and 1. |
| **PDF:** A random variable  has a **Beta** distribution if    Note: The values  and are the **parameters** of the distribution***.*** Often, the normalizing constant is expressed as a function of the beta function |
| **MGF:** Not useful |
| **Mean:** |
| **Variance:** |
| **Fun Fact:** If  the beta distribution is a Uniform distribution. |
| **In R:** To get the density, use:  *dbeta(x,* *, β*)  To find P(X ≤ x) use: *pbeta(x,* *, β*)  To find x\* such that P(X ≤ x\*) = c, use: *qbeta(c,* *, β*)  To simulate *m* random draws, use: *rbeta(m,* *, β*) |

How can the Beta distribution be applied to a random variable defined over the interval  where  and 

**Cauchy Distribution**

|  |
| --- |
| **When Used?** Surprisingly more often than one would expect. Usually this distribution arises when comparing ratios of standard normal random variables. |
| **PDF:** A random variable  has a **Cauchy** distribution if    Note: The value  is the **parameter** of the distribution***.*** This parameter is the median of the distribution. |
| **MGF:** Does not exist |
| **Mean:** Does not exist |
| **Variance:** Does not exist |
| **Fun Facts:** The Cauchy distribution is a symmetric, bell-shaped distribution.  No moments of the Cauchy distribution exist.  The ratio of two standard normal random variables has a Cauchy distribution. |
| **In R:** To get the density, use: dcauchy(x, *θ*)  To find P(X ≤ x) use: pcauchy(x, *θ*)  To find x\* such that P(X ≤ x\*) = c, use: qcauchy(c, *θ*)  To simulate *m* random draws, use: rcauchy(m, *θ*) |
| **Another Version:**  The following is a two-parameter version of a Cauchy distribution:    Similar to the one-parameter version, the moments and the mgf do not exist for this distribution. |

**Lognormal Distribution**

|  |
| --- |
| **When Used?** When the variable of interest is skewed to the right, and the natural log (log transformation) of the data is normally distributed, allowing the use of normal-theory statistical procedures. Examples include income, movement data, and electrical measurements. |
| **PDF:** If  the  follows a **Lognormal** distribution:    Note: The values  and are the **parameters** of the distribution***.*** |
| **MGF:** Does not exist |
| **Mean:** |
| **Variance:** |
| **In R:** To get the density, use: dlnorm(x,*,*)  To find P(X ≤ x) use: plnorm (x,*,*)  To find x\* such that P(X ≤ x\*) = c, use: qlnorm (c,*,*)  To simulate *m* random draws, use: rlnorm (m,*,*) |

**Double Exponential / Laplace Distribution**

|  |
| --- |
| **When Used?** Great question! We should investigate this more. |
| **PDF:** A random variable  has a **Double Exponential** distribution if    Note: The values  and are the **parameters** of the distribution***.*** |
| **MGF:** |
| **Mean:** |
| **Variance:** |
| **Fun Facts:** The double exponential distribution is also called the Laplace distribution.  The double exponential distribution is formed by reflecting the exponential distribution around its mean. |