# Mathematical Statistics

Gregg Waterman
Oregon Institute of Technology

©2016 Gregg Waterman

This work is licensed under the Creative Commons Attribution 4.0 International license. The essence of the license is that

**You are free to:**

- **Share** copy and redistribute the material in any medium or format
- **Adapt** remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

**Under the following terms:**

- **Attribution** You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

**Notices:**

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to the web page below.

To view a full copy of this license, visit https://creativecommons.org/licenses/by/4.0/legalcode.

# Contents

# 4 More Distributions

**Performance Criteria:**

4. (a) Apply the hypergeometric distribution to solve applied problem.
   (b) Apply the negative binomial distribution to solve applied problem.
   (c) Apply the Poisson distribution to solve applied problem.
   (d) Approximate a binomial probability with the Poisson distribution, when appropriate.
   (e) Apply the exponential distribution to solve applied problem.
   (f) Apply the gamma distribution to solve applied problem.
   (g) Choose and apply the appropriate distribution(s) to solve an applied problem.

In the last chapter we worked with the binomial and normal distributions, the standard examples of discrete (the binomial) and continuous (the normal) distributions. In this chapter we will introduce and work with several other commonly encountered distributions.

## 4.1 Hypergeometric Distribution

**Performance Criteria:**

4. (a) Apply the hypergeometric distribution to solve applied problem.

1. An urn contains 40 marbles, exactly 15 of which are blue. Three marbles are drawn at random, *without replacement*. Drawing a blue will be considered a success, and the number of successes in the three draws will be observed.

   (a) Why is this experiment not a Bernoulli process?

   (b) One could draw a tree for this experiment. Draw just those branches of the tree for which there are exactly two successes. Determine the probability of ending up at the end of each of those branches, indicating clearly on your tree the conditional probabilities used.

   (c) Each of your probabilities from (b) should be the same. Write that probability in terms of "partial factorials" (things like $7 \cdot 6 \cdot 5$).

   (d) The number of branches with two successes can be written as a combination. What combination is it?

   (e) Write the probability of two successes in terms of your answers to (b) and (d).

2. Suppose again the urn containing 40 marbles with 15 of them being blue, but now the experiment is to draw five marbles without replacement. Find the probability of drawing exactly three blue marbles, writing your answer in a form like that of part 1(e) above.

In general, suppose there are $N$ objects from which $n$ are to be drawn without replacement, and $k$ of the $N$ objects are considered successes when drawn. Let $X$ be the random variable that assigns to each set of $n$ objects drawn the number of successes. The probability distribution function for the random variable $X$ is the

**Hypergeometric Distribution**

$$h(x\,;n,k,N) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \qquad x = 0,\ 1,\ 2,\ ...,\ \min\{n,k\}.$$

Here it is assumed that $n, k \leq N$, and the notation $\min\{n,k\}$ means the smaller of $n$ and $k$. *Note that the family of hypergeometric distributions is a THREE parameter family of distributions.*

3. You keep a jar of change at home, and (unbeknownst to you) it contains 274 coins, 87 of which are quarters. You are looking for quarters, so you take 10 coins from the jar.

   (a) What is the probability that exactly four of those coins are quarters? *Give your answer by first giving the distribution with the values of the variable and the parameters, then give it in decimal form, rounded to the thousandths place.*

(b) What is the probability that none of the coins are quarters? *Give your answer in the same way as you did in part (a).*

4. Suppose that an urn contains 10 marbles, exactly 6 of which are blue. An experiment consists of drawing three marbles without replacement.

   (a) Writing $h(x) = h(x\,;3,6,10)$, give the probability distribution $h$. *Give all probabilities in fraction form, reduced as far as possible while keeping the same denominator for all probabilities.* **NOTE:** Remember that this is a discrete distribution, so it only takes values at the values of $x$ that are in the range of the random variable $X$ that assigns to each outcome of the experiment the number of blue marbles drawn.

   (b) Give the cumulative distribution $H(x\,;3,6,10)$.

5. For the experiment of the previous exercise, find the probability of selecting

   (a) exactly one blue marble.
   (b) fewer than three blue marbles.
   (c) zero blue marbles.
   (d) at least one blue marble.

6. Consider again the jar of change containing 274 coins, 87 of which are quarters, and the experiment of randomly selecting 10 coins from the jar. Use Excel or some other technology to determine the probability of selecting *at least* four quarters. Indicate how you obtained your answer, in terms of either the probability function $h$ or the cumulative probability function $H$. *Give your answer in decimal form, rounded to the thousandths place.*

Since the experiments of drawing marbles from an urn with or without replacement are similar, we might expect there to be a relationship between the binomial distribution and the hypergeometric distribution. The next exercise illustrates this relationship.

7. (a) A certain kind of part is claimed to have a defective rate of 2%. Assuming this holds for every shipment, what is the probability that three out of 100 randomly selected (without replacement) parts from a shipment of 10,000 will be defective? *Give your answer in decimal form, rounded to six places past the decimal.*

   (b) Suppose that 100 parts from the shipment are selected at random, but *WITH* replacement. What is the probability that three out of the 100 are defective? *Give your answer in decimal form, rounded to six places past the decimal.*

   (c) Compare your answers to (a) and (b) and explain what you see.

It should be clear that what you saw in the last exercise is due to the fact that the number $n$ of parts drawn is much smaller than the total number $N$ of parts. (Sometimes we write this as $n << N$.) A general rule of thumb is as follows:

> If $N \geq 20n$, then $h(x\,;n,k,N) \approx b(x\,;n,\frac{k}{N})$.

The condition $N \geq 20n$ should not be taken to be firm; the approximation will be fairly good if, for example, $N = 18.3n$. What we *can* deduce is that for something like $N = 10n$ the approximation is probably not very good, and if $N = 50n$ it is probably quite good.

8. Consider the results of Exercise 7.

   (a) The error in using the binomial distribution to approximate the hypergeometric distribution is the absolute value of the difference between your answers to 7(a) and (b). Find the error, to six places past the decimal.

   (b) The percent error is the error divided by the correct value. (This of course gives the error in decimal form.) Find the percent error, to the nearest tenth of a percent.

## 4.2 Negative Binomial Distribution

**Performance Criteria:**

4. (b) Apply the negative binomial distribution to solve applied problem.

1. Suppose we have an urn with three red marbles and two yellow marbles, and consider the experiment of drawing marbles *WITH replacement* until two yellow marbles have been drawn.

   (a) Draw a tree diagram for this experiment, stopping whenever two yellows have been drawn or at the fourth draw, whichever comes first.

   (b) Compute the probability that the second yellow is drawn on the third draw.

2. Suppose now that for the same urn, you will draw until seven yellow marbles have been obtained. In this exercise you will determine the probability that the seventh yellow marble is drawn on the tenth draw. This really boils down to two things: (1) Finding the probability of any single outcome for which the seventh yellow is obtained on the tenth draw. (2) Determining how many ways the seventh yellow can be obtained on the tenth draw.

   (a) Determine the probability of obtaining any one outcome for which the seventh yellow is drawn on the tenth draw. Give your answer as the product of two numbers to a power.

   (b) Now to obtain the seventh yellow on the tenth draw, six yellows must have been drawn on the first nine draws. How many ways can this happen?

   (c) Of course the probability of getting the seventh yellow on the tenth draw is the product of your answers to (a) and (b). What is that probability?

3. Consider now a Bernoulli process with probability of success $p$, and let $q = 1 - p$, as before. Fix a number $n$ of successes, and draw until the $k$th success. Let $X$ be the random variable that assigns to each outcome the number of the draw on which the $n$th success occurs. Give an expression for $P(X = x)$.

**Negative Binomial Distribution**

Consider conducting a Bernoulli process, with the goal of obtaining $k$ successes. Let the random variable $X$ be the number of trials needed to obtain $k$ successes. Then the probability distribution function is

$$b^*(x\,;k,p) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} p^k\, q^{x-k}, \qquad x = k,\, k+1,\, k+2, ...,$$

where $q = 1 - p$.

4. Mr. W figures that the probability that any e-mail received is "spam" is 0.13. It is assumed that whether or not an e-mail received is spam is independent of all previous e-mails received. What is the probability that Mr. W receives his fourth spam e-mail for the day as his 15th overall e-mail of the day? Give your answer to four places past the decimal.

## 4.3 Poisson Distribution

**Performance Criteria:**

4. (c) Apply the Poisson distribution to solve applied problem.
   (d) Approximate a binomial probability with the Poisson distribution, when appropriate.

A **Poisson process** is an experiment consisting of counting the number of successes in a given period of time or region of space. It must meet the following conditions:

- The number of successes in two disjoint time periods or regions are independent of each other.
- The probability of a single success in a very short time interval or region is proportional to the length of time or size (length, area, volume) of the region.
- The probability of more than one success in such a short time interval or small region is negligible.

**Poisson Distribution**

Consider a Poisson process with average number $\lambda$ of successes per unit of time or space. Let the random variable $X$ be the number of successes in a fixed period of time $t$. Then

$$P(X = x) = p\,(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \qquad x = 0,\ 1,\ 2,\ 3, ...$$

1. Suppose that an average of 315 semi-trucks pass through a weigh station every day, and assume that their arrival at the weigh station is a Poisson process.

   (a) Determine the rate $\lambda$ at which trucks arrive, using whatever time units you wish, *but giving units with your answer*. (You may wish to read the next part of this exercise before deciding what time units to use - *whatever you use, keep all decimal places.*) Note that since we are assuming this is a Poisson process, this rate is good 24 hours a day. (This may be a reasonable assumption, since many truckers travel at night to avoid traffic.)

   (b) Determine the probability that exactly two trucks will arrive in a five minute period. Give your answer to the hundredth's place.

2. Consider the situation from the previous exercise, and suppose that we wish to know the probability that more than two trucks will arrive in a five minute period.

   (a) Write an infinite sum, or an expression involving an infinite sum, whose value is the desired probability.

   (b) Write a *finite sum*, or an expression involving a finite sum, whose value is the desired probability.

(c) Rewrite your answer to (b) in terms of the *cumulative* probability function $P(x; \lambda t)$.

(d) Find the desired probability, to the ten-thousandth's place.

3. In the "real world" it is often the case that a person will not necessarily know whether (or how well) a situation satisfies the conditions necessary for it to be modeled by a particular distribution. In these cases one often gathers some data and finds its distribution, then sees how well it matches the results given by a known distribution. You will investigate this process in this exercise.

   A machine is designed to spread fertilizer evenly over an area. You spread some fertilizer over a 10 square foot area that is marked off in a grid of one inch by one inch squares. The number of fertilizer pellets in each square is counted, giving these results:

| **No. of pellets in a square:** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | $\cdots$ |
|---|---|---|---|---|---|---|---|---|
| **Number of squares:** | 731 | 494 | 169 | 42 | 4 | 0 | 0 | $\cdots$ |

   The numbers of squares containing the different numbers of pellets will be referred to as the *actual counts*. What you will now do is see what values you would *expect* to get in the second row if the spread of fertilizer followed a Poisson distribution. Those vales are what we call the *expected counts*

   (a) Find the average number of pellets per square. (This is simply a weighted average of the numbers 0, 1, 2, 3, ..., with the weights being the number of squares having those numbers of pellets in them.) The value you obtain is $\lambda$ for the Poisson distribution. *Round to the nearest ten-thousandth, and give units with your answer.*

   (b) Determine the probabilities of 0, 1, 2, ... particles per square inch from the Poisson distribution. Round these values to the nearest ten-thousandth as well, and record your results in a table with two columns.

   (c) Add a third column to your table, giving the number of squares (out of the total in the 10 square foot area) that should contain the given number of pellets if the spreading of fertilizer really follows a Poisson distribution.

   (d) Compare the expected values from your table with the actual results of the experiment. Do you think that the Poisson distribution models the spread of the fertilizer well?

> If $p$ is small, $b(x\,; n, p) \approx p(x\,; np)$.

**NOTE:** This is used to approximate binomial probabilities with the Poisson distribution, since it can easily be computed as an exponential function.

4. A certain kind of part is claimed to have a defective rate of 2%. Assuming this holds for every shipment, what is the probability that 3 out of 100 randomly selected (without replacement) parts from a shipment of 10,000 will be defective? *Round all answers to the parts of this exercise to the ten-thousandth's place.*

   (a) Compute the desired probability, using the appropriate distribution.

   (b) This is not a Bernoulli experiment. However, it "almost" is; compute an approximation of the desired probability using the binomial distribution.

   (c) Approximate the binomial probability from part (b) using the Poisson distribution.

## 4.4 The Exponential Distribution

**Performance Criteria:**

4. (e) Apply the exponential distribution to solve applied problem.

Consider a Poisson process with parameters $\lambda$ and $t$. Here $\lambda$ is a constant, and the time $t$ is a parameter to be fixed. If $X$ is the random variable that assigns to an outcome the number of successes in a time period of length $t$, then $P(X = x) = p(x\,;\lambda t)$, where $p$ is the Poisson distribution.

Now $p(0\,;\lambda t)$ represents the probability of zero successes in a time period of length $t$. If we think of fixing the value of zero for $x$ and letting $t$ be the variable, we can look at $p(0\,;\lambda t)$ as representing the probability that it will take at least time $t$ to obtain one success. If we then take the new random variable $X$ to be the length of time to the first success, we have

$$P(X \geq t) = p(0\,;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}\,.$$

So we now have a Poisson process with rate parameter $\lambda$, and we are considering the experiment of waiting until the first success and recording the time to that success. We define the random variable $X$ to be the function that assigns to each outcome the time $x$ to that outcome. (Note that we are replacing $t$ with $x$ in the above.) By the above, the cumulative probability function for this random variable is

$$F(x) = P(X \leq x) = 1 - P(X \geq x) = 1 - e^{-\lambda x}\,.$$

By Theorem 3.7, the probability density function $f$ for this distribution is given by $f(x) = F'(x) = \lambda e^{-\lambda x}$. It is customary to then let $\lambda = \frac{1}{\beta}$, giving us the following.

**Exponential Distribution**

The **exponential distribution** is the continuous probability density function $f$ given below, along with it's cumulative distribution function:

$$f(x) = \begin{cases} \frac{1}{\beta}\, e^{-\frac{x}{\beta}} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\frac{x}{\beta}} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

This distribution is a good model for time between "successes" for a Poisson process. The parameter $\beta$ is the average time between successes, and is the reciprocal of the parameter $\lambda$ from the Poisson distribution, the average number of successes per unit time. Note that

$$\beta = \frac{1}{\lambda} \quad \Rightarrow \quad \lambda = \frac{1}{\beta}.$$

1. Suppose that customers arriving at a drive-up ATM machine during lunch hour is a Poisson process (not a bad assumption). The average time between customers during this time is 8 minutes.

(a) A customer has just arrived at the machine. What is the probability that the next customer will arrive in exactly five minutes? (Remember that this is a *continuous* distribution!)

(b) What is the probability that the next customer will arrive sometime between three and five minutes from now?

(c) What is the probability that the next customer will arrive in less than ten minutes?

(d) What is the probability that the next customer will arrive in more than ten minutes?

2. Consider the same scenario as Exercise 1.

(a) Suppose that you just arrived at the ATM machine during the lunch hour. What is the probability that 3 people will arrive in the next 15 minutes? *Note that the random variable has changed! You need to use a different distribution here.*

(b) You just arrived at the machine. What is the probability that no customers will arrive in the next ten minutes?

(c) Compare your answer to (b) with your answer to 1(d). Think about this!

3. The exponential distribution models things like times to failure of electronic components, where $\beta$ is the average time to failure. Suppose that the average time to failure for a certain component is 3.7 years. What is the probability that a randomly selected component will fail in less than 2 years?

4. In this exercise you will find the mean and variance of the exponential distribution, using the facts (which are obtained using integration by parts) that

$$\int u\,e^u\,du = e^u(u-1)+C \qquad \text{and} \qquad \int u^2\,e^u\,du = e^u(u^2-2u+2)+C\,.$$

(a) Make the substitution $u = -\dfrac{x}{\beta}$ to show that $E(X) = \beta \displaystyle\int_0^{-\infty} u\,e^u\,du$. Then evaluate this integral, using one of the above. (What is $e^{-\infty}(-\infty - 1)$, and why?)

(b) Find $E(X^2)$ in the same manner, using the same substitution.

(c) Find $\sigma^2$.

> For the exponential distribution, $\mu = \beta$ and $\sigma^2 = \beta^2$.

5. The result of this exercise may surprise you a bit! Consider an exponential distribution with $\beta = 3$.

(a) What do you think $P(X \le \mu)$ is?

(b) Find $P(X \le \mu)$, as a decimal to the nearest hundredth. Does your answer agree with your conjecture from (a)?

(c) For a probability distribution, the value of $x$ such that $P(X \le x) = \frac{1}{2}$ is called the **median**. Find the median of the exponential distribution with $\beta = 3$.

## 4.5 The Gamma Distribution

**Performance Criteria:**

4. (f) Apply the gamma distribution to solve applied problem.

Consider again the distribution

$$f(x) = \begin{cases} 0 & \text{if} \quad x < 0\,, \\ k\,x^{\alpha-1}\,e^{-\frac{x}{\beta}} & \text{if} \quad x \geq 0 \end{cases} \tag{1}$$

for parameters $\alpha > 0$ and $\beta > 0$, with $k$ some constant to be determined.

1. (a) Make the substitution $y = \dfrac{x}{\beta}$ to show that

$$\int_0^\infty k\,x^{\alpha-1}\,e^{-\frac{x}{\beta}}\,dx = k\,\beta^\alpha\,\Gamma(\alpha)\,.$$

   (b) Determine what the value of $k$ must be in order for the above function to be a probability density function.

**The Gamma Distribution**

For $\alpha > 0$ and $\beta > 0$, the continuous probability distribution with probability distribution function

$$f(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ \dfrac{1}{\beta^\alpha \Gamma(\alpha)}\,x^{\alpha-1}\,e^{-\frac{x}{\beta}} & \text{for} \quad x > 0 \end{cases}$$

is called the **gamma distribution**.

Note that when $\alpha = 1$ the gamma distribution becomes the exponential distribution, which sometimes models time to failure of an electronic component, or time until arrival of a person or vehicle at some point. One interpretation of the gamma distribution is that it models time to failure of *several* electronic components, or time to arrival for multiple people/vehicles. In these cases, $\alpha$ represents the number of components/people/vehicles, and $\beta$ is again the average time between failures/arrivals.

2. Consider the weigh station of Exercise 1, Section 4.3, where the average time between trucks passing through the weigh station was 0.0762 hour. The scales used to weigh the trucks have to be recalibrated every fifty trucks. What is the probability of having to recalibrate less than four hours after the previous recalibration? Use your calculator to evaluate the appropriate integral, and give your answers to the thousandth's place.

3. Now we return to our ATM machine, with an average time between customers of 8 minutes (during the lunch hour). During the lunch hour, what is the probability that

   (a) That the amount of time for three customers to arrive (after the customer preceding all of them) will be ten to fifteen minutes?

   (b) That the amount of time for five customers to arrive (again, after some preceding customer) will be twenty minutes or less?

4. Now consider electrical components that have an average time to failure of 3.7 years. What is the probability it will be five years or longer before two of them fail?

5. (a) Use the substitution $y = \dfrac{x}{\beta}$ to show that the expected value of the gamma distribution is

$$\mu = \frac{\beta}{\Gamma(\alpha)} \int_0^\infty y^\alpha \, e^{-y} \, dy \, .$$

   (b) Use the fact that $y^\alpha = y^{(\alpha+1)-1}$ along with the definition of the gamma function to simplify the integral from part (a), getting the mean for the gamma distribution. (Your answer should be just an algebraic expression involving $\alpha$ and $\beta$.)

6. (a) Use a procedure like that of the previous exercise to find $E(X^2)$.

   (b) Use your answer to (a) along with the result of 3(b) to find the variance of the gamma distribution.

> For the gamma distribution, $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$.

It is not always the case that $\alpha$ and $\beta$ can be determined as they were for Exercises 1 and 2. In some situations an experiment is performed to determine $\alpha$ and $\beta$ and the validity of the resulting gamma distribution model is checked against the data before using the model. In a previous Exercise you did this for a situation that might have involved a Poisson process. (The spread of fertilizer pellets.) Using the information given, you created a Poisson distribution to model the situation, and checked your model against the real data. In the second exercise below you will do the same thing again, but with a gamma distribution.

7. Suppose that you suspect some data follows a gamma distribution, and you want to determine $\alpha$ and $\beta$ for the distribution. You find the mean and variance of the data to be 35 and 245, respectively.

   (a) The fact that $\mu = \alpha\beta$ means that $\alpha\beta = 35$. Multiply both sides of this equation by $\beta$, then substitute the value of $\sigma^2$ for $\alpha\beta^2$. Solve for $\beta$.

   (b) Substitute the value you obtained for $\beta$ into the equation for the mean to find $\alpha$.

   (c) Repeat the same process to find $\alpha$ and $\beta$ if $\mu = 18.3$ and $\sigma^2 = 23.8$.

8. (Somewhat True) Story: When $H_2O$man was young, he and his father used to go out early in the morning and catch unfortunate grasshoppers to use as fishing bait. (If you try this yourself, the early in the morning part is important - the grasshoppers are cold, so they don't move as fast!) Once when doing this process, older $H_2O$man watched the young lad catching grasshoppers, and he (dad) recorded the times, in seconds, between when successive grasshoppers were caught. Here are the times he recorded:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 13 | 29 | 58 | 14 | 40 | 17 | 9 | 27 | 36 | 21 |
| 24 | 11 | 26 | 33 | 48 | 19 | 63 | 35 | 23 | 47 | 12 |
| 15 | 24 | 32 | 52 | 19 | 14 | 44 | 21 | 36 | 27 | |

(a) Make a stem-and-leaf plot of this data, with split stems. (The top entry I found when doing a search for *stem-and-leaf plot split stems* has a nice explanation of this.)

(b) Enter the data in Excel and use Excel commands to find the mean and variance of the data. There are several variances - you want the one with a P in it's function call. *Note that the mean and variance are two parameters which characterize the distribution.*

(c) We will now assume that the data can be modelled with a gamma distribution. The gamma distribution should then have the same mean and variance as the data itself. Use the equations for the mean and variance of the gamma distribution to determine the values of the parameters $\alpha$ and $\beta$.

9. Now you need to check to see how well a gamma distribution with the parameters you just found models the data. You will find probabilities that the data falls in certain intervals, then find the predicted probabilities for the same intervals, using the gamma distribution.

(a) If you randomly select one of the data values above, what is the probability that it will lie in the interval $[15, 20)$? (Note that this interval includes 15, but not 20.) Put that value in the column for true probabilities in the spreadsheet. Then figure out how to use the spreadsheet functions to get all the probabilities for that column.

(b) Find the Excel gamma distribution and use it to compute the values for the modeled probability column. *Remember that you are dealing with a continuous distribution, so find the probabilities appropriately.* (Can you say "cumulative distribution?")

(c) To get a visual of how well the model matches the real data, you might want to plot the values from the two columns side by side. I think you can easily figure out how to do this. How does the model look?

## 4.6 A Summary of Distributions

**Performance Criteria:**

4. (g) Choose and apply the appropriate distribution(s) to solve an applied problem.

When solving a problem there are two decisions to be made:

1) Which probability distribution to use.

2) Whether to use the probability distribution/density function (pdf), or the cumulative distribution function (cdf).

Let's address the second decision first:

- If the distribution to be used it is discrete, one should likely use the pdf if a probability is needed for only one, or maybe just a few, value(s) of the random variable. Otherwise the cdf will likely be easier to use.
- If the distribution to be used is continuous, the following guidelines will likely be effective:
  - ◇ In the case of the normal distribution, one would likely use the cdf in the form of the normal distribution tables, or as a tool in *Excel*. The pdf could also be used directly with a tool that can approximate integrals of it using numerical integration.
  - ◇ In the case of the exponential distribution, the CDF has a simple form that can easily be applied.
  - ◇ The Poisson and gamma distributions are easy to work with in *Excel*, using their cdfs.

This now leaves us with the question of which specific distribution (or distributions) should be chosen for a given problem. Here are some guidelines for making a selection:

- First identify the random variable. If it consists of a count, then the distribution of interest will be discrete. If it consists of a measurement of a continuous (in theory) quantity, then the distribution of interest will be continuous.
- If one is determining the probability of a certain number of "successes" in a fixed number of trials, either the binomial or hypergeometric distribution should be used. The binomial distribution is used when items are selected with replacement, and the hypergeometric distribution is used when items are selected without replacement.
- The binomial distribution is also appropriate when determining the probability of a number of successes in a sample where the probability of success is the same for every item in the sample.
- If the number of successes is fixed in a Bernoulli process (like choosing with replacement), and trials are to be continued until that number of successes is attained, the negative binomial distribution should be used.
- If one is determining the probability of a number of successes in a fixed length of time, or in a fixed length, area or volume, the Poisson distribution is used.

- If the random variable is continuous and normally distributed, we use (of course!) the normal distribution.
- When calculating the probability of a given range of lengths of time to the first success for a Poisson process, the exponential distribution is used. If instead we are calculating the probability of a given range of lengths of time to a certain fixed number of successes in Poisson process, the gamma distribution is used.

All of the distributions we have encountered are summarized in the table on the next page. In the Chapter Exercises you will find a selection of scenarios for which you will select and apply the appropriate distribution.

# Some Commonly Used Distributions

| Distribution | Assumptions/ Conditions | Random Variable | Parameters | Notation | Formula |
|---|---|---|---|---|---|
| Binomial | Bernoulli Process - sampling with replacement | Number of successes in a fixed number of trials | Number $n$ of trials, probability $p$ of success, $q = 1 - p$ | $b(x; n, p)$ $B(x; n, p)$ | $b(x; n, p) = \binom{n}{x} p^x q^{n-x}$ |
| Negative Binomial | Repeated independent trials with constant probability of success | Number of trials to a fixed number of successes | Number $k$ of successes, probability $p$ of success, $q = 1 - p$ | $b^*(x; k, p)$ $B^*(x; k, p)$ | $b^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k}$ |
| Hypergeometric | Sampling without replacement | Number of successes in a fixed number of trials | Number $N$ of objects to be drawn from, number $k$ of the $N$ objects that are considered successes, number $n$ of trials | $h(x; n, k, N)$ $H(x; n, k, N)$ | $h(x; n, k, N) = \dfrac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$ |
| Poisson | Poisson Process - successes uniformly distributed in time or space | Number of successes in a fixed length $t$ of time or a fixed length/area/volume (also denoted by $t$) | Average number $\lambda$ of successes per unit of time/length/area/volume, $t$ units of time/length/ area/volume | $p(x; \lambda t)$ $P(x; \lambda t)$ | $p(x; \lambda t) = \dfrac{e^{-\lambda t}(\lambda t)^x}{x!}$ |
| Normal | Normally distributed continuous data | Data value | Mean $\mu$ and standard deviation $\sigma$ | $n(x; \mu, \sigma)$ $N(x; \mu, \sigma)$ | $n(x; \mu, \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ |
| Exponential | Poisson Process | Time to first success | Average length of time $\beta$ between successes | $f(x)$, $F(x)$ | $f(x) = \dfrac{1}{\beta} e^{-\frac{x}{\beta}}$ for $x > 0$ <br> $F(x) = 1 - e^{-\frac{x}{\beta}}$ for $x > 0$ |
| Gamma | Poisson Process | Time to some number of successes | Average length of time $\beta$ between successes, number $\alpha$ of successes | $f(x)$, $F(x)$ | $f(x) = \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$ for $x > 0$ |

## 4.7 Chapter 4 Exercises

For each exercise, provide the combination of the following that its asked for.

($i$) A probability statement giving the event of interest, in terms of the random variable $X$.

($ii$) An expression in terms of the pdf, with all parameters and the variable substituted, whose value is the desired probability.

($iii$) An expression in terms of the *formula for the pdf*, with all parameters and the variable substituted, whose value is the desired probability.

($iv$) An expression in terms of the cdf, with all parameters and the variable substituted, whose value is the desired probability.

($v$) The desired probability, as a decimal rounded to four places past the decimal.

**DO NOT label these things with the letters above.** They should be strung together in order, with equal signs between them.

1. A couple wants to have two boys, but have also decided that they wish to have no more than four children. The probability in the general population of having a boy is 0.52. Assume that this probability holds for the couple, and that the gender of any child born is independent of previous children born to the same parents.

   (a) What is the probability that the couple will need to have four children in order to get two boys? Provide *i*, *ii*, *iii*, *iv*, *v*.

   (b) What is the probability that the couple will not obtain two boys by the time they have their fourth child? Provide *i*, *ii*, *iii*, *iv*, *v*.

2. Consider again the couple from the previous exercise. Suppose that they had decided they were going to have (exactly) three children, What is the probability that will get (at least) two boys? Provide *i*, *ii*, *iii*, *iv*, *v*.

3. A particular highway has an average of 3 fatal accidents per year. What is the probability of two fatal accidents over a six month period? Provide *i*, *ii*, *iii*, *iv*, *v*.

4. The manager of the Rubber Hits The Road tire shop, Bud, knows that on the first day the snow flies, customers arrive to have their snow tires put on at a rate of 13 per hour. Given that Rubber Hits The Road is open from 8:00 AM to 5:00 PM, what is the probability that they will change 125 sets of tires on the first day that it snows? Provide *i*, *ii*, *iii*, *iv*, *v*.

5. GW likes to have strawberries on his cereal in the morning, and he occasionally finds that some of the strawberries he bought are moldy. If (unbeknownst to him) 4 of the fifty strawberries in his refrigerator are moldy and he selects ten of them for breakfast, what is the probability that

   (a) exactly one of them is moldy? (b) two or more of them are moldy?

   Provide *i*, *ii*, *iii*, *iv*, *v* for each.

6. An assembly line worker's job is to install a particular part in a device, a task which they can do with a probability of success of 0.78 on each attempt. (Assume that a success on one attempt is independent of success on all other previous or future attempts.)

(a) Suppose that on a particular day the worker needs to install five such parts. What is the probability that it will take them exactly 8 attempts to do so? Provide *i*, *ii*, *iii*, *v*.

(b) Suppose that on another day they need to install seven such parts. What is the probability that it will take them between 10 and 15 (inclusive, meaning including both of those values) attempts to do so? Provide *i*, *ii*, *iv*. Excel does not provide the cdf for this distribution, so an answer must be obtained using the pdf. We'll go over the most efficient way to get a value for this in class on Wednesday.

7. When backpacking in the Wind River Mountains of Wyoming, one evening mosquitoes were landing on me at a rate of 12 per minute. *Think carefully about what the variable is in each of parts (b) and (c)!*

   (a) What was the average length of time between mosquitoes landing?

   (b) What is the probability that, from the time one mosquito landed, the next mosquito would land in between 4 and 7 seconds? Tell which distribution you use and give some indication of how it is used to obtain your answer. Round to four places past the decimal.

   (c) What is the probability that fewer than 20 mosquitoes would land in a two minute period?

8. A manufacturer of a certain part claims that only one part in 50 will fail under a pressure of 500 psi. You are going to test this claim by randomly selecting five parts from a batch of 300 and testing them for failure. A part will not be replaced in the batch after testing.

   (a) What is the probability that exactly one of the five parts will fail? Provide *i*, *ii*, *iii*, *v*.

   (b) What is the probability that *at least* one of the five parts will fail? Provide *i*, *ii*, *v*. *Give ii without using a summation.*

9. A rigged deck of cards has only 10 each of hearts, diamonds and clubs, with 9 extra spades to make up for the missing cards.

   (a) If ten cards are selected at random, with replacement, what is the probability that exactly five of them are spades? Provide *i*, *ii*, *iii*, *v*.

   (b) 100 cards are selected at random, with replacement. What is the probability that between 30 and 60 of them, inclusive, are spades? Provide *i*, *ii*, *iv*, *v*.

   (c) Cards are to be drawn, with replacement, until the third spade is obtained. What is the probability that exactly seven cards must be drawn to obtain the third spade? Provide *i*, *ii*, *iii*, *v*.

10. A device being made contains spacers that must be of thicknesses between 0.05 inches and 0.07 inches. In the past you have found that thicknesses of spacers from a supplier are normally distributed, with mean 0.0608 inches and standard deviation 0.0047. What is the probability that a randomly selected spacer meets the requirement of being between 0.05 inches and 0.07 inches in thickness? Provide *i*, *ii*, *iv*, *v*.

11. Give the probability asked for in the previous exercise in terms of the standard normal distribution (mean zero and standard deviation one) cdf, with values rounded to the hundredth's place, Then determine the probability using the standard normal distribution table. Your answer should be close to what you got in the previous exercise.

12. (a) During summer mornings, the average time between vehicle arrivals at the entrance to Lava Beds National Monument is 18 minutes. If the attendant wants to read a magazine article that takes 15 minutes to read without being interrupted, what is the probability that they will have time to do this? (*Remember that you are dealing with a continuous distribution for this situation.*)

    (b) At the same time but about 100 miles away, the average time between vehicles arriving at Crater Lake National Park is 43 seconds. After the arrival of one carload of visitors, what is the probability that the next carload will arrive sometime between 30 and 50 seconds later?

13. Recall the following situation, from a previous exercise: An assembly line worker's job is to install a particular part in a device, a task which they can do with a probability of success of 0.78 on each attempt. (Assume that a success on one attempt is independent of success on all other previous or future attempts.) Suppose that they need to install ten such parts a day.

    (a) What is the probability that it will take them 12 or more tries to install ten such parts? Provide *i*, *iv*, *ii*, *v* **in that order**. What you provide for *ii* should be a slight modification of what you provide for *iv*.

    (b) What is the probability that, during a five day work week, it will take the worker 12 or more tries to install all the parts on exactly three of the days? **You will need to use your result from (a), along with a different distribution.** Provide *i*, whichever of *ii* and *iv* is more appropriate, and *v*.

    (c) What is the probability that, during a five day work week, it will take the worker *less than* 12 tries to install all the parts on one or two of the days? Provide *i*, *ii*, *v*.

14. Consider again the spacers that must be of thicknesses between 0.05 inches and 0.07 inches, which are normally distributed, with mean 0.0608 inches and standard deviation 0.0047.

    (a) What is the probability that a randomly selected spacer has a thickness of 0.055 inches or more? Provide *i*, *ii*, *iv*, *v*.

    (b) In the last assignment you should have found that the probability of selecting a spacer meeting the desired specifications to be 0.9641. If you checked each spacer before installing it, what is the probability that you would select 20 satisfactory spacers before obtaining your first one that doesn't meet the specifications? Provide *i*, *ii*, *iii*, *v*.

15. A brand of solar panels has small "flaws" or "blemishes" at a rate of 0.037 flaws per square foot.

    (a) What is the probability that a 3 foot by 10 foot panel will have 4 or more flaws? Provide *i*, *ii*, *iv*, *v*.

    (b) How many flaws would you expect a panel to have, on average?

16. On average, 23 trucks pass through a weigh station each hour. Assuming that the arrival of trucks at the weigh station is a Poisson process, what is the probability that a person working the weigh station would see 175 or fewer trucks pass though during an eight hour day of work? Provide *i*, *ii*, *iv*, *v*.

# H Solutions to Exercises

## H.4 Chapter 4 Solutions

**Section 4.1:**

1. (a) Since the marbles are drawn without replacement, the probability of a success on each trial does not remain the same.

   (c) $\dfrac{15 \cdot 14 \cdot 35}{40 \cdot 39 \cdot 38}$ (d) $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ (e) $\begin{pmatrix} 3 \\ 2 \end{pmatrix} \dfrac{15 \cdot 14 \cdot 35}{40 \cdot 39 \cdot 38}$

2. $\begin{pmatrix} 5 \\ 3 \end{pmatrix} \dfrac{15 \cdot 14 \cdot 13 \cdot 35 \cdot 34}{40 \cdot 39 \cdot 38 \cdot 37 \cdot 36}$

3. (a) $h(4; 274, 10, 87) = 0.219$ (b) $h(0; 274, 10, 87) = 0.020$

4. (a)

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $h(x; 10, 3, 6)$ | $\frac{1}{30}$ | $\frac{3}{10}$ | $\frac{1}{2}$ | $\frac{1}{6}$ |

   (b)
$$H(x; 10, 3, 6) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{30} & \text{for } 0 \le x < 1 \\ \frac{10}{30} & \text{for } 1 \le x < 2 \\ \frac{25}{30} & \text{for } 2 \le x < 3 \\ 1 & \text{for } x \ge 3 \end{cases}$$

5. (a) $h(1; 10, 3, 6) = \frac{3}{10}$ (b) $H(2; 10, 3, 6) = \frac{25}{30}$ (c) $h(0; 10, 3, 6) = \frac{1}{30}$
   (d) $1 - H(0; 10, 3, 6) = 1 - \frac{1}{30} = \frac{29}{30}$

7. (a) 0.1837175 (b) 0.182276
   (c) The answers are very close. This is because the number of defective parts is so small relative to the total number of parts that the probability of selecting a defective part doesn't change much with or without replacement.

8. (a) The error is 0.000899. (b) The percent error is 0.5%.

**Section 4.2**

1. (a) $2\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right) = \frac{24}{125}$

2. (a) $\left(\dfrac{2}{5}\right)^7\left(\dfrac{3}{5}\right)^3$ (b) $\begin{pmatrix} 9 \\ 6 \end{pmatrix}$ (c) $\begin{pmatrix} 9 \\ 6 \end{pmatrix}\left(\dfrac{2}{5}\right)^7\left(\dfrac{3}{5}\right)^3 = \dfrac{290304}{9765625} = 0.02973$

1. $P(X = x) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix}(p)^k(q)^{x-k}$

2. $b^*(15; 4, 0.13) = 0.0225$

**Section 4.3**

1. (a) $\lambda = 315 \text{ trucks/day} \cdot \dfrac{1 \text{ day}}{1440 \text{ minutes}} = 0.21875 \text{ trucks/minute}$

   (b) $p(2; (0.21875)(5)) = p(2; 1.09375) = \dfrac{e^{-1.09375}(1.09375)^2}{2!} = 0.20$

2. $\displaystyle\sum_{x=3}^{\infty} p(x; 1.09375) = 1 - \sum_{x=0}^{2} p(x; 1.09375) = 1 - P(2; 1.09375) = 0.0983$

4. (a) 0.1832 (b) 0.1823 (c) 0.1804

**Section 4.4**

1. (a) $P(X = 5) = 0$

   (b) $P(3 \le X \le 5) = F(5) - F(3) = (1 - e^{-\frac{5}{8}}) - (1 - e^{-\frac{3}{8}}) = e^{-\frac{3}{8}} - e^{-\frac{5}{8}} = 0.152$

   (c) $P(X < 10) = 1 - e^{-\frac{10}{8}} = 0.713$

   (d) $P(x > 10) = 1 - P(X < 10) = 1 - 0.713 = 0.287$

2. (a) $\lambda = \frac{1}{8}, \quad p(3; , \dfrac{1}{8} \cdot 15) = \dfrac{e^{-\frac{15}{8}}(\frac{15}{8})^3}{3!} = 0.168$

   (b) $p(0; \frac{10}{8}) = e^{-\frac{10}{8}} = 0.287$ (c) The answers are the same.

3. $P(X \le 2) = \frac{1}{3.7} e^{-\frac{2}{3.7}} = 0.1574$

4. (b) $P(X \le 3) = \frac{1}{3} e^{-1} = 0.1226$

   (c) $\dfrac{1}{2} = \displaystyle\int_{-\infty}^{x} \frac{1}{3} e^{-\frac{t}{3}} = -e^{-\frac{t}{3}}\Big]_{-\infty}^{x} = e^{-\frac{x}{3}} \;\Rightarrow\; \ln\left(\frac{1}{2}\right) = \ln(e^{-\frac{x}{3}}) = -\frac{x}{3} \;\Rightarrow\; x = -\dfrac{\ln \frac{1}{2}}{3}$