Module 2

This is a single, concatenated file, suitable for printing or saving as a PDF for offline viewing. Please note that some animations or images may not work.

Module 2 Study Guide and Deliverables

Topics:	**Lecture 3:** Analyzing Risk: An Introduction to Modeling Uncertain Inputs **Lecture 4:** Analyzing Risk: incorporating Uncertainty into the Decisions of the Enterprise
Readings:	Lectures 3 and 4 online content
Discussions:	**Module 2 Discussion**
Assignments:	Team and case assignments **Tutorial:** 2 **Assignment 2**: Individual Assignment covering Lecture 3 and Lecture 4.
Assessments:	**Quiz 2**

Lecture 3

Learning Objectives

After you complete this lecture, you will be familiar with the following:

- Commonly used discrete and continuous probability distributions
- Relationships among probability distributions
- Probability density function and probability mass function

- Cumulative distribution function
- R functions for commonly used probability distributions

Probability and Uncertainty

We start this lecture by reviewing probability concepts and their relation to uncertainty.

Probability provides a scale for the likelihood of an event.

- Probability 1 means the event is certain.
- Probability 0 means it will not happen.
- A probability of 0.8 means that over many identical situations, the event will happen 80% of the time.

The uncertain cells in our spreadsheets do not contain random numbers but probability distributions. A *probability distribution* literally distributes the probability that something happens among a collection of possible outcomes.

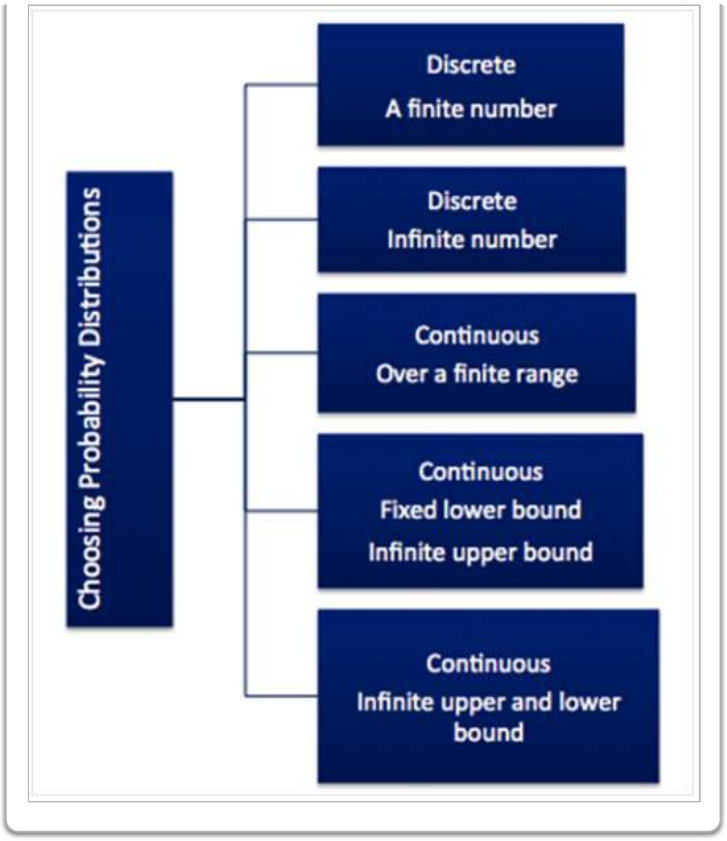
As discussed in the previous modules, we use probability distributions to model uncertain situations. Our goal in this lecture is to get familiar with common probability distributions and learn which distribution to use to model a particular uncertain input in our model.

A probability distribution is either discrete or continuous. A discrete probability distribution is used when the uncertain quantity can take only discrete values; e.g., the demand for a particular product during the next month. This can only take discrete values such as 5, 10, 15… A continuous probability distribution is used when the uncertain quantity can take all values in a range; e.g., time to finish an activity. This could go from 0 to infinity and any value is possible.

Classification of Probability Distributions Based on Possible Outcomes

When trying to model an uncertain situation using a probability distribution, the first question is: What is the collection of possible outcomes? Therefore, in this lecture we will classify probability distributions according to the possible values the outcomes they represent may take. Under each category, there are several probability distributions. Our goal in this lecture is to learn the physical basis of the distributions, i.e., what distribution best represents the situation we are trying to model.

Figure 3.1



Possible Outcomes - Discrete, and a finite number

Examples of situations where we have discrete outcomes and there are only a finite number of these outcomes are:

- Number of workers absent out of 70 in a shift
- Number of defective parts in a lot of 100

R can generate random variables from several such distributions. We will cover the binomial and the discrete uniform distribution.

The Binomial Distribution

This models the number of successes in *t** trials, when the trials are independent with success probability *p*. *t* must be a natural number, and *p* must be between 0 and 1.

Example[1]:

In a portfolio of 20 similar stocks, each of which has the same probability of increasing in value of p = 0.6, the total number of stocks that increase in value can be described by a binomial distribution with parameters t – 20 and p = 6.

To generate random variables from a binomial distribution, we can use the rbinom function. rbinom takes three arguments: the size of the vector you want to generate, the number of trials *for the binomial experiment being run*, and the probability of success for each trial. For example, rbinom(15, 100, .6) would generate 15 binomially distributed random variables, where each experiment has 100 trials and a 60% probability of success.

* More often, you'll see the number of trials referred to as *n*. Since R uses *n* as the size of the vector to be generated, we avoid confusion here by calling this *t*.

Figures 3.2-3.4 all show histograms for different binomial random variables. To get a good idea of what each distribution looks like, we set the number of trials to 1,000,000.

Figure 3.2: a~*bin*(20, .8)

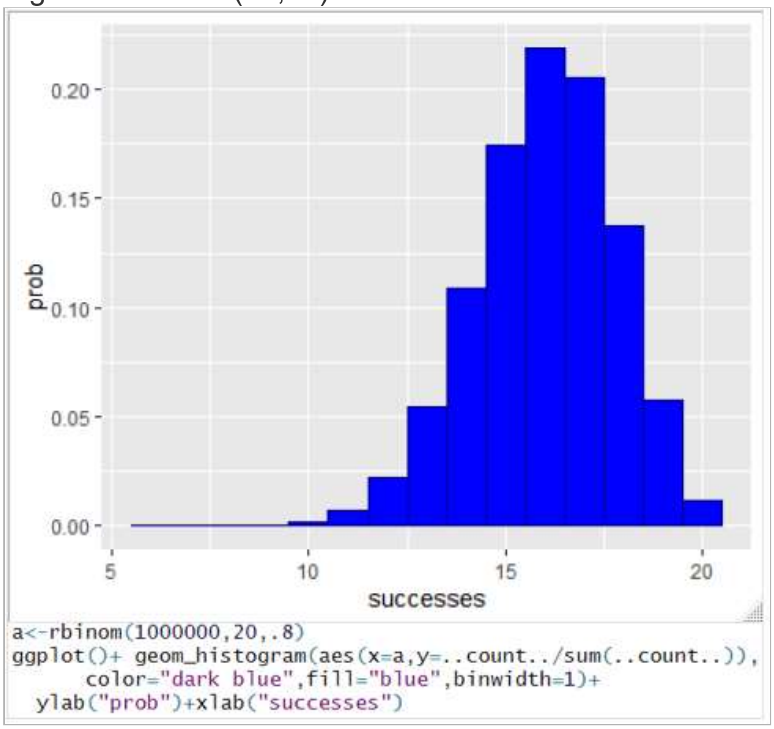
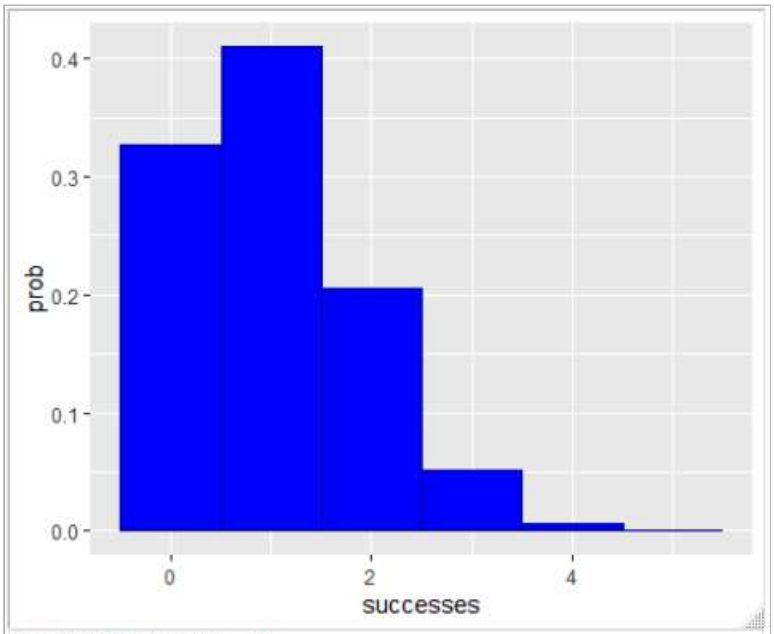
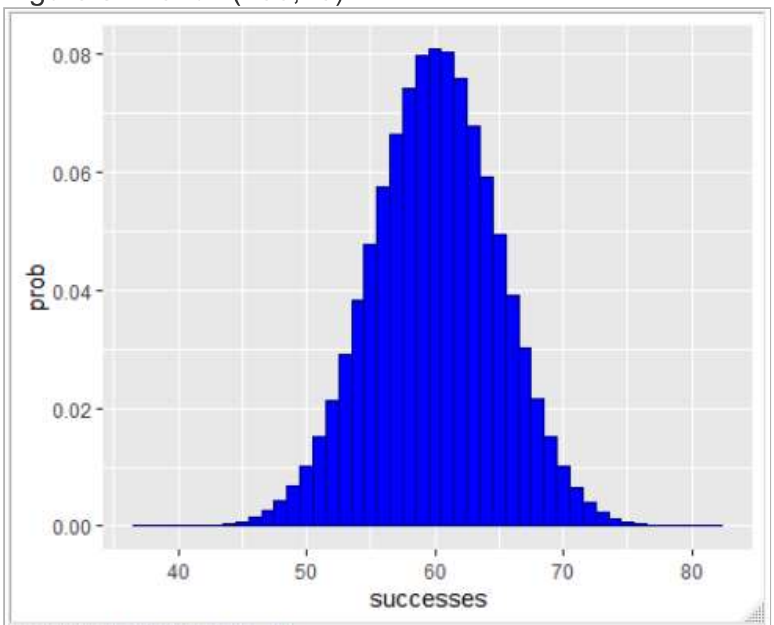


Figure 3.3: b~*bin*(5, .2)




```
b<-rbinom(1000000,5,.2)
ggplot()+ geom_histogram(aes(x=b,y=..count../sum(..count..)),
                         color="dark blue",fill="blue",binwidth=1)+
  ylab("prob")+xlab("successes")
```

Figure 3.4: c~*bin*(100, .6)




```
c<-rbinom(1000000,100,.6)
ggplot()+ geom_histogram(aes(x=c,y=..count../sum(..count..)),
                         color="dark blue",fill="blue",binwidth=1)+
  ylab("prob")+xlab("successes")
```

Discrete Uniform

Discrete Uniform distributions complete uncertainty, since all outcomes are equally likely. The function to simulate Discrete Uniform distribution takes two parameters: the minimum possible outcome (min) and the maximum possible outcome (max). Max must be greater than min.

Example:

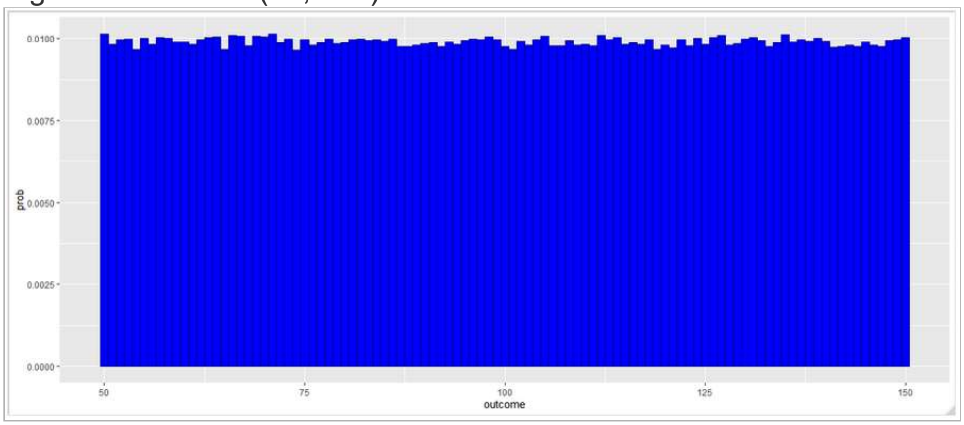
The demand for a particular shirt is expected to vary between *min* = 50 and *max* = 150. Each number between the quantities 50 and 150 is equally likely.

To generate random variables from a discrete uniform distribution, we can slightly modify the uniform distribution. (The uniform distribution is covered later in this lecture.) We start with a uniform distribution with the same minimum, but add one to the maximum. We then round the number *down*. The result will have a discrete uniform distribution. The function can be seen in Figure 3.5, and a histogram of the results of one such random variable (with 1,000,000 numbers generated) is shown in Figure 3.6.

Figure 3.5

```
rdunif<- function(n,min=1,max=100){
  floor(runif(n,min,max+1))
}
```

Figure 3.6: a~*dunif*(50, 150)



Possible Outcomes - Discrete, and infinite number

In some situations, we know the outcome will be an integer, but there are nonetheless infinite possible outcomes (the result is always some finite number, but it can be arbitrarily large). Examples include:

- Annual demand for a product
- Number of accidents during the year

R can generate numbers from these distributions too. We will cover geometric, Poisson, and negative binomial distributions.

Geometric Distribution

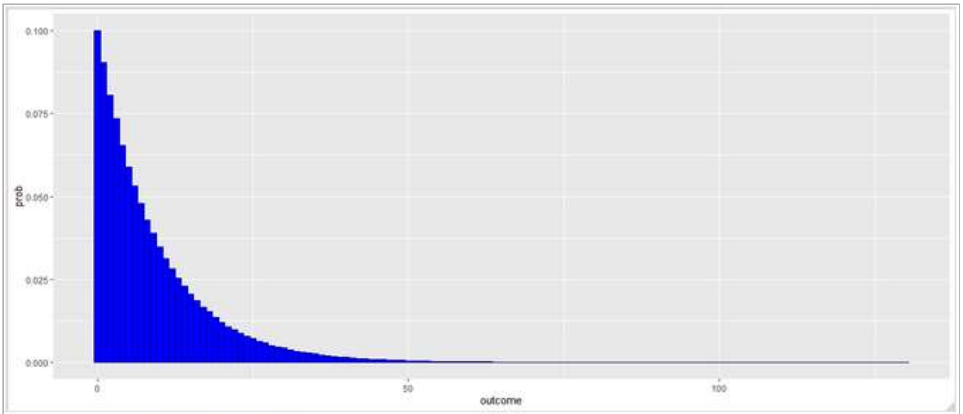
Geometric distribution models the number of failures before the first success in a sequence of independent trials with probability *p* of success on each trial. The only parameter for a geometric distribution is *p*. *p* must be between 0 and 1.

Example:[2]

An R&D division of a company may invest in several projects that fail before investing in a project that succeeds. If each project has a probability of success *p*, the number of projects that fail before a successful project occurs is a geometric random variable with parameter *p*.

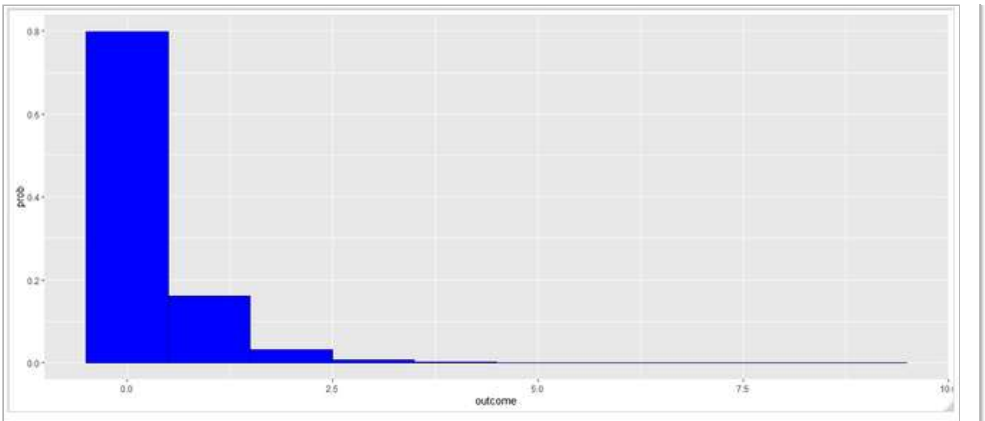
For this, we can use the rgeom function, native to R. Figures 3.7 and 3.8 show histograms from randomly generated geometric variables with $p = 0.1$ and $p = 0.8$.

Figure 3.7: *a~geom*(.1)




```
a<-rgeom(1000000,.1)
ggplot()+ geom_histogram(aes(x=a,y=..count../sum(..count..)),
                         color="dark blue",fill="blue",binwidth=1)+
  ylab("prob")+xlab("outcome")
```

Figure 3.8: *a~geom*(.8)




```
b<-rgeom(1000000,.8)
ggplot()+ geom_histogram(aes(x=b,y=..count../sum(..count..)),
                         color="dark blue",fill="blue",binwidth=1)+
  ylab("prob")+xlab("outcome")
```

Negative Binomial Distribution

This models the number of failures before the *sth* success in a sequence of independent trials with probability *p* of success on each trial. *s* and *p* are the only parameters. *s* must be a natural number, and $0 < p < 1$.

Example:

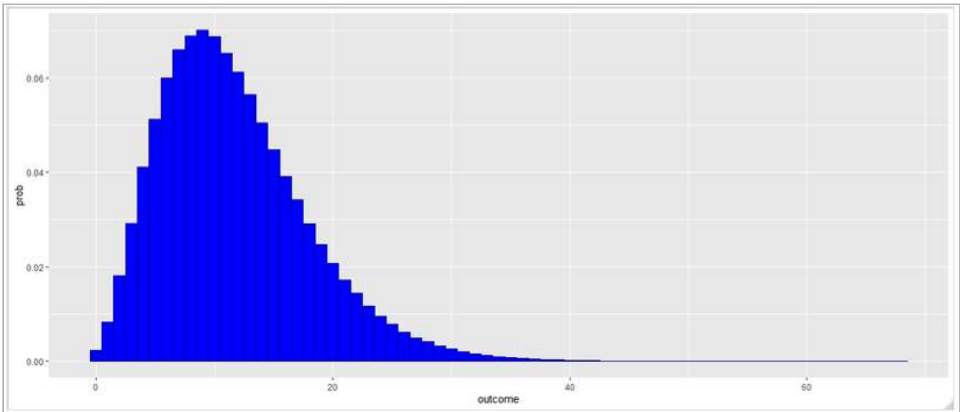
If each project in an R&D department has the same chance of success, the number of projects that fail in the R&D division before experiencing three successful projects is a negative binomial random variable.

For this, we can use the rnbinom function, native to R. Figure 3.9 shows a histogram from a randomly generated negative binomial distribution with $s = 5$ and $p = 0.3$.

Individual Exercise:

How could you set the parameters of a negative binomial distribution such that the result would also be a geometric distribution?

Figure 3.9: *a~negbin*(5, .3)




```
a<-rnbinom(1000000,5,.3)
ggplot()+ geom_histogram(aes(x=a,y=..count../sum(..count..)),
                         color="dark blue",fill="blue",binwidth=1)+
  ylab("prob")+xlab("outcome")
```

Poisson Distribution

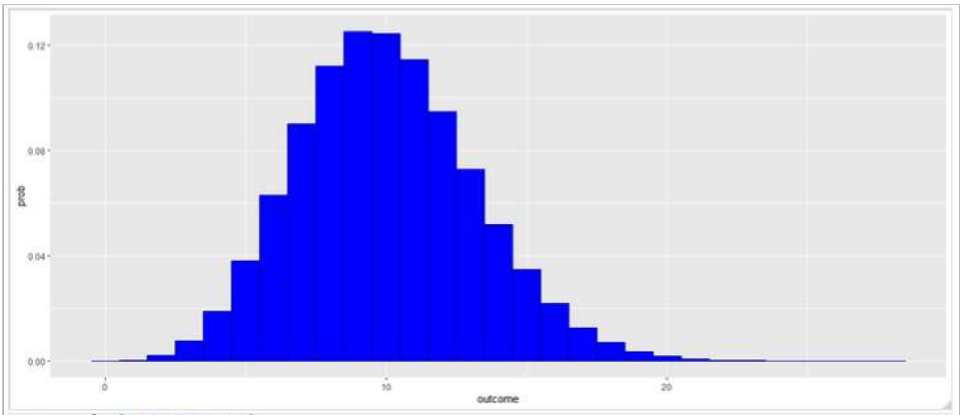
The Poisson distribution models the number of independent events that occur in a fixed period of time. It's fully parametrized by *m*, the mean (or expected) number of times the event will occur. *m* must be greater than 0.

Example:

If, on average, 10 customers arrive at a store during one hour, then the number of customers arriving at the store in an hour can be modeled as a Poisson random variable with *m* = 10.

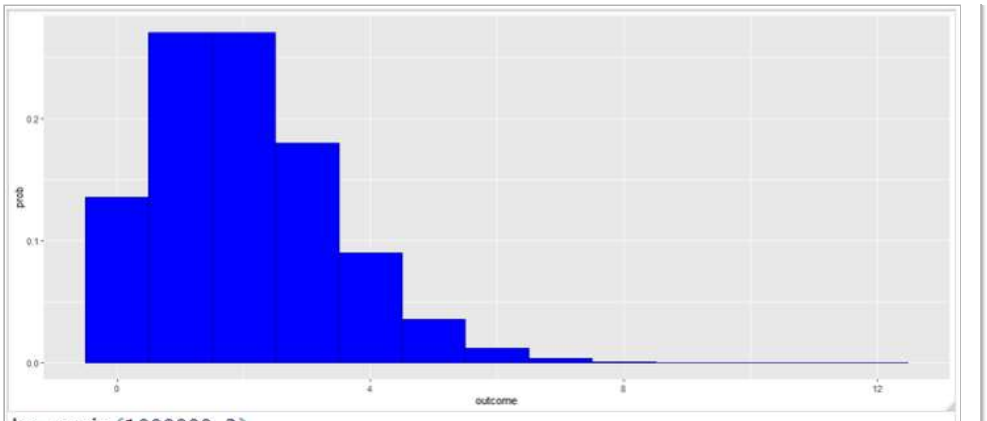
For this, we can use the rpois function, native to R. Figures 3.10 and 3.11 show Poisson distributions with *m* = 10 and *m* = 2.

Figure 3.10: *a~Pois*(10)




```
a<-rpois(1000000,10)
ggplot()+ geom_histogram(aes(x=a,y=..count../sum(..count..)),
   color="dark blue",fill="blue",binwidth=1)+ ylab("prob")+ xlab("outcome")
```

Figure 3.11: b~*Pois*(2)




```
b<-rpois(1000000,2)
ggplot()+ geom_histogram(aes(x=b,y=..count../sum(..count..)),
   color="dark blue",fill="blue",binwidth=1)+ ylab("prob")+ xlab("outcome")
```

Possible Outcomes - continuous over a finite range

Examples of situations where we have continuous outcomes over a finite range are:

- Number of productive hours during a day
- Interest rate change next month

Of distributions that fit this description, we will cover uniform, triangular, beta, and pert distributions.

For continuous distributions, I will use the geom_density function rather than the geom_histogram function, since geom_density better captures the notion of a continuous distribution.

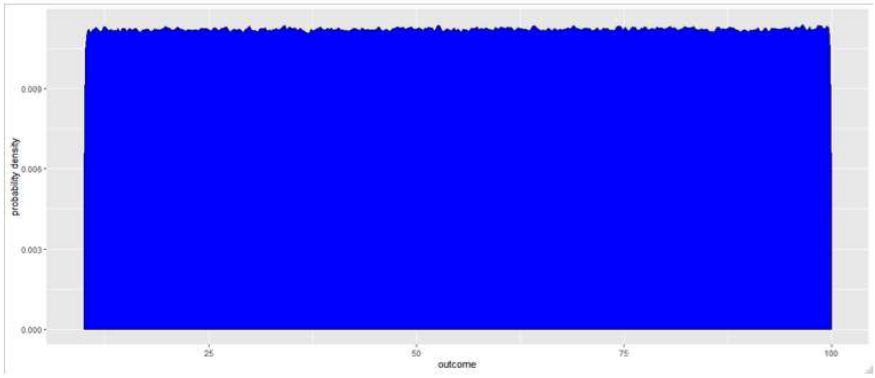
Uniform

A uniform distribution models complete uncertainty, since all outcomes are equally likely. It takes two parameters: the minimum possible outcome (min) and the maximum possible outcome (max). Min must be less than max.

Example:
A commercial pool supplystore sells chlorine by the ton. They anticipate a minimum demand of 10 tons and a maximum demand of 100 tons, with every intermediate amount equally likely.

To model a uniform distribution, we can use the runif function in R. (Here the adjust argument is used to mitigate R's density algorithm's inaccuracy at either end of the interval. This causes the random fluctuation over the rest of the interval. The result of the above example is displayed in Figure 3.12.

Figure 3.12: a~*Unif*(10, 100)




```
a<-runif(10000000,10,100)
ggplot()+ geom_density(aes(x=a),adjust=.1,
                       fill="blue")+ ylab("probability density")+ xlab("outcome")
```

Triangular Distribution

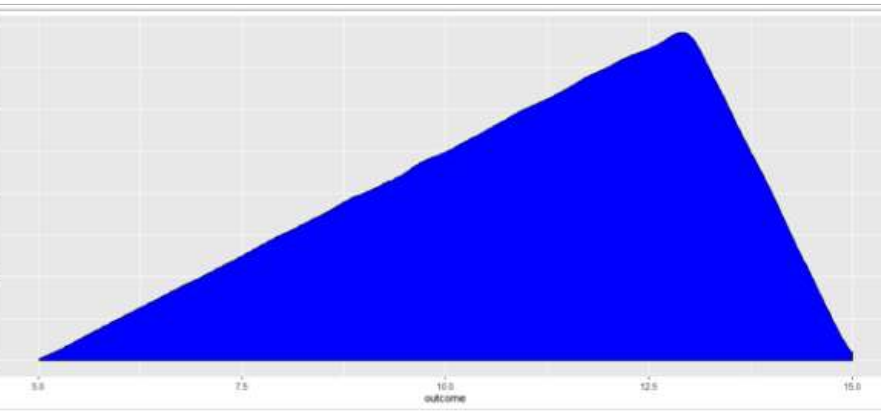
A triangular distribution models a process when **only** the minimum, most likely and maximum values of the distribution are known. The parameters are min, ml, and max, with *min* < *ml* < *max.*

Example:

If we are given the minimum, most likely, and maximum inflation rate we will have this year, then the random variable inflation rate can be modeled with a triangular distribution.

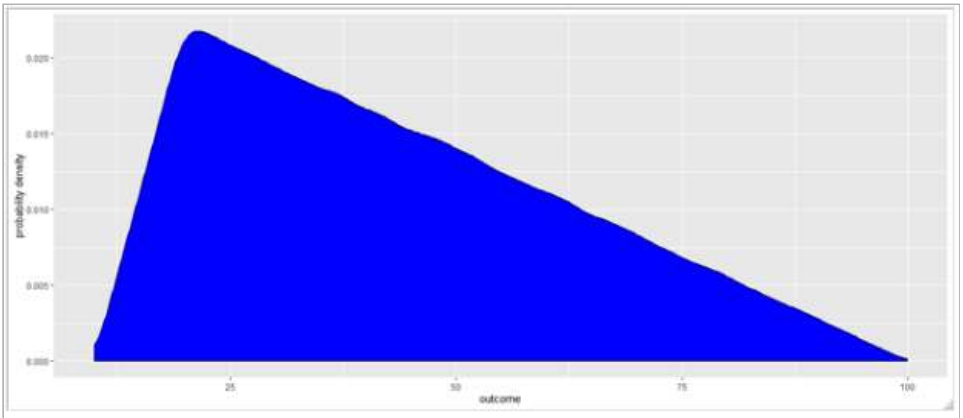
As in Lecture 2, since R has no triangular distribution in its base package, we use our own code to make one. We use the same syntax as before. Figures 3.13 and 3.14 show different triangular distributions.

Figure 3.13*: a~Tri*(5, 13, 15)




```
a<-rtri(1000000,5,13,15)
ggplot()+ geom_density(aes(x=a),fill="blue")+
  ylab("probability density")+ xlab("outcome")
```

Figure 3.14: *b~Tri*(10, 30, 100)




```
b<-rtri(1000000,10,20,100)
ggplot()+ geom_density(aes(x=b), fill="blue")+
  ylab("probability density")+ xlab("outcome")
```

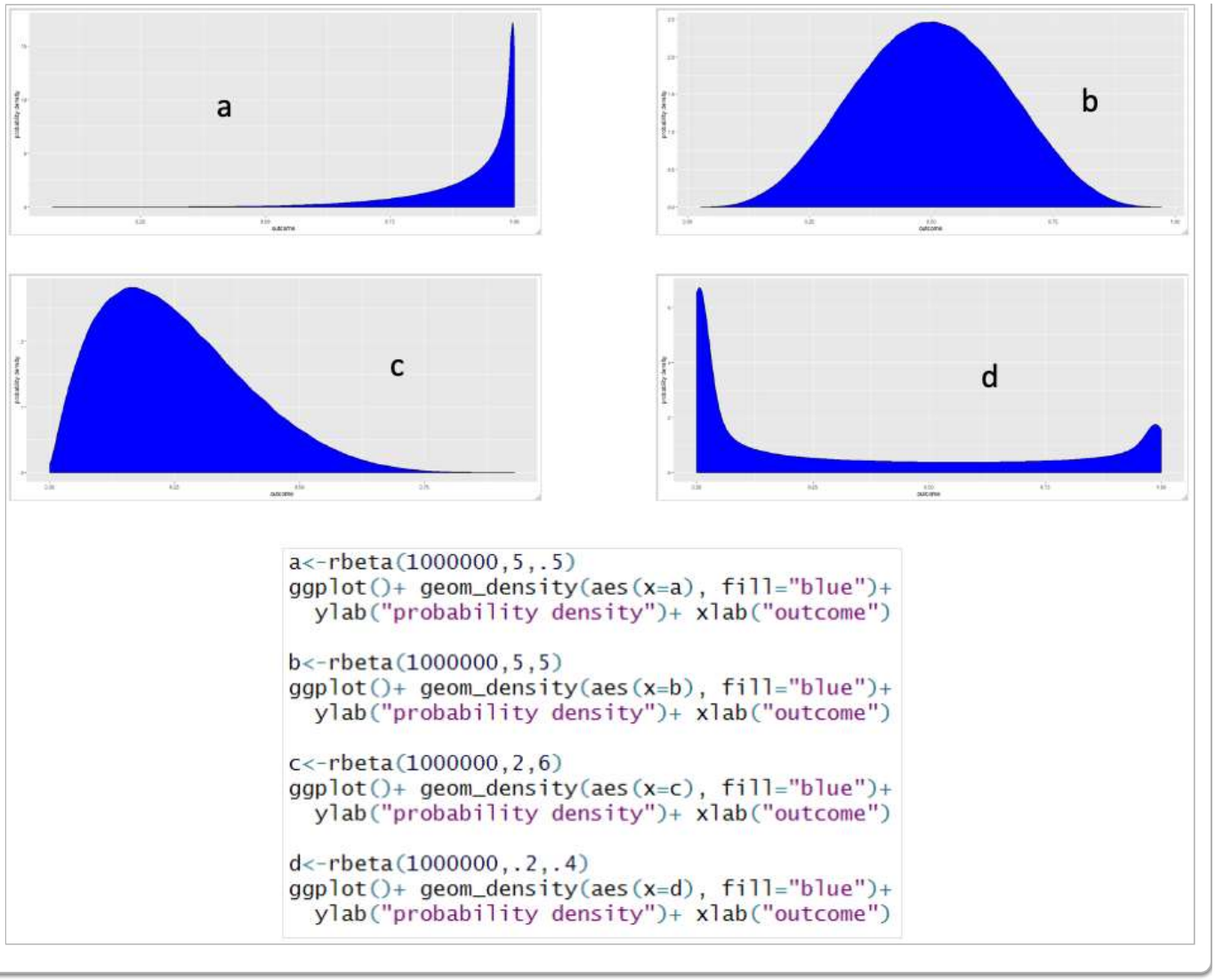
Beta Distribution

A beta distribution is an extremely flexible distribution used to model bounded (fixed upper and lower limits) random variables in the absence of data. It has two parameters: α and β, (alpha and beta) both of which must be greater than 0.

Example:
Proportion of defective items in a shipment.

A beta distributed random variable will only take on values between 0 and 1, but the shape can change drastically depending on α and β. The rbeta function can generate beta rv's. Figure 3.15 contains just some examples.

Figure 3.15




```
a<-rbeta(1000000,5,.5)
ggplot()+ geom_density(aes(x=a), fill="blue")+
  ylab("probability density")+ xlab("outcome")

b<-rbeta(1000000,5,5)
ggplot()+ geom_density(aes(x=b), fill="blue")+
  ylab("probability density")+ xlab("outcome")

c<-rbeta(1000000,2,6)
ggplot()+ geom_density(aes(x=c), fill="blue")+
  ylab("probability density")+ xlab("outcome")

d<-rbeta(1000000,.2,.4)
ggplot()+ geom_density(aes(x=d), fill="blue")+
  ylab("probability density")+ xlab("outcome")
```

Pert Distribution

The PERT distribution is used to model the activity times in project management problems and is defined by three point estimates, minimum value, most likely value and maximum value to complete an activity. Like the triangular distribution, *min* < *ml* < *max*.

Example:

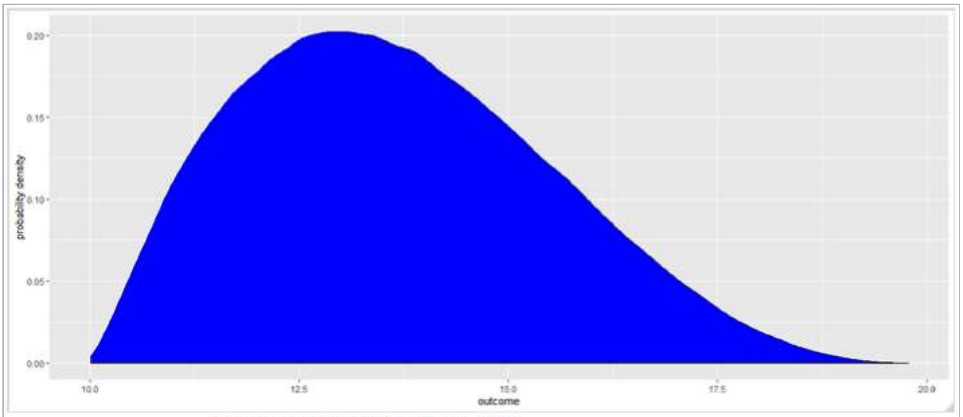
Time to complete a task in a PERT network.

Like the triangular distribution, the PERT distribution isn't in R's base package. It can be developed as a transformation of the Beta distribution. The code to do so is in Figure 3.16. An example of a PERT distribution is in Figure 3.17.

Figure 3.16

```
rpert<-function(n,min,ml,max,gam=4){
  y<-rbeta(n,1+gam*(ml-min)/(max-min),1+gam*(max-ml)/(max-min))
  min+y*(max-min)
}
```

Figure 3.17




```
a<-rpert(1000000,10,13,20)
ggplot()+ geom_density(aes(x=a), fill="blue")+
  ylab("probability density")+ xlab("outcome")
```

Possible Outcomes - Continuous with fixed lower bound and infinite upper bound

Examples of situations where we have continuous outcomes with fixed lower bound and infinite upper bound are (there is always some finite upper bound, but it may be quite large and unknown):

- Total time to complete a work order
- Time to failure of a machine or component

Next, we will cover the Exponential, Gamma, and Lognormal distributions.

Exponential Distribution

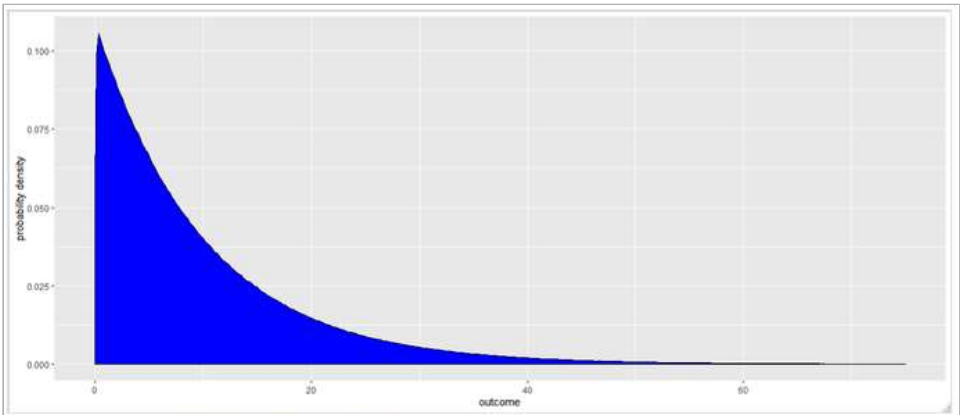
An exponential distribution models the time between independent events. It takes one parameter, which is the mean (expected value) of the distribution. It is sometimes parameterized as the multiplicative inverse of the mean (1/mean): this is the frequency of the distribution.

Example:

The time to failure for a system that has a constant failure rate over time is represented as an exponential random variable.

R generates random variables using the rexp function. It is parameterized by the frequency (in R, the “rate”) rather than the mean. Figure 3.18 contains an example of an exponential distribution with mean 10.

Figure 3.18




```
a<-rexp(10000000,1/10)
ggplot()+ geom_density(aes(x=a), fill="blue",adjust=.2)+
  ylab("probability density")+ xlab("outcome") +xlim(0,75)
```

Gamma Distribution

The gamma distribution is an extremely flexible distribution used to model nonnegative random variables. The gamma distribution has shape parameters: α and scale parameter β, with $\alpha, \beta > 0$.

Example:

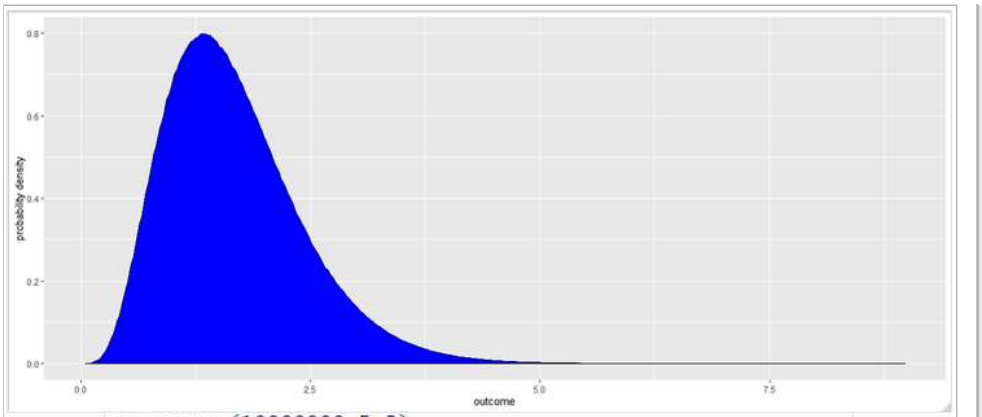
Time to complete some task, e.g., customer service or machine repair.

You can generate gamma distributed random variables using the rgamma function. Figure 3.19 contains an example of a gamma distribution with $\alpha = 5$, $\beta = 3$.

Individual Exercise:

Can you find discrete analogs for the exponential and gamma distributions?

Figure 3.19




```
a<-rgamma(10000000,5,3)
ggplot()+ geom_density(aes(x=a), fill="blue",adjust=.2)+
  ylab("probability density")+ xlab("outcome")
```

Lognormal Distribution

The lognormal distribution models the distribution of a process that can be thought of as the product of a number of component processes. The lognormal distribution is widely used for modeling service times, such as appointments. Note that the numbers generated by using lognormal distribution arealways positive. It is parameterized by its mean (m) and standard deviation (s), both of which must be greater than 0.

Example:

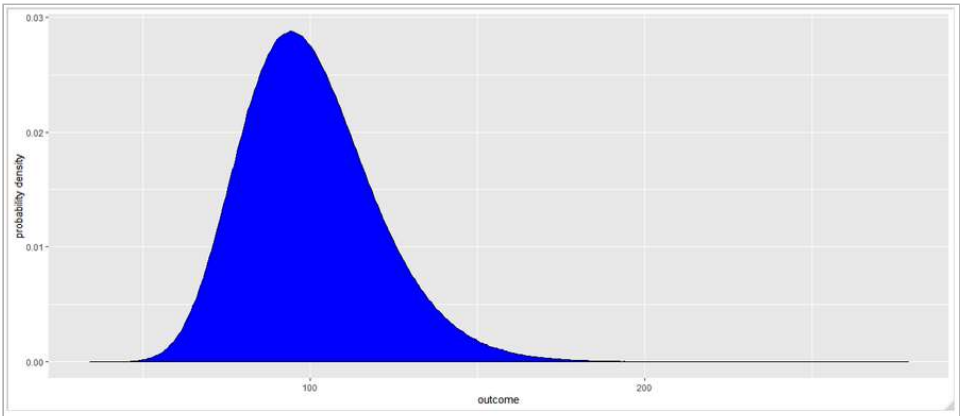
The rate of return on an investment, when interest is compounded, is the product of the returns for a number of periods. Therefore, the rate of return can be modeled as a lognormal random variable with a given mean and standard deviation

R parameterizes lognormal distributions by the mean and standard deviation of the *underlying normal distribution* rather than the lognormal distribution itself. This relationship will be described in further detail in class. In order to parameterize the lognormal distribution the way we desire, we write our own function, whose code is provided in Figure 3.20:

Figure 3.20

```
rlnorm2<-function(n,mean,sd){
  rlnorm(n,log(mean*(1+sd^2/mean^2)^-.5),log(1+sd^2/mean^2)^.5)
}
```

Figure 3.21: A lognormal distribution with mean 100 and standard deviation 20




```
a<-rlnorm(10000000,100,20)
ggplot()+ geom_density(aes(x=a), fill="blue",adjust=.2)+
  ylab("probability density")+ xlab("outcome")
```

Possible Outcomes - Continuous with fixed lower bound and infinite upper bound

Examples of situations where we have continuous outcomes with infinite upper **and** lower bounds are (there are always some finite bounds, but they may be quite large and unknown):

- Deviation of demand from forecast
- Deviation of an employee's performance from the average performance of all employees

Now we will cover the normal and Student distributions.

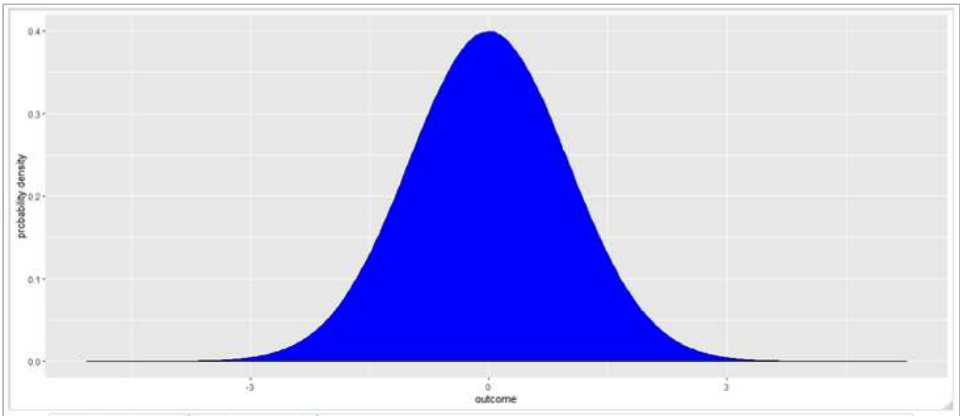
Normal Distribution

The normal distribution models errors of various types and quantities that are the sum of a large number of other quantities. Normal distributions are parameterized by their mean and standard deviation.

Example:
In human resource management, employee performance is often represented with a normal distribution.

As we've already seen, R generates random variables with normal distribution using the rnorm function. Figure 3.22 shows an example of the **standard normal distribution**, which has mean 0 and standard deviation 1.

Figure 3.22

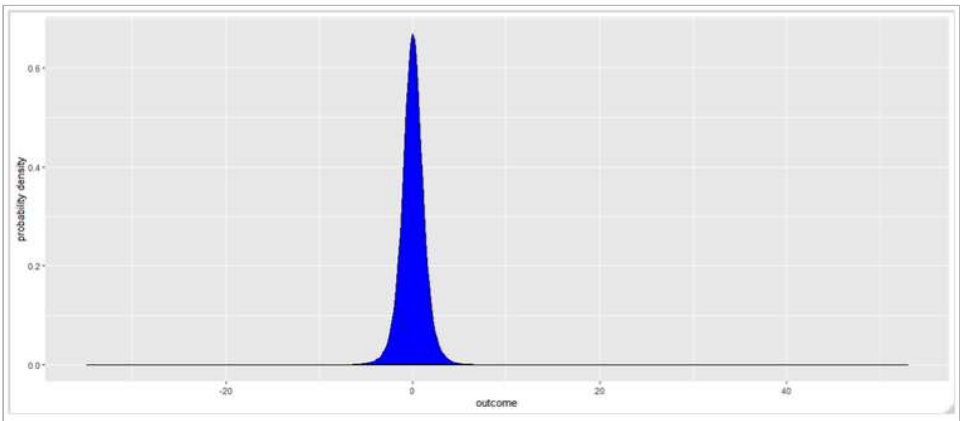



```
a<-rnorm(10000000)
ggplot()+ geom_density(aes(x=a), fill="blue",adjust=.2)+
  ylab("probability density")+ xlab("outcome")
```

Student Distribution

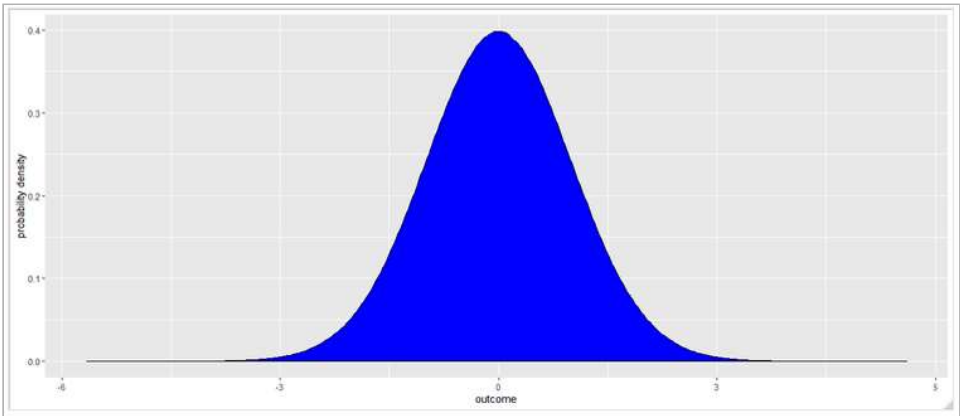
The Student distribution is very similar to normal, but with longer tails. It is also known as Student's t-distribution. Its only parameter is *v* the *degrees of freedom* of the distribution. It is always centered at 0. Figure 3.23 and Figure 3.24 present shapes of the Student distribution with *v* = 5 and 100. Note how the shape of the Student distribution resembles the normal distribution as the degrees of freedom (df) increases.

Figure 3.23




```
a<-rt(10000000,5)
ggplot()+ geom_density(aes(x=a), fill="blue")+
  ylab("probability density")+ xlab("outcome")
```

Figure 3.24




```
b<-rt(10000000,100)
ggplot()+ geom_density(aes(x=b), fill="blue")+
  ylab("probability density")+ xlab("outcome")
```

Relationships Among Distributions

In this part of the lecture, we will examine how different distributions relate to each other.

Binomial and Poisson

Individual Exercise 1:

We seek to demonstrate that as n gets high and p gets low for a binomial random variable, the resulting distribution approaches a Poisson distribution with $m = np$. Using the steps learned in this module, generate a histogram for a binomial distribution with t = 100,000 and p = .0001. Conduct 100,000 trials. Now generate a histogram for a Poisson distribution with m = 10. Compare the results.

Beta and Uniform

Individual Exercise 2:

We seek to demonstrate that a beta distribution with $\alpha = \beta = 1$ is identical to a uniform distribution with min 0 and max 1. Generate a density plot or histogram for a beta distribution with these parameters. Compare it to the plot for a uniform distribution. (Hint: as previously stated, the geom_density and geom_histogram functions do a poor job of modeling the end points of a distribution's intervals. You can mitigate this by setting the adjust argument or your bin width to a low number, but this will make your distribution appear less smooth. This effect can, in turn, be mitigated with a large number of trials.

Gamma and Exponential

Individual Exercise 3:

We seek to demonstrate that, when the shape parameter $\alpha = 1$ in a gamma distribution, the result is an exponential distribution with mean equal to the scale parameter. Create histograms or density plots for both distributions to test this theory.

Student and Normal

Individual Exercise 4:

Create histograms or density plots for Student distributions with increasing degrees of freedom to test the theory that, as the degrees of freedom increase, the Student distribution begins to resemble a normal distribution.

Exponential and Poisson

We will discuss this further in class.

Lognormal and Normal

If you take the logarithm of a lognormal distribution, the result will be normally distributed. We will discuss this relationship further in class.

Understanding Distributions

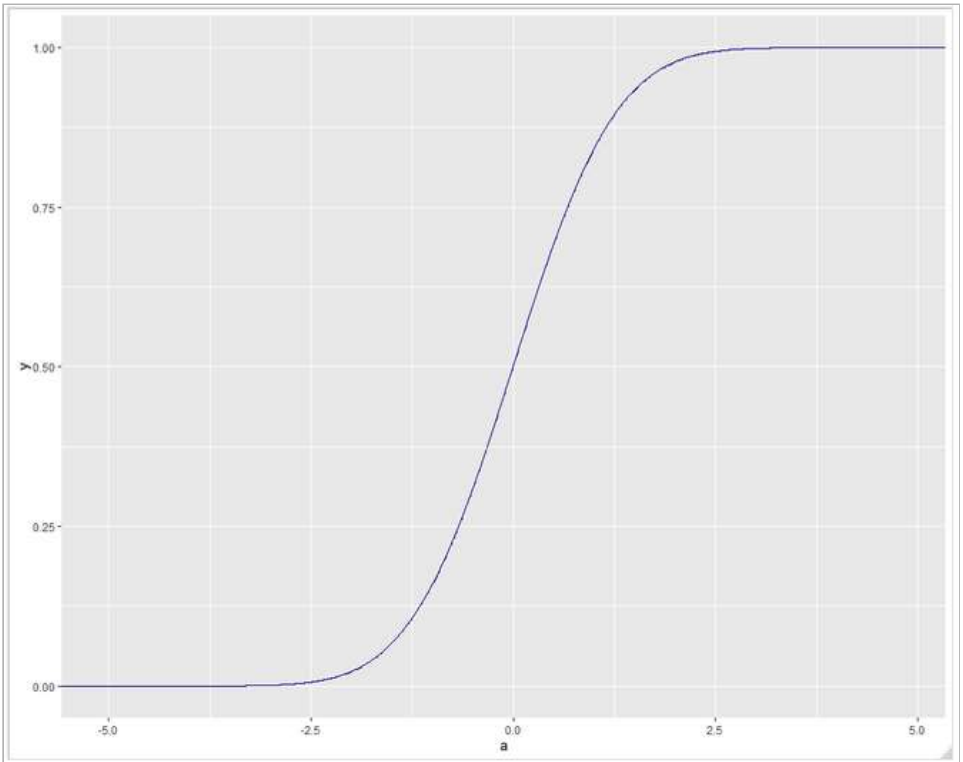
The **probability density function (pdf)** (for continuous outcomes) or **probability mass function (pmf)** (for discrete outcomes) gives the probability of specific values. The mass function allows us to answer questions of the type: What is the probability that the number of customers will be exactly 4? In the continuous case, this relationship is more complicated, since the probability of hitting any outcome exactly is held to be 0.

The **cumulative distribution function (cdf)** gives the probability of being less than or equal to a value. With a cdf, for instance, we can find the probability that an interest rate will be ≤ 8.3%. We can also use the cdf to determine the likelihood that an interest rate falls between two values, by taking the difference of two cdfs.

It is important to note that so far in this module, we have created pdfs and pmfs. However, it is not difficult to obtain a cdf.

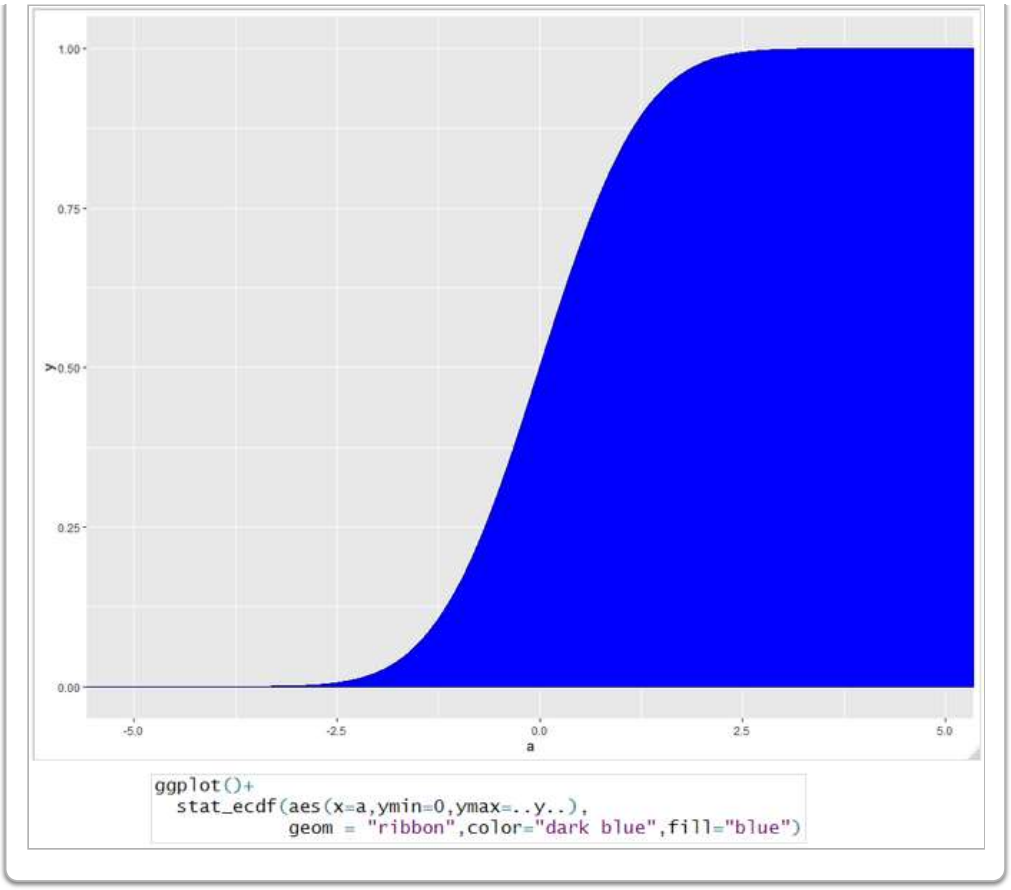
In Figures 3.25 and 3.26, you can see two different aesthetic approaches to creating a cdf for a standard normal distribution, using the stat_ecdf function from ggplot2. Use whichever you find easier or more visually appealing.

Figure 3.25




```
ggplot()+
  stat_ecdf(aes(x=a),geom = "step",color="dark blue")
```

Figure 3.26



Summary of Distributions

Distribution	R Function to randomly generate	Discrete or Continuous?	Bounded?
Beta	rbeta (n,alpha, beta)	Continuous	Yes
Binomial	rbinom(n,trials, probability)	Discrete	Yes
Discrete Uniform*	rdunif(n,min,max)	Discrete	Yes
Exponential	rexp(n,mean)	Continuous	Below
Gamma	rgamma(n, alpha, beta)	Continuous	Below
Geometric	rgeom(n,probability)	Discrete	Below
Lognormal*	rlnorm2 (n,mean, standard deviation)	Continuous	Below
Negative Binomial	rnbinom(n, successes, probability)	Discrete	Below

Normal	rnorm (n, mean, standard deviation)	Continuous	No
Pert*	rpert(n, min, most likely, max)	Continuous	Yes
Poisson	rpois (n, mean)	Discrete	Below
Student t	rt(n, degrees of freedom)	Continuous	No
Triangular*	rtri(n,min, most likely, max)	Continuous	Yes
Uniform	runif (n, min, max)	Continuous	Yes

* user defined function in R

Lecture 3 Footnotes

[1] Camm, J. D., Cochran, J. J., Fry, M. J., Ohlmann, J. W., Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2015). *Essentials of business analytics* (1st edition). Cengage Learning, p. 548.

[2] Camm, J. D., Cochran, J. J., Fry, M. J., Ohlmann, J. W., Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2015). *Essentials of business analytics* (1st edition). Cengage Learning, p. 548.

Lecture 3 References

Camm, J. D., Cochran, J. J., Fry, M. J., Ohlmann, J. W., Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2015). *Essentials of business analytics* (1st edition). Cengage Learning.

Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis* (3rd edition). McGraw-Hill Higher Education, pp. 299–318. (Note: This is a good reference book for the concepts discussed in this module.)

Lecture 3 Summary Questions

1. What is the basis of categorizing probability distributions as “discrete” or “continuous”?

2. In representing an uncertain variable in a simulation study, how do we choose among different probability distributions?
3. Give examples of continuous and discrete probability distributions.
4. Evaluate the following statements:
    a. If we know only the minimum, maximum, and most likely values of a process, then we can use the triangular distribution to model this process.
    b. If we know only the mean and standard deviation of a process, then normal distribution can be used to model this process.
    c. PERT distribution is particularly useful when modeling project activity times.
5. What is the difference between a “probability density function” and a “cumulative distribution function”?