Chapter 6

Distributions

6.1 Introduction

Now that we've seen many examples of estimated distributions when we plotted data. In what follows, we can introduce the probability distributions these graphs may have been estimating. We will cover many types of random variables, and their named distributions. For each distribution, we will discuss the types of scenarios they model and important aspects of the distribution itself. Luckily, many of the functions we'll discuss are cataloged by `R` and so probabilities associated with these models can be calculated and graphed easily. For distributions that don't have available functions in `R` we will discuss how to write code that evaluates probabilities of interest.

Definition 6.1. We are interested in defining an event in terms of a random variable, X. Let X be a random variable that takes on values in $\mathcal{X} \subset \mathbb{R}$. We can consider the following events around a known constant x.

$$\begin{aligned}
X = x &= \text{“}X \text{ is exactly } x\text{”} \\
X \neq x &= \text{“}X \text{ is not equal to } x\text{”} \\
X \leq x &= \text{“}X \text{ is at most } x\text{”} \\
X \geq x &= \text{“}X \text{ is at least } x\text{”} \\
X < x &= \text{“}X \text{ is less than } x\text{”} \\
X > x &= \text{“}X \text{ is greater than } x\text{”}
\end{aligned}$$

Additionally, we can consider the following events around known constants $x_1 < x_2$.

$$\begin{aligned}
x_1 < X < x_2 &= \text{“}X \text{ is between than } x_1 \text{ and } x_2 \text{ (exclusive)”} \\
x_1 < X \leq x_2 &= \text{“}X \text{ is greater than } x_1 \text{ and at most } x_2\text{”} \\
x_1 \leq X < x_2 &= \text{“}X \text{ is at least } x_1 \text{ and less than } x_2\text{”} \\
x_1 \leq X \leq x_2 &= \text{“}X \text{ is between than } x_1 \text{ and } x_2 \text{ (inclusive)”}
\end{aligned}$$

Definition 6.2. The **cumulative distribution function** or CDF of a random variable X, denoted by $F_X(x)$, is defined, for all x, by

$$F_X(x) = P_x(X \leq x).$$

Definition 6.3. The p^{th} percentile of a distribution is the observation $x \in \mathcal{X}$ such that p percent of observations are below $X = x$; e.g.,

$$p^{th}\text{percentile} = \inf\left\{x \in \mathbb{R} : F_X(x) \geq \frac{p}{100}\right\}.$$

Remark: With discrete distributions it is sometimes necessary to take the smallest observation $x \in \mathcal{X}$ such that at least p percent of observation are below $X = x$ as the p^{th} percentile. The infimum (inf) of a set is the greatest lower bound of that set; e.g., m is an infimum of a set if no member of the set is less than m.
Recall: The CDF, for discrete and continuous random variables, returns the percentile ($P(X \leq x)$) for an observation $x \in \mathcal{X}$.

Definition 6.4. The **inverse cumulative distribution function** or inverse CDF of a random variable X, denoted by $F_X^{-1}(p)$, is defined, for all $p \in [0, 1]$, by

$$F_X^{-1}(p) = \inf\left\{x \in \mathbb{R} : F_X(x) \geq \frac{p}{100}\right\}.$$

We note that the p^{th} percentile can be calculated by

$$F_X^{-1}\left(\frac{p}{100}\right) = x.$$

While, with this notation, finding a percentile appears to be simple, we note that many of the named distributions we cover in the next section do not have a closed form CDF. Due to these circumstances, percentiles are often calculated numerically; we will lean on `R` to do such calculations.

Definition 6.5. The **probability mass function (PMF)** of a discrete random variable X is given by

$$f_X(x) = P(X = x) \text{ for all x.}$$

The **CDF** of a discrete random variable X is given by

$$F_X(x) = P(X \leq x) = \sum_{-\infty}^{x} f_X(x).$$

For this reason, for a PMF to be valid the following two things must be true.

1. $0 \leq f_X(x) \leq 1$ for all $x \in \mathbb{R}$
2. $\sum_{-\infty}^{\infty} f_X(x) = \sum_{\mathcal{X}} f_X(x) = 1$

Remark: Often our discrete random variables have support of $\mathbb{N}$, $\mathbb{Z}$, or some subset of these values. In this case, we can consider the CDF evaluated at $x.x_0x_1x_2$ where x_i represents the value in the ith decimal place as being equal to the CDF evaluated at x; e.g.,

$$P(X \leq x.x_0x_1x_2) = P(X \leq x).$$

This is true because there are no observable values on $(x, x+1)$ and thus these events are the same. Noting that $x = \lfloor x.x_0x_1x_2 \rfloor$, we can evaluate the CDF as

$$F_X(x) = P(X \leq \lfloor x \rfloor) = \sum_{-\infty}^{\lfloor x \rfloor} f_X(x)$$

Definition 6.6. The **probability density function (PDF)** of a continuous random variable X is given by $f_X(x)$ such that

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

Note: The PDF is unlike the PMF in that $f_X(x) \neq P(X = x)$ because for continuous random variables $P(X = x) = 0$ for all $x \in \mathbb{R}$.

The **CDF** of a continuous random variable X is given by

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

For this reason, for a PMF to be valid the following two things must be true.

1. $0 \geq f_X(x)$ for all $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x) = \int_{\mathcal{X}} f_X(x) = 1$

Remark: Here, the PDF can be greater than 1 because the values do not represent probabilities; the constraint is that the area under $f_X(x)$ is 1.

Definition 6.7. A random variable, X, is **continuous** if $F_X(x)$ is a continuous function of x and **discrete** if $F_X(x)$ is a discontinuous step function of x.

Definition 6.8. Random variables, X and Y, are **identically distributed** if $F_X(x) = F_Y(x)$ for all x and $F_X(y) = F_Y(y)$ for all y.

6.2 Discrete Distribution Functions

Finding probability for the events described in Definition 6.1 requires some clever usage of our tools for calculating probability. For a discrete random variable, these tools are

$$f_X(x) = P(X = x) \qquad \textbf{[PMF]}$$
$$F_X(x) = P(X \leq x). \qquad \textbf{[CDF]}$$

For a known constant $x_1 \in \mathcal{X} = \mathbb{Z}$,

$$\begin{aligned}
P(X = x_1) &= f_X(x_1) && \textbf{[Definition of PMF]}\\
P(X \neq x_1) &= 1 - f_X(x_1) && \textbf{[Complement rule]}\\
P(X \leq x_1) &= F_X(x_1) && \textbf{[Definition of CDF]}\\
P(X \geq x_1) &= 1 - P(X < x_1) && \textbf{[Complement rule]}\\
&= 1 - P(X \leq x_1 - 1) && [(x_1 - 1) \textbf{ is the next smallest observation}]\\
&= 1 - F_X(x_1 - 1) && \textbf{[Definition of CDF]}\\
P(X < x_1) &= P(X \leq x_1 - 1) && [(x_1 - 1) \textbf{ is the next smallest observation}]\\
&= F_X(x_1 - 1) && \textbf{[Definition of CDF]}\\
P(X > x_1) &= 1 - P(X \leq x_1). && \textbf{[Complement rule]}\\
&= 1 - F_X(x_1) && \textbf{[Definition of CDF]}
\end{aligned}$$

Remark: To write these probabilities in terms of a PMF or CDF requires us to consider $P(X < x_1) = P(X \leq x_1 - 1)$. This is true when $x_1 \in \mathcal{X}$ because the $(x_1 - 1)$ is the next smallest observation and so the event of interest can be written as

$$A = \{x \in \mathcal{X} : x < x_1\} = \{..., x - 2, x - 3, x - 1\}.$$

Additionally, we can consider the following events around known constants $x_1 < x_2 \in \mathcal{X}$.

$$\begin{aligned}
P(x_1 < X < x_2) &= P(X < x_2) - P(X \leq x_1)\\
&= P(X \leq x_2 - 1) - P(X \leq x_1)\\
&= F_X(x_2 - 1) - F_X(x_1) && \textbf{[Definition of CDF]}\\
P(x_1 < X \leq x_2) &= P(X \leq x_2) - P(X \leq x_1)\\
&= F_X(x_2) - F_X(x_1) && \textbf{[Definition of CDF]}\\
P(x_1 \leq X < x_2) &= P(X < x_2) - P(X < x_1)\\
&= P(X \leq x_2 - 1) - P(X \leq x_1 - 1)\\
&= F_X(x_2 - 1) - F_X(x_1 - 1) && \textbf{[Definition of CDF]}\\
P(x_1 \leq X \leq x_2) &= P(X \leq x_2) - P(X < x_1)\\
&= P(X \leq x_2) - P(X \leq x_1 - 1)\\
&= F_X(x_2) - F_X(x_1 - 1) && \textbf{[Definition of CDF]}
\end{aligned}$$

Generalization: To use this technique more broadly, e.g. when $x \in \mathbb{R}$, we use the following relationships in the techniques shown above.

$$P(X \leq x) = F_X(\lfloor x \rfloor)$$
$$P(X < x) = \begin{cases} P(X \leq x - 1) = F_X(x - 1) & \text{if } \lfloor x \rfloor = x \\ P(X \leq \lfloor x \rfloor) = F_X(\lfloor x \rfloor) & \text{if } \lfloor x \rfloor > x \end{cases}$$

6.2.1 Named Distribution Functions

There are many named distribution functions which describe probabilities about random variables. Each named distribution covered here describes a particular type of random experiment. Often of interest is an event determined by the value of a random variable X that describes a set of sample points.

If Ω is countable we have a discrete random variable, X, with countable support $\mathcal{X}$. This random variable can be modeled by a PMF and a step-wise CDF. We say that X is distributed as $f_X(x)$, $X \sim f_X(x)$, if X has PMF $f_X(x)$. Below we denote several named discrete distributions which cover a range of random experiments. Each named distribution is specified by its PMF, and CDF.

Definition 6.9. The **Bernoulli distribution** is a discrete distribution used for the random variable, X, which can take on values one with probability p and zero with probability $(1-p)$. We often let the outcome $X = 1$ represent that an event of interest occurred, which we term a "success", and the outcome $X = 0$ represent that the event of interest did not occur, which we term a "failure." Graphs of the PMF and CDF are seen in Figure 6.2.1.

$$
\begin{aligned}
p &\in (0,1) && \textbf{[Parameter]}\\
\mathcal{X} &= \{\omega : \omega \in \{0,1\}\} && \textbf{[Support]}\\
f_X(x|p) &= p^x(1-p)^{1-x} I(x \in \{0,1\}) && \textbf{[PMF]}\\
F_X(x|p) &= P(X \leq \lfloor x \rfloor)\\
&= [(1-p)I(\lfloor x \rfloor = 0)] + I(\lfloor x \rfloor \geq 1) && \textbf{[CDF]}\\
F_X^{-1}(u|p) &= \begin{cases} 0 & \text{if } 0 < u \leq (1-p) \\ 1 & \text{if } (1-p) < u \leq 1 \end{cases}\\
E(X) &= p && \textbf{[Expected Value]}\\
var(X) &= p(1-p) && \textbf{[Variance]}
\end{aligned}
$$

Remark: The CDF and inverse CDF are simple to write out since X can only take on two values, zero and one.

R doesn't have the Bernoulli distribution distribution built in. We can, however, create functions to calculate these probabilities about X in R as follows.

```
> dbern<-function(x,prob){
+  if(prob<0 | prob>1){
+  errormsg<-"This function is only valid for probabilities between 0 and 1."
+  stop(errormsg)
+  }
+  indicator<-rep(0,length(x))
+  indicator[x==0]<-1 #indicator should be one if x=0
+  indicator[x==1]<-1 #indicator should be one if x=1
+  fx<-(prob^x*(1-prob)^(1-x))*indicator
+  return(fx)
+ }
> pbern<-function(q,prob){
+  if(prob<0 | prob>1){
+  errormsg<-"This function is only valid for probabilities between 0 and 1."
```

```
+  stop(errormsg)
+  }
+  indicator1<-rep(1,length(q))
+  indicator1[q!=0]<-0 #indicator should be zero if x!=0
+  indicator2<-rep(1,length(q))
+  indicator2[q<1]<-0 #indicator should be zero if x<1
+  Fx<-(1-prob)*indicator1 + indicator2
+  return(Fx)
+ }
> qbern<-function(p,prob){
+  if(prob<0 | prob>1){
+  errormsg<-"This function is only valid for probabilities between 0 and 1."
+  stop(errormsg)
+  }
+  if(any(p<0)|any(p>1)){
+     errormsg<-"This function is only valid for percentiles between 0 and 1."
+     stop(errormsg)
+   }
+  q<-rep(1,length(p))
+  q[p<=1-prob]<-0 #should return zero if u<=(1-p)
+  q[p>1-prob]<-1 #should return one if u>(1-p)
+  return(q)
+ }
> dbern(x=0,prob=0.25) #P(X=0|p=0.25)
[1] 0.75
> pbern(q=0,prob=0.25) #P(X<=0|p=0.25)
[1] 0.75
> qbern(p=0.7,prob=0.25)
[1] 0
```

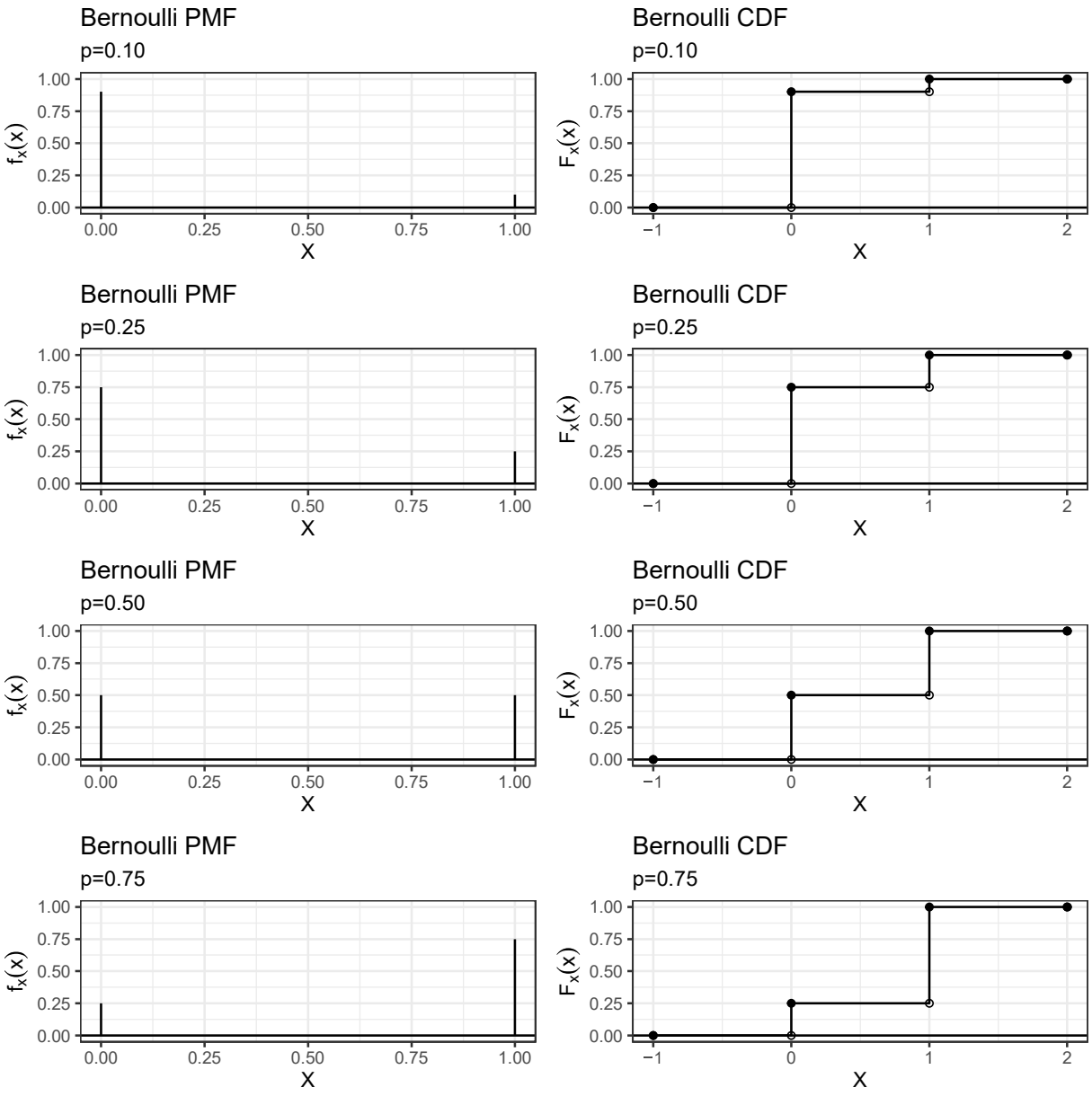

Figure 6.2.1: Bernoulli PMFs(left) and CDFs(right) for varying values of the parameter p.

The graph of the PMF and CDF for the binomial can be completed in R as follows. Below, we provide the code for graphing when $p = 0.10$, but we provide several examples in Figure 6.2.1.

```
> ggdat<-data.frame(x=(-1:2),
+ f1=dbern(x=(-1:2),prob=0.10),
+ F1=pbern(q=(-1:2),prob=0.10))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,1)+
+   ylim(0,1)+
```

```
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Bernoulli PMF",subtitle="p=0.10")
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=pbinom(ggdat$x-1,size=1,prob=0.10))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                              y=pbinom(ggdat$x,size=1,prob=0.10))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Bernoulli CDF",subtitle="p=0.10")
> grid.arrange(g1,g1.CDF,ncol=2)
```

Remark: We plot just before and just beyond the support of the distribution to see the "full picture."

For the Bernoulli distribution, we can randomly sample the zeros and ones according to the probability as evaluated via the PMF. To generate a random sample $X = (X_1, X_2, \ldots, X_n)$ from this distribution for fixed $p \in [0, 1]$ we can use the following function in R.

```
> rbern<-function(n,p){
+   support<-c(0,1)
+   x<-sample(x=support,size=n,replace=TRUE,prob=dbern(support,p=p))
+   return(x)
+ }
> rbern(n=10,p=0.25) #A random sample of 10 Xi~Bernoulli(p=0.25)
[1] 0 0 0 0 0 1 0 1 0 1
```

Definition 6.10. The **Binomial distribution** is a discrete distribution used for the random variable, X, which represents the number of "successes" in n identical and independent trials with binary output. This random variable represents a sum of n Bernoulli random variables, i.e. X is the number of n Bernoulli trials which yield an event of interest or "success;" e.g.,

$$Y_i \sim \text{Bernoulli}(p) \text{ for } i = 1, 2, \ldots, n$$
$$X = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, p).$$

Graphs of the PMF and CDF are seen in Figure 6.2.2.

$$
\begin{aligned}
p \in [0,1] \text{ and } n \in \mathbb{N} & \quad \textbf{[Parameters]} \\
\mathcal{X} = \{\omega : \omega \in \{0,1,\dots n\}\} & \quad \textbf{[Support]} \\
f_X(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x} I(x \in \{0,1,\dots n\}) & \quad \textbf{[PMF]} \\
F_X(x|n,p) = P(X \leq \lfloor x \rfloor) & \\
= \left[\sum_{x=0}^{\lfloor x \rfloor} \binom{n}{x} p^x (1-p)^{n-x} \right] I(x \in \{0,1,\dots n\}) + I(\lfloor x \rfloor > n) & \quad \textbf{[CDF]} \\
E(X) = np & \quad \textbf{[Expected Value]} \\
var(X) np(1-p) & \quad \textbf{[Variance]}
\end{aligned}
$$

We won't simplify the CDF any further here because doing so would require convoluted functions like the incomplete Beta function; instead, we will calculate the CDF and inverse CDF numerically using functions in R.

```
> dbinom(x=1,prob=0.25,size=10) #P(X=1|n=10,p=0.25)
[1] 0.1877117
> pbinom(q=1,prob=0.25,size=10) #P(X<=1|n=10,p=0.25)
[1] 0.2440252
> qbinom(p=0.2,prob=0.25,size=10) #The 20th percentile
[1] 1
> rbinom(n=10,prob=0.25,size=10) #A random sample of 10 Xi~Binomial(n=10,p=0.25)
[1] 3 1 0 0 3 5 3 3 3 3
```

The graphs of the PMF and CDF, in Figure 6.2.2, can be completed in R as follows.

```
> ggdat<-data.frame(x=(-1:11),
+                   f1=dbinom(x=(-1:11),size=10,prob=0.10),
+                   F1=pbinom(q=(-1:11),size=10,prob=0.10))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,10)+
+   ylim(0,0.50)+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Binomial PMF",subtitle="n=10, p=0.10")
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=pbinom(ggdat$x-1,size=10,prob=0.10))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                              y=pbinom(ggdat$x,size=10,prob=0.10))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
```

```
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Binomial CDF",subtitle="n=10, p=0.10")
> grid.arrange(g1,g1.CDF,ncol=2)
```

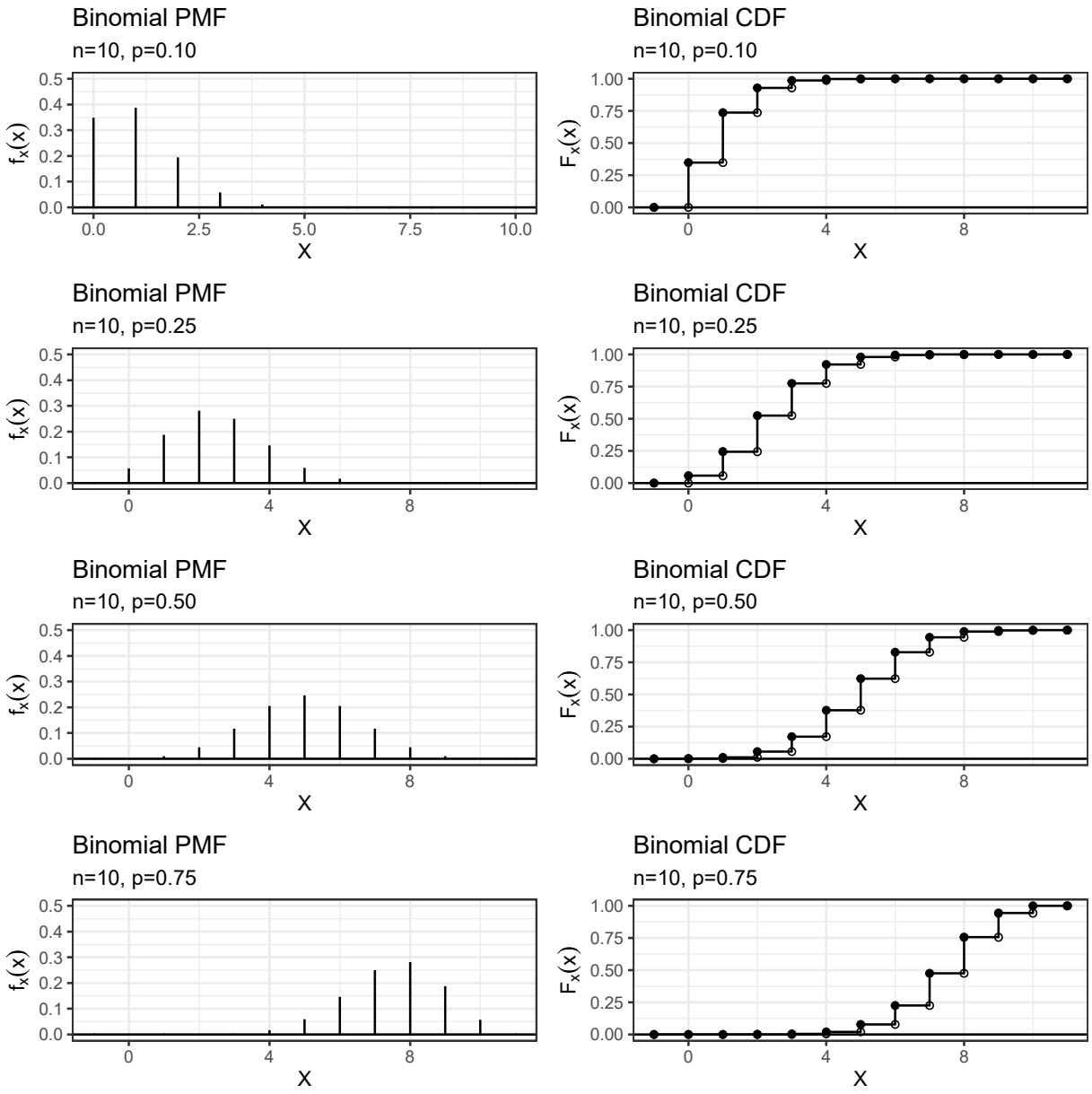

Figure 6.2.2: Binomial PMFs(left) and CDFs(right) for $n = 10$ and varying values of the parameter p.

Example 6.11. When manufactured items are tested, the long-run proportion of defective items is 0.05 for a particular item. Each item tested can be considered a trial, finding a defective item is considered a "success" and the probability of success is 0.05. Note, that here, a "success" is

something we might consider a failure in real life and it might be more appropriate to call it an event of interest instead of a "success." Suppose the quality analysis (QA) group at the plant manufacturing these items decides to test $n = 100$ items.

Question: What is the probability that the QA group finds exactly one defective item in the $n = 100$ items?
Answer:

$$\begin{aligned} P(X = 1) &= \binom{100}{1}(0.05)^1(1 - 0.05)^{100-1} && \textbf{[Binomial PMF]} \\ &= 100(.95)^{99}(0.05)^1 && \textbf{[Simplifying]} \\ &\approx 0.03116 \end{aligned}$$

This can be calculated in R as follows.

```
> dbinom(x=1,size=100,prob=0.05)#P(X=1|n=100,p=0.05)
[1] 0.03116068
```

We can also visualize this probability by graphing the PMF using R and highlighting the probability of interest in the graph as seen in Figure 6.2.3.

```
> ggdat<-data.frame(x=(-1:15), #a subset of the range
+                   fx=dbinom(x=(-1:15),size=100,prob=0.05))
> ggdat.highlight<-data.frame(x=1,fx=dbinom(x=1,size=100,prob=0.05))
> ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=fx), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(x=x,ymax=fx),color="red", ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,15)+
+   xlab("Defective Items (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Binomial PMF",subtitle = bquote(P(X==1)~"for n=100, p=0.05"))
```

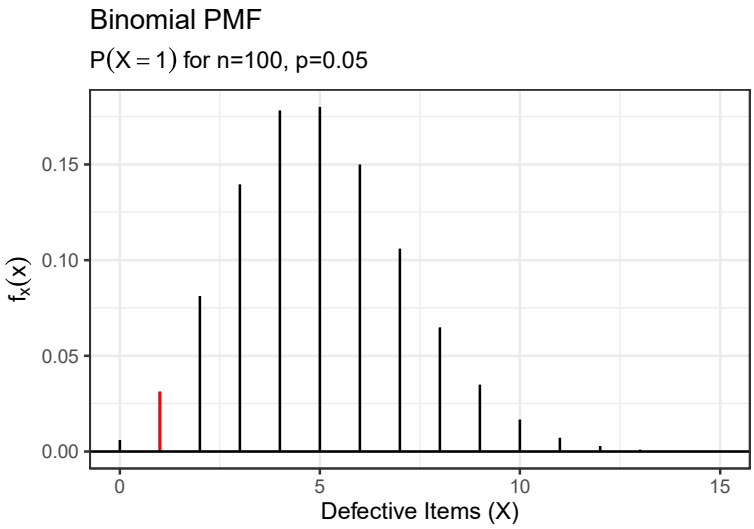

Figure 6.2.3: The Binomial PMF with $n = 100$ and $p = 0.05$ as in Example 6.11. The observation at $X = 1$ is highlighted in red; the height of this bar is equal to $P(X = 1|n = 100, p = 0.05)$.

Question: What is the probability that the QA group finds at most five defective items in the $n = 100$ items?
Answer:

$$P(X \leq 5) = \sum_{x=0}^{5} \binom{100}{x} (0.05)^x (1 - 0.05)^{100-x} \qquad \textbf{[Binomial CDF]}$$
$$\approx 0.6159991$$

This cumbersome sum can be calculated in R as follows.

```
> pbinom(q=5,size=100,prob=0.05)#P(X<=5|n=100,p=0.05)
[1] 0.6159991
```

We can also visualize this probability by graphing the PMF or CDF in R and highlighting the probability of interest in the graph as seen in Figure 6.2.4.

```
> ggdat<-data.frame(x=(-1:15), #a subset of the range
+                   fx=dbinom(x=(-1:15),size=100,prob=0.05),
+                   Fx=pbinom(q=(-1:15),size=100,prob=0.05))
> ggdat.highlight<-data.frame(x=0:5,fx=dbinom(x=0:5,size=100,prob=0.05))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=fx), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(x=x,ymax=fx),color="red", ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,15)+
+   xlab("Defective Items (X)")+
+   ylab(bquote(f[x](x)))+
```

```
+   ggtitle("Binomial PMF",subtitle = bquote(P(X<=5)~"for n=100, p=0.05"))
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=pbinom(ggdat$x-1,size=100,prob=0.05))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=pbinom(ggdat$x,size=100,prob=0.05))
> ggdat.highlight<-data.frame(x=5,y=pbinom(5,size=100,prob=0.05))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = Fx)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_point(data = ggdat.highlight, aes(x = x, y = y),color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Defective Items (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Binomial CDF",subtitle = bquote(P(X<=5)~"for n=100, p=0.05"))
> grid.arrange(g1,g1.CDF,ncol=2)
```

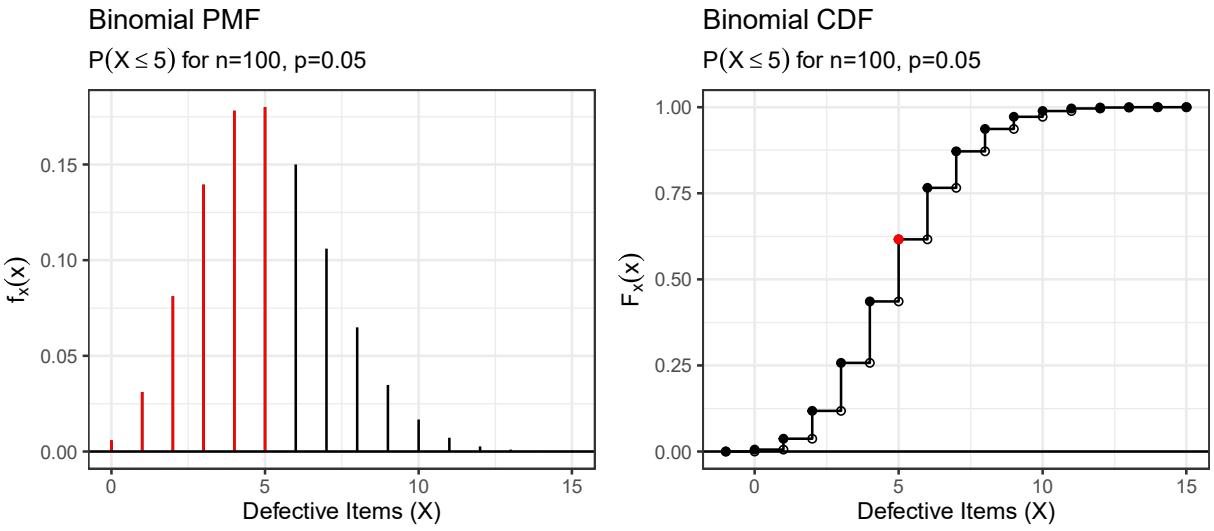

Figure 6.2.4: The Binomial PMF (left) and CDF (right) with $n = 100$ and $p = 0.05$ as in Example 6.11. The observations from $X = 0$ to $X = 5$ are highlighted in the PMF in red; the sum of the heights of these bars is equal to $P(X \leq 5|n = 100, p = 0.05)$. The observation at $X = 5$ is highlighted in the CDF with a red point; the corresponding $F_X(x)$ value is equal to $P(X \leq 5|n = 100, p = 0.05)$.

Question: What observation is at the 10^{th} percentile?
Answer:

$$x = F_X^{-1}(0.10|n = 100, p = 0.05)$$

We can ask `R` to solve this question since we don't have a closed form solution to the inverse CDF as follows.

```
> qbinom(p=0.10,size=100,prob=0.05)#The 10th percentile
[1] 2
```

Remark: This percentile is one such case where we're taking the minimum observation $x \in \mathcal{X}$ such that at least p percent of observations are below $X = x$. Due to the discrete nature of X, we take the smallest value x such that the CDF evaluate at x is at least p.

```
> pbinom(q=1,size=100,prob=.05)#P(X<=1|n=100,p=0.05)
[1] 0.03708121
> pbinom(q=2,size=100,prob=.05)#P(X<=2|n=100,p=0.05)
[1] 0.118263
```

We can also visualize this percentile by graphing the CDF in R and highlighting the percentile of interest in the graph as seen in Figure 6.2.5.

```
> ggplot(data=ggdat, aes(x = x, y = Fx)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_hline(yintercept=0)+
+   geom_hline(yintercept=0.10,color="red",linetype="dashed")+
+   theme_bw()+
+   xlab("Defective Items (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Binomial CDF",subtitle = bquote(bquote(F^-1*(0.10)==2 ~"for n=100, p=0.05"))
```

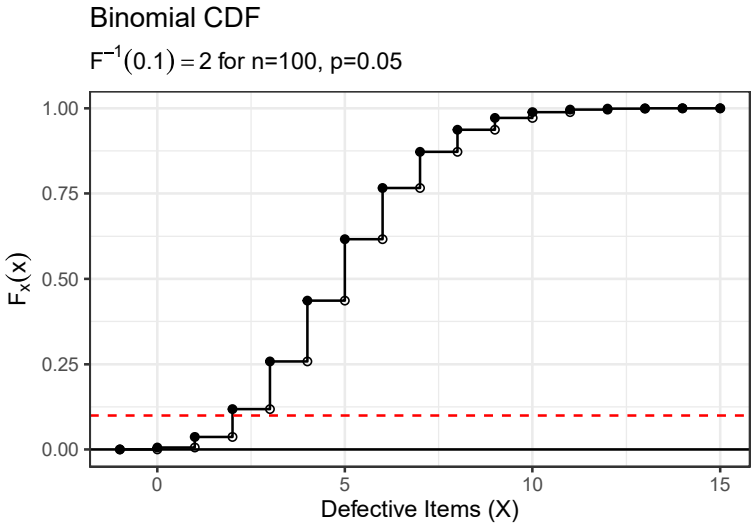

Figure 6.2.5: The Binomial CDF with $n = 100$ and $p = 0.05$ as in Example 6.11. There is a horizontal line drawn at 0.10 representing the tenth percentile. We take the percentile to be the smallest value for which the cumulative probability surpasses 0.10, here $x = 2$.

Definition 6.12. The **Hypergeometric distribution** is a discrete distribution used for the random variable, X, which represents the number of "successes" in k draws without replacement

from a population of size $N = n + m$, where n is the number of "failures" in the population and m is the number of "successes" in the population. This random variable is similar to the Binomial, the key difference being that with the Hypergeometric distribution sampling is done without replacement and so the trials are not independent. Graphs of the PMF and CDF are seen in Figure 6.2.6.

$$
\begin{aligned}
N &\in \mathbb{N} && \textbf{[Parameters]}\\
k &\in \{1, 2, \dots N\}\\
n, m &\in \{1, 2, \dots N\} \text{ such that } n + m = N\\
\mathcal{X} &= \{\omega : \omega \in \{\max(0, k + m - N), \dots, \min(k, m)\}\} && \textbf{[Support]}\\
f_X(x|N, n, m, k) &= \frac{\binom{m}{x}\binom{n}{(k-x)}}{\binom{N}{k}} I(x \in \mathcal{X}) && \textbf{[PMF]}\\
F_X(x|N, n, m, k) &= P(X \leq \lfloor x \rfloor)\\
&= \left[\sum_{x=\max(0,k+m-N)}^{\lfloor x \rfloor} \frac{\binom{m}{x}\binom{n}{(k-x)}}{\binom{N}{k}} \right] I(x \in \mathcal{X}) + I\left(\lfloor x \rfloor > \min(k, m)\right) && \textbf{[CDF]}\\
E(X) &= \frac{km}{m+n} && \textbf{[Expected Value]}\\
var(X) &= \frac{km}{m+n} \; \frac{-n}{m+n} \; \frac{m+n-k}{m+n-1} && \textbf{[Variance]}
\end{aligned}
$$

We won't simplify the CDF any further here because doing so would require convoluted functions like the generalized hypergeometric function; instead, we will calculate the CDF and inverse CDF numerically using functions in `R`.

```
> dhyper(x=1,m=20,n=90,k=5) #P(X=1|m=20,n=90,k=5)
[1] 0.4175436
> phyper(q=1,m=20,n=90,k=5) #P(X<=1|m=20,n=90,k=5)
[1] 0.7766311
> qhyper(p=0.9,m=20,n=90,k=5) #The 90th percentile
[1] 2
> rhyper(nn=10,m=20,n=90,k=5) #A random sample of 10 Xi~Hyper(m=20,n=90,k=5)
[1] 2 1 1 1 2 1 1 2 2 1
```

The graphs of the PMF and CDF in Figure 6.2.6 can be completed in `R` as follows.

```
> ggdat<-data.frame(x=(-1:21),
+                   f1=dhyper(x=(-1:21),m=20,n=90,k=5),
+                   F1=phyper(q=(-1:21),m=20,n=90,k=5))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,20)+
+   ylim(0,0.50)+
+   xlab("X")+
```

```
+   ylab(bquote(f[x](x)))+
+   ggtitle("Hypergeometric PMF",subtitle="m=20, n=90, k=5")
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=phyper(ggdat$x-1,m=20,n=90,k=5))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=phyper(ggdat$x,m=20,n=90,k=5))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Hypergeometric CDF",subtitle="m=20, n=90, k=5")
> grid.arrange(g1,g1.CDF,ncol=2)
```

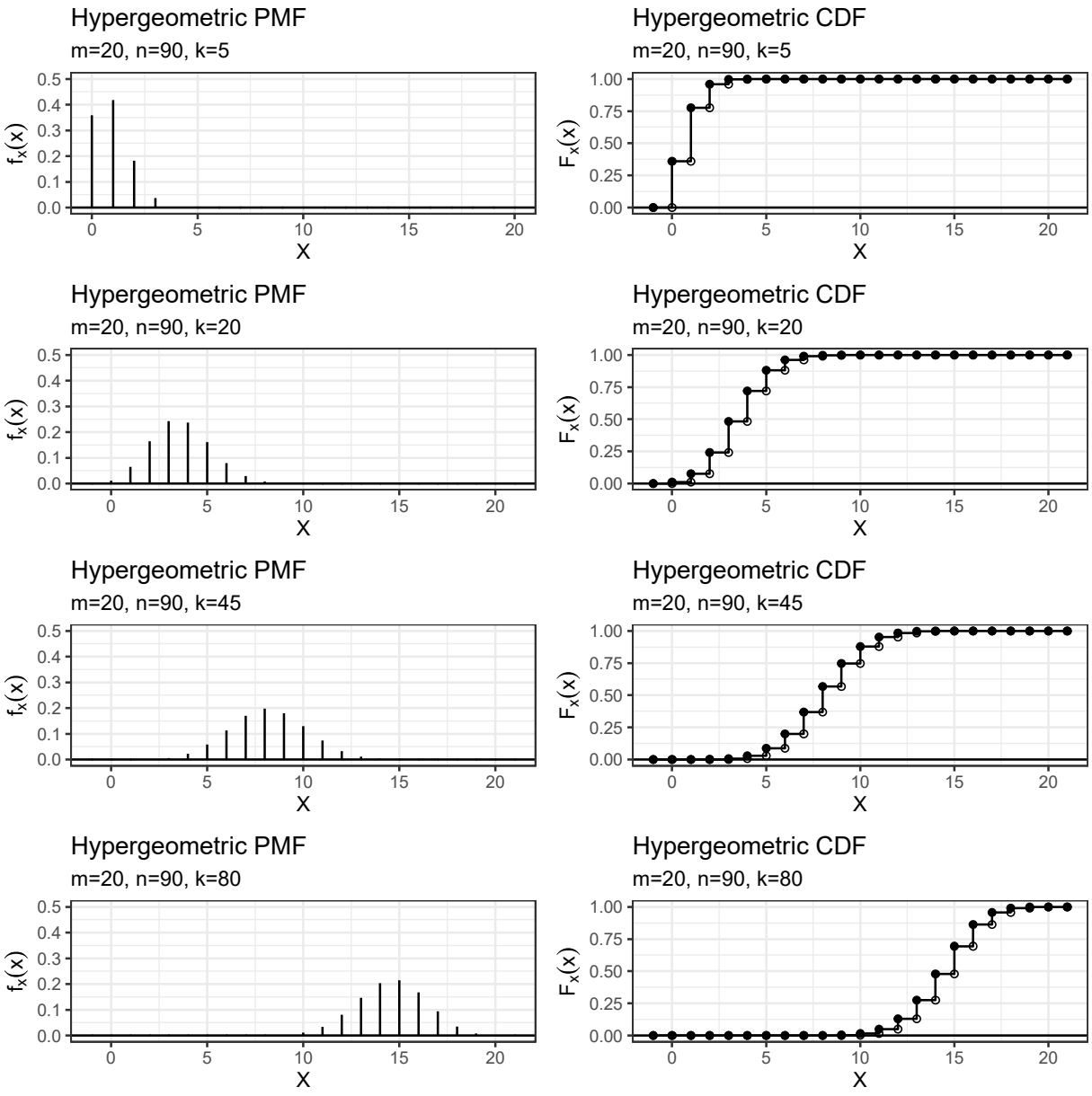

Figure 6.2.6: Hypergeometric PMF(left) and CDF(right) for $m = 20$, $n = 90$, and varying values of k.

Example 6.13. Consider a population of $N = 100$ parts being shipped to a company. The company has an acceptance sampling plan in which they sample five parts at random without replacement. If there are no defectives in the sample they accept the entire lot, otherwise they reject the entire lot and return it to the manufacturer for a new shipment. Let's consider the case where there are $m = 10$ defective items in the lot of $N = 100$; note that $n = N - m = 100 - 10 = 90$.

Question: What is the probability that the shipment will not be accepted?

Answer:

$$\begin{aligned}
P(X>0) &= 1 - P(X \leq 0) = 1 - P(X=0) \\
&= 1 - f_X(0|N=100, n=90, m=10, k=5) \\
&= 1 - \frac{\binom{10}{0}\binom{90}{(5-0)}}{\binom{100}{5}} && \textbf{[Hypergeometric PMF]} \\
&= 1 - \frac{1\binom{90}{5}}{\binom{100}{5}} \\
&= 1 - \frac{43949268}{75287520} \\
&= 0.416
\end{aligned}$$

Alternatively, this probability can be calculated in R as follows.

```
> 1-dhyper(x=0,m=10,n=90,k=5) #1-P(X=0|m=10,n=90,k=5)
[1] 0.4162476
```

We can also visualize this probability by graphing the PMF or CDF in R and highlighting the probability of interest in the graph as in Figure 6.2.7.

```
> ggdat<-data.frame(x=(-1:6),
+                   f1=dhyper(x=(-1:6),m=10,n=90,k=5),
+                   F1=phyper(q=(-1:6),m=10,n=90,k=5))
> ggdat.highlight<-data.frame(x=(1:6),
+                   f1=dhyper(x=(1:6),m=10,n=90,k=5))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(ymax=f1), ymin=0,color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Defective Items (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Hypergeometric PMF",subtitle=bquote(P(X>0)~"for m=10, n=90, k=5"))
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=phyper(ggdat$x-1,m=10,n=90,k=5))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=phyper(ggdat$x,m=10,n=90,k=5))
> ggdat.highlight<-data.frame(x=0,
+                                y=phyper(0,m=10,n=90,k=5))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_point(data = ggdat.highlight, aes(x = x, y = y),color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Defective Items (X)")+
+   ylab(bquote(F[x](x)))+
```

```
+   ggtitle("Hypergeometric CDF",subtitle="m=10, n=90, k=5")
> grid.arrange(g1,g1.CDF,ncol=2)
```

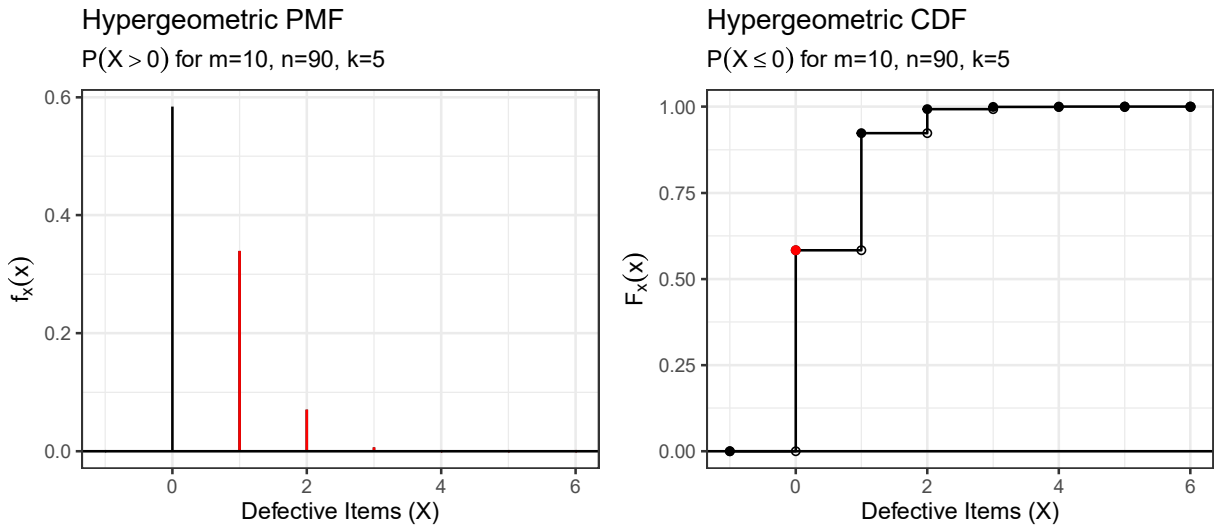

Figure 6.2.7: The Hypergeometric PMF (left) and CDF (right) with $n = 10$, $m = 90$ and $k = 5$ as in Example 6.13. The observations from $X = 1$ to $X = 5$ are highlighted in the PMF in red; the sum of the heights of these bars is equal to $P(X > 0|m = 10, n = 90, k = 5)$. The observation at $X = 0$ is highlighted in the CDF with a red point; the corresponding $F_X(x)$ value is equal to the $P(X = 0|m = 10, n = 90, k = 5)$ and the probability of interest can be calculated with the complement rule.

Question: What is the probability that at least three of the defective parts are sampled?
Answer:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - P(X \leq 2) \\ &= 1 - F_X(2|N = 100, n = 90, m = 10, k = 5) \\ &= 1 - \sum_{i=0}^{\lfloor 2 \rfloor} \frac{\binom{10}{i}\binom{90}{(5-i)}}{\binom{100}{5}} \\ &\approx 1 - 0.9933 \\ &= 0.0067 \end{aligned}$$

Alternatively, this probability can be calculated in R as follows.

```
> 1-phyper(q=2,m=10,n=90,k=5) #1-P(X<=2|m=10,n=90,k=5)
[1] 0.006637913
```

We can also visualize this probability by graphing the PMF or CDF in R and highlighting the probability of interest in the graph as in Figure 6.2.8.

```
> ggdat<-data.frame(x=(-1:6),
+                   f1=dhyper(x=(-1:6),m=10,n=90,k=5),
```

```
+                     F1=phyper(q=(-1:6),m=10,n=90,k=5))
> ggdat.highlight<-data.frame(x=(3:6),
+                             f1=dhyper(x=(3:6),m=10,n=90,k=5))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(ymax=f1), ymin=0,color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Defective Items (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Hypergeometric PMF",subtitle=bquote(P(X>=3)~"for m=10, n=90, k=5"))
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=phyper(ggdat$x-1,m=10,n=90,k=5))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=phyper(ggdat$x,m=10,n=90,k=5))
> ggdat.highlight<-data.frame(x=2,
+                             y=phyper(2,m=10,n=90,k=5))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_point(data = ggdat.highlight, aes(x = x, y = y),color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Defective Items (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Hypergeometric CDF",subtitle=bquote(P(X<=2)~"for m=10, n=90, k=5"))
> grid.arrange(g1,g1.CDF,ncol=2)
```

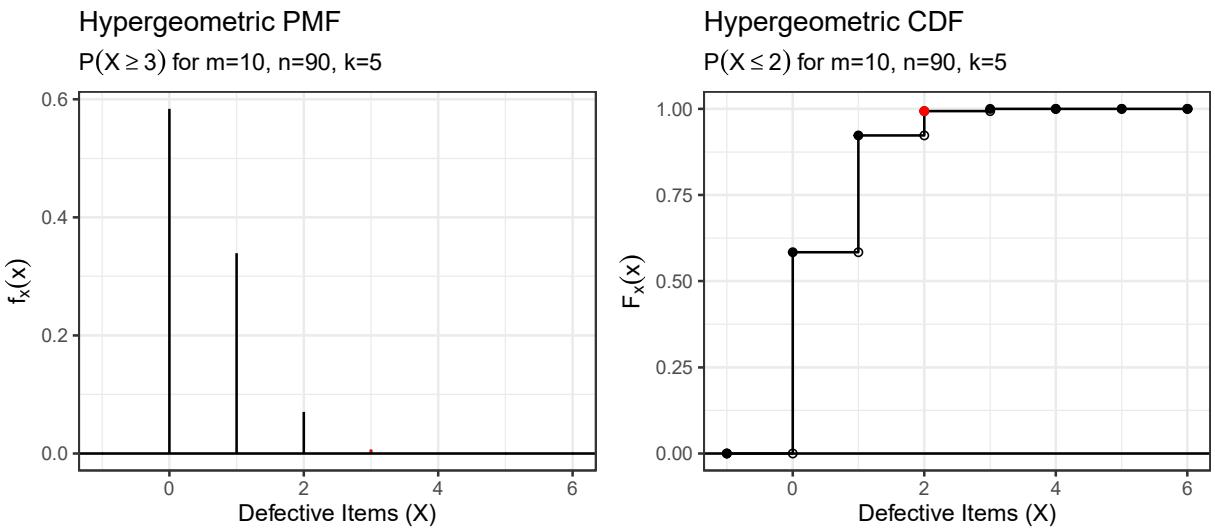

Figure 6.2.8: The Hypergeometric PMF (left) and CDF (right) with $n = 10$, $m = 90$ and $k = 5$ as in Example 6.13. The observations from $X = 3$ to $X = 5$ are highlighted in the PMF in red; the sum of the heights of these bars is equal to $P(X \geq 3|m = 10, n = 90, k = 5)$. The observation at $X = 2$ is highlighted in the CDF with a red point; the corresponding $F_X(x)$ value is equal to $P(X \leq 2|m = 10, n = 90, k = 5)$ and the probability of interest can be calculated with the complement rule.

Definition 6.14. The **Negative Binomial distribution** is a discrete distribution used for the random variable, X, which represents the number of failures to achieve in n successes in identical and independent trials with binary output. Graphs of the PMF and CDF are seen in Figure 6.2.9.

This is similar to the Binomial distribution, the difference being that the negative Binomial counts the number of failures to a fixed number of successes and the Binomial counts the number of successes in a fixed number of trials.

$$
\begin{aligned}
p \in (0,1) \text{ and } n \in \mathbb{N} & && \textbf{[Parameter]} \\
\mathcal{X} = \{\omega : \omega \in \{0, 1, \ldots\}\} & && \textbf{[Support]} \\
f_X(x|n,p) = \binom{n+x-1}{x} p^n (1-p)^x I(x \in \{0,1,\ldots\}) & && \textbf{[PMF]} \\
F_X(x|n,p) = P(X \leq \lfloor x \rfloor) & && \\
= \sum_{x=0}^{\lfloor x \rfloor} \left[\binom{n+x-1}{x} p^n (1-p)^x \right] I(x \in \mathcal{X}) & && \textbf{[CDF]} \\
E(X) = \frac{n(1-p)}{p} & && \textbf{[Expected Value]} \\
var(X) = \frac{n(1-p)}{p^2} & && \textbf{[Variance]}
\end{aligned}
$$

We won't simplify the CDF any further here because doing so would require convoluted functions like the hypergeometric function; instead, we will calculate the CDF and inverse CDF numerically using functions in `R`.

```
> dnbinom(x=2,size=5,prob=.35) #P(X=2|n=5,p=0.35)
[1] 0.03328574
> pnbinom(q=2,size=5,prob=.35) #P(X<=2|n=5,p=0.35)
[1] 0.05560754
> qnbinom(p=0.70,size=5,prob=.35) #The 70th percentile
[1] 11
> rnbinom(n=10,size=5,prob=.35) #A random sample of 10 Xi~Negbinom(n=5,p=0.35)
[1] 12 23 13  5  5  3  7  4 12 18
```

Additionally, the graphs in Figure 6.2.9 can be created in R as follows.

```
> ggdat<-data.frame(x=(-1:30),
+                   f1=dnbinom(x=(-1:30),size=5,prob=0.25),
+                   F1=pnbinom(q=(-1:30),size=5,prob=0.25))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,30)+
+   ylim(0,0.30)+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Negative Binomial PMF",subtitle="n=5, p=0.25")
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=pnbinom(ggdat$x-1,size=5,prob=0.25))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=pnbinom(ggdat$x,size=5,prob=0.25))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Negative Binomial CDF",subtitle="n=5, p=0.25")
```

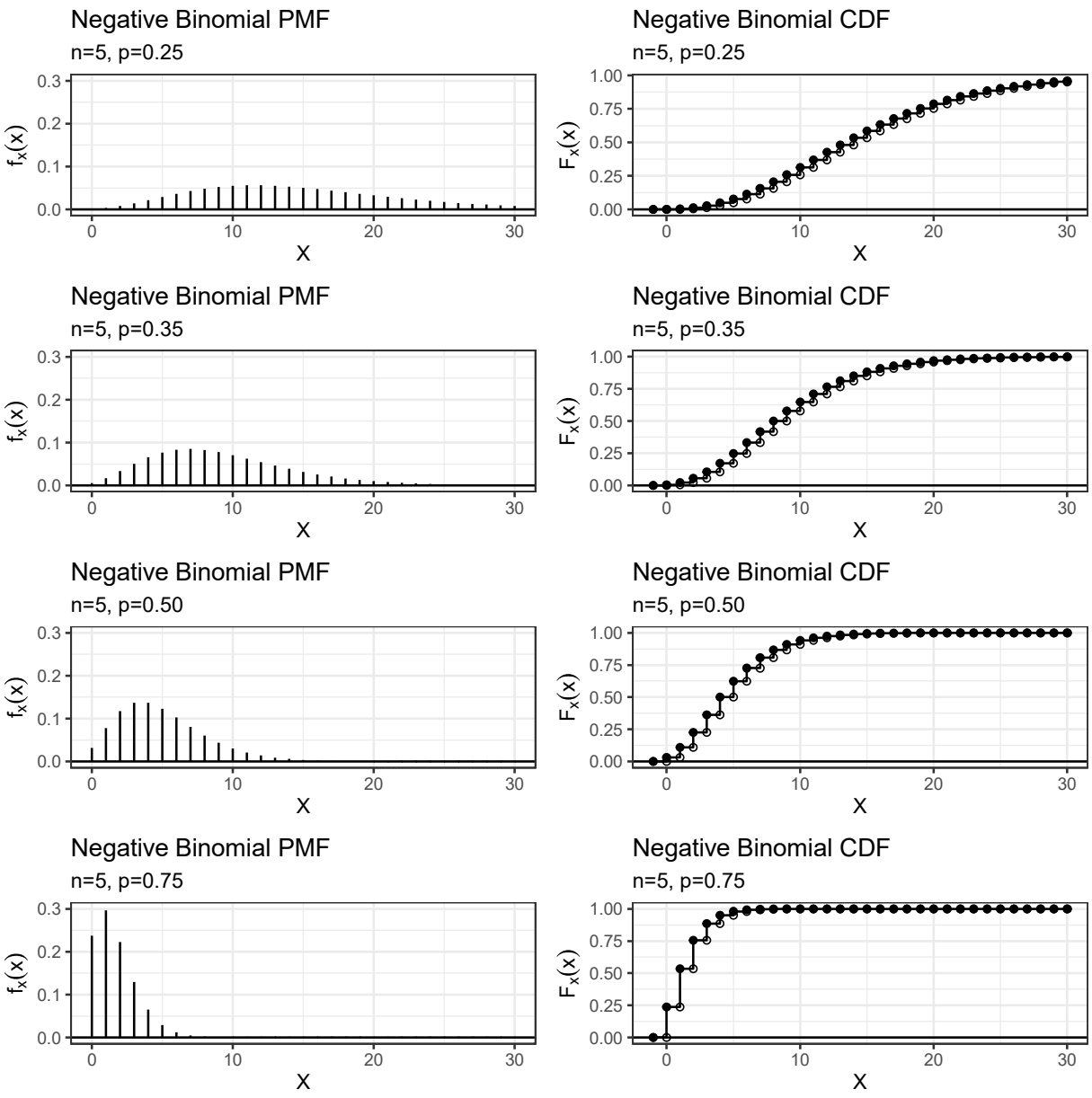

Figure 6.2.9: Negative Binomial PMF(left) and CDF(right) for $n = 5$ and varying values of p.

Example 6.15. A traditional apicoectomy is a dental procedure used to retreat an infected root canal. An apicoectomy involves a small incision near the infected tooth so that an oral surgeon can remove infected tissue and retreat the root tips in an attempt to preserve the tooth and avoid extraction. The success rate of traditional apicoectomy is reported to be 59%.

Question: What is the probability that an oral surgeon will perform her third successful apicoectomy on her fifth surgery?

Answer:

$$\begin{aligned} P(X=2) &= f_X(2|n=3, p=0.59) \\ &= \binom{3+2-1}{2} 0.59^3 (1-0.59)^2 \\ &= 0.2071 \end{aligned}$$

This probability can be calculated and visualized in R as follows. See Figure 6.2.10.

```
> dnbinom(x=2,size=3,prob=0.59)
[1] 0.2071453
> ggdat<-data.frame(x=(0:10),
+                   f1=dnbinom(x=(0:10),size=3,prob=0.59))
> ggdat.highlight<-data.frame(x=2,
+                   y=dnbinom(x=2,size=3,prob=0.59))
> ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(x=x,ymax=y), ymin=0,color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,10)+
+   ylim(0,0.30)+
+   xlab("Number of Failed Apicoectomies Until Third Success (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Negative Binomial PMF",subtitle=bquote(P(X==2)~"for n=5, p=0.25"))
```

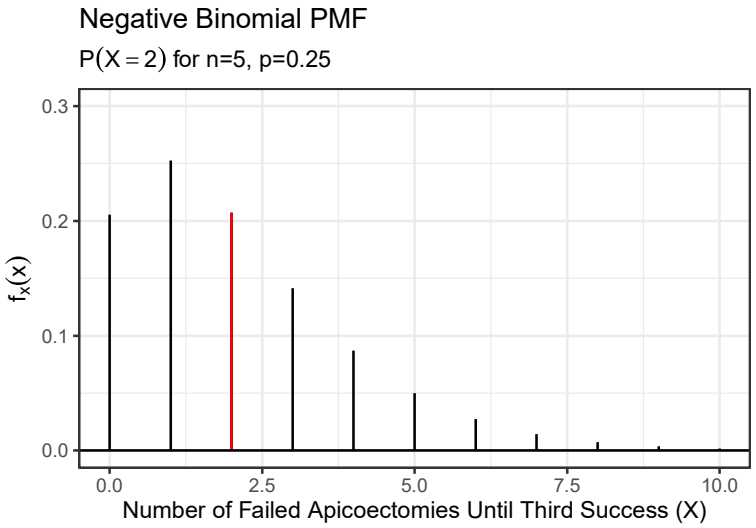

Figure 6.2.10: Negative Binomial PMF(left) and CDF(right) for $n = 5$ and $p = 0.59$ as in Example 6.15. The observation of $X = 2$ is highlighted in red; the height of this bar is equal to $P(X = 2|n = 3, p = 0.59)$.

Remark: There are other formulations of the negative binomial distribution; e.g., where X is the number of trials until n successes occur; i.e., $Y = X + n$.

Definition 6.16. The **Poisson distribution** is a discrete distribution used for the random variable, X, which represents the number of times a specific event occurs in a given amount of time or space. Graphs of the PMF and CDF are seen in Figure 6.2.11.

The Poisson distribution assumes that events occurring in a given units of time or space are independent and the probability of an event occurring in a given unit of time or space is the same across the units of time or space. The parameter of the model, λ, denotes the expected number of events in each unit of time or space.

$$\begin{aligned}
\lambda &\in (0,\infty) && \textbf{[Parameter]}\\
\mathcal{X} &= \{\omega : \omega \in \{0,1,\ldots\}\} && \textbf{[Support]}\\
f_X(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!}\, I(x \in \{0,1,\ldots\}) && \textbf{[PMF]}\\
F_X(x|\lambda) &= P(X \leq \lfloor x \rfloor)\\
&= \left[e^{-\lambda} \sum_{x=0}^{\lfloor x \rfloor} \frac{\lambda^x}{x!} \right] I(x \in \mathcal{X}) && \textbf{[CDF]}
\end{aligned}$$

Note: The indicator function isn't necessary during derivation here because the way we define the CDF will return zero for all values of $x < 0$, and the maximum value of X is unbounded.

We won't simplify the CDF any further here because doing so would require convoluted functions like the incomplete Gamma function; instead, we will calculate the CDF and inverse CDF numerically using functions in `R`.

```
> dpois(x=2,lambda=1) #P(X=2|lambda=1)
[1] 0.1839397
> ppois(q=2,lambda=1) #P(X<=2|lambda=1)
[1] 0.9196986
> qpois(p=0.90,lambda=1) #The 90th percentile
[1] 2
> rpois(n=10,lambda=1) #A random sample of 10 Xi~Poisson(lambda=1)
[1] 2 2 2 0 0 0 0 1 1 2
```

Additionally, the graphs can be completed in `R` as follows.

```
> ggdat<-data.frame(x=(-1:20),
+                   f1=dpois(x=(-1:20),lambda=1),
+                   F1=ppois(q=(-1:20),lambda=1))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,20)+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Poisson PMF",subtitle=bquote(lambda==1))
```

```
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=ppois(ggdat$x-1,lambda=1))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=ppois(ggdat$x,lambda=1))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Poisson CDF",subtitle=bquote(lambda==1))
```

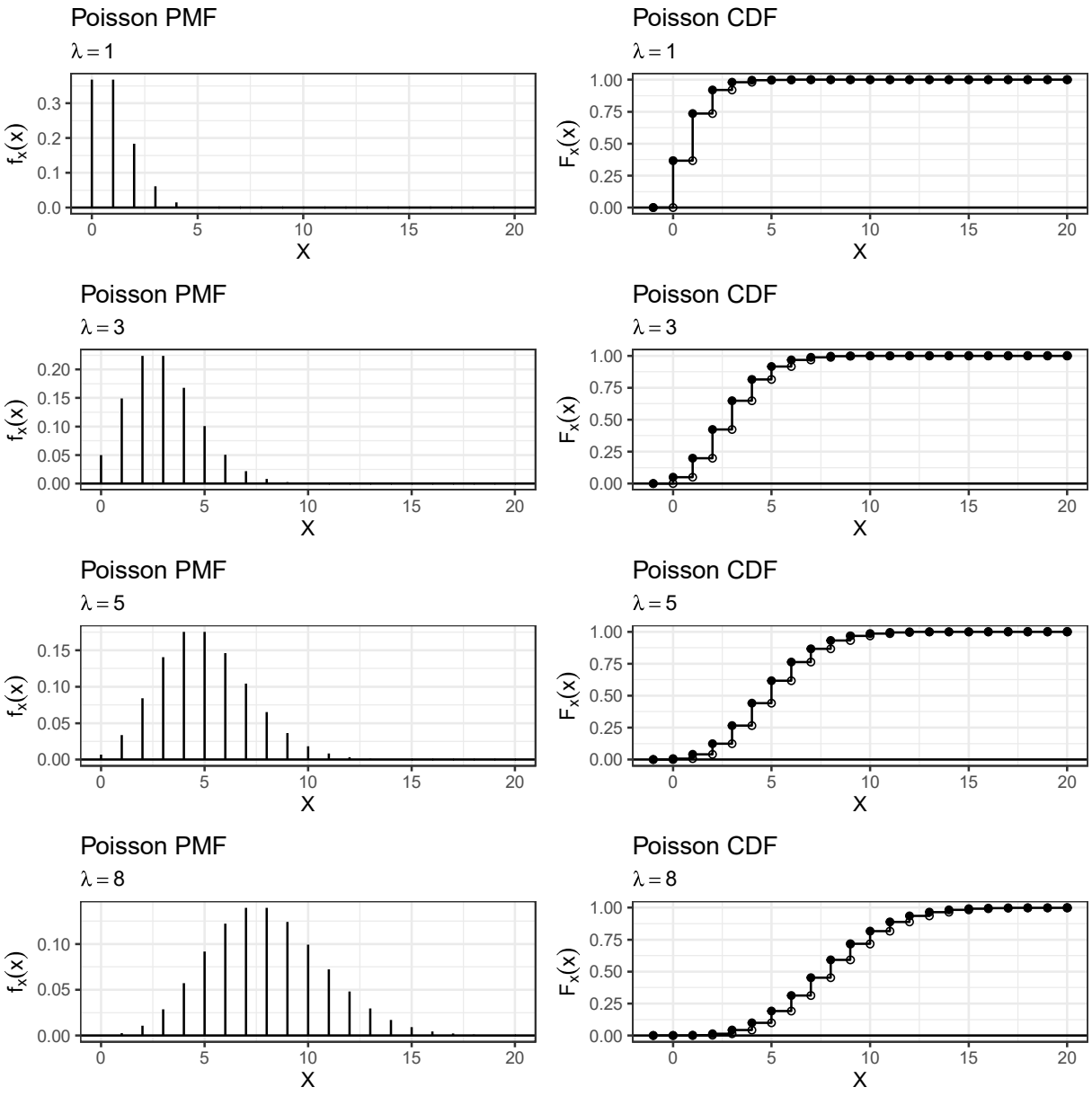

Figure 6.2.11: Poisson PMF(left) and CDF(right) for $\lambda = 1$.

Example 6.17. A study about the nesting and spawning of horseshoe crabs provides information about how unattached males, called satellites, interact with coupled horseshoe crabs. Satellite males form large groups around some couples and compete with the attached males for fertilizations and some couples have no satellites. Suppose it is hypothesized that each female horseshoe crab that takes satellites has 5.295 during the high-tide in which they mate, on average; i.e., take $\lambda = 5.295$.

Question: What is the probability that a randomly selected female horseshoe crab with satellites has exactly two?

Answer:

$$\begin{aligned} P(X=2) &= f_X(2|\lambda=5.295) \\ &= \frac{5.295^2 e^{-5.295}}{2!} \\ &\approx 0.0703 \end{aligned}$$

The probability can be calculated and visualized in R as follows.

```
> dpois(x=2,lambda=5.295) #P(X=2|lambda=5.295)
[1] 0.07032547
> ggdat<-data.frame(x=(-1:15),
+                   f1=dpois(x=(-1:15),lambda=5.295))
> ggdat.highlight<-data.frame(x=2,
+                             f1=dpois(x=2,lambda=5.295))
> ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(ymax=f1),ymin=0,color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,15)+
+   xlab("Number of Satellites (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Poisson PMF",subtitle=bquote(P(X==2)~"for"~lambda==5.295))
```

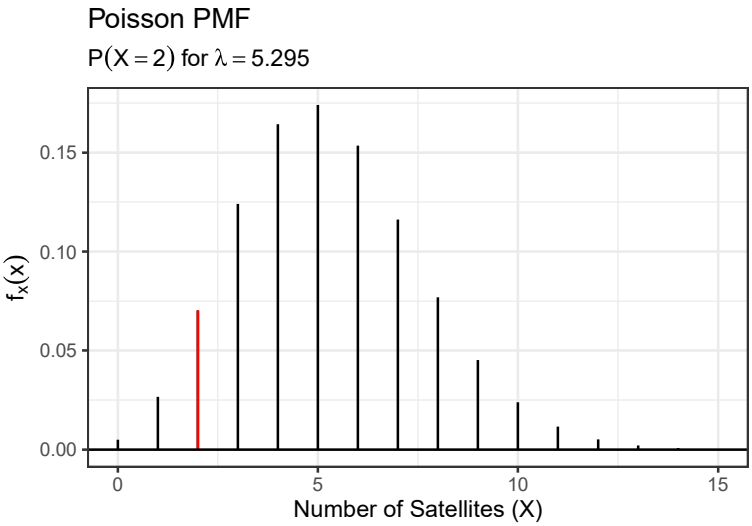

Figure 6.2.12: Poisson PMF for $\lambda = 0.5295$ as in Example 6.17. The observation of $X = 2$ is highlighted in red; the height of this bar is equal to $P(X = 2|\lambda = 5.295)$.

Question: What is the probability that a randomly selected female horseshoe crab has no more than two satellites?

Answer:

$$P(X \leq 2) = \sum_{x=0}^{2} \frac{5.295^x e^{-5.295}}{x!}$$
$$\approx 0.1019$$

The probability can be calculated and visualized in R as follows.

```
> ppois(q=2,lambda=5.295) #P(X<=2|lambda=5.295)
[1] 0.1019051
> ggdat<-data.frame(x=(-1:15),
+                   f1=dpois(x=(-1:15),lambda=5.295),
+                   F1=ppois(q=(-1:15),lambda=5.295))
> ggdat.highlight<-data.frame(x=0:2,
+                             f1=dpois(x=0:2,lambda=5.295))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_linerange(data=ggdat.highlight,aes(ymax=f1),ymin=0,color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,15)+
+   xlab("Number of Satellites (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Poisson PMF",subtitle=bquote(P(X<=2)~"for"~lambda==5.295))
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=ppois(ggdat$x-1,lambda=5.295))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=ppois(ggdat$x,lambda=5.295))
> ggdat.highlight<-data.frame(x=2,
+                             y=ppois(2,lambda=5.295))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_point(data = ggdat.highlight, aes(x = x, y = y),color="red") +
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Number of Satellites (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Poisson CDF",subtitle=bquote(P(X<=2)~"for"~lambda==5.295))
> grid.arrange(g1,g1.CDF,ncol=2)
```

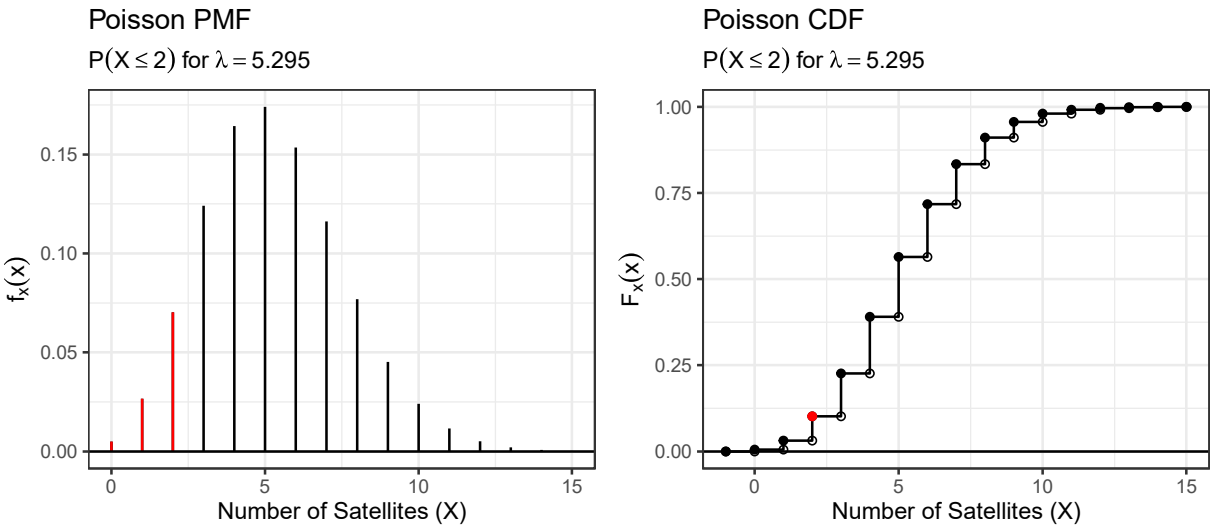

Figure 6.2.13: Poisson PMF (left) and CDF (right) for $\lambda = 0.5295$ as in Example 6.17. The observations from $X = 0$ to $X = 2$ are highlighted in the PMF in red; the sum of the heights fo these bars is equal to $P(X \leq 2|\lambda = 5.295)$. The observation at $X = 2$ is highlighted in the CDF with a red point; the corresponding $F_X(x)$ value is equal to $P(X \leq 2|\lambda = 5.295)$.

Definition 6.18. The probabilities $p_1, p_2, ..., p_k$ do not change from trial to trial and $\sum_{j=1}^{k} p_j = 1$. **Experiment:** Perform n mutually independent trials. Each trial results in one of k distinct categorical outcomes:

			Probability	Count
	$\longrightarrow$	Category 1	p_1	Y_1
	$\longrightarrow$	Category 2	p_2	Y_2
Trial outcome	$\longrightarrow$	Category 3	p_3	Y_3
		$\vdots$	$\vdots$	$\vdots$
	$\longrightarrow$	Category k	p_k	Y_k

Define

$$Y_j = \text{ number of outcomes in Category } j \text{ (out of } n \text{ trials)},$$

for $j = 1, 2, ..., k$. We call $\mathbf{Y} = (Y_1, Y_2, ..., Y_k)$ a **multinomial random vector**. The joint PMF of $Y_1, Y_2, ..., Y_k$ is

$$f_{\mathbf{Y}}(y_1, y_2, ..., y_{k-1}) = \frac{n!}{y_1! y_2! \cdots y_{k-1}! \left(n - \sum_{j=1}^{k-1} y_j\right)!} p_1^{y_1} p_2^{y_2} \cdots p_{k-1}^{y_{k-1}} \left(1 - \sum_{j=1}^{k-1} p_j\right)^{n - \sum_{j=1}^{k-1} y_j}$$

with the support $\mathcal{A} = \{(y_1, y_2, ..., y_k) : y_j = 0, 1, 2, ..., n; \sum_{j=1}^{k} y_j = n\}$. We write

$$\mathbf{Y} \sim \text{mult}(n, \mathbf{p}; \sum_{j=1}^{k} p_j = 1).$$

The parameter $\mathbf{p} = (p_1, p_2, ..., p_k)$ is k-dimensional. However, because $\sum_{j=1}^{k} p_j = 1$, only $k - 1$ of these parameters are "free to vary," as knowing the value of any $k-1$ parameters gives us the value of the k^{th}; this is also true for the random variables because $\sum_{j=1}^{k} Y_j = n$.

```
> library(extraDistr)
> #P(X1=4,X2=5,X3=1|p1=0.25,p2=0.625,p3=0.125)
> dmnom(x=c(4,5,1),size=10,pr=c(0.25,0.625,0.125))
[1] 0.058673324
> #A random sample of 5 multinomial observations (rows)
> rmnom(n=5,size=10,pr=c(0.25,0.625,0.125))
[,1] [,2] [,3]
[1,]    4    5    1
[2,]    2    6    2
[3,]    2    7    1
[4,]    3    6    1
[5,]    4    5    1
```

Illustration: A graph of a bivariate multinomial PMF for $(Y_1, Y_2)'$ with $n = 10$, $p_1 = 0.25$, and $p_2 = 0.75$, created with the following R code, is displayed in Figure 6.2.14.

```
> y1<-seq(0,10,1)
> y2<-seq(0,10,1)
> Y<-expand.grid(y1,y2) #all combinations of (y1,y2)
> colnames(Y)<-c("y1","y2")
> ggdat<-data.frame(y1=Y$y1,
+                   y2=Y$y2,
+                   fxy=dmnom(cbind(Y$y1,Y$y2),size=10,prob=c(0.25,0.75)))
> ggplot(data=ggdat,aes(x=y1,y=y2,fill=fxy))+
+     geom_tile()+
+     scale_fill_gradient2(low = "magenta", high="cyan",mid="white",
+                          midpoint = 0.15, limit = c(0,0.30),
+                          name="Probability")+
+     geom_text(aes(label=round(fxy,2)),size=3)+
+     theme_bw()+
+     xlab(bquote(Y[1]))+
+     ylab(bquote(Y[2]))
```

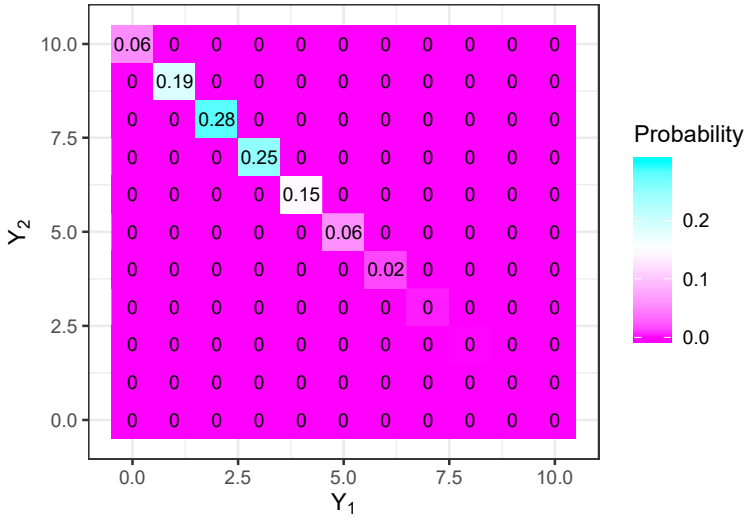

Figure 6.2.14: A graph of bivariate multinomial PMF with $n = 10$, $p_1 = 0.25$, and $p_2 = 0.75$.

Remark: The multinomial distribution is an extension of the binomial distribution to allow for more than 2 categories. Recall that in the binomial distribution, there were only 2 categories: "success" and "failure;" i.e.,

			Probability	Count
Trial outcome	$\longrightarrow$	Category 1 ("success")	p	Y
	$\longrightarrow$	Category 2 ("failure")	$1-p$	$n-Y$

We write $Y \sim \text{binomial}(n, p)$.

There is a clear connection to the values in the marginal PMFs in Figure 6.2.15 and the multinomial plot in Figure 6.2.14. Due to the fact that only $k-1$ of these parameters are "free to vary," it suffices to consider only the "success category;" e.g., the multinomial distribution in the bivariate case is the binomial distribution

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \frac{n!}{y_1!y_2!}\, p_1^{y_1} p_2^{y_2} && \textbf{[Multinomial PMF]} \\ &= \frac{n!}{y_1!(n-y_1)!}\, p_1^{y_1} p_2^{(n-y_1)} && \textbf{[Noting } y_2 = n - y_1\textbf{]} \\ &= \frac{n!}{y_1!(n-y_1)!}\, p_1^{y_1} (1-p_1)^{(n-y_1)}. && \textbf{[Noting } p_2 = 1 - p_1\textbf{]} \end{aligned}$$

The marginal (univariate) binomial PMFs in Figure 6.2.15 are created with the following R code.

```
> ggdat<-data.frame(x=0:10,
+                   f1=dbinom(0:10,size=10,prob=0.25))
> g1<-ggplot(data=ggdat,aes(x=x,y=f1))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   ylim(0,0.35)+
+   theme_bw()+
```

```
+   xlab(bquote(Y[1]))+
+   ylab(bquote(f[Y[1]](y)))+
+   ggtitle("Binomial PMF",subtitle="p=0.25")+
+   geom_text(aes(label=round(f1,2)),vjust=-0.50,size=3)
> ggdat<-data.frame(x=0:10,
+                   f1=dbinom(0:10,size=10,prob=0.75))
> g2<-ggplot(data=ggdat,aes(x=x,y=f1))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   ylim(0,0.35)+
+   theme_bw()+
+   xlab(bquote(Y[1]))+
+   ylab(bquote(f[Y[1]](y)))+
+   ggtitle("Binomial PMF",subtitle="p=0.75")+
+   geom_text(aes(label=round(f1,2)),vjust=-0.50,size=3)
> grid.arrange(g1,g2,ncol=2)
```

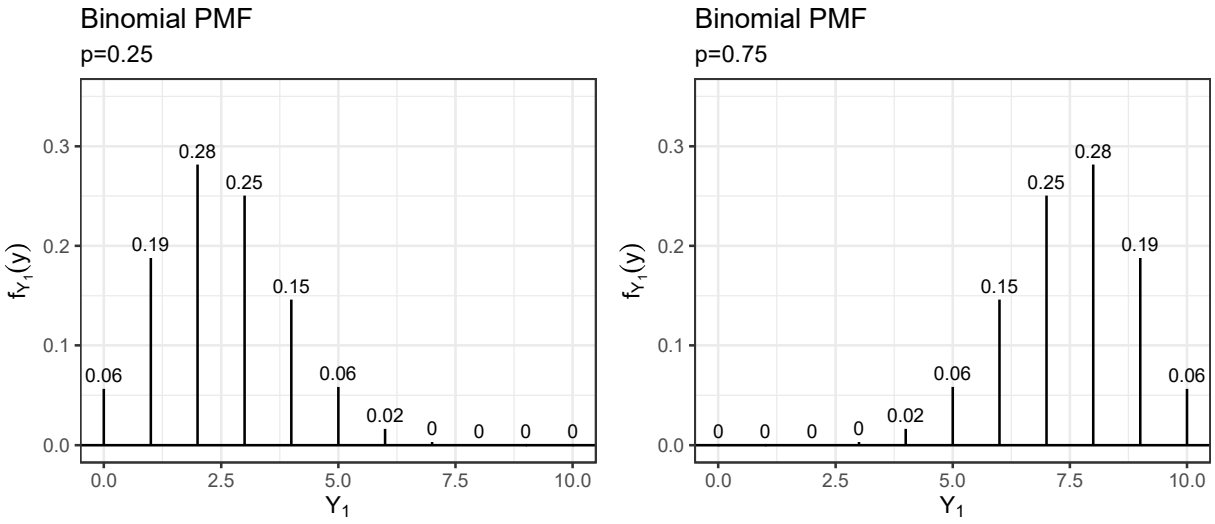

Figure 6.2.15: A graph of the marginal distributions of Y_1 and Y_2 for $(Y_1, Y_2) \sim$ multinomial$(n = 10, p_1 = 0.25, p_2 = 0.75)$.

Facts Suppose X_1, X_2 and X_3 follow Multinomial distribution with size n and probabilities p_1, p_2, and p_3.

1. Marginal distributions are Binomial; i.e., $X_i \sim$ **binomial**$(n = n, p = p_i)$, (univariate) Binomial distributions.
   Remark: This should make sense as in each marginal case i we simply define outcomes in the i^{th} category a "success" (which occur with probability p_i) and the other $k-1$ categories as "failure" (which occur with probability $1-p_i$, which preserves the interpretation of a binomial distributed random variable.

2. Conditional distributions are Binomial; i.e., $X_i|X_j \sim$ Binomial$(n = n - x_j, p = \frac{p_i}{1-p_j})$.
   Remark: This makes sense because we know that x_1 of the n observations are of category 1, leaving $n - x_1$ remaining observations for either category 2 ("success"), or category 3

("failure"), which preserves the interpretation of a binomial distributed random variable. The success probability is the conditional probability that an observation is of category 2 given it is not of category 1.

Example 6.19. The State Hygienic Laboratory at the University of Iowa tests thousands of Iowa residents each year for chlamydia (CT) and gonorrhea (NG). On a given day, suppose the lab receives $n = 100$ specimens to be tested. Define

$$\begin{aligned} &\text{Category 1:CT}-/\text{NG}- \quad (p_1 = 0.90) \\ &\text{Category 2:CT}+/\text{NG}- \quad (p_2 = 0.07) \\ &\text{Category 3:CT}-/\text{NG}+ \quad (p_3 = 0.02) \\ &\text{Category 4:CT}+/\text{NG}+ \quad (p_4 = 0.01) \end{aligned}$$

and let $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$ denote the category counts observed after testing. Envisioning each specimen as a "trial," regarding the specimens as mutually independent, and assuming the category probabilities in $\mathbf{p} = (p_1, p_2, p_3, p_4)$ are the same for each specimen, then

$$\mathbf{Y} \sim \text{multinomial}\left(n = 100, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right).$$

The PMF of $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, where nonzero, is given by

$$p_{\mathbf{Y}}(y_1, y_2, y_3, y_4) = \frac{100!}{y_1!y_2!y_3!y_4!}\,(0.90)^{y_1}(0.07)^{y_2}(0.02)^{y_3}(0.01)^{y_4}.$$

For example,

$$\begin{aligned} p_{\mathbf{Y}}(88, 10, 1, 1) &= P(Y_1 = 88, Y_2 = 10, Y_3 = 1, Y_4 = 1) \\ &= \frac{100!}{88!\ 10!\ 1!\ 1!}\,(0.90)^{88}(0.07)^{10}(0.02)^{1}(0.01)^{1} \approx 0.017. \ \square \end{aligned}$$

Remark: Many scenarios require more than two variables. While we can calculate these probabilities, graphing in more than three dimensions is problematic. We can, however, ask `R` to calculate this probability for us.

```
> dmnom(x=c(88,10,1,1),size=100,pr=c(0.90,0.07,0.02,0.01))
[1] 0.007366964
```

6.3 Continuous Distribution Functions

Finding probability for the events described in Definition 6.1 is simpler in the continuous case. For a continuous random variable

$$\begin{aligned} f_X(x) &\neq P(X = x) && \textbf{[PDF]} \\ P(X = x) &= 0 && \textbf{[Continuous property]} \\ F_X(x) &= P(X \leq x). && \textbf{[CDF]} \end{aligned}$$

For a known constant $x_1 \in \mathbb{R}$,

$$\begin{aligned}
P(X = x_1) &= 0 && \textbf{[Continuous Property]}\\
P(X \neq x_1) &= 1 && \textbf{[Complement rule]}\\
P(X \leq x_1) &= F_X(x_1) && \textbf{[Definition of CDF]}\\
P(X \geq x_1) &= 1 - P(X < x_1) && \textbf{[Complement rule]}\\
&= 1 - P(X \leq x_1) && \textbf{[Continuous Property]}\\
&= 1 - F_X(x_1) && \textbf{[Definition of CDF]}\\
P(X < x_1) &= P(X \leq x_1) && \textbf{[Continuous Property]}\\
&= F_X(x_1) && \textbf{[Definition of CDF]}\\
P(X > x_1) &= 1 - P(X \leq x_1). && \textbf{[Complement rule]}\\
&= 1 - F_X(x_1) && \textbf{[Definition of CDF]}
\end{aligned}$$

Remark: To write these probabilities in terms of a CDF requires us to consider $P(X < x_1) = P(X \leq x_1)$. This is true because the $P(X = x_1) = 0$ so incorporating $\{x_1\}$ in the event $X < x_1$ doesn't change the probability.

Additionally, we can consider the following events around known constants $x_1 < x_2 \in \mathbb{R}$.

$$\begin{aligned}
P(x_1 < X < x_2) &= P(X < x_2) - P(X \leq x_1)\\
&= P(X \leq x_2) - P(X \leq x_1)\\
&= F_X(x_2) - F_X(x_1) && \textbf{[Definition of CDF]}\\
P(x_1 < X \leq x_2) &= P(X \leq x_2) - P(X \leq x_1)\\
&= F_X(x_2) - F_X(x_1) && \textbf{[Definition of CDF]}\\
P(x_1 \leq X < x_2) &= P(X < x_2) - P(X < x_1)\\
&= P(X \leq x_2) - P(X \leq x_1)\\
&= F_X(x_2) - F_X(x_1) && \textbf{[Definition of CDF]}\\
P(x_1 \leq X \leq x_2) &= P(X \leq x_2) - P(X < x_1)\\
&= P(X \leq x_2) - P(X \leq x_1)\\
&= F_X(x_2) - F_X(x_1) && \textbf{[Definition of CDF]}
\end{aligned}$$

Remark: Since $P(X = x) = 0$ for all $x \in \mathbb{R}$, these four probabilities are equivalent.

6.3.1 Named Distribution Functions

Definition 6.20. The **uniform distribution** is used for the random variable, X, which can take on real values between a and b with equal probability. Graphs of the PDF and CDF are seen in

Figure 6.3.16.

$$
\begin{aligned}
a, b \in \mathbb{R} \text{ such that } a < b & \qquad \textbf{[Parameters]} \\
\mathcal{X} = \{\omega : \omega \in [a, b]\} & \qquad \textbf{[Support]} \\
f_X(x|a, b) = \frac{1}{b-a}\ I(x \in [a, b]) & \qquad \textbf{[PDF]} \\
F_X(x|a, b) = P(X \leq x) & \\
= p(b-a) + a & \qquad \textbf{[For } p \in [0, 1]\textbf{]} \\
E(X) = \frac{a+b}{2} & \qquad \textbf{[Expected Value]} \\
var(X) = \frac{(b-a)^2}{12} & \qquad \textbf{[Variance]}
\end{aligned}
$$

These calculations can be carried out in R as follows.

```
> dunif(x=0.20,min=0,max=1) #This the height of the PDF at x=0.20
[1] 1
> punif(q=0.20,min=0,max=1) #P(X<=0.20|a=0,b=1)
[1] 0.2
> qunif(p=0.90,min=0,max=1) #The 90th percentile
[1] 0.9
> runif(n=5,min=0,max=1) #A random sample of 5 Xi~Uniform(a=0,b=1)
[1] 0.84107160 0.27353372 0.64576933 0.54045498 0.05480477
```

Additionally, the graphs can be completed in R as follows.

```
> ggdat<-data.frame(x=seq(-1,6,0.001),
+                   f1=dunif(x=seq(-1,6,0.001),min=0,max=1),
+                   f2=dunif(x=seq(-1,6,0.001),min=0.25,max=3),
+                   f3=dunif(x=seq(-1,6,0.001),min=1.5,max=5),
+                   f4=dunif(x=seq(-1,6,0.001),min=2,max=4),
+                   f5=dunif(x=seq(-1,6,0.001),min=3.5,max=4.5))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1,color="a=0 b=1"))+
+   geom_line(aes(y=f2,color="a=0.25 b=3"))+
+   geom_line(aes(y=f3,color="a=1.5 b=5"))+
+   geom_line(aes(y=f4,color="a=2 b=4"))+
+   geom_line(aes(y=f5,color="a=3.5 b=4.5"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Uniform PDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",labels=c("a=0, b=1","a=0.25, b=3","a=1.5, b=5",
+                                    "a=2, b=4","a=3.5, b=4.5"))
> ggdat<-data.frame(x=seq(-1,6,0.001),
+                   F1=punif(q=seq(-1,6,0.001),min=0,max=1),
+                   F2=punif(q=seq(-1,6,0.001),min=0.25,max=3),
+                   F3=punif(q=seq(-1,6,0.001),min=1.5,max=5),
```

```
+                     F4=punif(q=seq(-1,6,0.001),min=2,max=4),
+                     F5=punif(q=seq(-1,6,0.001),min=3.5,max=4.5))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1,color="a=0 b=1"))+
+   geom_line(aes(y=F2,color="a=0.25 b=3"))+
+   geom_line(aes(y=F3,color="a=1.5 b=5"))+
+   geom_line(aes(y=F4,color="a=2 b=4"))+
+   geom_line(aes(y=F5,color="a=3.5 b=4.5"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Uniform CDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",labels=c("a=0, b=1","a=0.25, b=3","a=1.5, b=5",
+                                    "a=2, b=4","a=3.5, b=4.5"))
> grid.arrange(g1,g1.cdf,ncol=2)
```

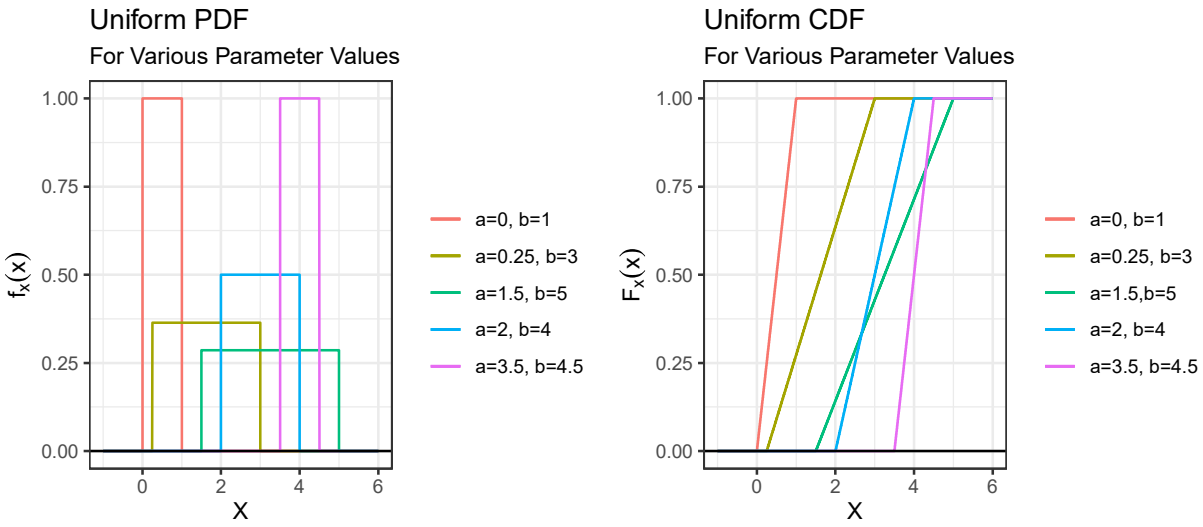

Figure 6.3.16: Uniform PDF(left) and CDF(right) for $a = 0$ and $b = 1$.

Example 6.21. A pseudo random number generator gives a random number between zero and one uniformly.

Question: What is the probability of outputting a random number greater than 0.80?
Answer:

$$\begin{aligned} P(X > 0.80|a = 0, b = 1) &= 1 - P(X \leq 0.80|a = 0, b = 1) \\ &= 1 - F_X(0.80|a = 0, b = 1) \\ &= 1 - \frac{0.80 - 0}{1 - 0} \qquad \textbf{[Definition of Uniform CDF]} \\ &= 1 - 0.80 = 0.20 \end{aligned}$$

We can also visualize this probability by graphing the PMF in R and highlighting the probability of interest in the graph.

```
> ggdat<-data.frame(x=seq(-0.5,1.5,0.001),
+                   f1=dunif(x=seq(-0.5,1.5,0.001),min=0,max=1))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x>0.8 & x<=1),aes(ymax=f1),ymin=0,
+                      fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Pseudo Random Number (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Uniform PDF",subtitle=bquote(P(X>0.80)~"for a=0, b=1"))
> ggdat<-data.frame(x=seq(-0.5,1.5,0.001),
+                   F1=punif(q=seq(-0.5,1.5,0.001),min=0,max=1))
> ggdat.highlight<-data.frame(x=0.80,
+                             y=punif(q=0.80,min=0,max=1))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Pseudo Random Number (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Uniform CDF",subtitle=bquote(P(X<=0.80)~"for a=0, b=1"))
> grid.arrange(g1,g1.cdf,ncol=2)
```

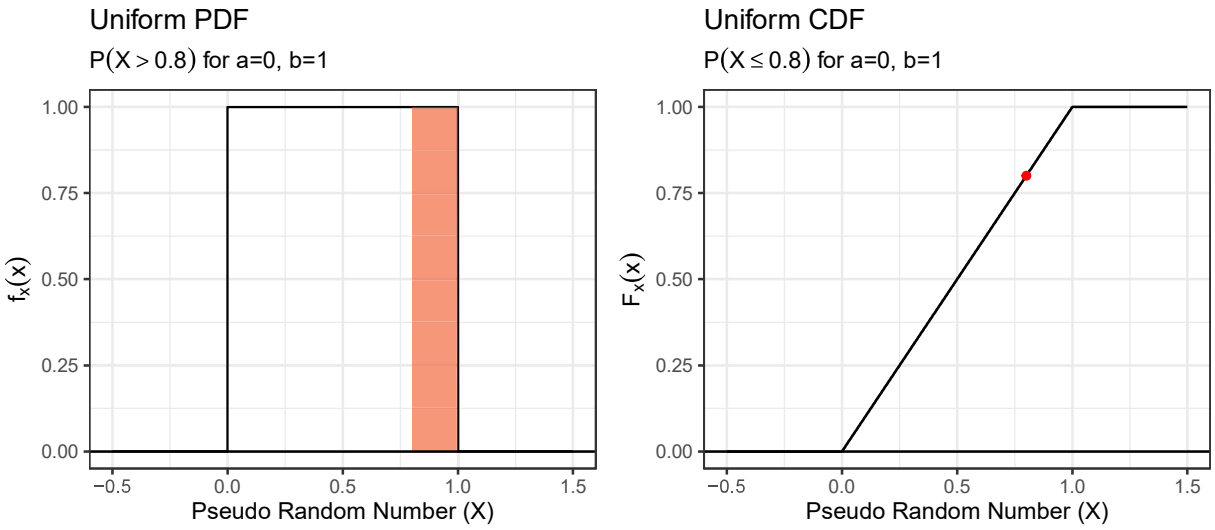

Figure 6.3.17: Uniform PDF(left) and CDF(right) for $a = 0$ and $b = 1$ as in Example 6.21. The area of the shaded in the PDF (left) is equal to $P(X > 0.80|a = 0, b = 1)$. The observation at $X = 0.80$ is highlighted in the CDF with a red point; the corresponding $F_x(x)$ value is equal to $P(X \leq 0.80|a = 0, b = 1)$ and the probability of interest can be calculated with the complement rule.

Definition 6.22. The **Gaussian distribution**, also known as the normal distribution or bell curve, is used for a random variable, X, which can take on real values. The Gaussian distribution is unimodal with a peak and line of symmetric at the mean which intimates that observations often

occur near the mean and less often occur the further away from the mean. Graphs of the PDF and CDF are seen in Figure 6.3.18.

The Gaussian distribution serves as a very good model for a wide range of measurements; reaction times, fill amounts, part dimensions, weights/heights, measures of intelligence/test scores, economic indicators, etc.

$$\mu \in \mathbb{R} \text{ and } \sigma \in \mathbb{R}^+ \qquad \textbf{[Parameters]}$$

$$\mathcal{X} = \{\omega : \omega \in \mathbb{R}\} \qquad \textbf{[Support]}$$

$$f_X(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \; I(x \in \mathbb{R}) \qquad \textbf{[PDF]}$$

$$\begin{aligned} F_X(x|\mu,\sigma) &= P(X \leq x) \\ &= \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \end{aligned} \qquad \textbf{[CDF Definition]}$$

$$E(X) = \mu \qquad \textbf{[Expected Value]}$$

$$var(X) = \sigma^2 \qquad \textbf{[Variance]}$$

We won't simplify the CDF any further here because doing so would require convoluted functions like the Gauss Error function; instead, we will calculate the CDF and inverse CDF numerically using functions in R.

```
> dnorm(x=0.20,mean=0,sd=1) #This the height of the PDF at x=0.20
[1] 0.3910427
> pnorm(q=2,mean=0,sd=1) #P(X<=2|mu=0,sigma=1)
[1] 0.5792597
> qnorm(p=0.90,mean=0,sd=1) #The 90th percentile
[1] 1.281552
> rnorm(n=5,mean=0,sd=1) #A random sample of 5 Xi~Gaussian(mu=0,sigma=1)
[1]  0.86441607  0.38121236 -1.55845605  1.28107061  0.05139334
```

Additionally, the graphs can be completed in R as follows.

```
> ggdat<-data.frame(x=seq(-6,6,0.001),
+                   f1=dnorm(x=seq(-6,6,0.001),mean=0,sd=1),
+                   f2=dnorm(x=seq(-6,6,0.001),mean=0,sd=3),
+                   f3=dnorm(x=seq(-6,6,0.001),mean=0,sd=0.5),
+                   f4=dnorm(x=seq(-6,6,0.001),mean=1,sd=1),
+                   f5=dnorm(x=seq(-6,6,0.001),mean=-1,sd=2))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1,color="m=0 sd=1"))+
+   geom_line(aes(y=f2,color="m=0 sd=3"))+
+   geom_line(aes(y=f3,color="m=0 sd=0.50"))+
+   geom_line(aes(y=f4,color="m=1 sd=1"))+
+   geom_line(aes(y=f5,color="m=-1 sd=2"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
```

```
+   ggtitle("Gaussian PDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("m=0 sd=1","m=0 sd=3","m=0 sd=0.50",
+                                    "m=1 sd=1","m=-1 sd=2"),
+                        labels=c(bquote(mu==0~","~sigma==0.50),
+                                 bquote(mu==0~","~sigma==1),bquote(mu==0~","~sigma==3),
+                                 bquote(mu==1~","~sigma==1),bquote(mu==-1~","~sigma==2)))
> ggdat<-data.frame(x=seq(-6,6,0.001),
+                   F1=pnorm(q=seq(-6,6,0.001),mean=0,sd=1),
+                   F2=pnorm(q=seq(-6,6,0.001),mean=0,sd=3),
+                   F3=pnorm(q=seq(-6,6,0.001),mean=0,sd=0.5),
+                   F4=pnorm(q=seq(-6,6,0.001),mean=1,sd=1),
+                   F5=pnorm(q=seq(-6,6,0.001),mean=-1,sd=2))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1,color="m=0 sd=1"))+
+   geom_line(aes(y=F2,color="m=0 sd=3"))+
+   geom_line(aes(y=F3,color="m=0 sd=0.50"))+
+   geom_line(aes(y=F4,color="m=1 sd=1"))+
+   geom_line(aes(y=F5,color="m=-1 sd=2"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Gaussian CDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("m=0 sd=1","m=0 sd=3","m=0 sd=0.50",
+                                    "m=1 sd=1","m=-1 sd=2"),
+                        labels=c(bquote(mu==0~","~sigma==0.50),
+                                 bquote(mu==0~","~sigma==1),bquote(mu==0~","~sigma==3),
+                                 bquote(mu==1~","~sigma==1),bquote(mu==-1~","~sigma==2)))
> grid.arrange(g1,g1.cdf,ncol=2)
```

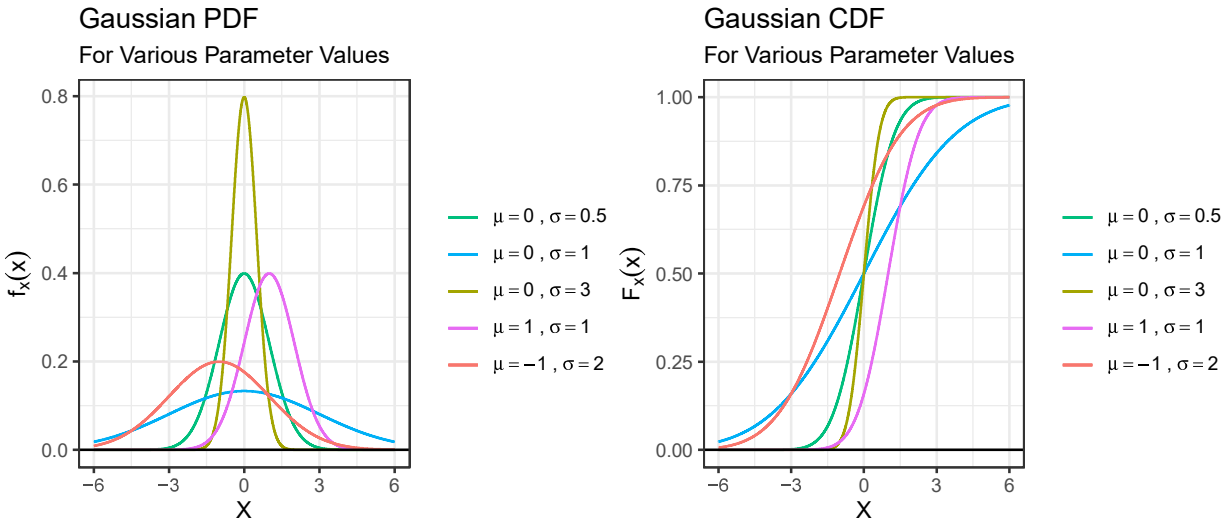

Figure 6.3.18: Gaussian PDF(top) and CDF(bottom) for $\mu = 0$ and $\sigma = 1$.

Example 6.23. Suppose that the wing length of houseflies are Gaussian distributed with a mean

wing length is 45.5mm and a standard deviation 3.92mm.

Question: What is the probability that a randomly selected housefly has wing length greater than 40mm?
Answer: We can calculate this probability using the Gaussian distribution.

$$\begin{aligned} P(X > 40|\mu = 45.5, \sigma = 3.92) &= 1 - P(X \leq 40|\mu = 45.5, \sigma = 3.92) \qquad \textbf{[Complement rule]} \\ &= 1 - F_X(40|\mu = 45.5, \sigma = 3.92) \\ &= 1 - \int_{-\infty}^{40} \frac{1}{3.92\sqrt{2\pi}} e^{\frac{-(x-45.5)^2}{2(3.92)^2}} dx \end{aligned}$$

This integral is troublesome and for a long time we were left searching through printed tables for an answer. Now, we can ask for this probability in R as follows.

```
> 1- pnorm(q=40,mean=45.5,sd=3.92)#Complement rule
[1] 0.9197007
>  pnorm(q=40,mean=45.5,sd=3.92,lower.tail=FALSE)#Ask directly
[1] 0.9197007
```

We can visualize this probability in R as follows.

```
> ggdat<-data.frame(x=seq(30,60,0.01),
+                   f1=dnorm(x=seq(30,60,0.01),mean=45.5,sd=3.92))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x>40),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Housefly Wingspan in mm (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Gaussian PDF",subtitle=bquote(P(X>40)~"for"~mu==45.5*","~sigma==3.92))
> ggdat<-data.frame(x=seq(30,60,0.01),
+                   F1=pnorm(q=seq(30,60,0.01),mean=45.5,sd=3.92))
> ggdat.highlight<-data.frame(x=40,
+                             y=pnorm(q=40,mean=45.5,sd=3.92))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Housefly Wingspan in mm (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Gaussian CDF",subtitle=bquote(P(X<=40)~"for"~mu==45.5*","~sigma==3.92))
> grid.arrange(g1,g1.cdf,ncol=2)
```

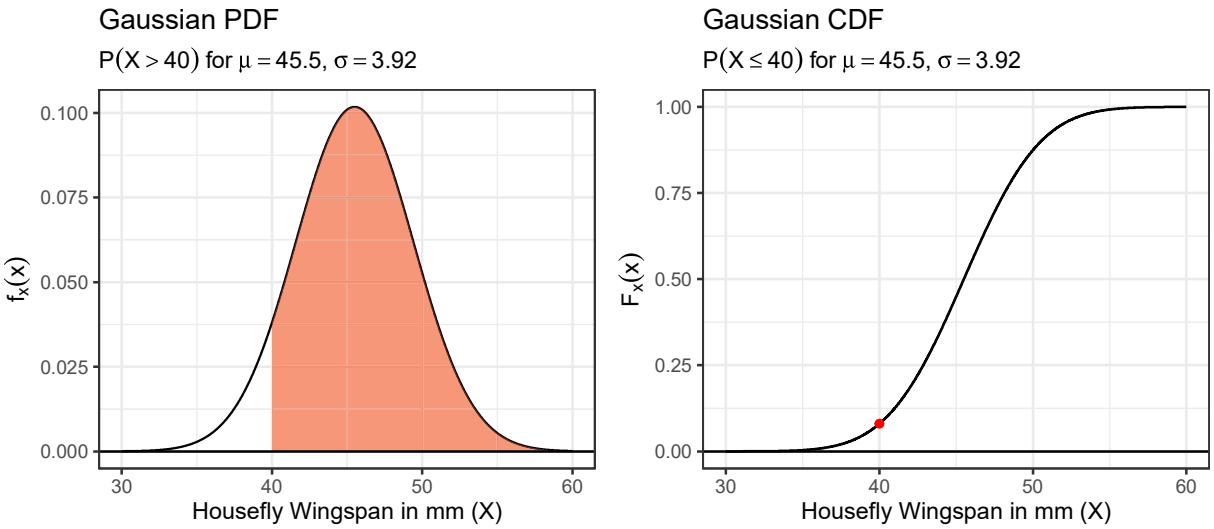

Figure 6.3.19: Normal PDF (left) and CDF (right) for $\mu = 45.5$ and $\sigma = 3.92$ as in Example 6.23 with area under the PDF shaded from 40mm to the right representing $P(X > 40)$. The observation at $X = 40$ is highlighted in the CDF with a red point; the corresponding $F_X(x)$ value is equal to $P(X \leq 40|\mu = 45.5, \sigma = 3.92)$ and the probability of interest can be calculated with the complement rule.

Question: What is the probability that a randomly selected housefly has wing length of at most 50mm?
Answer: We can calculate this probability using the Normal distribution.

$$\begin{aligned} P(X \leq 50|\mu = 45.5, \sigma = 3.92) &= F_X(50|\mu = 45.5, \sigma = 3.92) \\ &= \int_{-\infty}^{50} \frac{1}{3.92\sqrt{2\pi}} e^{\frac{-(x-45.5)^2}{2(3.92)^2}} dx \end{aligned}$$

we can ask for this probability in R as follows.

```
> pnorm(q=50,mean=45.5,sd=3.92)
```

We can visualize this probability in R as follows.

```
> ggdat<-data.frame(x=seq(30,60,0.01),
+                   f1=dnorm(x=seq(30,60,0.01),mean=45.5,sd=3.92))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x<=50),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Housefly Wingspan in mm (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Gaussian PDF",subtitle=bquote(P(X=40)+P(X>50)~"for"~mu==45.5*","~sigma==3.92))
```

```
> ggdat<-data.frame(x=seq(30,60,0.01),
+                   F1=pnorm(q=seq(30,60,0.01),mean=45.5,sd=3.92))
> ggdat.highlight<-data.frame(x=50,
+                             y=pnorm(q=50,mean=45.5,sd=3.92))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Housefly Wingspan in mm (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Gaussian CDF",subtitle=bquote(P(X<=50)~"for"~mu==45.5*","~sigma==3.92))
> grid.arrange(g1,g1.cdf,ncol=2)
```

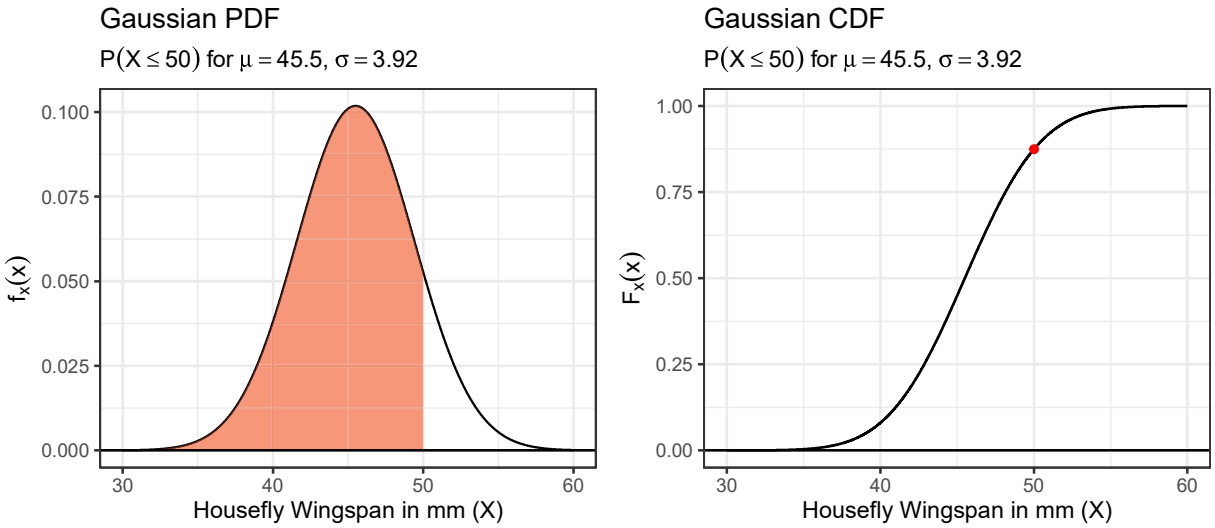

Figure 6.3.20: Normal PDF (left) and CDF (right) for $\mu = 45.5$ and $\sigma = 3.92$ as in Example 6.23 with area under the PDF shaded from 50mm to the left representing $P(X \leq 50)$. The observation at $X = 50$ is highlighted in the CDF with a red point; the corresponding $F_X(x)$ value is equal to $P(X \leq 40|\mu = 45.5, \sigma = 3.92)$.

Question: What is the probability that a randomly selected housefly has wing length less than 40mm or more than 50mm?
Answer: We can calculate this probability using the Normal distribution.

$$\begin{aligned} P(X < 40 \cup X > 50|\mu = 45.5, \sigma = 3.92) &= P(X < 40|\mu = 45.5, \sigma = 3.92) + P(X > 50|\mu = 45.5, \sigma = 3.92) \\ &= P(X \leq 40|\mu = 45.5, \sigma = 3.92) + (1 - P(X \leq 50|\mu = 45.5, \sigma = 3.92)) \\ &= F_X(40|\mu = 45.5, \sigma = 3.92) + (1 - F_X(50|\mu = 45.5, \sigma = 3.92)) \\ &= \int_{-\infty}^{40} \frac{1}{3.92\sqrt{2\pi}} e^{\frac{-(x-45.5)^2}{2(3.92)^2}} dx + \left(1 - \int_{-\infty}^{50} \frac{1}{3.92\sqrt{2\pi}} e^{\frac{-(x-45.5)^2}{2(3.92)^2}} dx\right) \end{aligned}$$

We can ask for this probability in R as follows.

```
> pnorm(q=40,mean=45.5,sd=3.92) + (1-pnorm(q=50,mean=45.5,sd=3.92))
[1] 0.205792
```

We can visualize this probability in R as follows.

```
> ggdat<-data.frame(x=seq(30,60,0.01),
+                   f1=dnorm(x=seq(30,60,0.01),mean=45.5,sd=3.92))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1))+
+   geom_ribbon(data=subset(ggdat,x<40),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,x>50),aes(ymax=f1),ymin=0,
+               fill="red",colour=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Housefly Wingspan in mm (X)")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Gaussian PDF",subtitle=bquote(P(X>40)~"for"~mu==45.5*","~sigma==3.92))
> ggdat<-data.frame(x=seq(30,60,0.01),
+                   F1=pnorm(q=seq(30,60,0.01),mean=45.5,sd=3.92))
> ggdat.highlight<-data.frame(x=c(40,50),
+                             y=pnorm(q=c(40,50),mean=45.5,sd=3.92))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1))+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Housefly Wingspan in mm (X)")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Gaussian CDF",subtitle=bquote(P(X<=40)~~"and"~P(X<=50)"for"~mu==45.5*
+   ","~sigma==3.92))
> grid.arrange(g1,g1.cdf,ncol=2)
```

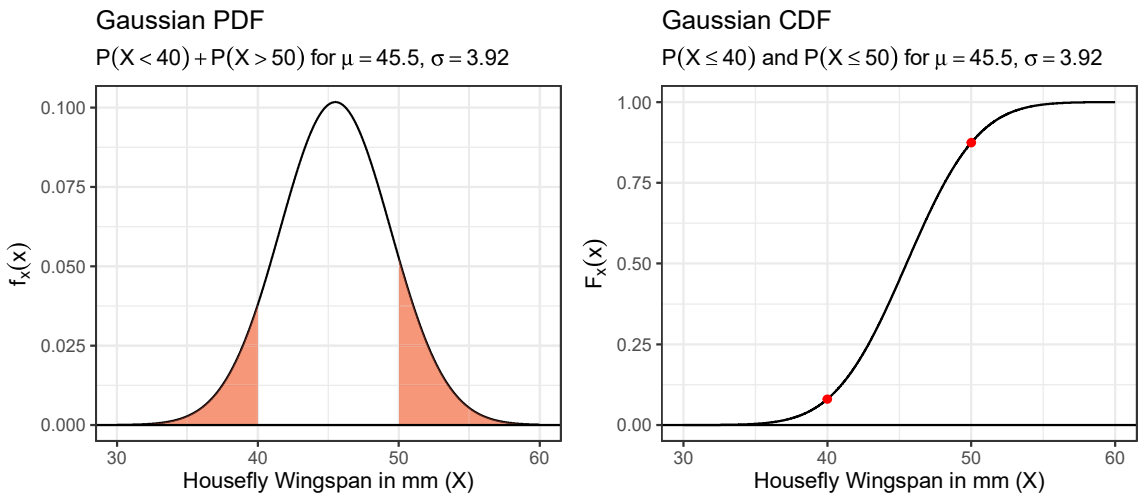

Figure 6.3.21: Normal PDF (left) and CDF (right) for $\mu = 45.5$ and $\sigma = 3.92$ as in Example 6.23 with area under the PDF shaded from 40mm to the left and from 50mm to the right representing $P(X < 40 \cup X > 50)$. The observations of $X = 40$ and $X = 50$ are highlighted in the CDF with red points; the corresponding $F_X(x)$ values are equal to $P(X \leq 40|\mu = 45.5, \sigma = 3.92)$ and $P(X \leq 50|\mu = 45.5, \sigma = 3.92)$ which are used to calculate the probability of interest.

Definition 6.24. The **Log-Normal distribution** is used for the random variable, X, which can take on positive real values. This distribution is a popular competitor to the Weibull distribution in reliability/engineering applications. Graphs of the PDF and CDF are seen in Figure 6.3.22.

$$\begin{aligned}
\mu \in \mathbb{R} \text{ and } \sigma \in \mathbb{R}^+ & \qquad \textbf{[Parameters]} \\
\mathcal{X} = \{\omega : \omega \in (0, \infty)\} & \qquad \textbf{[Support]} \\
f_X(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{(ln(x)-\mu)^2}{2\sigma^2}} \; I(x \in (0, \infty)) & \qquad \textbf{[PDF]} \\
F_X(x|\mu, \sigma) = P(X \leq x) & \\
= \int_0^x \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}} \; dx & \qquad \textbf{[CDF Definition]} \\
E(X) = e^{\mu+\sigma^2/2} & \qquad \textbf{[Mean]} \\
var(X) = e^{2\mu+\sigma^2} e^{\sigma^2-1} & \qquad \textbf{[Variance]}
\end{aligned}$$

We won't simplify the CDF any further here because doing so would require convoluted functions like the Gauss Error function; instead, we will calculate the CDF and inverse CDF numerically using functions in R.

```
> dlnorm(x=0.50,meanlog=0,sdlog=1) #This the height of the PDF at x=0.50
[1] 0.6274961
> plnorm(q=0.30,meanlog=0,sdlog=1) #P(X<=0.30|mu=0,sd=1)
[1] 0.1143
> qlnorm(p=0.90,meanlog=0,sdlog=1) #The 90th percentile
[1] 3.602224
```

```
> rlnorm(n=5,meanlog=0,sdlog=1) #A random sample of 5 Xi~lognormals(mu=0,sd=1)
[1] 2.2851451 0.7551318 1.5078784 0.6281679 0.3747440
```

Additionally, the graphs can be completed in R as follows.

```
> ggdat<-data.frame(x=seq(0,6,0.001),
+                   f1=dlnorm(x=seq(0,6,0.001),meanlog=0,sdlog=1),
+                   f2=dlnorm(x=seq(0,6,0.001),meanlog=0,sdlog=0.25),
+                   f3=dlnorm(x=seq(0,6,0.001),meanlog=0,sdlog=0.5),
+                   f4=dlnorm(x=seq(0,6,0.001),meanlog=1,sdlog=1),
+                   f5=dlnorm(x=seq(0,6,0.001),meanlog=-1,sdlog=2))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1,color="m=0 sdlog=1"))+
+   geom_line(aes(y=f2,color="m=0 sdlog=0.25"))+
+   geom_line(aes(y=f3,color="m=0 sdlog=0.50"))+
+   geom_line(aes(y=f4,color="m=1 sdlog=1"))+
+   geom_line(aes(y=f5,color="m=-1 sdlog=2"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Log-Normal PDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("m=0 sdlog=1","m=0 sdlog=0.25","m=0 sdlog=0.50",
+                                    "m=1 sdlog=1","m=-1 sdlog=2"),
+                        labels=c(bquote(mu==0~","~sigma==0.50),
+                                 bquote(mu==0~","~sigma==1),bquote(mu==0~","~sigma==0.25),
+                                 bquote(mu==1~","~sigma==1),bquote(mu==-1~","~sigma==2)))
> ggdat<-data.frame(x=seq(0,6,0.001),
+                   F1=plnorm(q=seq(0,6,0.001),meanlog=0,sdlog=1),
+                   F2=plnorm(q=seq(0,6,0.001),meanlog=0,sdlog=0.25),
+                   F3=plnorm(q=seq(0,6,0.001),meanlog=0,sdlog=0.5),
+                   F4=plnorm(q=seq(0,6,0.001),meanlog=1,sdlog=1),
+                   F5=plnorm(q=seq(0,6,0.001),meanlog=-1,sdlog=2))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1,color="m=0 sdlog=1"))+
+   geom_line(aes(y=F2,color="m=0 sdlog=0.25"))+
+   geom_line(aes(y=F3,color="m=0 sdlog=0.50"))+
+   geom_line(aes(y=F4,color="m=1 sdlog=1"))+
+   geom_line(aes(y=F5,color="m=-1 sdlog=2"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Log-Normal CDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("m=0 sdlog=1","m=0 sdlog=0.25","m=0 sdlog=0.50",
+                                    "m=1 sdlog=1","m=-1 sdlog=2"),
+                         labels=c(bquote(mu==0~","~sigma==0.50),
+                                 bquote(mu==0~","~sigma==1),bquote(mu==0~","~sigma==0.25),
+                                 bquote(mu==1~","~sigma==1),bquote(mu==-1~","~sigma==2)))
> grid.arrange(g1,g1.cdf,ncol=2)
```

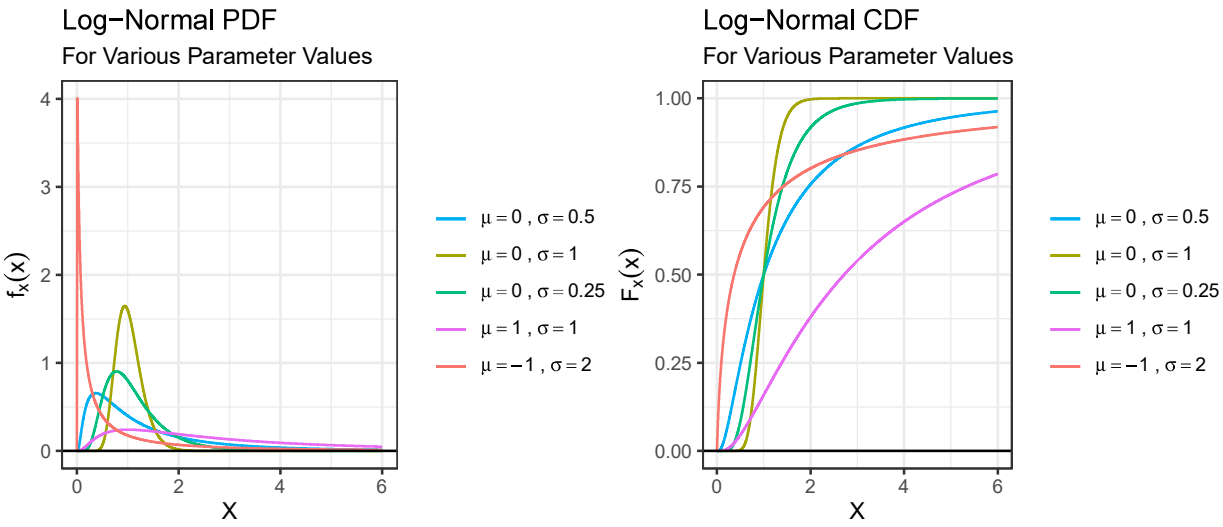

Figure 6.3.22: Log-Normal PDF(left) and CDF(right) for $\mu = 0$ and $\sigma = 1$.

Definition 6.25. the **chi-squared** $(\mathcal{X}^2_v)$ distribution. This distribution is right skewed and often used for various hypothesis testing methodologies, including the chi-squared independence hypothesis test.

$$v \in \mathbb{N} \qquad \textbf{[Parameters]}$$

$$S = \mathbb{R}^+ \qquad \textbf{[Support]}$$

$$f_X(x) = \frac{1}{\Gamma\left(\frac{v}{2}\right) 2^{v/2}} x^{\frac{v}{2}-1} e^{\frac{-x}{2}} \qquad \textbf{[PDF]}$$

$$F_X(x) = \int_0^x \frac{1}{\Gamma\left(\frac{v}{2}\right) 2^{v/2}} x^{\frac{v}{2}-1} e^{\frac{-x}{2}} dx \qquad \textbf{[CDF]}$$

$$E(X) = v \qquad \textbf{[Mean]}$$

$$var(X) = 2v \qquad \textbf{[Variance]}$$

Note: The parameter v is called the **degrees of freedom**.

We won't simplify the CDF any further here. Instead, we calculate the CDF and inverse CDF numerically using functions in `R`

```
> dchisq(x=0.50,df=10) #This the height of the PDF at x=0.50
[1] 6.337897e-05
> pchisq(q=0.250,df=10) #P(X<=0.0.250|v=10)
[1] 2.29191e-07
> qchisq(p=0.90,df=10) #The 90th percentile
[1] 15.98718
> rchisq(n=5,df=10) #A random sample of 5 Xi~chisquared(v=10)
[1] 13.093251 16.816490  7.987314  8.294768 10.044074
```

Graphs of the PDF and CDF for various parameter values are seen in Figure 6.3.23, which are created with the following `R` code.

```
> > ggdat<-data.frame(x=seq(0,10,0.001),
+                   f1=dchisq(x=seq(0,10,0.001),df=1),
+                   f2=dchisq(x=seq(0,10,0.001),df=3),
+                   f3=dchisq(x=seq(0,10,0.001),df=5),
+                   f4=dchisq(x=seq(0,10,0.001),df=8),
+                   f5=dchisq(x=seq(0,10,0.001),df=10))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1,color="df=1"))+
+   geom_line(aes(y=f2,color="df=3"))+
+   geom_line(aes(y=f3,color="df=5"))+
+   geom_line(aes(y=f4,color="df=8"))+
+   geom_line(aes(y=f5,color="df=10"))+
+   geom_hline(yintercept=0)+
+   ylim(0,1)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Chi-squared PDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("df=1","df=3","df=5","df=8","df=10"),
+                        labels=c(bquote(v==1),bquote(v==3),bquote(v==5),
+                                 bquote(v==8),bquote(v==10)))
> ggdat<-data.frame(x=seq(0,10,0.001),
+                   F1=pchisq(q=seq(0,10,0.001),df=1),
+                   F2=pchisq(q=seq(0,10,0.001),df=3),
+                   F3=pchisq(q=seq(0,10,0.001),df=5),
+                   F4=pchisq(q=seq(0,10,0.001),df=8),
+                   F5=pchisq(q=seq(0,10,0.001),df=10))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1,color="df=1"))+
+   geom_line(aes(y=F2,color="df=3"))+
+   geom_line(aes(y=F3,color="df=5"))+
+   geom_line(aes(y=F4,color="df=8"))+
+   geom_line(aes(y=F5,color="df=10"))+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Chi-squared CDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("df=1","df=3","df=5","df=8","df=10"),
+                        labels=c(bquote(v==1),bquote(v==3),bquote(v==5),
+                                 bquote(v==8),bquote(v==10)))
> grid.arrange(g1,g1.cdf,ncol=2)
```

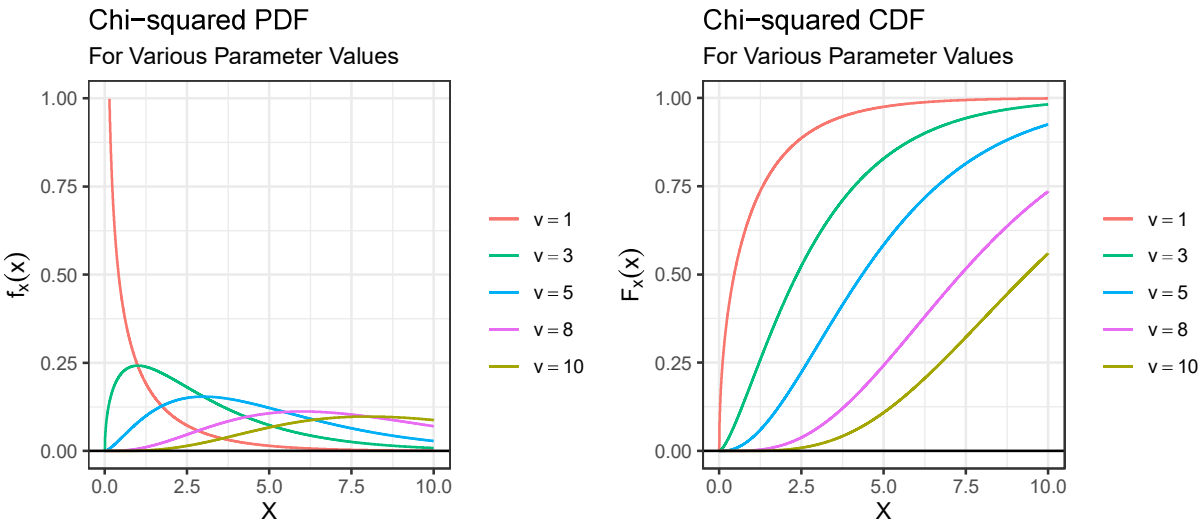

Figure 6.3.23: Chi-squared PDF(left) and CDF(right) for $\mu = 0$ and $\sigma = 1$.

Definition 6.26. *Student's t distribution:* Suppose that $U \sim \text{Gaussian}(0,1)$, $V \sim \chi^2_v$, and $U \perp\!\!\!\perp V$. The random variable

$$T = \frac{U}{\sqrt{V/v}} \sim t_v,$$

a t distribution with v degrees of freedom. The t **distribution** with $n-1$ degrees of freedom models a continuous quantitative random variable T.

$$v = \in \mathbb{N}^+ \qquad \textbf{[Parameters]}$$

$$\mathcal{X} = \mathbb{R} \qquad \textbf{[Sample space]}$$

$$f_T(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi}\,\Gamma(v/2)} \left(1 + \frac{t^2}{2}\right)^{-(v+1)/2} \qquad \textbf{[PDF]}$$

$$F_X(x) = \int_{-\infty}^{x} \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi}\,\Gamma(v/2)} \left(1 + \frac{t^2}{2}\right)^{-(v+1)/2} dx \qquad \textbf{[CDF]}$$

$$E(X) = 0 \qquad \textbf{[Mean for } v > 1\textbf{]}$$

$$var(X) = \frac{v}{v-2} \qquad \textbf{[Variance for } v > 2\textbf{]}$$

The PDF and CDF of a T distributed random variable can be calculated in R.

```
> > dt(x=0.50,df=10) #This the height of the PDF at x=0.50
[1] 0.3396951
> pt(q=0.250,df=10) #P(X<=0.0.250|v=10)
[1] 0.5961759
> qt(p=0.90,df=10) #The 90th percentile
[1] 1.372184
> rt(n=5,df=10) #A random sample of 5 Xi~t(v=10)
[1] -1.5680473  2.8102648 -2.6883653 -0.5348518  0.3863503
```

Graphs of the PDF and CDF for various parameter values are seen in Figure 6.3.24, which are created with the following R code.

```
> ggdat<-data.frame(x=seq(-4,4,0.001),
+                   f1=dt(x=seq(-4,4,0.001),df=1),
+                   f2=dt(x=seq(-4,4,0.001),df=3),
+                   f3=dt(x=seq(-4,4,0.001),df=5),
+                   f4=dt(x=seq(-4,4,0.001),df=15),
+                   f5=dt(x=seq(-4,4,0.001),df=30),
+                   fnorm<-dnorm(x=seq(-4,4,0.001),mean=0,sd=1))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=f1,color="df=1"))+
+   geom_line(aes(y=f2,color="df=5"))+
+   geom_line(aes(y=f3,color="df=15"))+
+   geom_line(aes(y=f4,color="df=30"))+
+   geom_line(aes(y=f5,color="df=50"))+
+   geom_line(aes(y=fnorm,color="Standard Normal"),linetype="dashed",size=1.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Student T PDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("df=1","df=5","df=15","df=30","df=50","Standard Normal"),
+                        labels=c(bquote(v==1),bquote(v==5),bquote(v==15),
+                                 bquote(v==30),bquote(v==50),"Standard Normal"))
> ggdat<-data.frame(x=seq(-4,4,0.001),
+                   F1=pt(q=seq(-4,4,0.001),df=1),
+                   F2=pt(q=seq(-4,4,0.001),df=3),
+                   F3=pt(q=seq(-4,4,0.001),df=5),
+                   F4=pt(q=seq(-4,4,0.001),df=15),
+                   F5=pt(q=seq(-4,4,0.001),df=30),
+                   Fnorm<-pnorm(q=seq(-4,4,0.001),mean=0,sd=1))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1,color="df=1"))+
+   geom_line(aes(y=F2,color="df=5"))+
+   geom_line(aes(y=F3,color="df=15"))+
+   geom_line(aes(y=F4,color="df=30"))+
+   geom_line(aes(y=F5,color="df=50"))+
+   geom_line(aes(y=Fnorm,color="Standard Normal"),linetype="dashed",size=1.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Student T CDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("df=1","df=5","df=15","df=30","df=50","Standard Normal"),
+                        labels=c(bquote(v==1),bquote(v==5),bquote(v==15),
+                                 bquote(v==30),bquote(v==50),"Standard Normal"))
> grid.arrange(g1,g1.cdf,ncol=2)
```

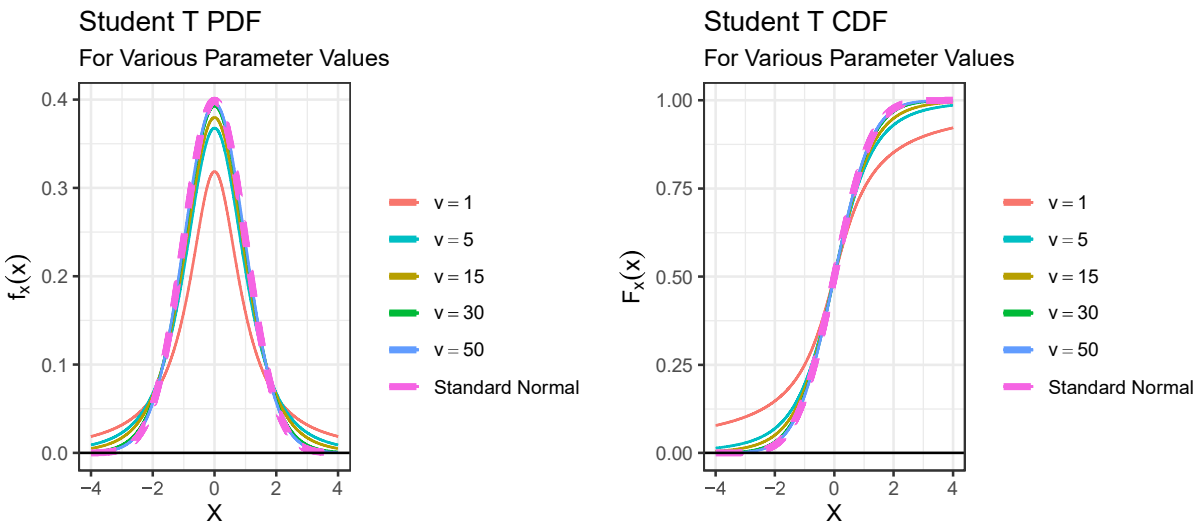

Figure 6.3.24: Student T PDF(left) and CDF(right) for various parameter values.

Some details about the T distribution:

- discovered in 1908 by William S. Gosset, a statistician at the Guinness brewing company
- continuous and symmetric about zero, like the normal distribution
- indexed by a parameter called the **degrees of freedom** which is related to the sample size (thus there are infinitely many t distributions indexed by the degrees of freedom)

$$v = \text{degrees of freedom} = n - 1$$

- As n increases the t distribution approaches the standard normal distribution, e.g. the normal distribution with $\mu_X = 0$ and $\sigma_X = 1$. See Figure 6.3.24.
- When compared to the standard normal distribution, the t distribution, in general, is less peaked, and has more area in the tails.

Note: If $v = 1$, then $T \sim \text{Cauchy}(0, 1)$.

Application: Suppose that $X_1, X_2, ..., X_n$ are *iid* Gaussian(μ, σ^2). We already know that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Gaussian}(0, 1).$$

The quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where S denotes the sample standard deviation of $X_1, X_2, ..., X_n$. To see why, note that

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sigma}{S}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)$$
$$= \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} \sim \frac{\text{“Gaussian}(0,1)\text{”}}{\sqrt{\frac{\text{“}\chi^2_{n-1}\text{”}}{n-1}}}.$$

Because $\bar{X} \perp\!\!\!\perp S^2$, the numerator and denominator are independent. Therefore, $T \sim t_{n-1}$.

Definition 6.27. *Snedecor's F distribution* Suppose that $U \sim \chi^2_u$, $V \sim \chi^2_v$, and $U \perp\!\!\!\perp V$. The random variable

$$W = \frac{U/u}{V/v} \sim F_{u,v},$$

an F distribution with (numerator) u and (denominator) v degrees of freedom.

$$u \in \mathbb{N}^+, v \in \mathbb{N}^+ \qquad \textbf{[Parameters]}$$
$$f_W(w) = \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})}\left(\frac{u}{v}\right)^{u/2}\frac{w^{\frac{u}{2}-1}}{[1+(\frac{u}{v})w]^{(u+v)/2}}\, I(w>0) \qquad \textbf{[PDF]}$$
$$F_W(w) = \int_0^x \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})}\left(\frac{u}{v}\right)^{u/2}\frac{w^{\frac{u}{2}-1}}{[1+(\frac{u}{v})w]^{(u+v)/2}}dw \qquad \textbf{[CDF]}$$
$$E(W) = \frac{v}{v-2} \qquad (\textbf{[Expected Value for } v > 2\textbf{]})$$
$$var(W) = \left(\frac{u-2}{u}\right)\left(\frac{v}{v+2}\right) \qquad (\textbf{[Variance]})$$

The PDF and CDF of a F distributed random variable can be calculated in R.

```
> df(x=0.50,df1=10,df2=5) #This the height of the PDF at x=0.50
[1] 0.648023
> pf(q=0.250,df1=10,df2=5) #P(X<=0.0.250|u=10,v=5)
[1] 0.0296753
> qf(p=0.90,df1=10,df2=5) #The 90th percentile
[1] 3.297402
> rf(n=5,df1=10,df2=5) #A random sample of 5 Xi~F(u=10,v=5)
[1] 0.4310192 1.6080171 0.6712107 1.8694693 4.2602548
```

Graphs of the PDF and CDF for various parameter values are seen in Figure 6.3.25, which are created with the following R code.

```
> ggdat<-data.frame(x=seq(0,5,0.001),
+                   f1=df(x=seq(0,5,0.001),df1=1,df2=1),
+                   f2=df(x=seq(0,5,0.001),df1=5,df2=1),
+                   f3=df(x=seq(0,5,0.001),df1=1,df2=5),
+                   f4=df(x=seq(0,5,0.001),df1=5,df2=10),
+                   f5=df(x=seq(0,5,0.001),df1=10,df2=10))
> g1<-ggplot(data=ggdat,aes(x=x))+
```

```
+   geom_line(aes(y=f1,color="df1=1 df2=1"))+
+   geom_line(aes(y=f2,color="df1=5 df2=1"))+
+   geom_line(aes(y=f3,color="df1=1 df2=5"))+
+   geom_line(aes(y=f4,color="df1=5 df2=10"))+
+   geom_line(aes(y=f5,color="df1=10 df2=10"))+
+   geom_hline(yintercept=0)+
+   ylim(0,1)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("F PDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("df1=1 df2=1","df1=5 df2=1","df1=1 df2=5",
+                                    "df1=5 df2=10","df1=10 df2=10"),
+                        labels=c(bquote(v[1]==1~","~v[2]==1),bquote(v[1]==5~","~v[2]==1),
+                                 bquote(v[1]==1~","~v[2]==5),bquote(v[1]==5~","~v[2]==10),
+                                 bquote(v[1]==10~","~v[2]==10)))
> grid.arrange(g1,g1.cdf,ncol=2)
> ggdat<-data.frame(x=seq(0,5,0.001),
+                   F1=pf(q=seq(0,5,0.001),df1=1,df2=1),
+                   F2=pf(q=seq(0,5,0.001),df1=5,df2=1),
+                   F3=pf(q=seq(0,5,0.001),df1=1,df2=5),
+                   F4=pf(q=seq(0,5,0.001),df1=5,df2=10),
+                   F5=pf(q=seq(0,5,0.001),df1=10,df2=10))
> g1.cdf<-ggplot(data=ggdat,aes(x=x))+
+   geom_line(aes(y=F1,color="df1=1 df2=1"))+
+   geom_line(aes(y=F2,color="df1=5 df2=1"))+
+   geom_line(aes(y=F3,color="df1=1 df2=5"))+
+   geom_line(aes(y=F4,color="df1=5 df2=10"))+
+   geom_line(aes(y=F5,color="df1=10 df2=10"))+
+   geom_hline(yintercept=0)+
+   ylim(0,1)+
+   theme_bw()+
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("F CDF",subtitle="For Various Parameter Values")+
+   scale_color_discrete("",breaks=c("df1=1 df2=1","df1=5 df2=1","df1=1 df2=5",
+                                    "df1=5 df2=10","df1=10 df2=10"),
+                        labels=c(bquote(v[1]==1~","~v[2]==1),bquote(v[1]==5~","~v[2]==1),
+                                 bquote(v[1]==1~","~v[2]==5),bquote(v[1]==5~","~v[2]==10),
+                                 bquote(v[1]==10~","~v[2]==10)))
> grid.arrange(g1,g1.cdf,ncol=2)
```

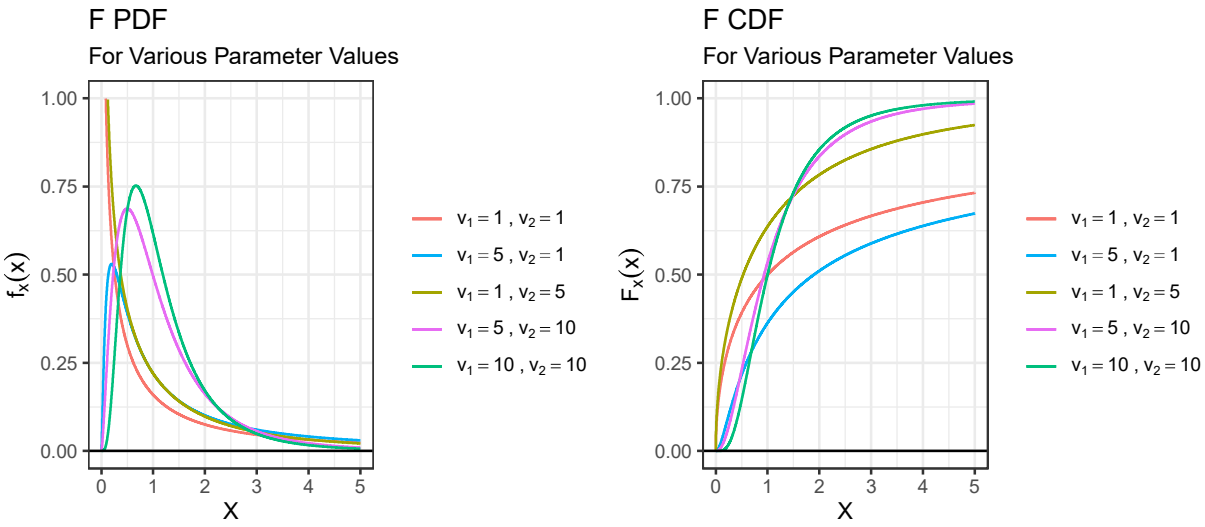

Figure 6.3.25: F PDF(left) and CDF(right) for various parameter values.