Lab 2. Many Distributions: Discrete Random Variables

Objectives

- Discrete distributions in R
- Related statistics, properties, and simulation

Distributions in R

R has all the functions to handle many probability distributions. The table below gives the names of the distribution functions.

Distributions	Percentile	Quantile	Density	Random numbers
β (beta)	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
χ^2 (chi-square)	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Γ (Gamma)	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student's t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibul
Wilcoxon Rank Sum	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank	psignrank	qsignrank	dsignrank	rsignrank

Every distribution has four functions. There is a "root" distribution name, and it's prefixed by one of the letters

- p for "probability", the cumulative distribution function (cdf) will be produced
- q for "quantile", (i.e., the inverse cdf)
- d for "density", the density function
- r for "random", random variables having the specified distribution will be produced

Uniform random variable – Discrete Case

$$X \sim \text{uniform}, \qquad x = 1, 2, \ldots, m$$

Then,

$$f(x) = \frac{1}{m}, \qquad x = 1, 2, \ldots, m$$

$$\mu = \frac{m+1}{2}, \qquad \sigma^2 = \frac{m^2 - 1}{12}$$

Uniform random variable

$$\mu = \sum_{i=1}^{m} x_i \cdot p(x_i) = \frac{1}{m} \sum_{i=1}^{m} i = \frac{1}{m} \cdot \frac{m(m+1)}{2} = \frac{m+1}{2}$$

$$E(X^2) = \sum_{i=1}^{m} x_i^2 \cdot p(x_i) = \frac{1}{m} \sum_{i=1}^{m} i^2 = \frac{1}{m} \cdot \frac{m(m+1)(2m+1)}{6} = \frac{(m+1)(2m+1)}{6}$$

$$\sigma^2 = E(X^2) - \{E(X)\}^2 = \frac{(m+1)(2m+1)}{6} - \frac{(m+1)^2}{4} = \frac{m^2-1}{12}$$

Here is how to simulate discrete uniform random numbers.

```
x <- sample(1:10, 1000, rep=T)
table(x)
mean(x)
sd(x)
var(x)
(100-1)/12
```

The last part is to check if the sample mean and sample variance are close to the theoretical mean and variance. For a discrete uniform random variable, $\mu = \frac{m+1}{2}$ and $\sigma^2 = \frac{m^2-1}{12}$.

Geometric random variable

$$Y \sim \text{geometric}\,(p),$$

where Y = "number" of "trials" (i.e., "failures") **before** the 1st "success," and p is the probability of "success" in a single trial. It's called a "fair" coin when $p = 0.5$.

For example,

- $Y \sim$ geometric (0.5) means you're flipping a "fair" coin and wondering about how many "times" you have to keep doing it before the first head shows up.
- $Y \sim$ geometric $\left(\frac{1}{6}\right)$ means you're rolling a "fair" die and record how many "times" you roll before the first 6 shows up, for example.

Here is how to simulate the number of times you have to flip a "fair" coin before the first "head" shows up. By the way, when the very first trial is a "success," then $Y = 0$ (not 1).

```
(y <- rgeom(10,0.5))
(y <- rgeom(100,0.5))
```

Now, let's simulate 1000 geometric random numbers with p=0.5.

```
table(y <- rgeom(1000,0.5))
mean(y); sd(y)
hist(y,breaks=c(0:10),prob=T)     # prob=TRUE plots 'density', NOT frequency
(0.5/0.5)
sqrt(0.5/(0.5^2))
```

The last part is to check if the sample mean and sample sd are close to the theoretical mean and sd. For a geometric random variable, $\mu = \frac{q}{p}$ and $\sigma^2 = \frac{q}{p^2}$.

We can also compute the probability of a geometric random variable. For a geometric random variable Y with a parameter p, we have the pmf as

$$f(y) = p \cdot (1-p)^y, \qquad y = 0, 1, 2, \ldots$$

In plain terms, this is the probability of "having" y many *trials* before you flip the "first" head of a coin, and p stands for the *quality* of a coin. Some more examples of a geometric random variable $Y \sim$ geometric(p) are shown below.

One note: Textbook (page. 64) uses a slightly different definition. There, $X =$ the trial number on which the 1st "success" is observed and it's related by $Y = X - 1$, $(x = 1, 2, \ldots)$.

Ex 1: Geometric

1. Let $Y \sim$ geometric with $p = 0.5$. Find the probability of having tried 4 "times" before the first "head."

```
dgeom(4,0.5)
[1] 0.03125
(0.5^4)*0.5                  #This is by plugging into the pmf equation.
[1] 0.03125
pgeom(4,0.5)-pgeom(3,0.5)    #P(x=4) can also be found by P(x<=4)-P(x<=3)
[1] 0.03125
```

2. Assume that the probability of a defective computer component is 0.02. Components are randomly selected.
   (a) Find the probability that the first defect is caused by the seventh component tested.
   (b) How many components do you expect to test until one is found to be defective? Let $X =$ the number of computer components tested until the first defect is found.

```
dgeom(6,0.02)
(0.98/0.02)
sqrt(0.98/(0.02^2))
```

3. Let $Y \sim$ geometric with $p = 0.5$. Construct the probability distribution table of X.

```
y <- 0:10
prob <- dgeom(y,0.5)
cdf <- pgeom(y,0.5)
round(cbind(y,prob,cdf),4)
plot(y,prob,type="h",col="red",ylim=c(0,1))
text(0:10,prob,labels=round(prob,4),pos=1,cex=0.6,offset=0.3)
lines(0:10,cdf,pch=16,type="o",col="blue")
text(0:10,cdf,labels=round(cdf,4),pos=3,cex=0.6,offset=0.3)
legend(locator(1),c("pmf","cdf"),pch=c(1,16),col=c("red","blue"))
```

Challenge! Geometric

A fair six-sided die is rolled until each face is observed at least once. On the average, how many rolls are needed? Can you simulate this in R?

- After the first roll, seeing a different face from the first roll is like a geometric random variable with $p = \frac{5}{6}$. So on the average it takes $\frac{1/6}{5/6} + 1 = 1.2$ rolls. (Because in $Y \sim$ geometric$(p = 5/6)$, Y can be zero, i.e., it's number of "failures" before "success", so you're adding 1 to the expected

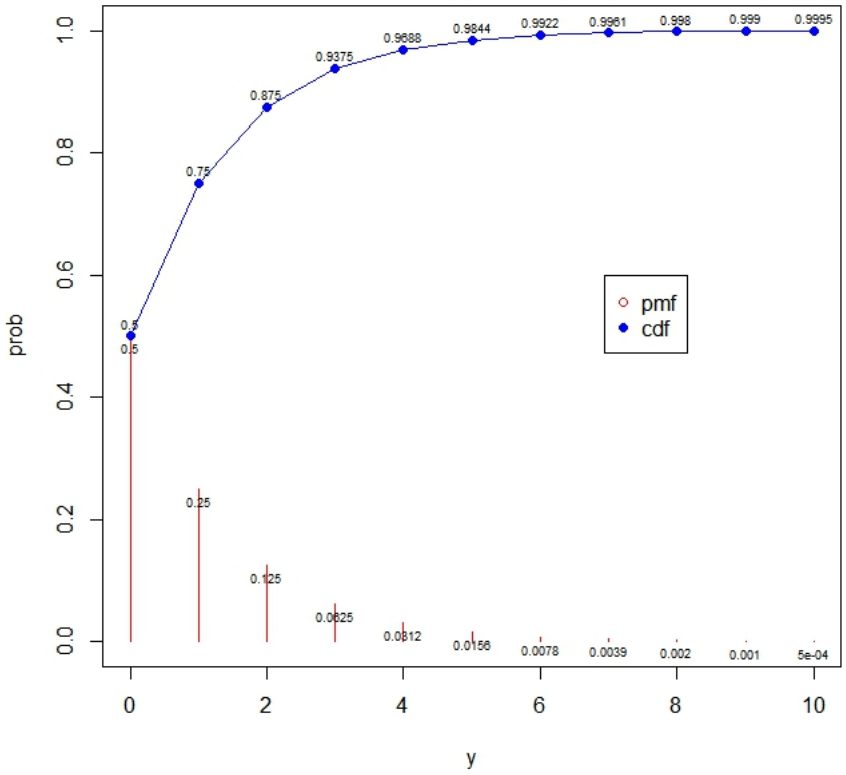

Figure 1: pmf and cdf of a geometric random variable with $p = 0.5$

value of q/p.) After two different faces, the probability of seeing a new face is $\frac{4}{6}$, so it will take $\frac{2/6}{4/6} + 1 = 1.5$ rolls. Continuing in this manner, the answer is

$$1 + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + \frac{6}{1} = \frac{147}{10} = 14.7$$

The following R codes smiluate this probblem 1,000 times.

```
n_rolls <- numeric(1000)
for(i in 1:1000){
n_rolls[i] = 0
     for (j in 1:50) {
     x <- sample(1:6,j,rep=T)
     if(length(unique(x)) == 6) { break; }
     }
     n_rolls[i]=length(x)
}
hist(n_rolls)
mean(n_rolls)
```

Binomial random variable

$$X \sim \text{binomial}\,(n, p),$$

where X = "total" number of "successes," n is the total number of trails, and p is the probability of "success' in a single trial. It's called a "fair" coin when $p = 0.5$.

For example,

- $X \sim$ binomial $(10, 0.5)$ means you're flipping a "fair" coin 10 times and wondering about how many "heads" you've got.
- $X \sim$ binomial $\left(100, \frac{1}{6}\right)$ means you're rolling a "fair" die 10 times and wondering about how many "1's" you've got.

Here is how to simulate flipping a "fair" coin 10 times and how many "heads" you've got.

```
(x <- rbinom(1,10,0.5))
(x <- rbinom(1,10,0.5))
```

Now, let's simulate 1000 binomial random numbers with n=10, p=0.5.

```
x <- rbinom(1000,10,0.5)
table(x)
hist(x,breaks=c(0:10),prob=T)     # prob=TRUE plots 'density', NOT frequency
mean(x)
sd(x)
10*0.5
[1] 5
sqrt(10*0.5*0.5)
[1] 1.581139
```

The last part is to check if the sample mean and sample sd are close to the theoretical mean and sd. For a binomial random variable, $\mu = n \cdot p$ and $\sigma = \sqrt{np(1-p)}$.

We can also check the probability of a binomial random variable. For a binomial random variable X with two parameters n and p, we have the pmf as

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

In plain terms, this is the probability of "having" x many *heads* when you flip a coin n times, where p stands for the *quality* of a coin. Some more examples of a binomial random variable $X \sim$ binomial (n, p) are shown below.

Ex 2: Binomial

1. Let $X \sim$ binomial with $n = 10$ and $p = 0.5$. Find the probability of having 4 "successes."

   ```
   dbinom(4,10,0.5)
   [1] 0.2050781
   pbinom(4,10,0.5)-pbinom(3,10,0.5)    #P(x=4) can also be found by P(x<=4)-P(x<=3)
   [1] 0.2050781
   ```

2. In the 2013 Jerry's Artarama art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.
   (a) What values does X take on?
   (b) What is the probability distribution? Find the following probabilities:
      i. the probability that two pages feature signature artists
      ii. the probability that *at most* six pages feature signature artists
      iii. the probability that *more than* three pages feature signature artists.

iv. calculate the mean and standard deviation of X.

```
dbinom(2,100,8/560)
(factorial(100))/((factorial(2))*(factorial(98))) * ((8/560)^2) * ((1-8/560)^98)
dbinom(0,100,8/560)+dbinom(1,100,8/560)+dbinom(2,100,8/560)+dbinom(3,100,8/560)+
          dbinom(4,100,8/560)+dbinom(5,100,8/560)+dbinom(6,100,8/560)
pbinom(6,100,8/560)
1-(dbinom(0,100,8/560)+dbinom(1,100,8/560)+dbinom(2,100,8/560)+dbinom(3,100,8/560))
1-pbinom(3,100,8/560)
```

In (c), $P(X > 3)$ was calculated by $1 - P(X \leq 3)$.

3. Let $X \sim$ binomial with $n = 10$ and $p = 0.5$. Construct the probability distribution table of X.

```
x <- 0:10
prob <- dbinom(x,10,0.5)
cdf <- pbinom(x,10,0.5)
round(cbind(x,prob,cdf),4)
plot(x,prob,type="o",col="red",ylim=c(0,1))
text(0:10,prob,labels=round(prob,4),pos=1,cex=0.6,offset=0.3)
lines(0:10,cdf,pch=16,type="o",col="blue")
text(0:10,cdf,labels=round(cdf,4),pos=3,cex=0.6,offset=0.3)
legend(locator(1),c("pmf","cdf"),pch=c(1,16),col=c("red","blue"))
```

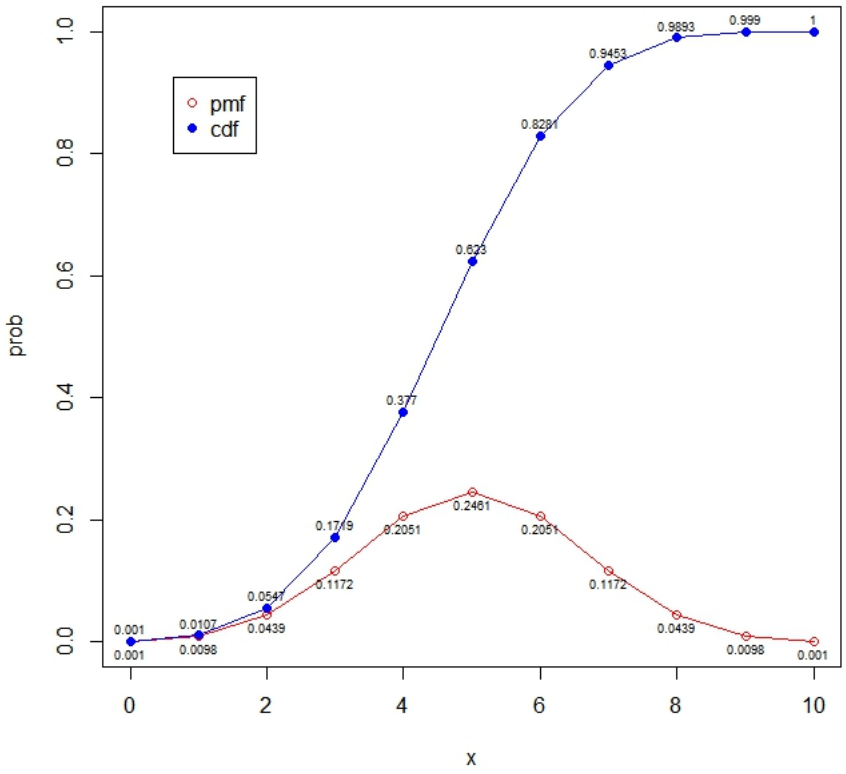

Figure 2: pmf and cdf of a binomial random variable with $n = 10$ and $p = 0.5$

Negative binomial random variable

$$Y \sim \text{negative binomial}\,(r, p),$$

where Y = "total" number of "failures" before the rth "success", and p is the probability of "success' in a single trial.

For example,

- $Y \sim$ negative binomial $(5, 0.5)$ means you're flipping a "fair" coin many times and wondering about how many times you had to flip until you have the fifth "head."
- $X \sim$ negative binomial $\left(10, \frac{1}{6}\right)$ means you're rolling a "fair" die many times and wondering about how many times you rolled until the tenth "6" was shown.

Here is how to simulate flipping a "fair" coin many times until you had five "heads." The random varaible is the total number of flips excluding the time when the fifth head appeared.

```
y <- rnbinom(100,5,0.5)
table(y)
```

Now, let's simulate 1000 negative binomial random numbers with r=5, p=0.5.

```
y <- rnbinom(1000,10,0.5)
table(y)
hist(y,breaks=c(0:35),prob=T)     # prob=TRUE plots 'density', NOT frequency
mean(y)
sd(y)
var(y)
```

The last part is to check if the sample mean and sample sd are close to the theoretical mean and sd. For a negative binomial random variable, $\mu = r\frac{q}{p}$ and $\sigma^2 = r\frac{q}{p^2}$.

We can also check the probability of a negative binomial random variable. For a negative binomial random variable Y with two parameters r and p, we have the pmf as

$$f(y) = \binom{y+r-1}{r-1} p^r (1-p)^y, \qquad y = 0, 1, 2, \ldots$$

In plain terms, this is the probability of having y many *failures* before you flip head r times, and p stands for the *quality* of a coin. Some more examples of a negative binomial random variable $Y \sim$ negative binomial (r, p) are shown below.

One note: Textbook (page. 64) uses a slightly different definition. There, X = the trial number on which the rth "success" is observed and it's related by $Y = X - r$, $(x = r, r+1, \ldots)$.

Ex 3: Negative Binomial

1. Let $Y \sim$ negative binomial with $r = 10$ and $p = 0.5$. Find the probability of having 6 "failures."

   ```
   dnbinom(6,10,0.5)
   pnbinom(6,10,0.5)-pnbinom(5,10,0.5)
   (factorial(15))/(((factorial(9))*((factorial(6)))))*(0.5^16)
   ```

2. A school newspaper reporter decides to survey 10 students who will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. She wonders how many people she has to contact to gather 10 students who

will attend Tet festivities. Let X = the number of people she has to contact until finding the 10 people who will attend Tet festivities.

(a) What values does X take on?

(b) What is the probability distribution? Find the following probabilities:

i. the probability that she finds all she needs after asking 30 people

ii. the probability that *at most* 30 people are to be asked

iii. the probability that *more than* 30 people are needed

iv. calculate the mean and standard deviation of X.

```
dnbinom(29,10,0.18)
pnbinom(29,10,0.18)
1-pnbinom(30,10,0.18)
```

In (c), $P(X > 30)$ was calculated by $1 - P(X \leq 30)$.

3. Let $X \sim$ negative binomial with $r = 10$ and $p = 0.5$. Construct the probability distribution table of X.

```
y <- 0:25
prob <- dnbinom(y,10,0.5)
cdf <- pnbinom(y,10,0.5)
round(cbind(y,prob,cdf),4)
plot(y,prob,type="o",col="red",ylim=c(0,1))
text(0:25,prob,labels=round(prob,4),pos=1,cex=0.6,offset=0.3)
lines(0:25,cdf,pch=16,type="o",col="blue")
text(0:25,cdf,labels=round(cdf,3),pos=3,cex=0.6,offset=0.3)
legend(locator(1),c("pmf","cdf"),pch=c(1,16),col=c("red","blue"))
```

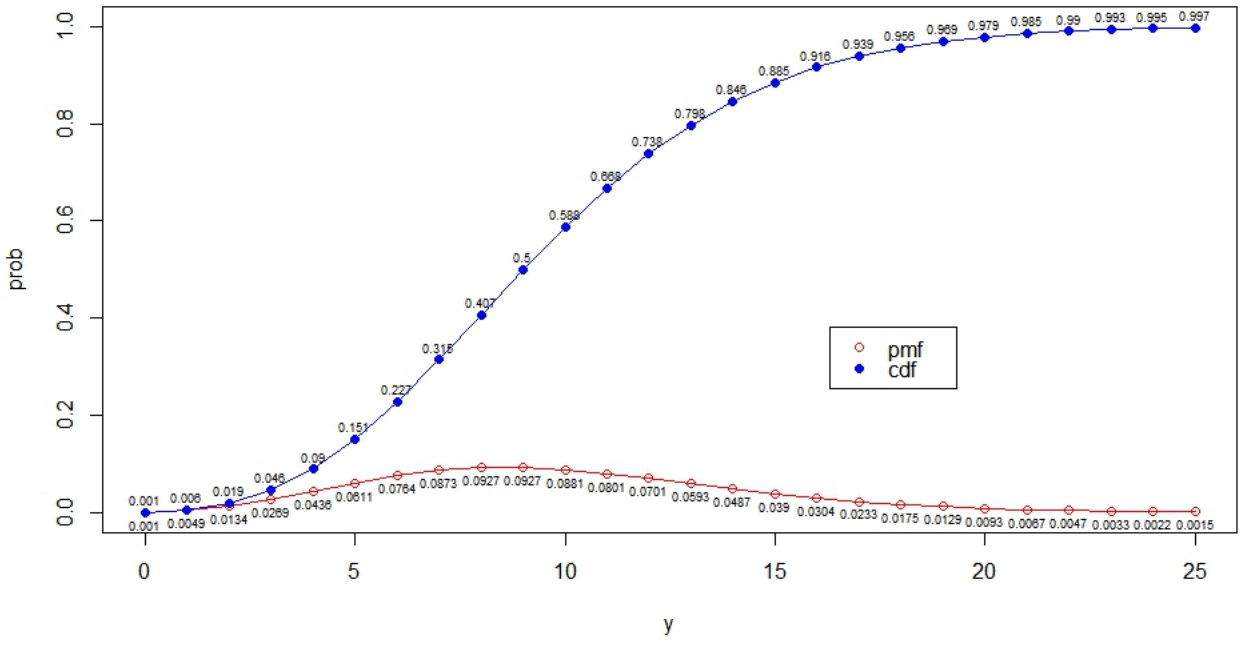

Figure 3: pmf and cdf of a negative binomial random variable with $r = 10$ and $p = 0.5$

Poisson random variable

The Poisson probability distribution gives the probability of a number of events occurring in a fixed interval of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.

The Poisson distribution may be used to approximate the binomial if p (the probability of success) is "small" (such as 0.01) and n (the number of trials) is "large" (such as 1,000).

$$X \sim \text{Poisson}\,(\lambda)\,,$$

where $\lambda =$ the "mean" number for the interval of interest.

For example,

- Suppose Jenny's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. Let $X =$ the number of calls Jenny receives in 15 minutes. (The interval of interest is 15 minutes or $\frac{1}{4}$ hour.) We write $X \sim \text{Poisson}\,(0.75)$.
- Suppose an email user gets, on average, 53 emails per day. Let $X =$ the number of emails an email user receives per day. The discrete random variable X has a Poisson distribution with $\lambda = 53$ and we write $X \sim \text{Poisson}\,(\lambda = 53)$.

Here is how to simulate Poisson random variables with "average" number of occurrences = 5. The random variable is the total number of occurrences during the interval of interest.

```
x <- rpois(100,5)
table(x)
```

Now, let's simulate 1000 Poisson random numbers with $\lambda = 5$.

```
x <- rpois(1000,5)
table(x)
hist(x,breaks=c(0:20),prob=T)     # prob=TRUE plots 'density', NOT frequency
mean(x)
sd(x)
var(x)
```

The last part is to check if the sample mean and sample sd are close to the theoretical mean and sd. For a Poisson random variable, $\mu = \lambda$ and $\sigma^2 = \lambda$.

We can also compute the probability of a Poisson random variable. For a Poisson random variable X with a parameter λ, we have the pmf as

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots$$

In plain terms, this is the probability of having x many *occurrences* for an event that typically occurs λ times. Some more examples of a Poisson random variable $X \sim \text{Poisson}\,(\lambda)$ are shown below.

Ex 4: Poisson

1. Let $X \sim$ Poisson with $\lambda = 5$. Find the probability of 10 "occurrences."

```
dpois(10,5)
ppois(10,5)-ppois(9,5)
(5^10)*(exp(-5))/(factorial(10))
```

2. According to Baydin, an email management company, an email user gets, on average, 53 emails per day. Let X = the number of emails an email user receives per day. The discrete random variable X takes on the values $x = 0, 1, 2, \ldots$. The random variable X has a Poisson distribution with $\lambda = 53$. The mean is 53 emails.
   (a) What is the probability that an email user receives exactly 60 emails per day?
   (b) What is the probability that an email user receives *at most* 60 emails per day?
   (c) What is the standard deviation of X?

```
dpois(60,53)
ppois(60,53)
```

3. Let $X \sim$ Poisson with $\lambda = 10$. Construct the probability distribution table of X.

```
x <- 0:20
prob <- dpois(x,10)
cdf <- ppois(x,10)
round(cbind(x,prob,cdf),4)
plot(x,prob,type="h",col="red",ylim=c(0,1))
text(0:20,prob,labels=round(prob,4),pos=1,cex=0.6,offset=0.3)
lines(0:20,cdf,pch=16,type="o",col="blue")
text(0:20,cdf,labels=round(cdf,3),pos=3,cex=0.6,offset=0.3)
legend(locator(1),c("pmf","cdf"),pch=c(1,16),col=c("red","blue"))
```

Selected Problems

1. The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Let X = the number of people you ask until one says he or she has pancreatic cancer. Then X is a discrete random variable.
   (a) What is the distribution of X? Don't forget to identify relevant parameter(s).
   (b) What is the probability of that you ask ten people before one says he or she has pancreatic cancer?
   (c) What is the probability that you must ask 20 people?
   (d) Find the (i) mean and (ii) standard deviation of X.

2. The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.
   (a) What is the probability distribution for X? Don't forget to identify any relevant parameter(s).
   (b) Calculate the (i) mean and (ii) standard deviation of X.
   (c) Find the probability that *at most* eight people develop pancreatic cancer.
   (d) Are five people more likely to develop develop pancreatic cancer than six people? Justify your answer numerically.

3. According to a recent article, when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. During winter months people keep calling in claiming to have the flu, we are interested in how many calls are needed to gather 10 patients who truly have the flue.

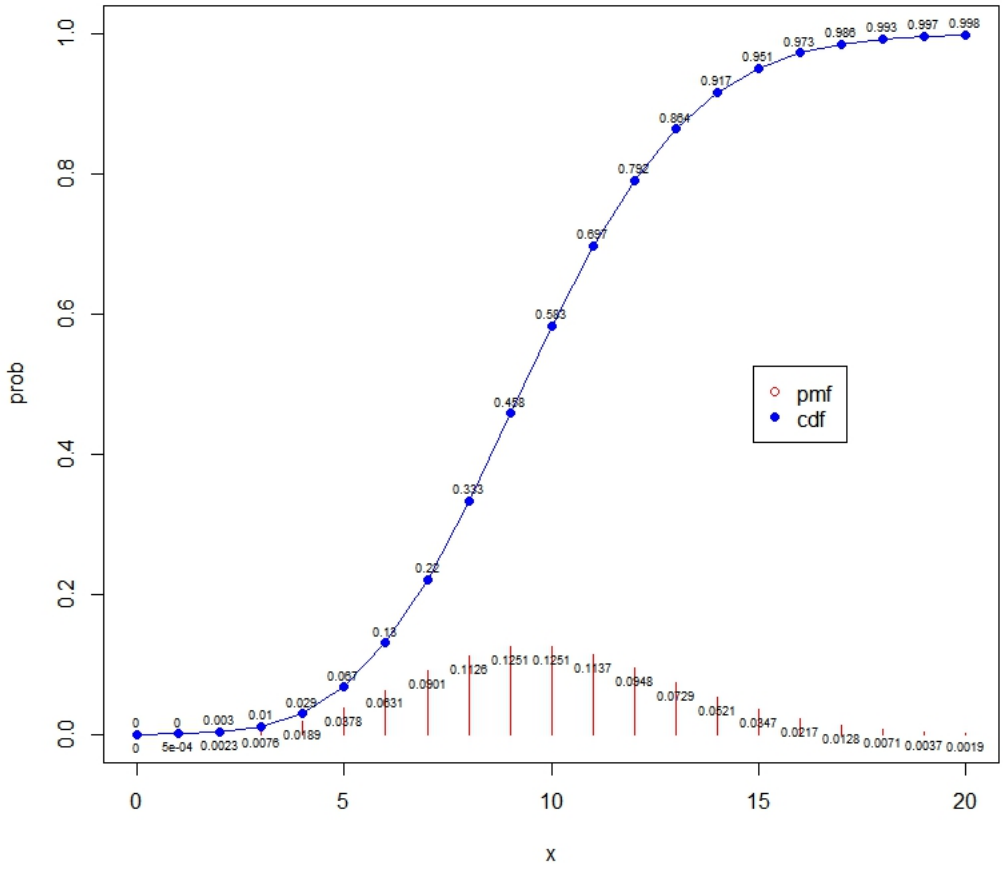

Figure 4: pmf and cdf of a Poisson random variable with $\lambda = 10$

(a) Define the random variable X and list its possible values.
(b) State the distribution of X and its relevant parameters.
(c) Find the probability that *at least* 50 calls are needed to have 10 patients who actually have the flu.
(d) On average, how many calls do you expect to answer to find 10 patients who actually have the flu?

4. The maternity ward at Dr. Jose Fabella Memorial Hospital in Manila in the Philippines is one of the busiest in the world with an average of 60 births per day. Let X = the number of births in an hour.
   (a) State the distribution of X and its relevant parameter(s).
   (b) Find the mean and standard deviation of X.
   (c) Plot a graph of the pmf and cdf of X.
   (d) What is the probability that the maternity ward will deliver three babies in one hour?
   (e) What is the probability of delivering *at most* three babies in one hour?
   (f) What is the probability of delivering *more than* five babies in one hour?