Statistics 213 – 2.2: Discrete Probability Models

Textbook:

4.4, 4.5, 4.6

Objectives:

- Be able to identify ***binomial***, ***hypergeometric***, and ***Poisson*** random variables and provide the appropriate parameter values given a specific situation
- Be able to create a probability distribution graph in R
- Be able to calculate probabilities for binomial, hypergeometric, and Poisson random variables, both by hand and using R
- Be able to calculate the ***expected value***, ***variance***, and ***standard deviation*** of binomial, hypergeometric, and Poisson random variables

Motivation:

In the 2.1 notes, we introduced the concept of random variables. Recall that a ***random variable*** is a quantity whose values are the numerical values associated with the random outcomes of an experiment. Often times, we represent a random variable with a letter.

Example

An experiment is performed in which two fair coins are tossed. Let X be the number of heads resulting from an iteration of this experiment.

Here, X is a random variable that can take on values: 0, 1, or 2.

We also focused on the concept of a probability distribution for discrete random variables. Recall that a ***discrete random variable*** is a variable that can take on a countable number of values in a given range. That is, a discrete random variable is one for which you could list all its possible values in a finite amount of time.

Examples

- The number of heads resulting from flipping two coins
- The number of houses on a block
- The amount of change in someone's pocket

A ***probability distribution*** for a discrete random variable is a table, graph, or formula that specifies every possible value that a random variable can assume along with the probabilities associated with each of these values.

Example

An experiment is performed in which two fair coins are tossed. Let X be the number of heads that result. The following is a probability distribution table for X.

x	0	1	2
P(X = x)	0.25	0.50	0.25

In the previous notes, we saw probability distributions defined using tables (as in the example above) or using graphs. But as the definition of a probability distribution states, a probability distribution and thus the behavior of the corresponding random variable can also be defined using a formula. In this set of notes, we will focus on variables that can be described (or modeled) by common types of probability distributions. These probability distributions, as we'll see, are defined using formulas. Specifically, we will be focusing on three different discrete probability models and the random variables that can be described by them:

1. Binomial ⟶ coin or die repeated
2. Hypergeometric ⟶ candy dish selection probability changes everytime you pick a candy
3. Poisson ⟶ birth rates, population

For each of these three models, we will discuss the following:

- The definition or type of random variable that can be described using the model
- The probability distribution equation (probability mass function)
- Calculations for the mean, variance, and standard deviation

Binomial Random Variables

To help understand our first probability model, let's start with an example scenario.

Example

A national poll reveals that 81% of Canadians consume coffee on a given day. Suppose you randomly ask three Canadians about their coffee consumption. Let X represent the number of Canadians you asked who said that they have, in fact, consumed coffee that day.

↳ $x = \{0, 1, 2, 3\}$

Domain, support of x — what options are possible

Definition

A scenario or experiment with the following characteristics produces a ***binomial random variable***:

- The experiment consists of n identical, independent trials and this number of trials is fixed
- The experiment involves the same dichotomous (two option) response (yes/no, pass/fail, etc.) for each trial
- The probability of "success," p, remains the same from trial to trial — independent
- The binomial variable itself is defined as the number of successes out of the total n trials

Think of a binomial random variable as counting the number of successes in a total of n trials of an experiment.

The values that define a random variable's probability distribution are called ***parameters***. To be able to compute probabilities for a binomial random variable, you need to know two parameters:

n – the number of trials in the experiment

p – the probability of success

Probability Mass Function (PMF) for Binomial Random Variables

A ***probability mass function (PMF)*** is the equation that is used to find the probability that a random variable X assumes a particular value x. The following is the PMF for a binomial random variable X.

$X \sim Bin(n=3, p=0.81)$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

let x be # of successes out of 3

$$P(X = x) = \binom{3}{x} (0.81)^x (1-0.81)^{n-x}$$

where:

n = the number of trials p = the probability of success on any given trial x = the number of successes in n trials

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Note: a special case of the binomial random variable is a Bernoulli random variable. A Bernoulli random variable is defined as a binomial random variable when $n = 1$ (only one trial is performed).

Expected Value, Variance, and Standard Deviation

In the 2.1 notes, we discussed the general equations used to calculate the expected value (mean), variance, and standard deviation of discrete random variables. These calculations can be done in a simpler way for a binomial random variable!

Let X be a binomial random variable. Then:

$\mu_X = E(X) = np$

$\sigma_X^2 = VAR(X) = np(1-p)$

$\sigma_X = SD(X) = \sqrt{np(1-p)}$

Some examples can help us get familiar with recognizing situations where we've got binomial random variables. You'll notice that we've seen some of these types of questions before!

Example 2.2.1

A national poll reveals that 81% of Canadians consume coffee on a given day. Suppose you randomly ask three Canadians about their coffee consumption. Let X represent the number of Canadians you asked who said that they have, in fact, consumed coffee that day.

1. Fill in the missing values for the probability distribution table. Calculate the probabilities by hand.

x	0	1	2	3
P(X = x)	0.006859	0.087723	0.373977	0.531441

NOTES

$X \sim Bin(n=3, p=0.81)$

2. What is $E(X)$? What is $VAR(X)$?

$E(X) = np = 3(0.81) = 2.43 = \text{sum}(x \cdot \#\text{prob})$

$V(X) = np(1-p) = npq = 3(0.81)(1-0.81) = 0.4617$

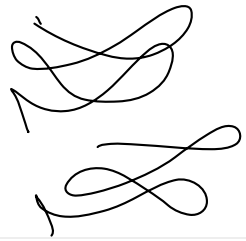
Example 2.2.2

A nation-wide poll of university graduates suggests that 66% of them had changed their majors after their first year of university. A sample of 25 U of C graduates are selected. Let Y represent the number of these students who had changed their major after their first year. $Y \sim Bin(n=25, p=0.66)$

1. What is the probability that 20 of the selected graduates had changed their major after their first year of university? Calculate this probability by hand and using R.

$P(Y=20) = \binom{25}{20} 0.66^{20} (1-0.66)^5$

≈ 0.0594



To calculate this probability using R, use the package `mosaic`. We will be using the `pdist()` function for a binomial distribution. This function has the following arguments:

`pdist("binom", size, prob, q)`

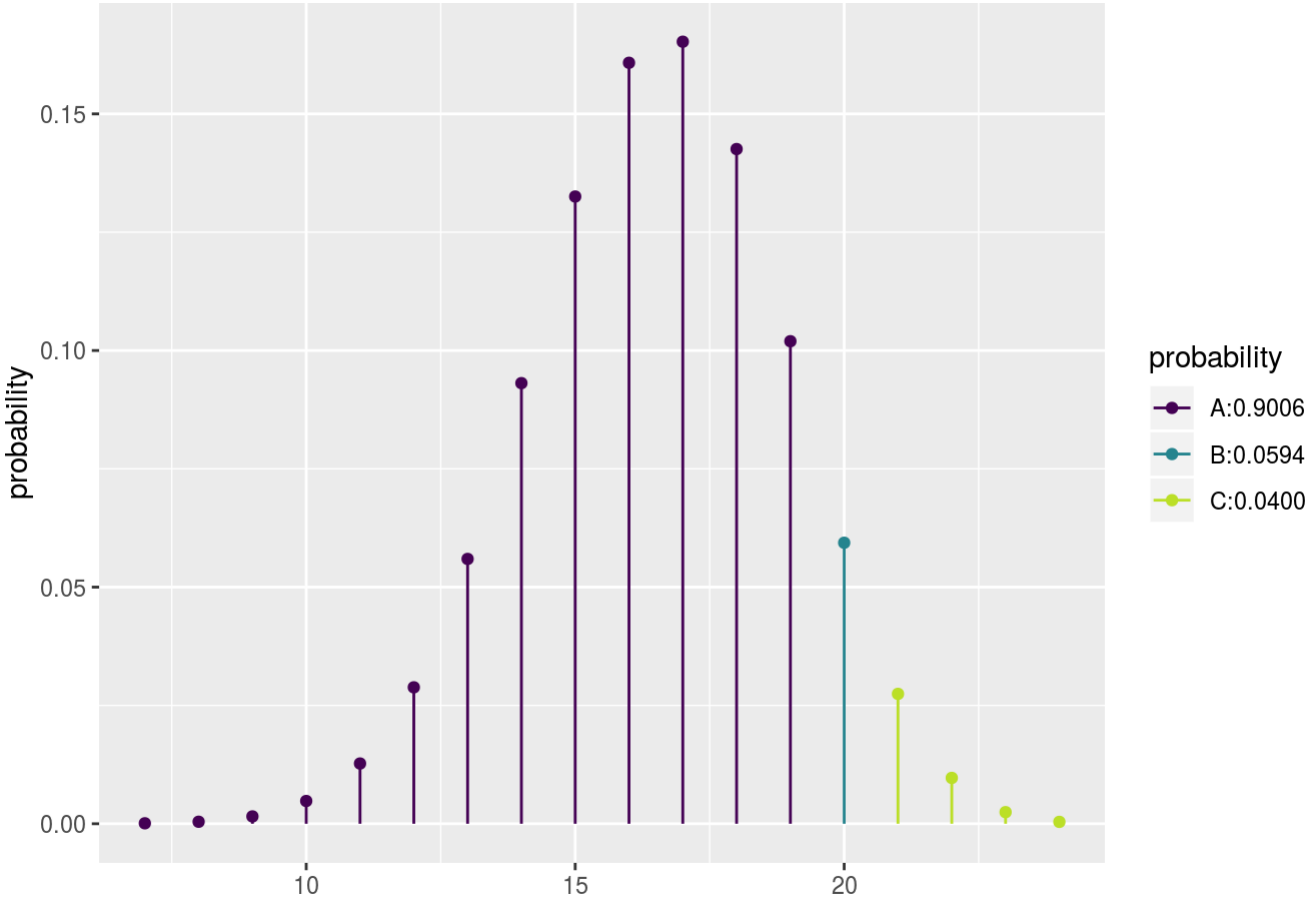
where `size` is your number of trials n, `prob` is your probability of success p, and `q` relates to x, the number of successes.

When we are trying to calcuate an "exactly equals" probability like $P(X = x)$, we need to enter `c(x-1,x)` as our `q`.

So for $P(Y = 20)$...

Studio

```
library(mosaic)
pdist("binom", size = 25, prob = 0.66, q = c(19,20), digits = 4)
```




```
## [1] 0.9006153 0.9599916
```

2. What is the probability that more than 23 of the selected graduates had changed their major after their first year of university? Use R to compute this probability.

$X = \{0, 1, \ldots, 23, 24, 25\}$

$P(X > 25) = P(X \geq 24) = P(X = 24) + P(X = 25)$

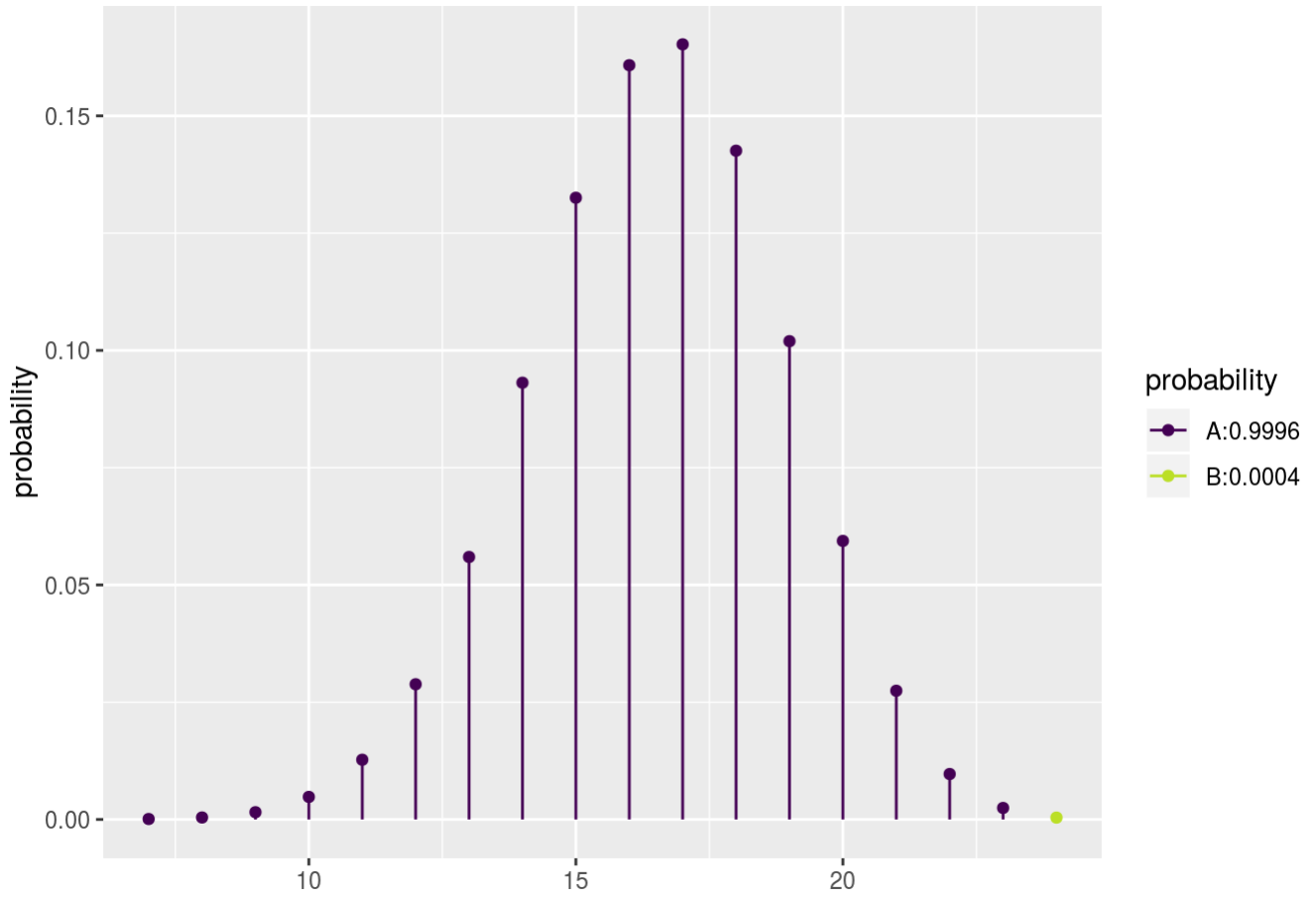
$1 - P(X \leq 23)$

When we are trying to calcuate a "left tail" probability of $P(X \leq x)$, we need to enter `x` as our `q`.

So for $P(Y > 23) = 1 - P(Y \leq 23)$...

R Studio

```
1-pdist("binom", size = 25, prob = 0.66, q = c(23), digits = 4)
```



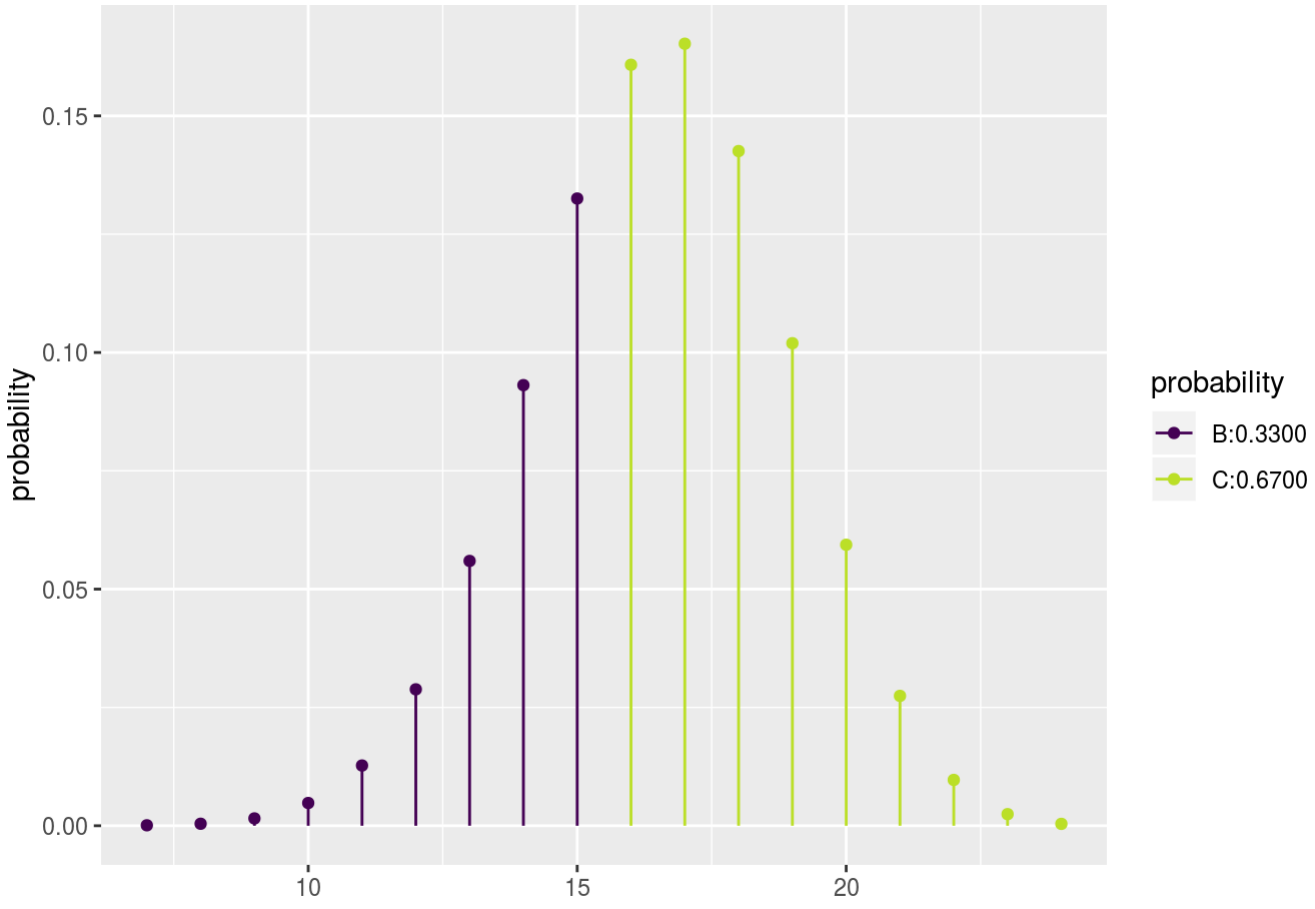

```
## [1] 0.0004275136
```

3. What is the probability that between 5 and 15 (inclusive) of the selected graduates had changed their major after their first year of university? Use R to compute this probability.

When we are trying to calcuate a "middle" probability like $P(a \leq X \leq b)$, we need to enter `c(a-1, b)` as our `q`.

So for $P(5 \leq Y \leq 15)$...

```
1-pdist("binom", size = 25, prob = 0.66, q = c(4,15), digits = 4)
```




```
## [1] 0.9999996 0.6699544
```

$X \sim Bin(n=35, p=0.82)$ P(cancelling)

Example 2.2.3

A taxi company keeps a fleet of 30 taxis ready to be reserved by customers. The company owner, based on experience, knows that 18% of customers who reserve a taxi actually end up canceling the reservation. Thus, he is comfortable over-reserving his 30 taxis. Suppose that one evening he ends up with 35 taxi reservations. Let X represent the number of reservations that are not cancelled on a given evening.

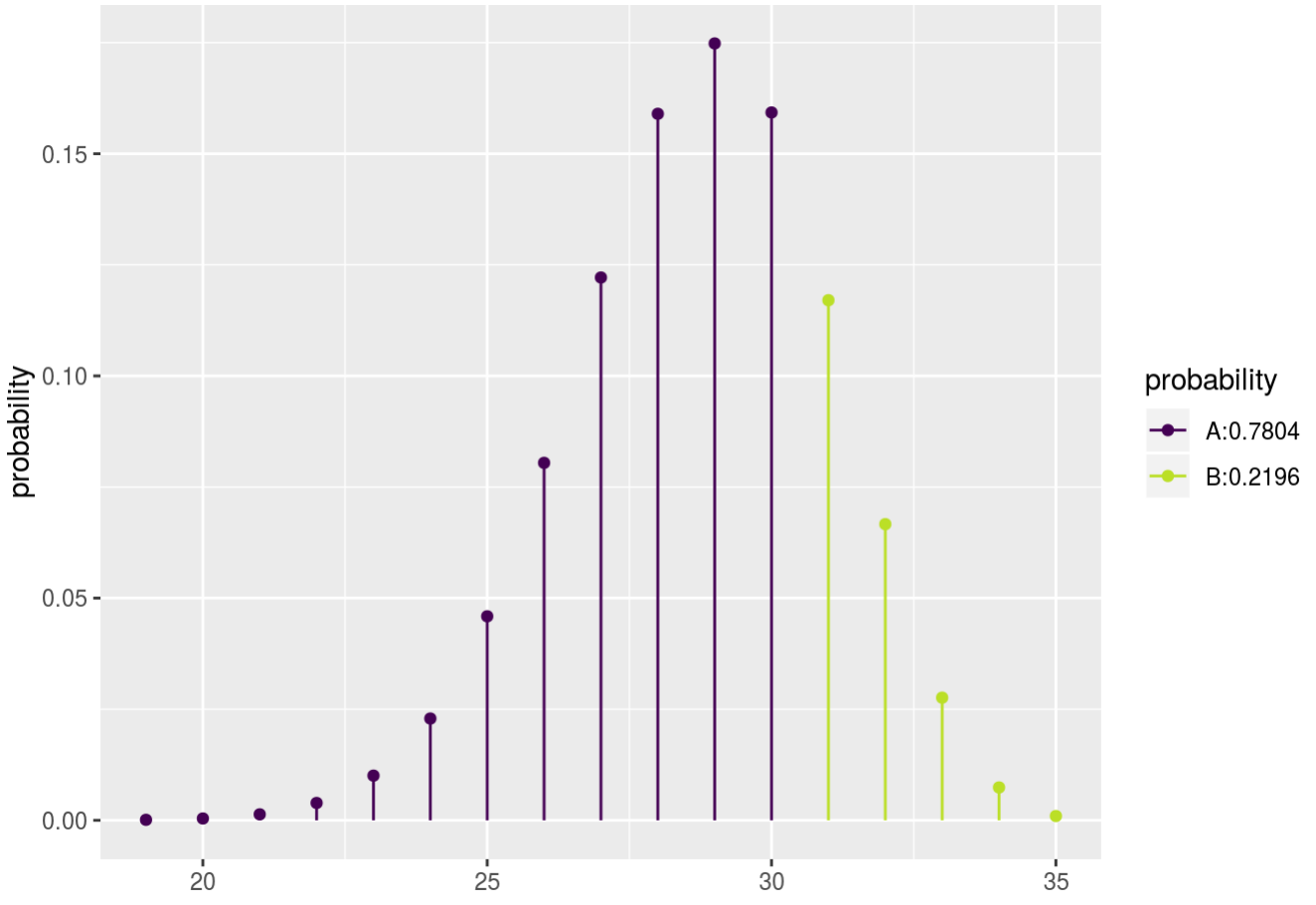
1. What is the probability that everyone who has reserved a taxi will be provided with one on a given evening? In other words, what is the probability of not having too many reservations? Use R to compute this probability.

$P(X \leq 30) = sum(dbinom(0:30, size=35, prob=.82)) = pbinom(30, 35, .82) = 0.7804$

= ≤

Studio

```
1-pdist("binom", size = 35, prob = 0.82, q = 30, digits = 4)
```




```
## [1] 0.2196409
```

2. How many people can the owner expect to keep their reservations on a given evening?

$E[x] = np = 35(0.82) = 28.7$

$SD[x] = \sqrt{np(1-p)} = \sqrt{35(0.82)(0.18)} = \sqrt{5.166} \approx 2.27288$

expect 28.7 people will keep reservations give or take 2.272

Hypergeometric Random Variables

To help understand our second probability model, let's start with an example scenario.

Example

An assignment has a total of 15 questions. Suppose that you hadn't started the assignment yet, but decided to randomly select four problems to complete tonight. Of the 15 questions, you know how to solve eight of them without using your notes. Let X represent the number of problems solved without using your notes.

$$\frac{\text{know}\binom{8}{x}\ \text{know}^{C}\binom{7}{4-x}}{\binom{15}{4}}$$

Definition

A scenario or experiment with the following characteristics produces a ***hypergeometric random variable***:

- The experiment involves randomly selecting (sampling) n elements, without replacement, from a set or population of N total elements
- The experiment involves the same dichotomous (two option) response for each element (yes/no, pass/fail, etc.)
- The number of total "successes" in the population *N* is known to be a certain number *r* • The hypergeometric variable itself is defined as the number of successes in the set of n elements that are drawn from N

Think of a hypergeometric random variable as counting the number of successes in a total of n trials of an experiment where the n trials are not independent from one another. Recall that for binomial random variables, the trials are considered to be independent (the same probability of success for each trial). Here, they are not, as you can think of each element as being selected without replacement.

To be able to compute probabilities for a hypergeometric random variable, you need to know three parameters:

N – the total number of elements (the "population")

n – the number of elements drawn from N

r – the total number of "successes" in the N elements

Probability Mass Function (PMF) for Binomial Random Variables

The following probability mass function can be used to find the probability that a hypergeometric random variable X assumes a particular value x.

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

where:

N = the total number of elements

n = the number of elements drawn (sampled) from N

r = the total number of “successes” in the N elements

x = the number of “successes” in the n elements

Expected Value, Variance, and Standard Deviation

Just as was the case for binomial random variables, we have specific equations for the mean, variance, and standard deviation for hypergeometric random variables.

Let X be a hypergeometric random variable. Then:

$\mu_X = E(X) = \frac{nr}{N}$

$\sigma_X^2 = VAR(X) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$

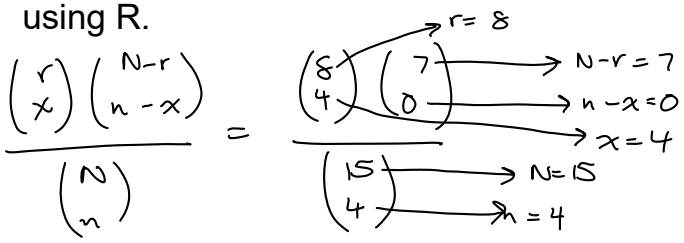
$\sigma_X = SD(X) = \sqrt{n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)}$

Let’s look at some examples involving hypergeometric random variables.

Example 2.2.4

An assignment has a total of 15 questions. Suppose that you hadn't started the assignment yet, but decided to randomly select four problems to complete tonight. Of the 15 questions, you know how to solve eight of them without using your notes. Let X represent the number of problems solved without using your notes.

1. What is the probability that you would be able to solve all four of the selected problems without consulting your notes? Calculate this probability by hand and using R.



The `mosaic` package in R does not currently have the capabilities to display a graph for a hypergeometric distribution. However, we can still use a different function to obtain hypergeometric-related probabilities.

When we are trying to calcuate an "exactly equals" probability like $P(X = x)$ for a hypergeometric random variable, we will be using the `dhyper()` function. This function has the following arguments:

```
dhyper(x, m, n, k)
```

where `x` is the number of success of interest x, `m` is the total number of "successes" in the population, M, `n` is the total number of "failures" in the population, $N - M$, and `k` is the size of the sample being selected from the population, n.

So for $P(X = 4)$...

RStudio

```
dhyper(4,8,7,4)
```

```
## [1] 0.05128205
```

2. What is the probability that you solve at least half of the selected problems without consulting your notes? Calculate this probability using R.

$$\frac{\binom{8}{2}\binom{7}{2}}{\binom{15}{4}} + \frac{\binom{8}{3}\binom{7}{1}}{\binom{15}{4}} + \frac{\binom{8}{4}\binom{7}{0}}{\binom{15}{4}} = P(x=2) + P(x=3) + P(x=4)$$

When we are trying to calcuate "tail" probability like $P(X \leq x)$ or $P(X \geq x)$ for a hypergeometric random variable, we will still be using the `phyper()` function. This function has the following arguments:

```
phyper(x, m, n, k)
```

It is used just like the `dhyper()` function, except it returns the probability $P(X \leq x)$ (the "left-tail" probability) for whatever value you enter as `x`.

So for $P(X \geq 2)$...

RStudio

```
1-phyper(1,8,7,4)
```

```
## [1] 0.7692308
```

Example 2.2.5

A company that manufactures remote control cars ships them out to stores in boxes of 100. During the production of a set of 100 cars, an automated machine malfunctioned, causing 15 of the 100 cars to be packaged without remote controls. The store that purchased the box of remote control cars decides to display 10 of them on a shelf in their toy department. Let V represent the number of cars displayed without a remote.

1. How many cars would we expect to be displayed without a remote?

$$E[V] = n\frac{r}{N} = \frac{(10)(15)}{100} = 1.5$$

2. What is the probability that three of the 10 displayed cars will not have a remote control? Find this probability using R.

$$P(V=3) = \frac{\binom{15}{3}\binom{85}{7}}{\binom{100}{10}} \approx 0.1297383$$

R Studio

```
dhyper(3,15,85,10)
```

```
## [1] 0.1297383
```

3. Suppose the shelf-stocker notices that some of the cars are missing the remotes and decides to inform her boss. However, she can't remember the exact amount that were missing the remotes, so she tells her boss that "between two and five cars" (inclusive) were missing remotes. Given this information, what is the probability that the shelf-stocker saw exactly three cars missing their remotes? Use R to help calculate this probability.

RStudio

```
phyper(5,15,85,10)-phyper(1,15,85,10)
```

```
## [1] 0.4618274
```

Poisson Random Variables

→ 2 options
→ based on a rate

To help understand our third probability model, let's start with an example scenario.

of 'success' in a given time or space

Example

Burgrass is a weed that grows in the type of soil typically used in potato fields. Suppose a farmer has determined that the number of weeds in a crop of potatoes fluctuates from acre to acre, with an average number of weeds per acre of 5.5. Let X represent the number of weeds in a given acre.

Definition

A scenario or experiment with the following characteristics produces a ***Poisson random variable***:

- The experiment involves an event occurring during a given amount of time/area
- The probability that an event occurs in the given amount of time/area is the same for all other equal amounts of time/areas
- The number of events that occur in one given amount of time/area is independent of the number of events that occur in other amounts of time/area
- There is a known average or expected number of events, λ, that occur during/in the amount of time/area

rate = success/time or success/area

- The random variable itself is defined as the number of times an event has occurred in a given amount of time/area

Think of a Poisson random variable as counting the number of times an event occurs in a given time/area, where the average number of events per unit of time/area is known.

To be able to compute probabilities for a Poisson random variable, you only need to know one parameter:

λ – the average number of events during a given amount of time or area

Probability Mass Function (PMF) for Binomial Random Variables

Poisson

The following probability mass function can be used to find the probability that a Poisson random variable X assumes a particular value x.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where:

λ = the average number of events during a given amount of time or area

e = 2.71828…

x = the number of times an event occurs in the given time or area

Expected Value, Variance, and Standard Deviation

The rate parameter, λ, is very important for Poisson random variables and is used in all three "summary value" calculations.

Let X be a Poisson random variable. Then:

$\mu_X = E(X) = \lambda$

$\sigma_X^2 = VAR(X) = \lambda$

$\sigma_X = SD(X) = \sqrt{\lambda}$

Reparametrizing λ

You can think of λ as a rate, or the average number of times an event occurs given a specific amount of time/area.

Examples

$\lambda = 10.2$ represents the average number of shoppers using an escalator in an hour

$\lambda = 0.34$ represents the average number of typos on a page

The value of λ relies on the amount of time/area used to define λ itself. Sometimes, however, you might be interested in the probability of an event happening in a time/area that is different than the time/area used to originally define λ. In such cases, you need to reparametrize λ so that it is in terms of your new time/area.

Example 2.2.6

$\lambda = 10.2$ tells you the average number of shoppers using an escalator in an hour. You are interested in the probability that 15 shoppers use the escalator in two hours. You have to reparametrize λ so that it tells you the average number of shoppers using an escalator in two hours. What is the new λ?

λ = usually 10.2 shoppers/hr $\rightarrow 2(10.2)$ shoppers/2hr $= 20.4$

let x be # of shoppers in 2 hrs $\quad X \sim Pois(\lambda = 20.4)$

$$P(x=15) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(20.4)^{15} \cdot (e^{-20.4})}{15!} \approx 0.0965$$

Example 2.2.7 $\lambda = 0.34$ typos/page $\beta = \frac{1}{0.34}$ pages/typo } area

$\lambda = 0.34$ tells you the average number of typos on a page. You are interested in finding the probability of observing one or more typos in half a page. What is the reparametrized λ you need to use?

$\lambda = 0.34\left(\frac{1}{2}\right)$ typos/$\frac{1}{2}$ pg $= 0.17$ typos/$\frac{1}{2}$ pg

let w be # of typos per 1/2 page

$P(w \geq 1) = 1 - P(w = 0) = 1 - \frac{0.17^0 e^{-0.17}}{0!}$

Example 2.2.8

Burgrass is a weed that grows in the type of soil typically used in potato fields. Suppose a farmer has determined that the number of weeds in a crop of potatoes fluctuates from acre to acre, with an average number of weeds per acre of 5.5. Let X represent the number of weeds in a given acre.

1. What is the expected number of weeds per acre? The standard deviation of weeds per acre?

$E[x] = \lambda = 5.5$ $\quad SD[x] = \sqrt{5.5}$

$V[x] = \lambda = 5.5$

2. An acre of the farmer's crop is randomly selected and inspected for burgrass. What is the probability that exactly two weeds are found? Calculate this probability by hand and using R.

We will be using the `pdist()` function for the Poisson distribution. This function has the following arguments:

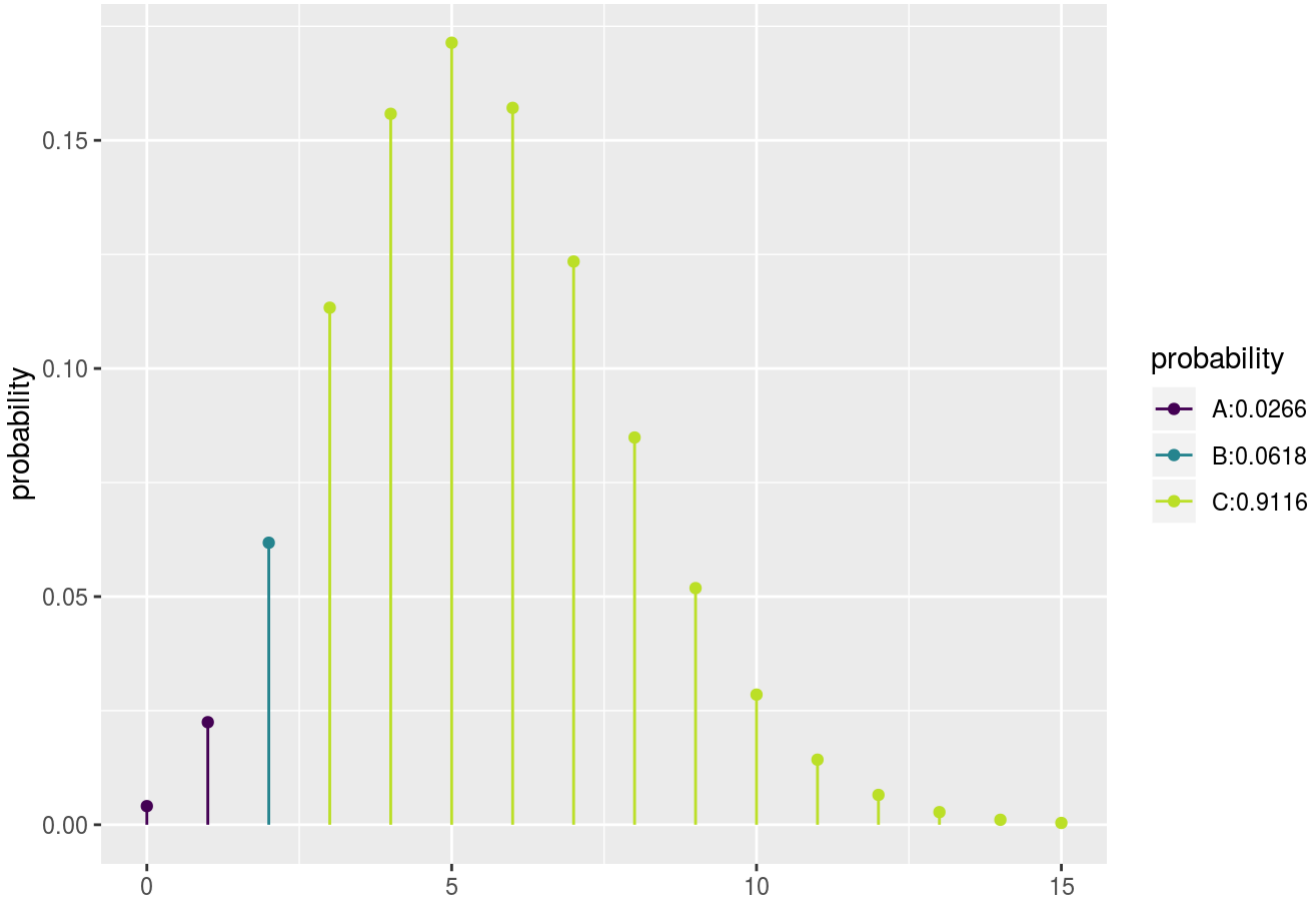
`pdist("pois", lambda, q)`

where `lambda` is your lambda λ and `q` relates to x, the number of successes.

> So for $P(X = 2)$...

R Studio

```
pdist("pois", lambda = 5.5, q=c(1,2), digits = 4)
```

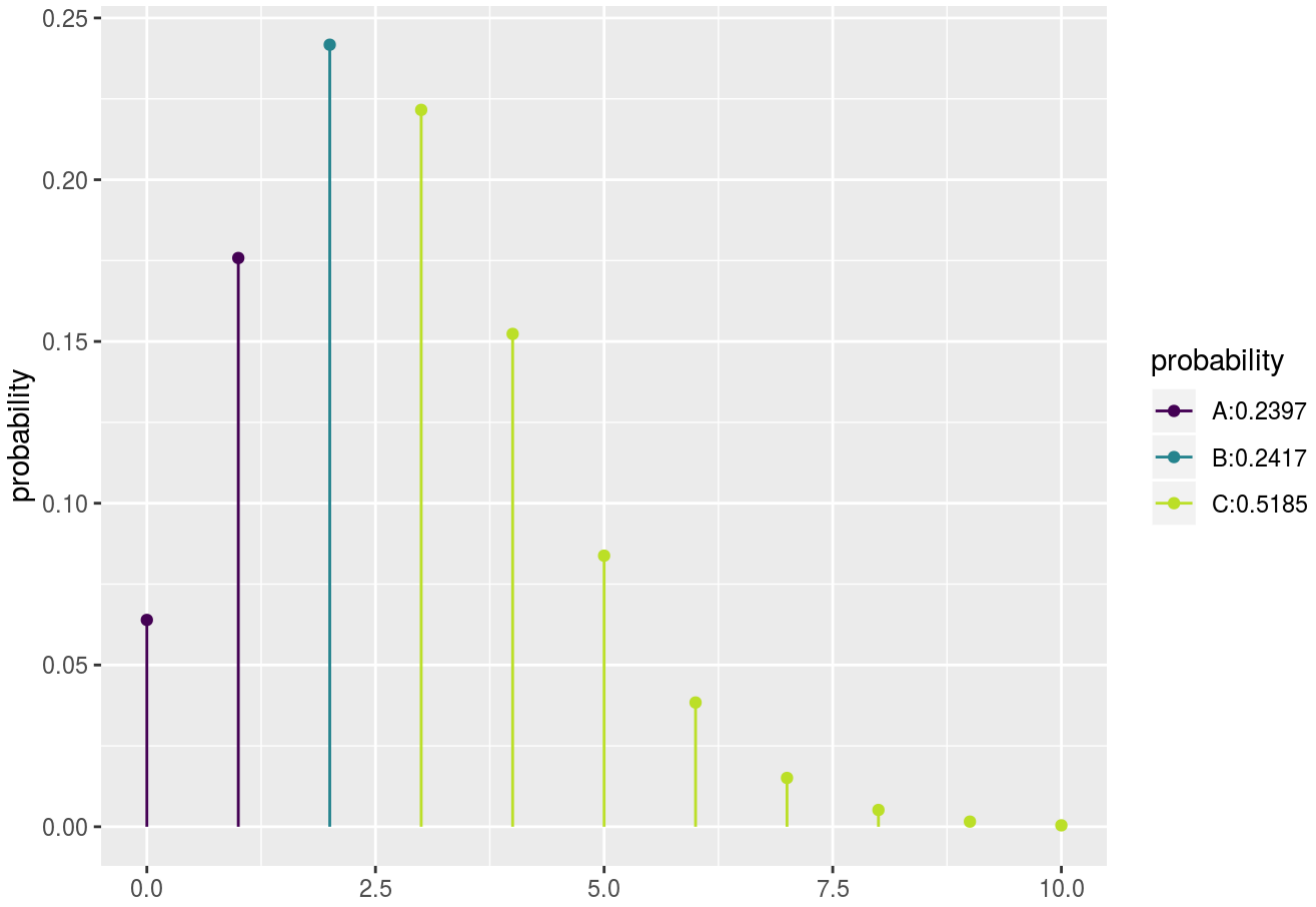



```
## [1] 0.02656401 0.08837643
```

4. Suppose, due to time constraints, that the farmer can only inspect a half-acre of his crop. What is the probability that exactly two weeds are found in the half-acre? Use R to calculate this probability.

RStudio

```
pdist("pois", lambda = 2.75, q=c(1,2), digits = 4)
```




```
## [1] 0.2397295 0.4814567
```

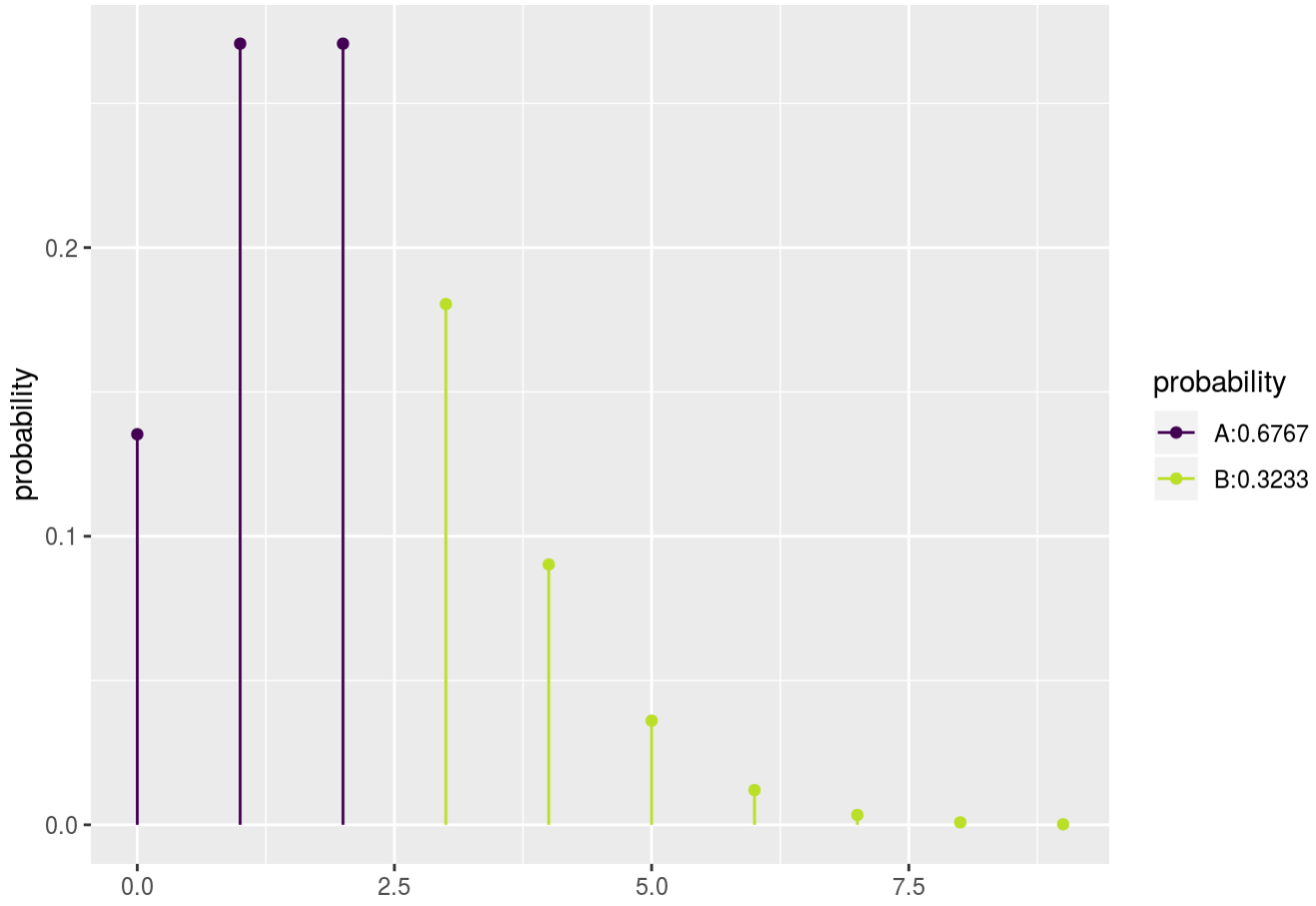
Example 2.2.9

A call center receives an average of two calls per minute. Suppose you were to count the number of calls received today between 14:00 and 14:01. Let X represent the number of calls between 14:00 and 14:01.

1. What is the probability that the call center receives at least three calls between 14:00 and 14:01? Use R to compute this probability.

Studio

```
1-pdist("pois", lambda = 2, q=2, digits = 4)
```

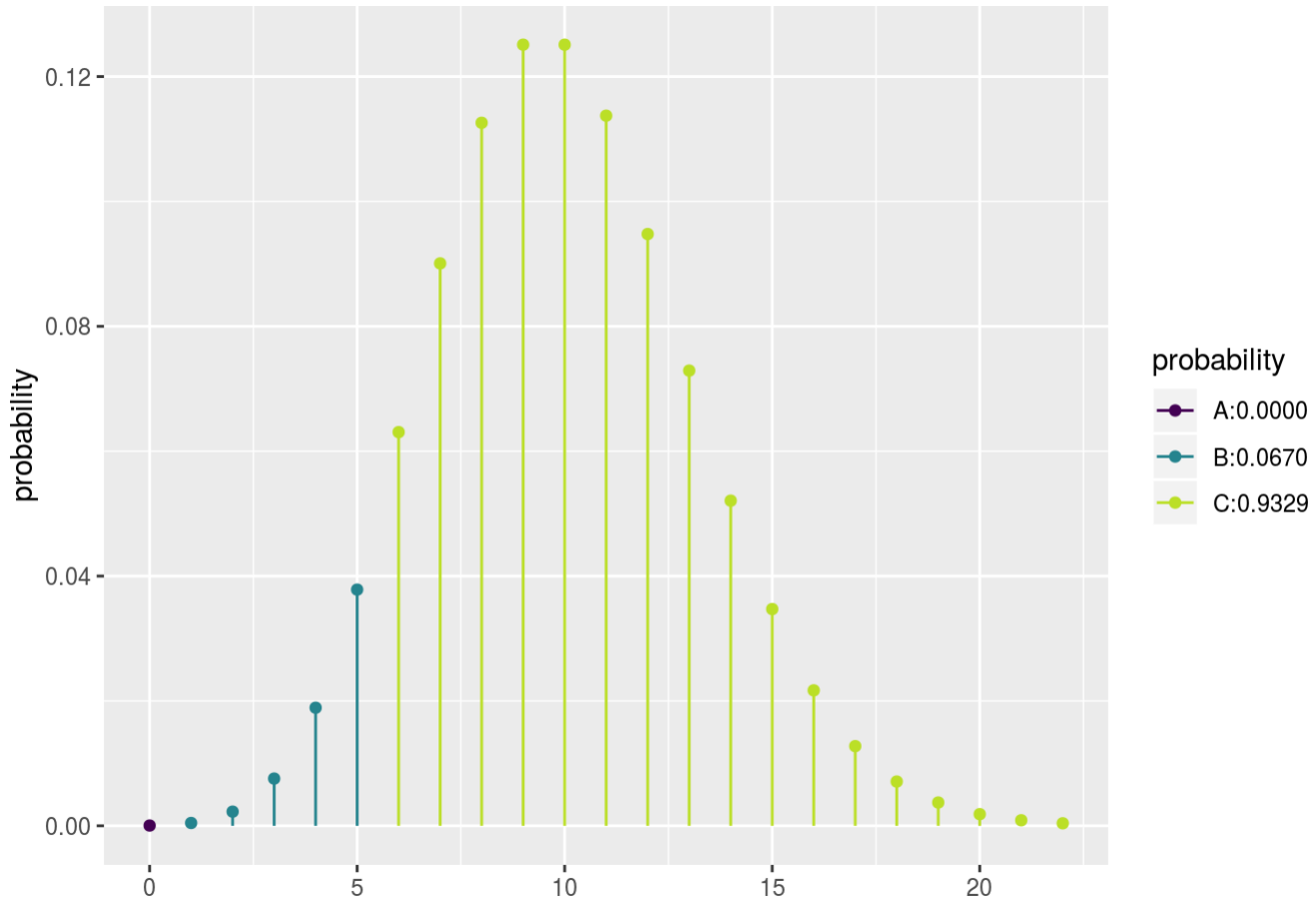



```
## [1] 0.3233236
```

2. What is the probability that the call center receives between one and five calls (inclusive) between 14:00 and 14:05 (a five-minute interval)? Use R to compute this probability.

RStudio

```
1-pdist("pois", lambda = 10, q=c(0,5), digits = 4)
```




```
## [1] 0.9999546 0.9329140
```