STAT 631 ROS Chapter3: Probability Basics

Prof. Kapitula

9/14/2021

```
# Load all packages here:
library(tidyverse)
library(dplyr)
library(ggplot2)
# Set seed value of random number generator to get "replicable" random numbers.
# The choice of seed value of 76 was an arbitrary one on my part.
set.seed(76)
```

Weighted Averages

We can enter vectors using code like below in R.

```
stratum=1:3
country=c('United States','Mexico','Canada')
population=c(310,112,34)
averageAge=c(36.8,26.7,40.7)
#put vectors in a tibble to display
tibble(stratum,country,population,averageAge)
```

```
## # A tibble: 3 x 4
##   stratum country       population averageAge
##     <int> <chr>              <dbl>      <dbl>
## 1       1 United States        310       36.8
## 2       2 Mexico               112       26.7
## 3       3 Canada                34       40.7
```

To compute a weighted average in R, we can do the below:

```
#the wrong way
averageNA=(population/sum(population))*averageAge
averageNA
```

```
## [1] 25.017544  6.557895  3.034649
```

```
averageNA=(population/sum(population))%*%averageAge
round(averageNA,1)
```

```
##      [,1]
## [1,] 34.6
```

So the average age is 34.6 million.

##Vectors and Matrices

```
x1=c(-1,0,1)
x0=rep(1,3)
```

```
X = matrix(c(x0,x1),3,2)
X
```

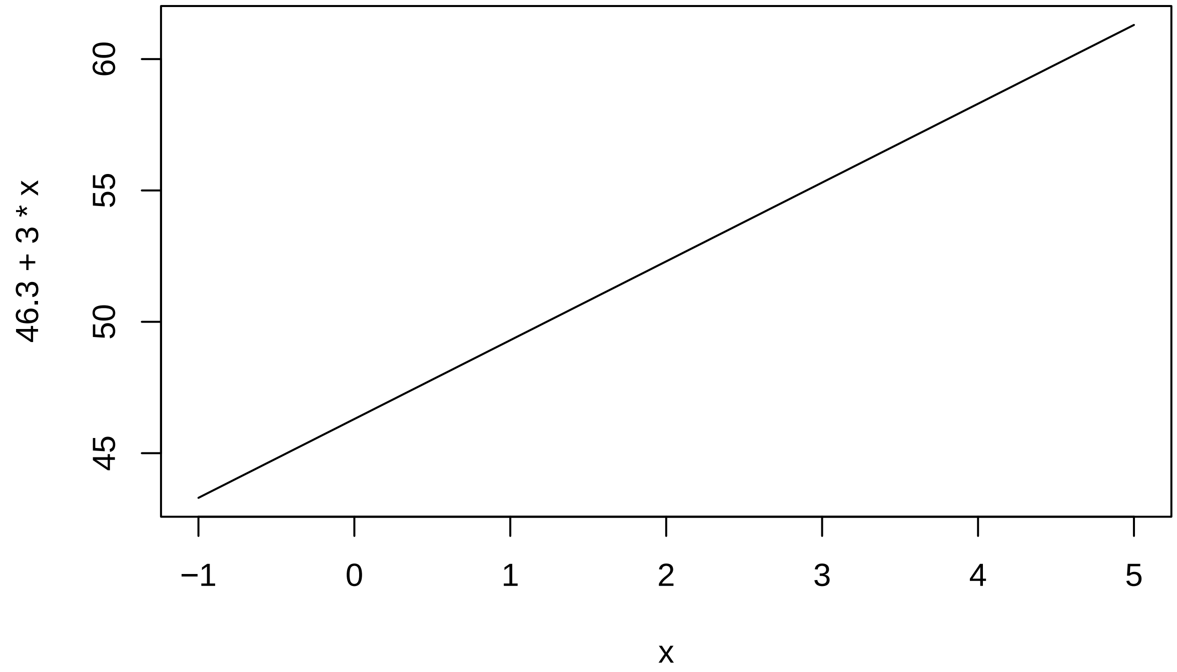
```
##      [,1] [,2]
## [1,]    1   -1
## [2,]    1    0
## [3,]    1    1
```

```
betahat=c(46.3,3.0)
yhat=X%*%betahat
yhat
```

```
##      [,1]
## [1,] 43.3
## [2,] 46.3
## [3,] 49.3
```

##Graphing a Function such as a Line in R

```
curve(46.3+3.0*x, from=-1, to=5)
```



Introducing Probability

Today we will introduce some probability basics.

Five Key Facts About Probability

1. Probabilities are between 0 and 1.
2. Something that happens with probability 1 is a sure thing.
3. If something has no chance of occurring, it has probability 0.
4. If something has probability p of happening it is has probability $1 - p$ of not happening.
5. If two items are mutually exclusive, the probability one or the other of them happen is the sum of their probabilities.

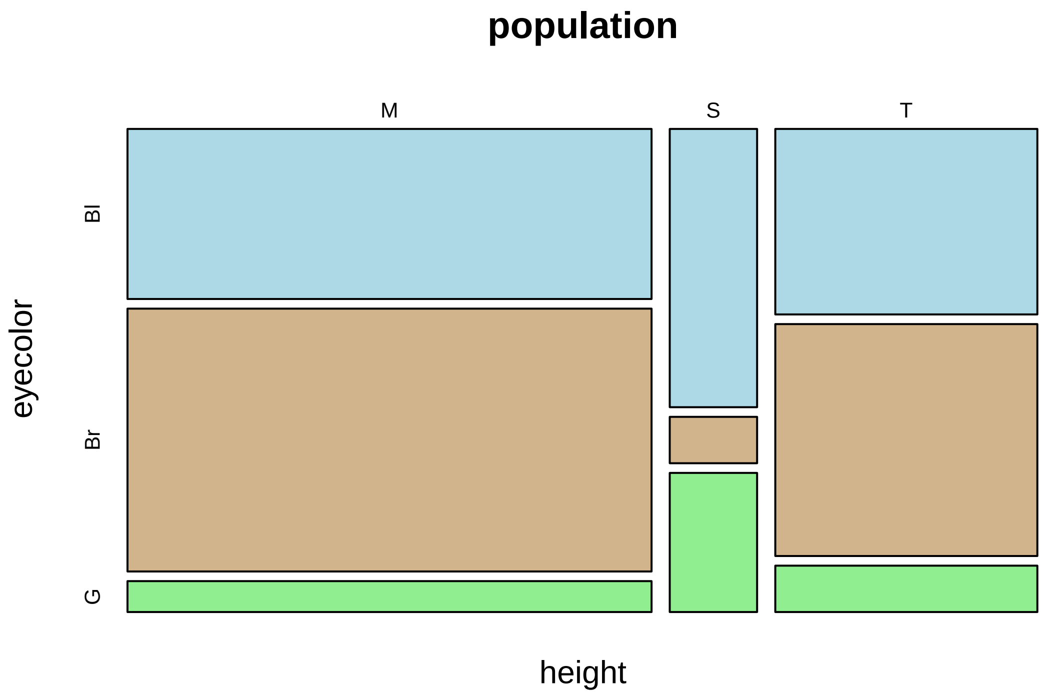
note: We have to assume something always occurs in our random experiment, and that all the probabilities add up to one.

Independence

Example: consider a population with the proportions below, S=small height, M=medium height, T=tall, Bl=blue eyes, Br=brown eyes, G=Green eyes.

```
poptable=tribble(
  ~height, ~eyecolor,  ~p,
 "T", "Bl",  .12,
"T", "Br", .15,
"T", "G", .03,
"M", "Bl", .22,
"M", "Br", .34,
"M", "G",  .04,
"S", "Bl",  .06,
"S", "Br",  .01,
"S", "G",  .03)
poptable <- poptable %>%
  mutate(p1000=p*1000)

population = poptable %>%
  uncount(p1000) #uncount takes a frequency variable and
#replicates the data that many times and keeps all the #other variables in the data
mosaicplot(height~eyecolor, data=population,
           color=c('lightblue','tan','lightgreen'))
```



Two events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

Are the events a person is medium height and a person has blue eyes independent? Why or why not?

** answer here

Random Variables

A random variable is a function that equates outcomes from a random experiment with numbers. For example the number of dots on the upside if we roll a die. Lets see how we can simulate from such a random variable.

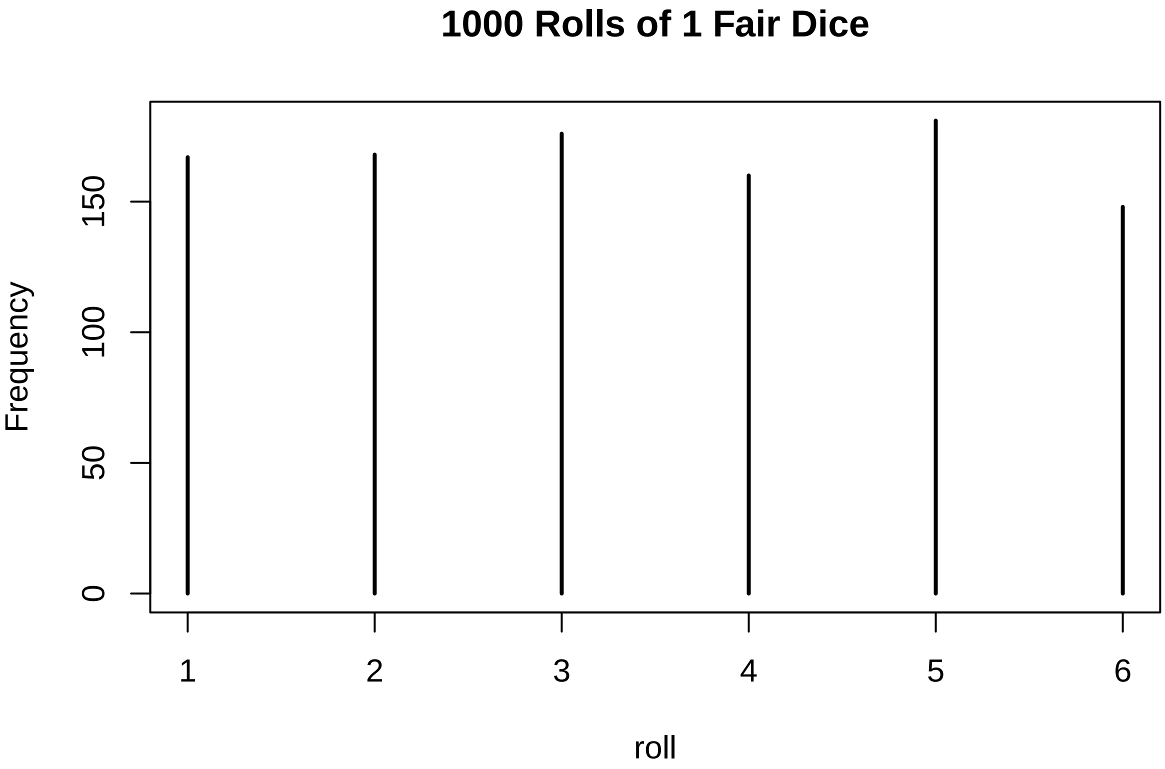
```
dice = 1:6
sample(x = dice, size = 1, replace = TRUE)
```

```
## [1] 5
```

```
sims=sample(x = dice, size = 1000, replace = TRUE)
table(sims)/length(sims)
```

```
## sims
##     1     2     3     4     5     6
## 0.167 0.168 0.176 0.160 0.181 0.148
```

```
plot(table(sims), xlab = 'roll', ylab = 'Frequency', main = '1000 Rolls of 1 Fair Dice')
```



This is an example of a discrete random variable, as it has countable values. We can use our simulation to get an idea of what would happen in the long run.

The expected value of a Random Variable (aka the population mean) is what we expect the mean value to be if we take an infinite sample. It is defined to be:

Similarly the variance of a RV (population variance) is defined to be:

and the standard deviation is the square root of the variance.

Calculate the mean and variance for the random variable X, the value on the top of a 4 sided die.

Covariance and Correlation

The covariance is a measure of the strength of the linear relationship between two variables U and V, it is defined to be:

It can be rescaled into a correlation, a value inbetween -1 and 1, where 0 means no linear relationship and 1 and -1 is a perfect linear relationship.

Note that if two random variables are independent it implies that the the covariance is necessarily 0. However, a covariance of 0 does not necessarily imply two variables are independent, remember covariance and correlation only measure the strength of the linear relationship.

Rules for linear combinations of correlated random variables:

If two random variables U and V have mean μ_u and μ_v and standard deviation σ_u and σ_v and correlation ρ then the mean and variance of $U + V$ is:

and the mean and variance of $aU + bV$ is:

Rules for linear combinations of independent random variables:

We can use these rules to find the expected value and variance of the sample mean for a set of n, independent identically distributed(IID) random variables (RVs).

Continuous Random Variables

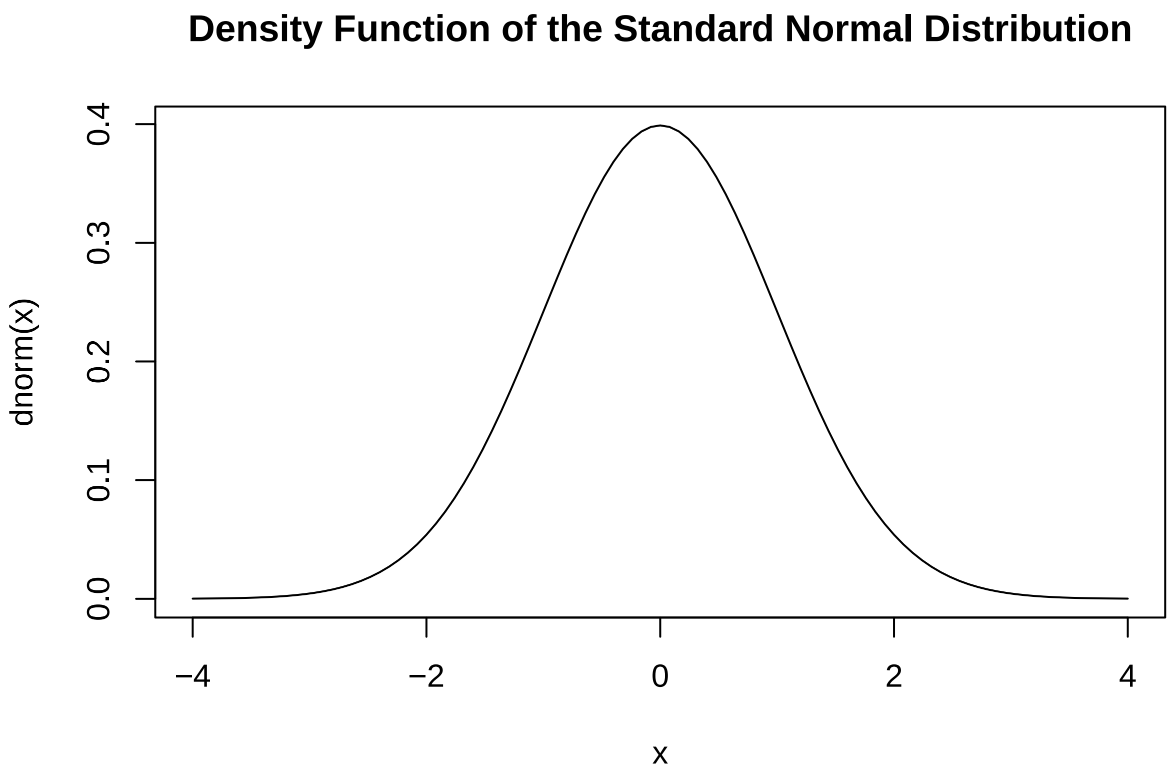
Many variables can not be exactly measured, like height, weight, etc. We can only measure them to within a certain interval. So theoretically we also need continuous random variables. The most famous of the continuous RVs is the Normal Distribution.
\

The Standard Normal Distribution

The probability density function for the standard normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

```
curve(dnorm(x), from=-4, to=4, main="Density Function of the Standard Normal Distribution")
```



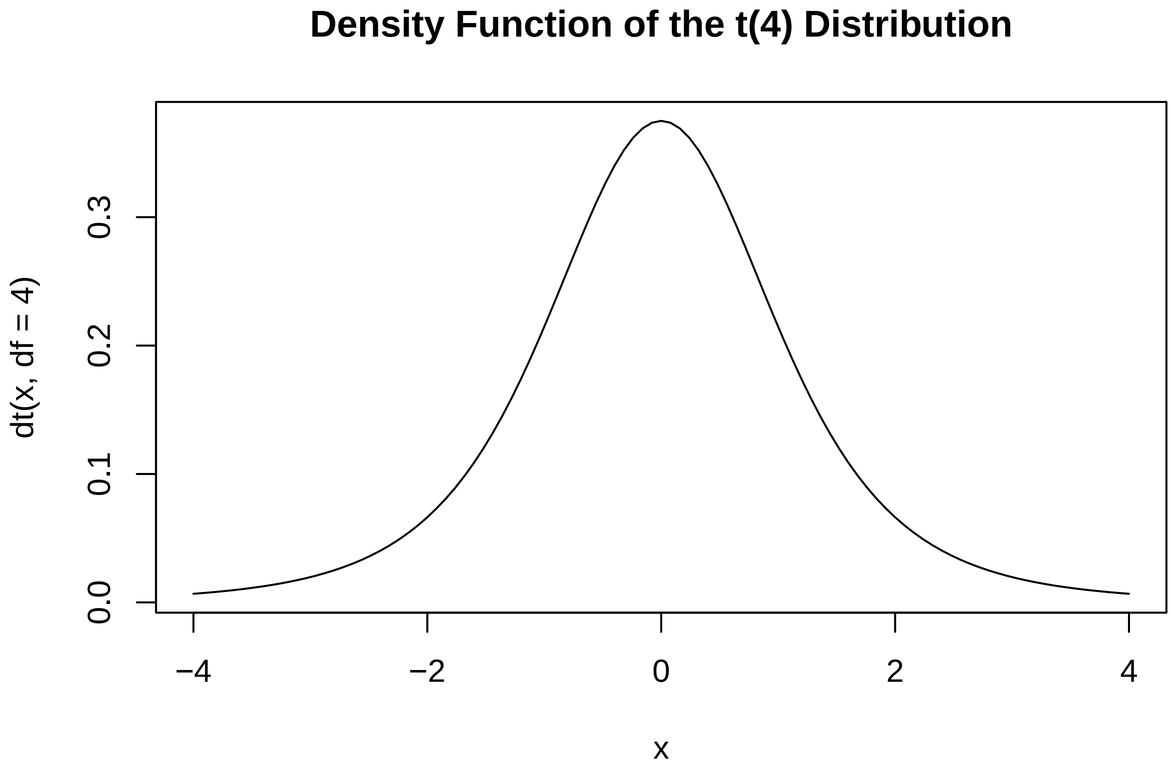
The T-distribution

The distribution of the random variable:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

where $\bar{X}$ is the mean of n independent random variables from a normal distribution with mean μ and standard deviation σ and S is the sample standard deviation. has a t-distribution with n-1 degrees of freedom. The T-distribution is also symmetric, but it has heavier tails than the normal distirbution.

```
curve(dt(x,df=4), from=-4, to=4, main="Density Function of the t(4) Distribution")
```



The exponential distribution

The exponential distribution is another named continuous distributions. It is often used to model the time elapsed between events.

A continuous random variable X is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim Exponential(\lambda)$, if its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The standard deviation and expected value for an exponential RV are:

If X is exponential with parameter $\lambda > 0$, then X is a memoryless random variable, that is

$$P(X > x + a | X > a) = P(X > x), \text{for } a, x \geq 0.$$

From the point of view of waiting time until arrival of a customer, the memoryless property means that it does not matter how long you have waited so far. If you have not observed a customer until time a, the distribution of waiting time (from time a) until the next customer is the same as when you started at time zero.

For the exponential distribution you can write down the cumulative distribution function (CDF)

$$Pr(X \leq x) = F_X(x) = \begin{cases} 0, x \leq 0 \\ 1 - e^{-\lambda x}, \quad x > 0 \end{cases}$$

Example: What is the median of an exponential random variable with $\lambda = 0.5$

Discrete Random Variables

Binomial Random Variables

The Binomial model is helpful when we want to model the number of successes when we draw a random sample from a really big population (by big I mean big enough where it is really unlikely someone would be sampled twice) or when we have a random process with independent trials and two possible outcomes, such as coin flipping or in a manufacturing process where each product is either "in spec" or "out of spec". Probabilities were being calculated with this distribution as far back as the early 18th century. It is attributed to Bernoulli. Interestingly the normal distribution was discovered by Abraham de Moivre later in the 18th century as he was needing to calculate probabilities from the Binomial for gambling problems. The binomial model is NOT always the right model to use as is discussed in the book.

Suppose we have N independent trials that can result in success or failure and the probability of success is constant, p, then the random variable X that is the number of successes in the N trials can be modeled using the Binomial distribution. Notation: The distribution of X is:

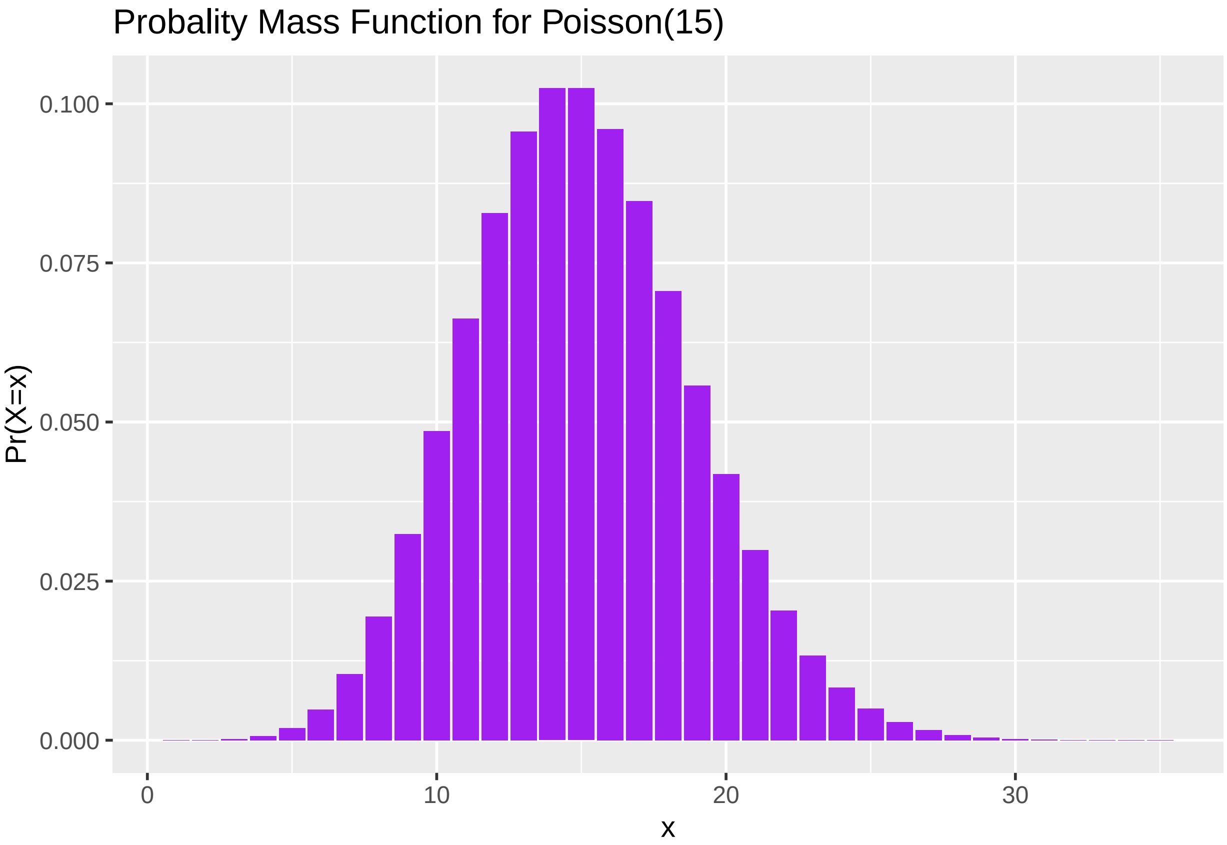
Poisson Distribution

The Poisson distribution is another famous distribution that is frequently used in practice. It is usually used in scenarios where we are counting the occurrences of certain events in an interval of time or space. Or it is used when as an approximation to the Binomial when n is large and p is very small. The PMF of the Poisson is given by:

$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

```
poissonpmf <-
  tibble (x=1:35, prob=dpois(x,15))

ggplot(data=poissonpmf, mapping=
         aes(x=x,y=prob))+
  geom_col(fill='purple') +
  ggtitle("Probality Mass Function for Poisson(15)") +   xlab("x") + ylab("Pr(X=x)")
```



Suppose that we are counting the number of customers who visit a certain store from 1pm to 2pm. Based on data from previous days, we know that on average $\lambda = 15$ customers visit the store. Of course, there will be more customers some days and fewer on others. Here, we may model the random variable X showing the number customers as a Poisson random variable with parameter $\lambda = 15$

example Suppose, the number of emails that I get in a weekday can be modeled by a Poisson distribution with an average of 0.2 emails per minute.

1. What is the probability that I get no emails in an interval of length 5 minutes?
2. What is the probability that I get more than 3 emails in an interval of length 10 minutes?