**Poisson probability distribution**

The binomial probability distribution allows one to work with a count response when there is a number of successes out of a particular number of trials. Count responses can also arise from other mechanisms that have nothing to do with success/failure like events. Examples include:

* The number of credit cards an individual owns
* The number of arrests for a city per year
* The number of people arriving at an airport on a given day
* The number of cars stopped at an intersection of streets
* The number of people standing in line at Starbucks.

For these settings, a Poisson probability distribution can be used to model the count responses.

One may also want to control for other items that affect how the count is taken.

* If the purpose of the number of arrests example was to make comparisons across cities in the United States, we would want to account for city sizes. It would not make sense to compare Omaha directly to New York City without taking into account their vastly different sizes.
* If the purpose of the airport example was to examine it over different lengths of time periods (say, one day or one week), it would not make sense to make a direct comparison.

We will discuss a simple fix for this shortly.

The Poisson PDF for a random variable Y is

 for y = 0, 1, 2, …

for a parameter λ > 0. Once E(Y) is given, you will see an easy way to estimate λ!

Notes:

* Y does not have an upper bound!
* dpois() in R calculates f(y)
* The Poisson CDF is



The ppois() calculates F(y).

The mean and variance for the Poisson PDF are:

 and 

Note: The mean and the variance is the same! This can cause problems though with the PDF being truly representative of a population.

proof:

E(Y) = 

= 

=  since y=0 does not contribute to sum

and e-λ does not contain a y value

=

= 

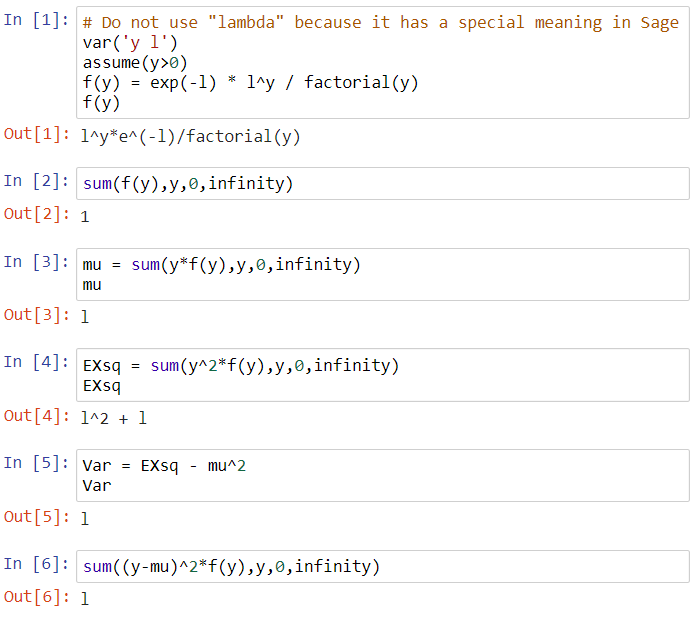
=  where x = y - 1

=  using the result of 

= λ

One can also derive E(Y2) and use it to obtain Var(Y). Instead, let’s to do this in Sage!

Example: Poisson symbolic calculations (Poisson.ipnyb)



Example: Poisson numerical calculations (Poisson\_plots.R)

Calculate probabilities with λ = 3:

# f(2)

> dpois(x = 2, lambda = 3)

[1] 0.2240418

> exp(-3)\*3^2 / factorial(2)

[1] 0.2240418

> # F(2)

> ppois(q = 2, lambda = 3)

[1] 0.4231901

> sum(dpois(x = 0:2, lambda = 3))

[1] 0.4231901

> # f(y) and F(y) up to y = 15

> y <- 0:15

> data.frame(y, fy = round(dpois(x = y, lambda = 3),4),

Fy = round(ppois(q = y, lambda = 3),4))

y fy Fy

1 0 0.0498 0.0498

2 1 0.1494 0.1991

3 2 0.2240 0.4232

4 3 0.2240 0.6472

5 4 0.1680 0.8153

6 5 0.1008 0.9161

7 6 0.0504 0.9665

8 7 0.0216 0.9881

9 8 0.0081 0.9962

10 9 0.0027 0.9989

11 10 0.0008 0.9997

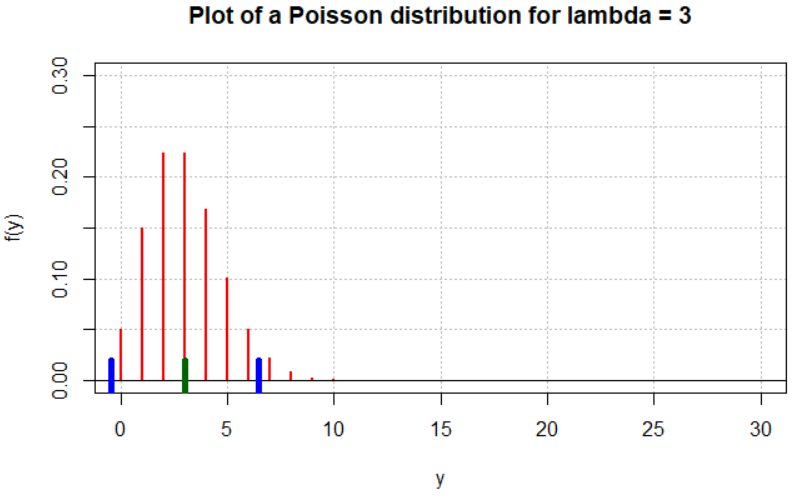
12 11 0.0002 0.9999

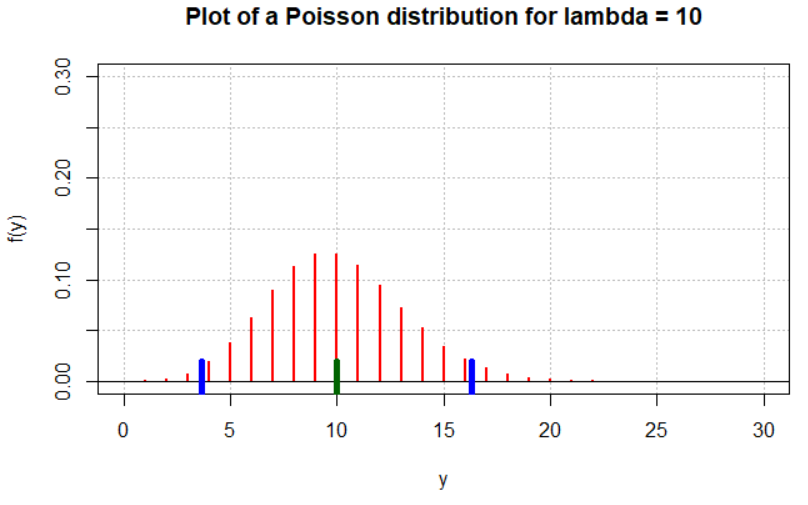
13 12 0.0001 1.0000

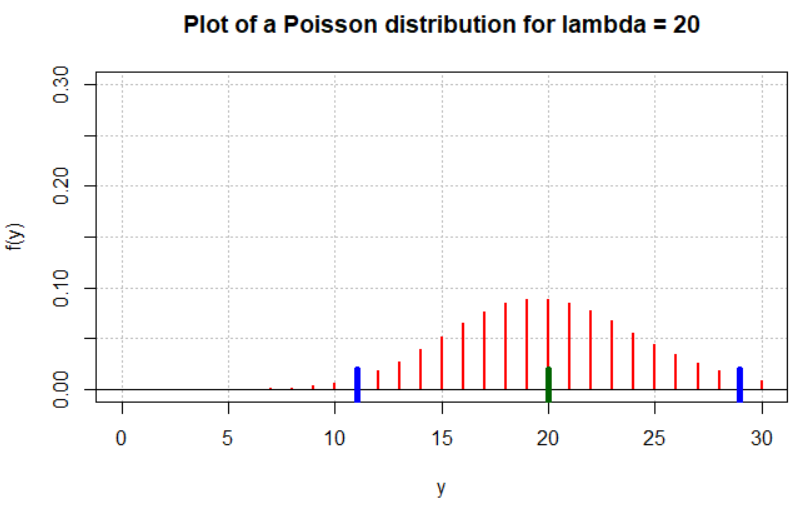
14 13 0.0000 1.0000

15 14 0.0000 1.0000

Below are plots of the PDF (green line is μ and green like is rule of thumb). Please see the program for the code.







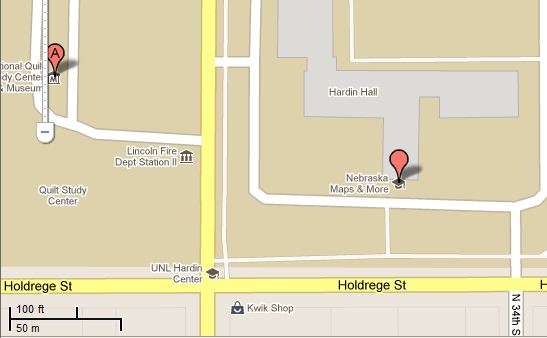
Examine the following with these plots:

* When are the plots “symmetric” and when are the plots “skewed”?
* Where is the largest probability?
* Notice the μ ± 2σ lines.
* Given the results of these plots, why do you think it is important to have a good value for λ relative to an actual application?

On your own, simulate samples from a population characterized by a Poisson distribution using rpois() in R. Compare these samples to what you would expect using the same methods as we did for the binomial PDF.

Example: 33rd and Holdrege streets in Lincoln, NE (Stoplight.R, Stoplight.csv)

The intersection at 33rd and Holdrege Streets is a typical north-south/east-west, 4-way intersection.



Below is a picture taken from a location northeast of the intersection:



Approximately 150 feet north of the intersection is a fire station located on the west side of the street. A back-up of vehicles at the stoplight waiting to go south could block the fire station's driveway, which would prevent emergency vehicles from exiting the station.

To examine this more closely, I took a sample of 40 consecutive stoplight cycles from 3:25PM to 4:05PM on a non-holiday weekday, and the number of vehicles stopped at the stoplight going south were counted. Below is part of the data:

> stoplight <- read.csv(file = "stoplight.csv")

> head(stoplight)

Observation vehicles

1 1 4

2 2 6

3 3 1

4 4 2

5 5 3

6 6 3

Note that there were no vehicles remaining in the intersection for more than one stoplight cycle. Why is this important to know?

Is a Poisson PDF appropriate for this data? Compare what we would expect to happen with a Poisson PDF to what was actually was observed.

> # Numerical summaries

> mean(stoplight$vehicles)

[1] 3.875

> var(stoplight$vehicles)

[1] 4.317308

> #Frequencies

> table(stoplight$vehicles)

0 1 2 3 4 5 6 7 8

1 5 7 3 8 7 5 2 2

> rel.freq <- table(stoplight$vehicles) /

length(stoplight$vehicles)

> rel.freq2 <- c(rel.freq, rep(0, times = 7))

> #Poisson calculations

> y <- 0:15

> prob <- round(dpois(x = y, lambda =

mean(stoplight$vehicles)), 4) # Examine lambda used

> data.frame(y, prob, rel.freq = rel.freq2)

y prob rel.freq

1 0 0.0208 0.025

2 1 0.0804 0.125

3 2 0.1558 0.175

4 3 0.2013 0.075

5 4 0.1950 0.200

6 5 0.1511 0.175

7 6 0.0976 0.125

8 7 0.0540 0.050

9 8 0.0262 0.050

10 9 0.0113 0.000

11 10 0.0044 0.000

12 11 0.0015 0.000

13 12 0.0005 0.000

14 13 0.0001 0.000

15 14 0.0000 0.000

16 15 0.0000 0.000

> plot(x = y - 0.1, y = prob, type = "h", ylab =

"Probability", xlab = "Number of vehicles", lwd = 2,

xaxt = "n")

> axis(side = 1, at = 0:15)

> lines(x = y + 0.1, y = rel.freq2, type = "h", lwd = 2,

lty = "solid", col = "red")

> abline(h = 0)

> legend(x = 9, y = 0.15, legend = c("Poisson",

"Observed"), lty = c("solid", "solid"), lwd = c(2,2),

col = c("black", "red"), bty = "n")



If the Poisson PDF was appropriate, what could it be used for?

Questions:

* Why examine PDFs? They can be used to help model real life events. Remember we are making ASSUMPTIONS about the population. Rarely (if ever) will these assumptions be totally satisfied! Often, these assumptions will be satisfied “close enough” to justify their use.
* What if data is obtained over different time or unit sizes. For example, suppose we are interested in the number of people arriving at an airport. One can use t×λ, where t is the unit of time, rather than λ alone in the Poisson PDF. Thus, we could work with the number of people arriving over t hours of time. We could let λ be the mean number for one hour. If we are interested in two hours, use 2λ.