Exploring Data

Jiapeng Zhang

Our goals here are:

- Look at data, test fit to some reasonable distribution on a per decade bases (Hint hurricanes are rare).
- Then do further research on hurricanes, see what you can find out.

But the data source is Wikipedia. Or to be more specific, the so-called data is from several wikipedia tables. The data isn't very "clean" and consistant in terms of format. We have to somehow first clean the data a little bit, with my familar Python Pandas package.

```
import pandas as pd
df=pd.read_excel("Category5.xlsx")
for i in range(len(df)):
    df.loc[i, "wind speeds"]=int(df.loc[i, "wind speeds"][-9:-6])
    df.loc[i, "Pressure"]=int(df.loc[i, "Pressure"][:3])
df.rename(columns={"wind speeds":"wind speeds(km/h)", "Pressure": "Pressure(hPa)"}, inplace=True)
df.to_excel("Cat5.xlsx")
```

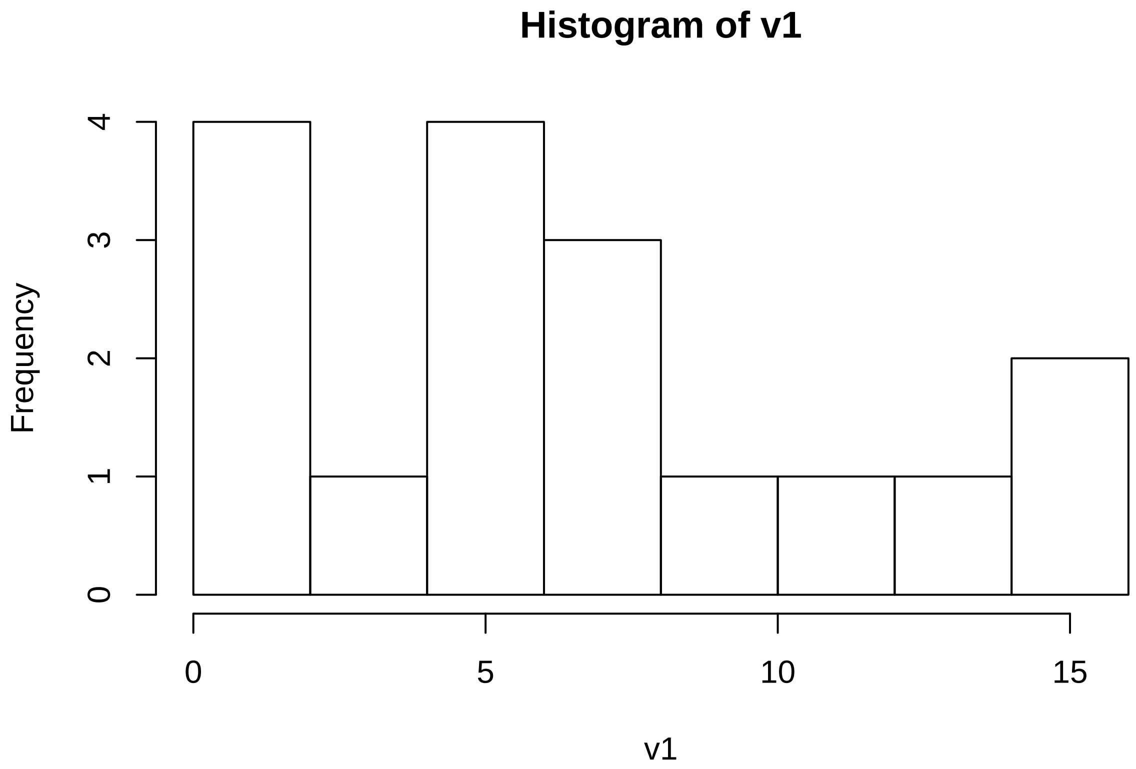
This time I will going only foucus on the year feature. In the future if I have time I will dig further. I have posted my dataset on the Kaggle platform. Maybe we will see some more in-depth analysis there.

The data starts from year 1853 to this year(2018), spanning 2018-1853+1=166 years. We combine every 10 years since 1853 as a row and count Category 4 Hurricanes occurences during each decade.

```
cat4<-read.csv("Cat4.csv")
v1<-NULL
i<-1853
while(i<2018){
  v1<-c(v1, nrow(subset( cat4,Season>=i & Season<(i+10) ) ) )
  i=i+10
}
```

Now we get the vector v1 containing # of hurricanes for each decade since 1853. What are we going to do next? By no means the distribution is going to be Guassian, or Uniform, etc. We think about Poisson, since the Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space.

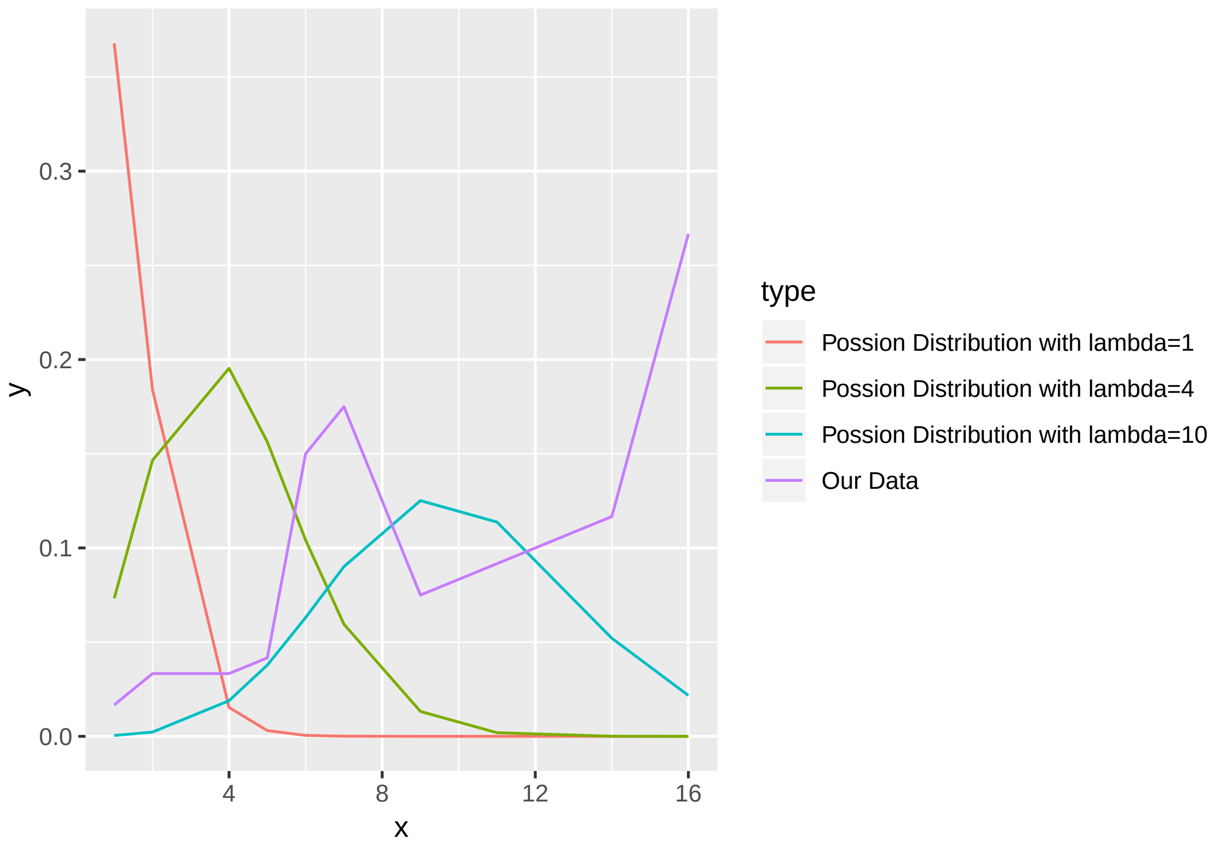
```
hist(v1)
```



Can our variable *v1* (which stands for # of hurricane of category 4 occurance) be of Possion distribution? We cannot get much from a single histogram. Can draw some historgrams for some given Poisson distribution? No, since Possion distribution only gives you probability for any number of occurence rather than concrete #number of occurance. In fact any number of occurance is possible though bigger ones tend to assoicate with smaller possibity.

```r
p1<-NULL
for(c in unique(v1)){
  p1<-c(p1, c*sum(v1==c)/sum(v1))
}
```

```r
poi1=dpois(unique(v1), lambda = 1)
df1=data.frame(x=unique(v1), y=poi1, type="Possion Distribution with lambda=1")
poi2=dpois( unique(v1), lambda = 4)
df2=data.frame(x=unique(v1), y=poi2, type="Possion Distribution with lambda=4")
poi3=dpois( unique(v1), lambda = 10)
df3=data.frame(x=unique(v1), y=poi3, type="Possion Distribution with lambda=10")
df4=data.frame(x=unique(v1), y=p1, type="Our Data")
df<-rbind(df1, df2, df3, df4)
library(ggplot2)
ggplot(df)+geom_line(aes(x,y,colour=type))
```



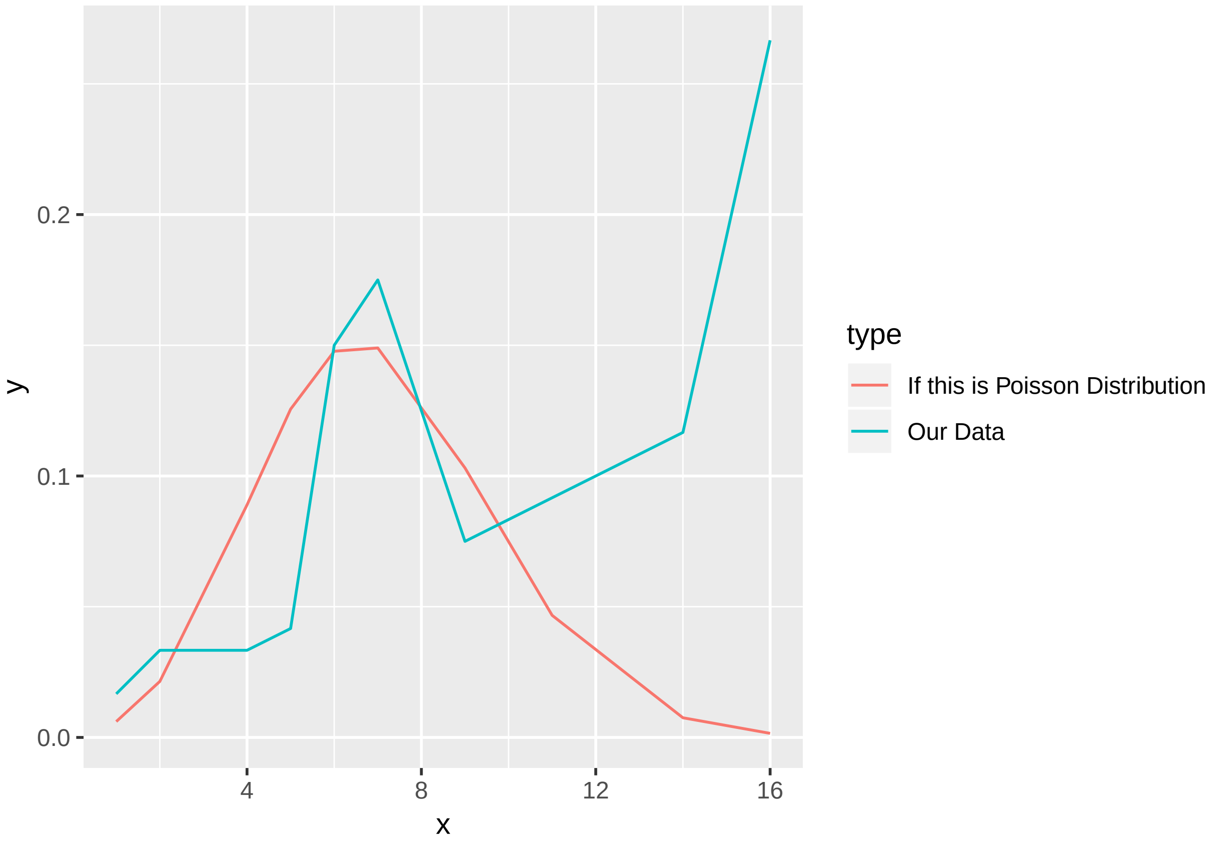
After plotting the density against three *real* poisson distributions. From the graph, you can see several peaks and bottoms, which makes the density doesn't look like Poisson. If it is, probably the its lambda value is near 10.

Recall from *Parameter Estimation* knowledge, if the distribution follows Poisson distribution, the *Maximum Likelihood Estimation* gives us the formula

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} k_i$$

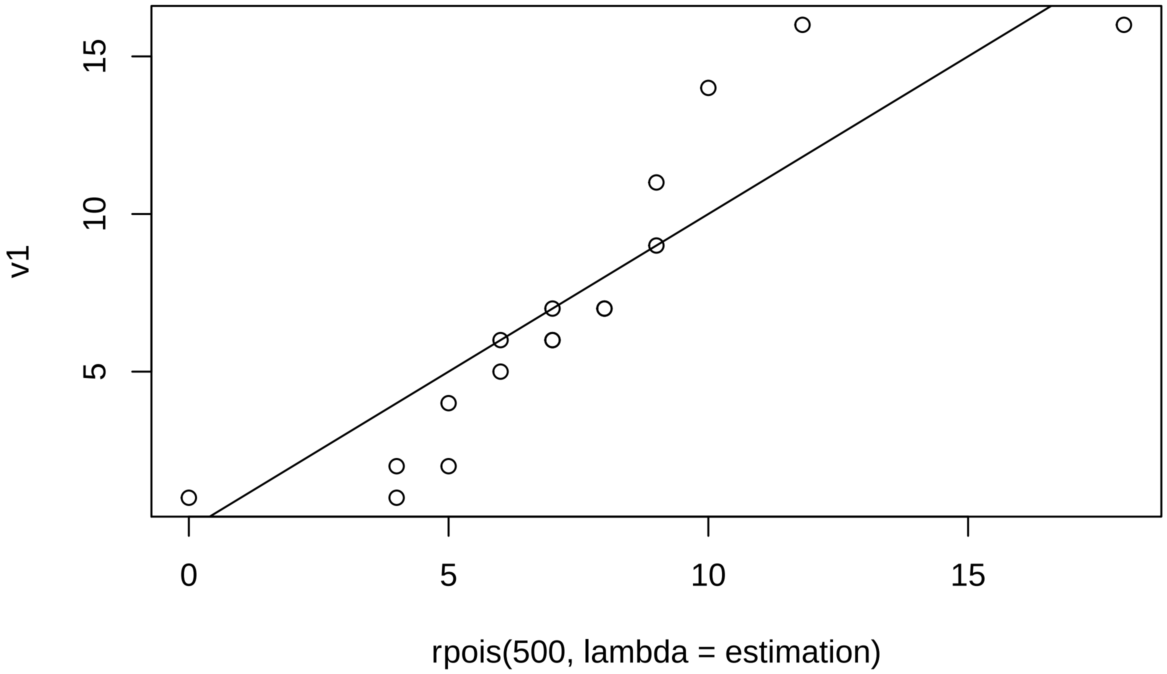
, where k_i stands for each occurence in datasets.

```
estimation=sum(v1)/length(v1)
df_lambda=data.frame(x=unique(v1), y=dpois( unique(v1), lambda = estimation), type="If this is Poisson
df_real=data.frame(x=unique(v1), y=p1, type="Our Data")
df<-rbind(df_lambda, df_real)
ggplot(df)+geom_line(aes(x,y,colour=type))
```



Uhhhh, doesn't look very like that…
Another way to compare distribution is to look at the so-called *Quantile-Quantile (Q-Q) plot*, as is described in here and here. In QQplot, we plot against two datasets' quantiles agianst each other.

```
qqplot(rpois(500, lambda = estimation), v1)
abline(0,1)
```

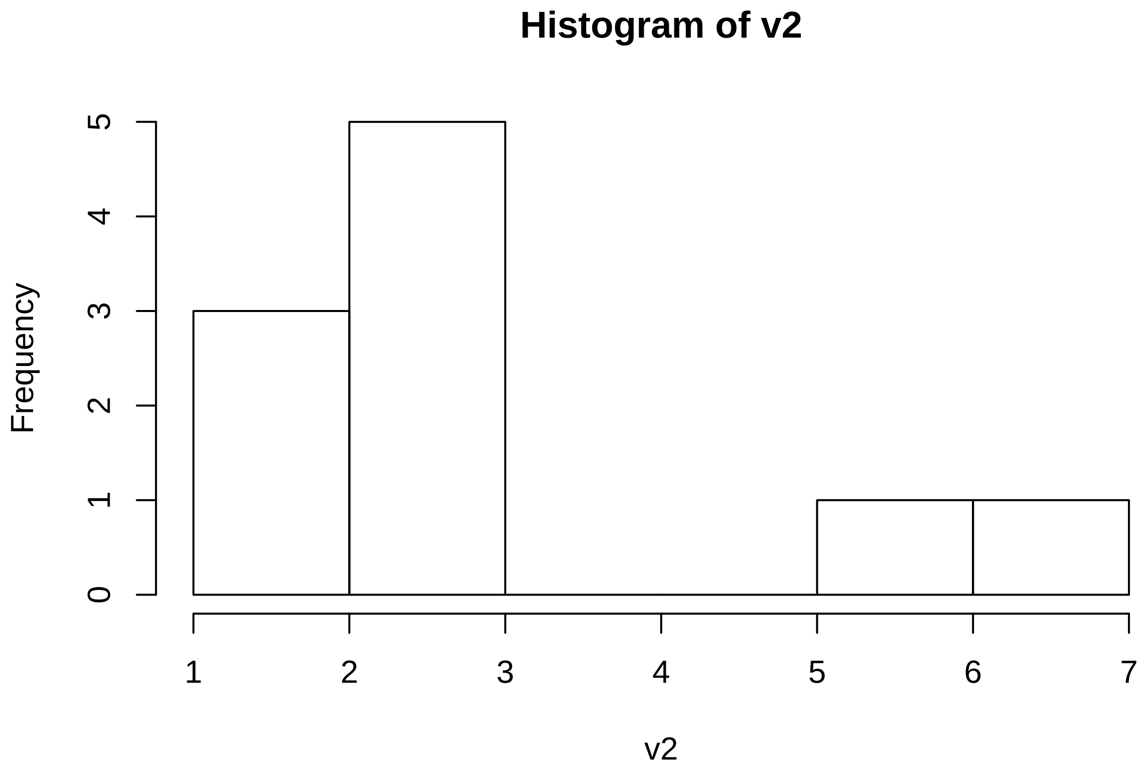


Again, in qqplot, we confirm our idea that our hurricane data doesn't follow Poisson distrition.

Forget about this, let's Category 5 dataset.

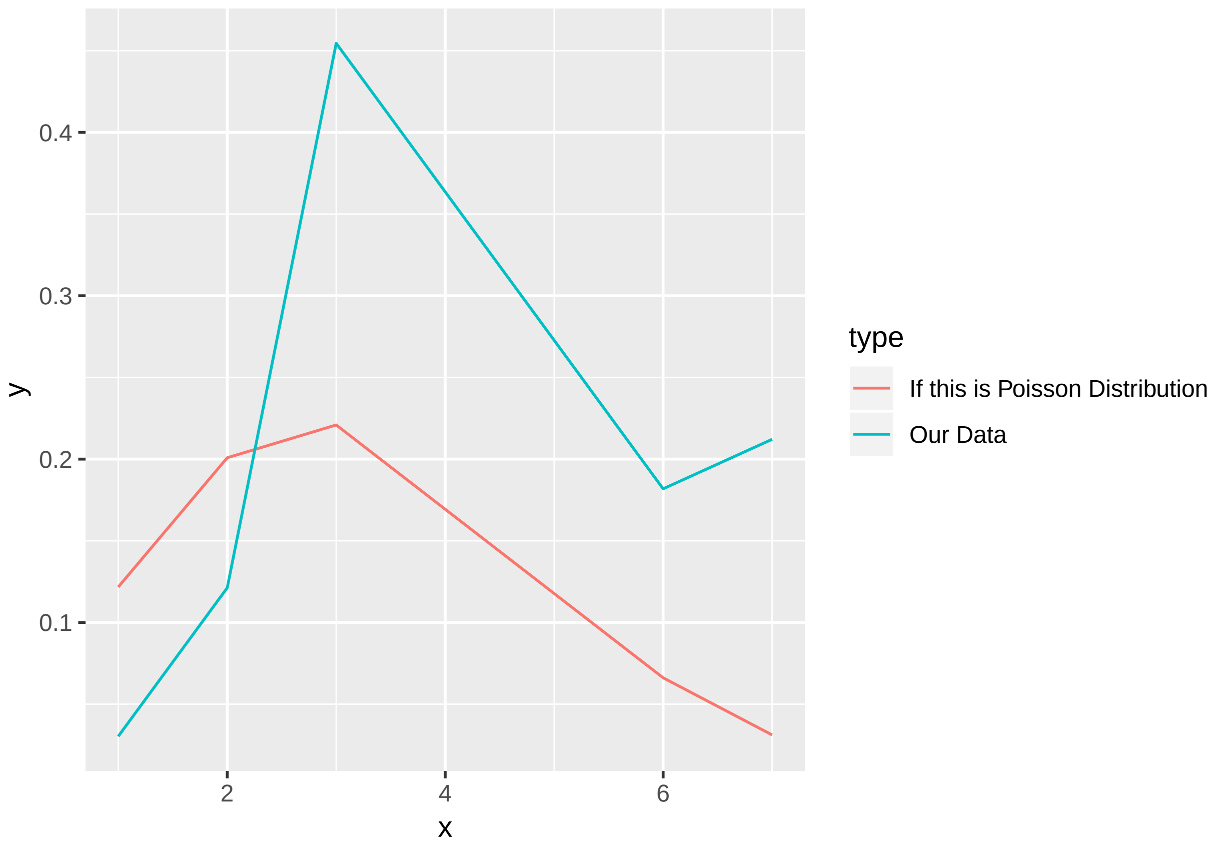
```
DF<-read.csv("Cat5.csv")
v2<-NULL
for(i in 1:nrow(DF)){
  DF$year[i]=as.integer( substr(DF$Dates[i], regexpr("[0-9]{4}", DF$Dates[i]), regexpr("[0-9]{4}", DF$Da
}
 DF<-DF[order(DF$year),]
 i<-DF$year[1]
while(i<DF$year[nrow(DF)]){
  v2<-c(v2, nrow(subset( DF, year>=i & year<(i+10) ) ) )
  i=i+10
}
```

```
hist(v2)
```

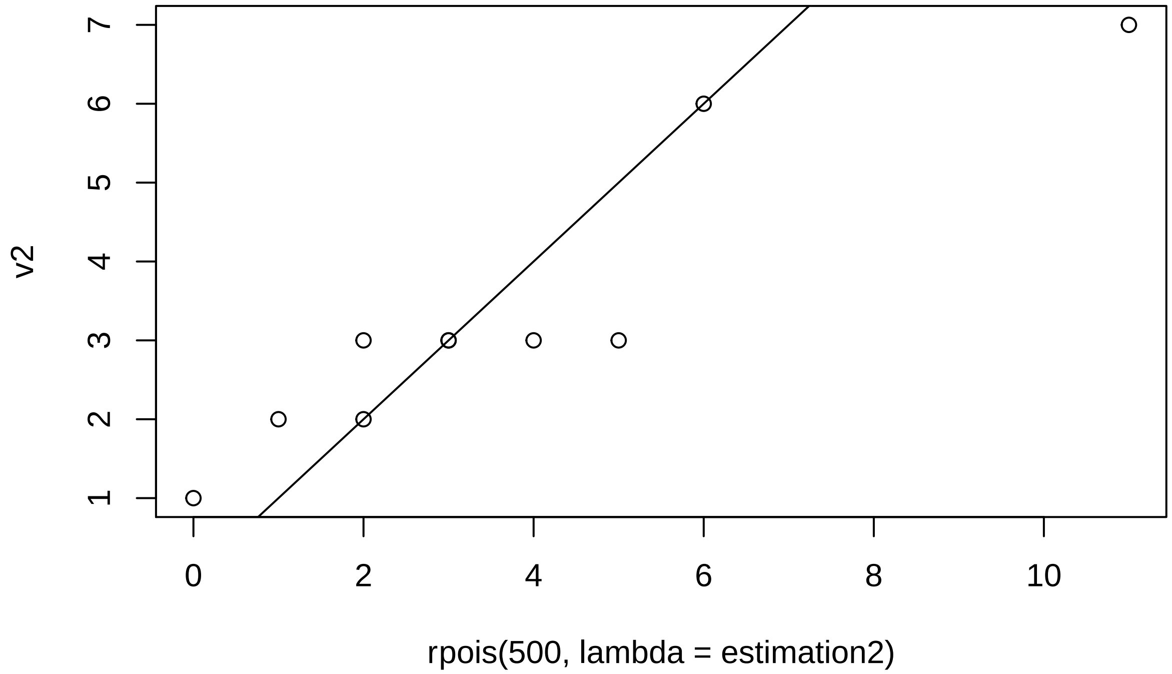


```
p2<-NULL
for(c in unique(v2)){
  p2<-c(p2, c*sum(v2==c)/sum(v2))
}
```

```
estimation2=sum(v2)/length(v2)
df_lambda2=data.frame(x=unique(v2), y=dpois( unique(v2), lambda = estimation2), type="If this is Poisson
df_real_2=data.frame(x=unique(v2)
                     , y=p2, type="Our Data")
df2<-rbind(df_lambda2, df_real_2)
ggplot(df2)+geom_line(aes(x,y,colour=type))
```



```r
qqplot(rpois(500, lambda = estimation2), v2)
abline(0,1)
```



Neither of the datasets look lke follow Possion. What distribution do they follow? Hell knows. Quote from stackexchange here

> The thing is that real data doesn't necessarily follow any particular distribution you can name ... and indeed it would be surprising if it did.
>
> So while I could name a dozen possibilities, the actual process generating these observations probably won't be anything that I could suggest either. As sample size increases, you will likely be able to reject any well-known distribution.
>
> Parametric distributions are often a useful fiction, not a perfect description.

From my past experience, in machine learning people are doing the same thing too. But they usually don't try to fit the data to these well-known distributions now, since the complexity of nature fat exceeds common probability distributions can express. Maybe some general models like random forest or gradient boosting decesion tree may help. In the future it's worth a try.