954:581 - Probability and Statistical Inference
Assignment 1
Explore Hurricane Data
Due: 19th September 2019

Keya Desai
RUID: 197001879

1. Data and distribution

The data has been taken from Wikipedia page on the occurrence of hurricanes (Category 4 and Category 5) throughout the years: 1950 till present. We compiled the data and categorized the occurrence frequency per decade. The data compiled is as shown below:

decade	frequency
1850–1860	2
1860–1870	1
1870–1880	3
1880–1890	2
1890–1900	5
1900–1910	2
1910–1920	5
1920–1930	7
1930–1940	9
1940–1950	12
1950–1960	12
1960–1970	9
1970–1980	5
1980–1990	7
1990–2000	14
2000–2010	17
2010–2019	8

decade	frequency
1930–1940	6
1940–1950	0
1950–1960	2
1960–1970	4
1970–1980	4
1980–1990	2
1990–2000	2
2000–2010	8
2010–2019	5

Fig: Category - 4 and 5 data: (Source: Wikipedia Category-4 Atlantic hurricanes)

Here we are considering the random variable (X) as the number of hurricanes per decade. Since it is a discrete variable, the possible distributions that can fit the data are discrete uniform, Bernoulli, binomial, geometric, Poisson distribution, etc. The data is not uniform and thus discrete uniform distribution is not an option in this data. Since we only have the occurrence frequency of the hurricane distributed over the years and not the probability of occurrence of hurricane, binomial distribution, Bernoulli, and geometric distribution would not give an accurate distribution for the data. Since Poisson distribution is a good fit for data which occur less frequently or rarely, we have chosen Poisson distribution to fit the data.

Poisson distribution has one parameter- λ (lambda) which is the average/mean of the data. In this data,

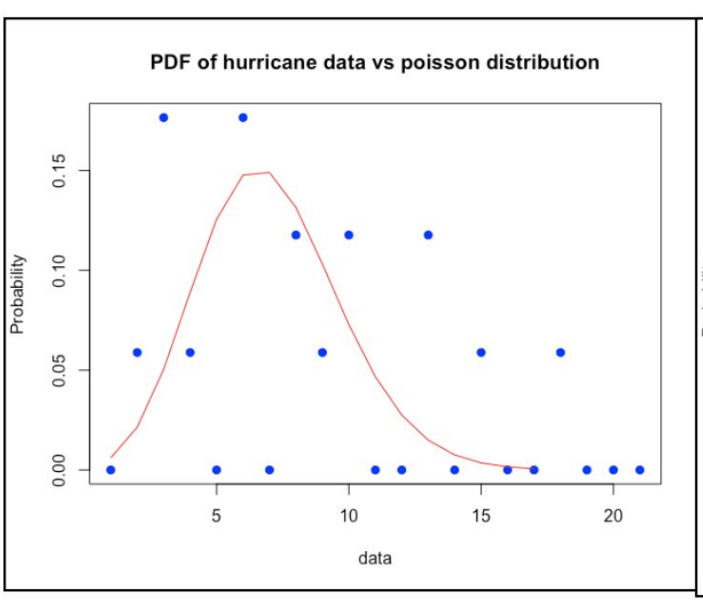
$$\lambda = 7.05 \text{ (for Category 4)}$$
$$\lambda = 3.66 \text{ (for Category 5)}$$

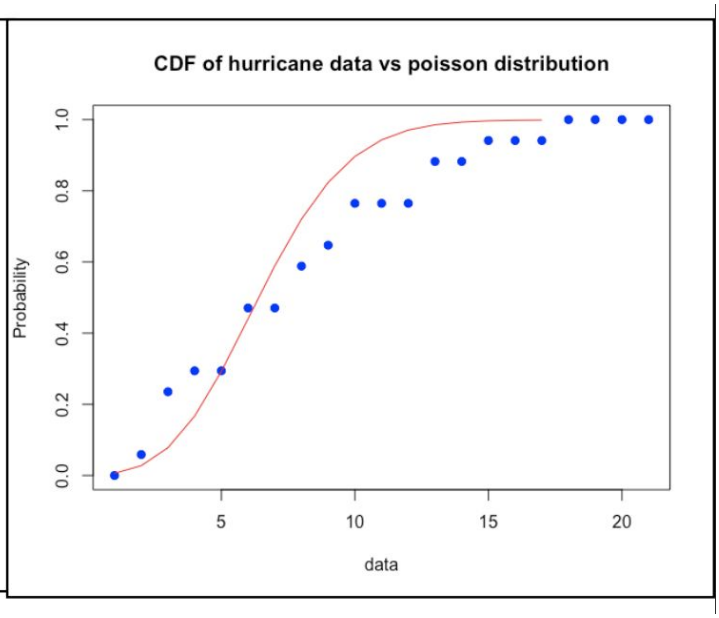
2. Approach and Results

In order to check whether the distribution of the number of hurricanes per decade follows a Poisson distribution, we look at the data's PDF and CDF and see if it is similar to Poisson's PDF and CDF. And then we also the Q-Q plot to determine the data's distribution.

a. PDF/CDF

Probability Distribution Function should be the same if the samples are picked from a similar distribution.





As we can observe from the data, the PDF and CDF of the given data deviate from the PDF and CDF of theoretical Poisson distribution with similar parameter λ.

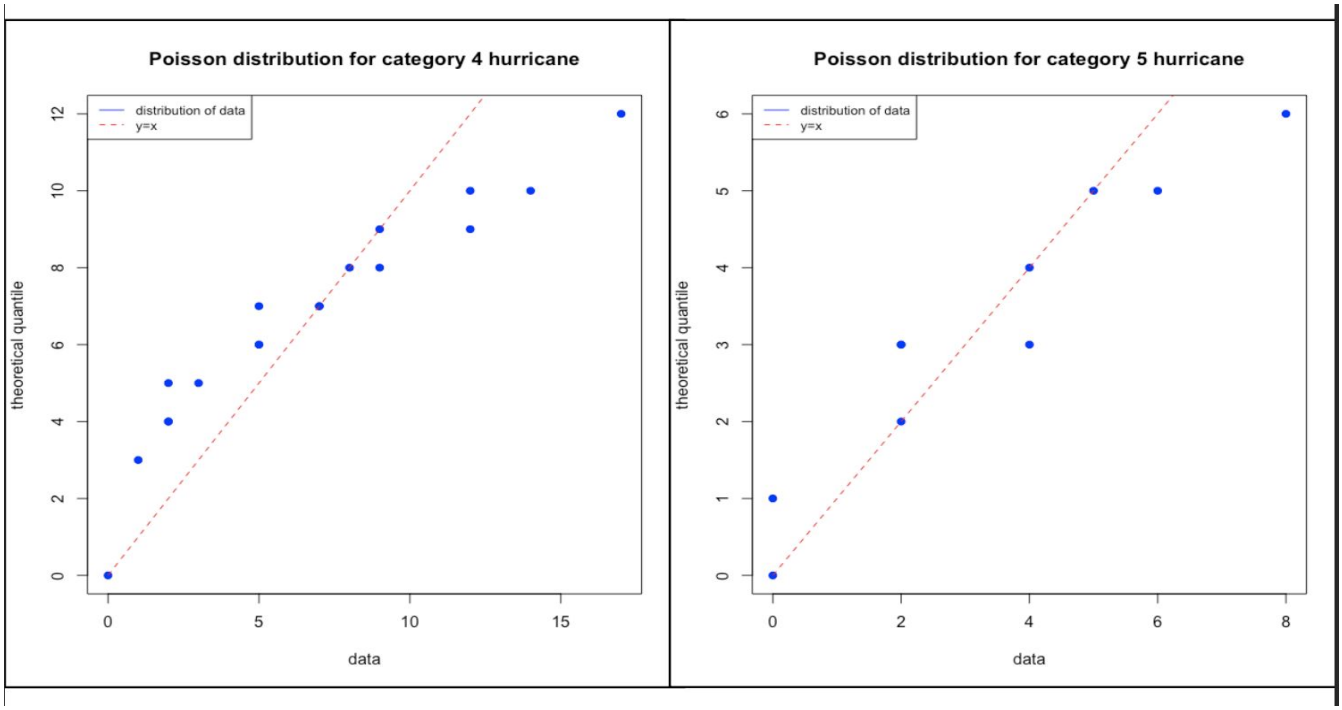
b. Q-Q plot

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Here, instead of comparing two samples, we are comparing the sample with the actual Poisson distribution.

Method:

1. x is the sample data vector => **N = length(x)**
2. Calculate λ from sample data x => **λ = mean(x)**
3. Compute N quantiles from the theoretical distribution using λ calculated in previous step => **p = c(1:N)/(N+1)**, **q = qpois(p, λ)**
4. Plot sorted samples **sort(x)** versus **q.**

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.



As we can see from the plots, for both category 4 and category 5 hurricanes, the qqplot obtained by fitting Poisson distribution to the data does not produce a straight line. This implies that the **hurricane data does not follow the Poisson distribution.**

3. Analysis

Assumptions for Poisson distribution are:

1) Homogeneity - The probability of hurricanes occurring over every decade is constant. The parameter λ in the poisson distribution is the average number of hurricanes in a decade calculated from the given data.
2) Independence - The probability of hurricanes occurring over one decade is independent of hurricanes occurring over the next or any other decade.

From the plots, we can clearly see that the **Poisson distribution does not justify** the frequency of hurricanes in any decade. The reason for the same lies in the underlying assumptions of Poisson distribution and the way hurricanes are originated. Hurricanes originate when the warm, moist air over the ocean rises upward from near the surface. Low pressure is generated over the sea as the temperature of water rises. As a result, air from high-pressure regions flows there giving rise to turbulence in air. Hence, SST (Surface Sea Temperature) plays a key role in the formation of a hurricane.

The Atlantic Multi-decadal Oscillation (AMO) is a climate cycle affecting the surface temperature of the Atlantic Ocean with an estimated period of 60-80 years. As a result of this, our **assumption of Homogeneity is violated.** Hence, the Poisson variable is not a good fit for the data.

4. References

QQ-plot

1. https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm
2. https://data.library.virginia.edu/understanding-q-q-plots/
3. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Poisson.html

Hurricane Information

1. https://en.wikipedia.org/wiki/List_of_Category_4_Atlantic_hurricanes
2. https://climatedataguide.ucar.edu/climate-data/atlantic-multi-decadal-oscillation-amo