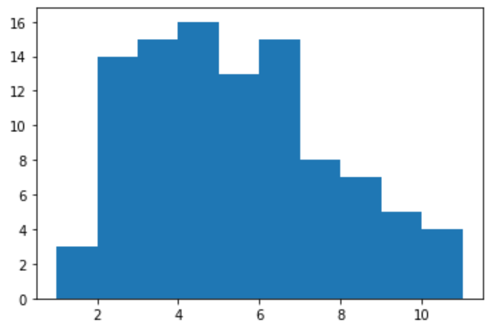
2.4 Guessing the Distribution of Dataset 1

Sunday, 09 October 2022 14:09

	Poisson Distribution

Histogram



- Since there are two peaks, it could be a non-Gaussian distribution.
- But, we cannot say whether it's a symmetric or a non-symmetric distribution.

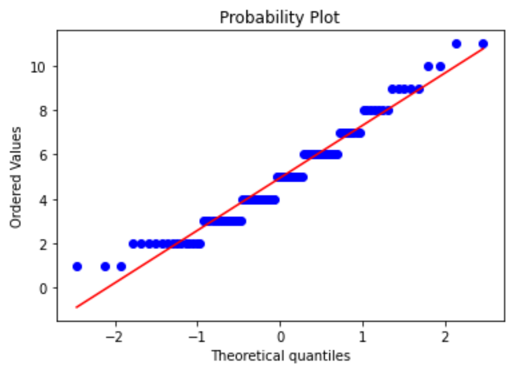
Descriptive Statistics

count	100.000000
mean	4.940000
std	2.381834
min	1.000000
25%	3.000000
50%	5.000000
75%	6.000000
max	11.000000

{('Variance Observed', 5.67), ('Mean Observed', 4.94)},
('Skew Observed', 0.51), ('Kurt Observed', -0.38)

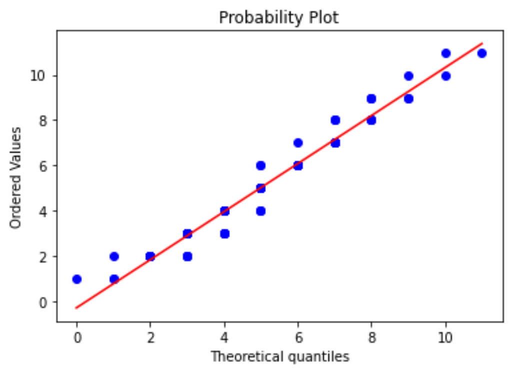
- Variance and mean are close to each other. Think about what distributions have mean and variance close to each other.
  Ans: Poisson Distribution
- Observing the skewness and the kurtosis, we can conclude that we can eliminate certain symmetric distributions such as normal distribution, uniform distribution.
- Mean = Median = 5. So, 50% points are lying below 5 and 50% lying above 5. So it's still a symmetric distribution. But, Poisson is not a symmetric distribution.

P-P Plot



- Points are not lying on the red line. It means the data is not coming from the normal.
- Since P-P plot deals with the CDF, and we have data similar to a step function, which also hints that this **distribution could be discrete in nature.**

P-P w.r.t Poisson distribution with lambda = 4.94



- Now the points are almost around the red line. So, with a certain degree of certainty, we can say that the cumulative distribution from which this data is coming is a Poisson distribution which has a lambda of 4.94.

Statistical tests

obs	frequency	OBS_PROBA	POISSON_PMF	POISSON_FREQ
1	3	0.03	0.035344	3.53
2	14	0.14	0.087299	8.73
3	15	0.15	0.143752	14.38
4	16	0.16	0.177534	17.75
5	13	0.13	0.175404	17.54
6	15	0.15	0.144416	14.44
7	8	0.08	0.101916	10.19
8	7	0.07	0.062933	6.29
9	5	0.05	0.034543	3.45
10	2	0.02	0.017064	1.71
11	2	0.02	0.007663	0.77

- What we first did here, we made a frequency chart (shown in frequency column).

	• OBS_PROB $= P(x_i = k) = \frac{frquency\ of\ k}{total\ frquency}$. ◦ *e.g.,* Total frequency $= 100$ $P(x_i = 1) = \frac{3}{100} = 0.03$. • Then we calculate Poisson pmf ◦ $P(x_i = k) = e^{-\lambda}\frac{\lambda^k}{k!}$. ◦ Here, $\lambda = 4.94$. • After this step, we calculate Poisson frequency, ◦ $Possion\ freq = Poisson\ pmf \times total\ freq$ $\Longrightarrow \quad = Poisson\ pmf \times 100$.
Observed and Expected frequency	• Observed probability tells you the probability of an observation occurring from the sample. • The Poisson pmf tells you, the probability of getting the same observation from the population (assuming that the population is Poisson) with the mean of 4.94. • So now we have both, the observed frequency and the expected frequency (Poisson freq).
Hypothesis	NULL HYPOTHESIS: The given data follows Poisson distribution. ALTERNATE HYPOTHESIS: The given data does not follow Poisson distribution
Chi-square Test	Calculated chi square statistic = 7.92 p-value = 0.64 • p-value > α This is said to be almost coming from the distribution. So we can accept the null.
Degrees of freedom	$df = k - p - 1 = 11 - 1 - 1 = 9$. $k = 11 \Longrightarrow$ total number of classes.
Tabulated Chi-Square	$= 16.92$ • Tabulated value > Calculated value » We accept the null hypothesis
Business Cases	• suppose the given data is from a traffic signal in a city, where the number represents the number of times the signal was violated on a given day, i.e., on day #1 the signal was violated 5 times, on day #2 the signal was violated 4 times, and so on. This is a count information. • Another example would be that let's say everyday you're manufacturing a thousand products. Out of those thousand products, how many are defective. So, 5 could be the number of defective items on day 1, 4 could be on day #2, and so on. • **Wherever you have count information, those places could be ideal for Poisson distributions to happen.** So, a number of discrete events happening in continuous space is Poisson distribution. • Other examples could include customer arrivals and so on.

•