Problem 2

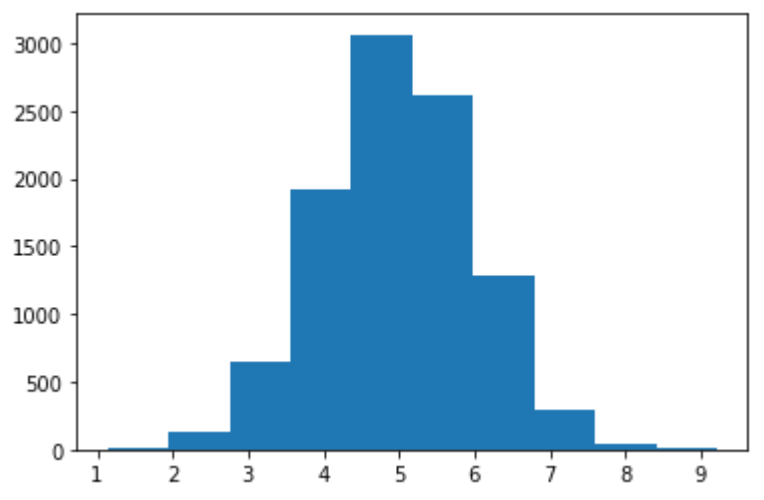
In this part of the homework, we will be inspecting different QQ plots that verify different sets of distributed data. In other words, we will be observing the nature of distribution through plotting.

In [23]:
```
# importing packages
from helper import getData

import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np
```

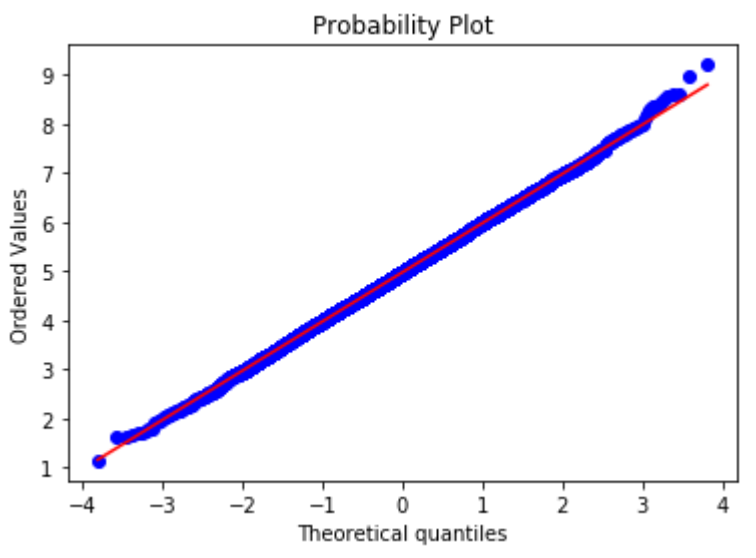
For DistA:

In [29]:
```
dataA = getData('distA.csv')
plt.hist(dataA)
plt.show()
```



The distribution of dataset A visualized by a histogram shows a symmetric histogram.

In [25]:
```
stats.probplot(dataA, dist = 'norm', plot=plt)
plt.show()
plt.clf()
```

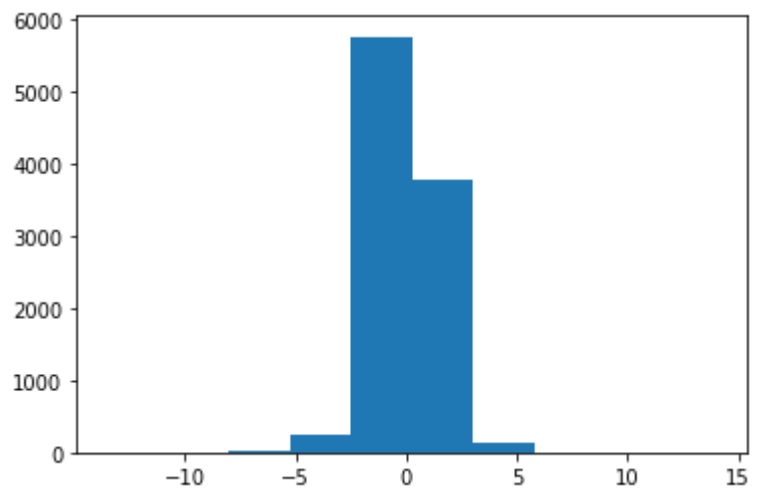



```
<Figure size 432x288 with 0 Axes>
```

Looking at the scatter plot and the trendline, we can interfere that the two sets of data plotted (ordered VS theoretical) are distributed the same. Therefore, the distribution that best matches the curve of the A dataset is probably normal distribution.

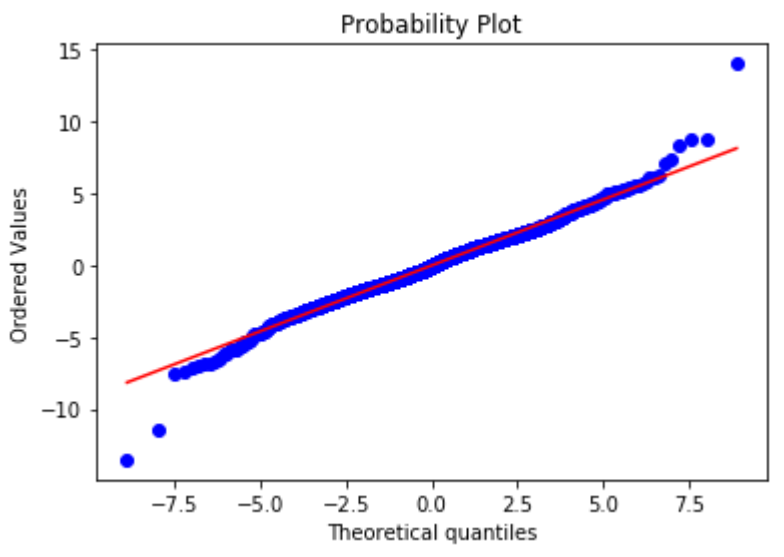
For DistB:

In [30]:
```
dataB = getData('distB.csv')
plt.hist(dataB)
plt.show()
```



The distribution of dataset A visualized by a histogram shows a two peak in the middle histogram or a havily tailed distribution.

In [27]:
```
stats.probplot(dataB, dist = 'laplace', plot=plt)
plt.show()
plt.clf()
```

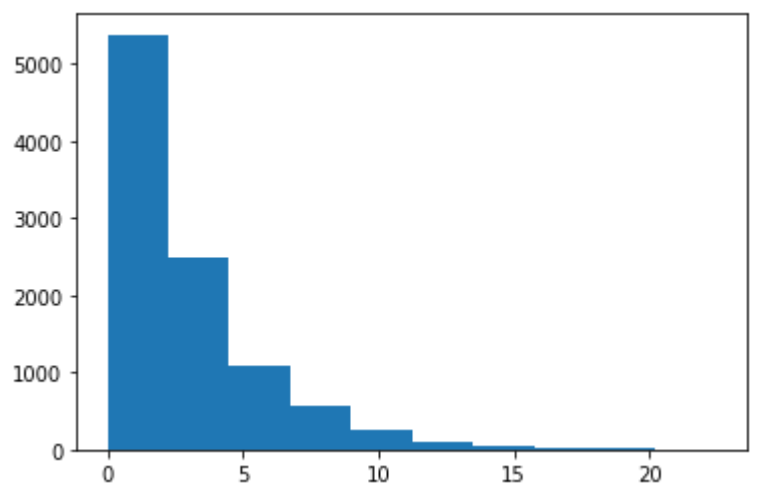



```
<Figure size 432x288 with 0 Axes>
```

Looking at the scatter plot and the trendline, we can interfere that the two sets of data (ordered VS theoretical) are differently distributed. Therefore, In order to linearize the graph and get the best matching curve, Laplace distribution was used as a double exponential to the ordered values.

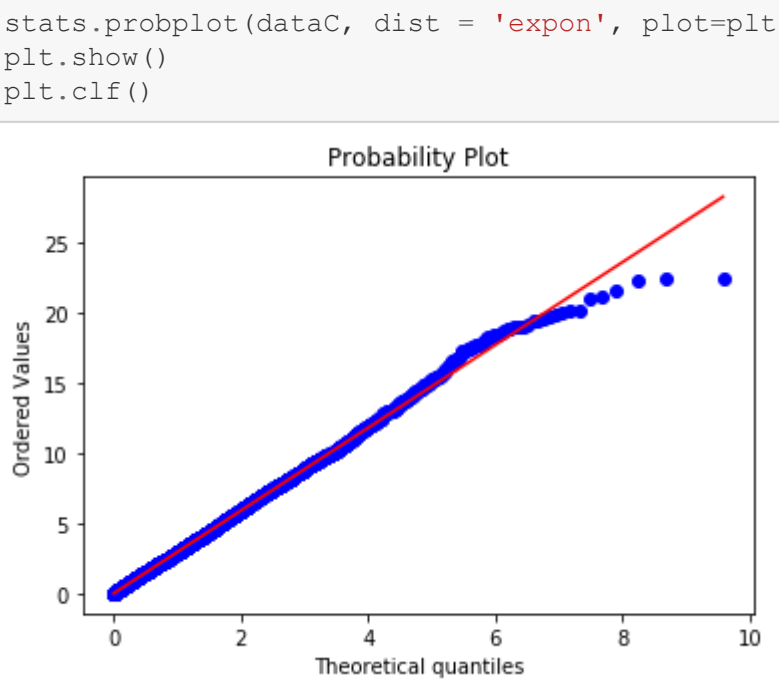
For DistC:

In [34]:
```
dataC = getData('distC.csv')
plt.hist(dataC)
plt.show()
```



The distribution of dataset A visualized by a histogram shows a right tail histogram or a positively skewed histogram.

In [35]:
```
stats.probplot(dataC, dist = 'expon', plot=plt)
plt.show()
plt.clf()
```




```
<Figure size 432x288 with 0 Axes>
```

Looking at the scatter plot and the trendline, we can interfere that the two sets of data (ordered VS theoretical) are differently distributed. Therefore, In order to linearize the graph and match the curves. exponential values were assigned to the ordered data, which resulted in an exponential distribution for dataset C.

Conclusion

QQ plot can be used to compare any two distribution and can be used to verify an unknown distribution by comparing it with a known distribution. Hence, the best matching curves in this problem were found through the trial and error process.

In [ ]:
```

```