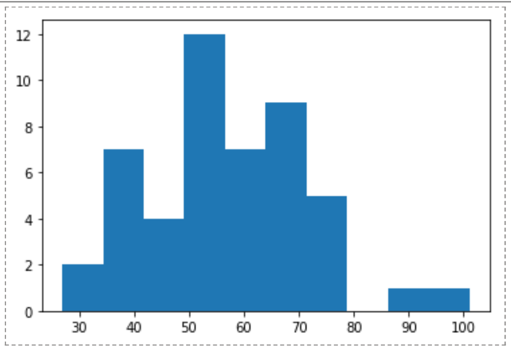
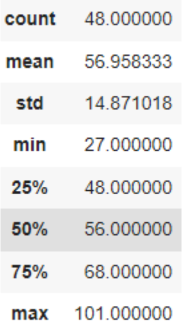
2.6 Guessing the Distribution of Dataset 3

Sunday, 02 October 2022 8:22

Normal Distribution



- Somewhat Gaussian in nature.
- Side bit is sitting between 90 to 100, so it's inclining towards something. It's telling us that there are significant number of observations that are towards one side. Meaning the distribution has a tail which is on the right side. So it could be a Gaussian with the right tail.



count	48.000000
mean	56.958333
std	14.871018
min	27.000000
25%	48.000000
50%	56.000000
75%	68.000000
max	101.000000

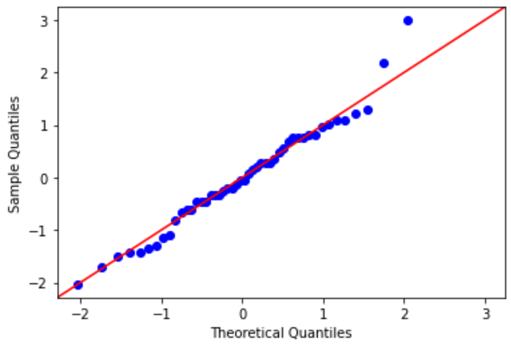
{('Mean Observed', 56.96), ('Variance Observed', 221.15), ('Skew Observed', 0.37), ('Kurt Observed', 0.6)}

- The difference between the minimum and the 25th quantile around 20.
- Between 25th and 50th is around 10.
- 50th to 75th is again around 10.
- 75th to the maximum is more than 30, that means it has a tail on the right side.

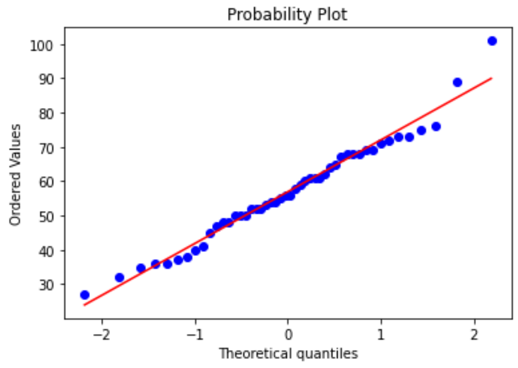
- Also, mean and the median are almost the same. Meaning, it's symmetric. So, huge symmetry with a slight tail.

- Positive skew, means that the distribution is right tailed.

Q-Q Plot



P-P Plot



- Both plots suggest that it's a normal distribution.

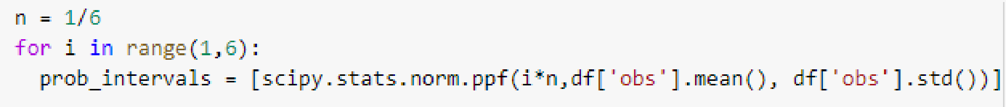
NULL HYPOTHESIS: The given data follows Normal distribution.

ALTERNATE HYPOTHESIS: The given data does not follow Normal distribution

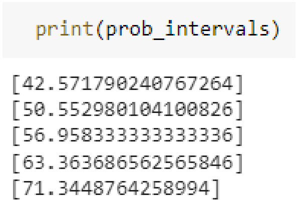
We split the data into 6 equal intervals (you can split it in as many intervals you want).

As the area under the curve is one, the area under each bucket will be $\frac{1}{6} = 0.167$.

This the command to do that:



```
n = 1/6
for i in range(1,6):
  prob_intervals = [scipy.stats.norm.ppf(i*n,df['obs'].mean(), df['obs'].std())]
```



```
  print(prob_intervals)

[42.571790240767264]
[50.552980104100826]
[56.958333333333336]
[63.363686562565846]
[71.3448764258994]
```

So the 6 buckets we have now are:

1st bucket : 42.57 and below
2nd bucket : 42.57 to 50.55
3rd bucket : 50.55 to 56.95
4th bucket : 56.95 to 63.36
5th bucket : 63.36 to 71.34
6th bucket : 71.34 and above

	Since, we're splitting into six buckets, and we had 48 observations, so the expected frequency under each bucket $\frac{48}{6} = 8$. • Expected freq = [8, 8, 8, 8, 8, 8] • Observed freq = [9, 7, 9, 7, 9, 7] • Observed frequencies can be obtained by first sorting the dataset in excel in ascending order, and then count the number of values fall in each buckets. ○ e.g., here, there are 9 values in the 1st bucket, i.e., 9 values in the dataset below 42.57
	Calculated chi square statistic = 0.75 p-value = 0.98 • p-value > α We accept the null.
	$df = k - p - 1 = 6 - 2 - 1 = 3$ $k = 6$: number of buckets $p = 2$: for normal distribution
	Tabulated Chi Square value $= 7.81$ Tabulated value $>$ Calculated value We accept the null.
Business Cases	• Most of our life events can be assumed to be normal. • e.g., Our test scores over the years could be a normal distribution. • Or test scores of an entire class will be normally distributed. • The data that we were working on comes from automotive sales. There were around 48 salesman who have sold tractors on a particular month. So, ideally you will see a couple of people who are performing extremely well, and there will be some salesman who may not be performing well, and in between you will have the majority.