Statistical Inference

Testing Assumption of Normality

Contents

Normality test

- An assessment of the normality of data is a prerequisite for many statistical tests because normal distribution is an underlying assumption in parametric testing.

- Normality can be assessed using two approaches: graphical and numerical.
  - Graphical approach
    - Box-Whisker plot (It is used to asses symmetry rather than normality.)
    - Quantile-Quantile plot (Q-Q plot).

  - Numerical (Statistical) approach
    - Shapiro-Wilk test (Used generally for **small sample**)
    - Kolmogorov-Smirnov test (Used generally for **large sample**)

* Box-Whisker plot is used to asses symmetry rather than normality. Hence, only Q-Q plot method is explained.

Case Study

To assess normality of data in Python, we shall consider the below case as an example.

Background

Data has 2 variables recorded for 80 guests in a large hotel.
Customer Satisfaction Index (csi) & Total Bill Amount in thousand Rs. (billamt)

Objective

To check if variables follow normal distribution

Sample Size

Sample size: 80
Variables: id, csi, billamt

Data Snapshot

Normality Testing Data

Variables

id	csi	billamt
1	38.35	34.85
2	47.02	10.99
3	36.96	24.73
4	43.07	7.9
5	38.77	9.38

ervations

Column	Description	Type	Measurement	Possible Values
id	Customer ID	Numeric	-	-
csi	Customer Satisfaction Index	Numeric	-	Positive value
billamt	Total Bill Amount in thousand Rs.	Numeric	Rs.	Positive value

Quantile-Quantile plot

- Very powerful graphical method of assessing Normality.
- Quantiles are calculated using sample data and plotted against expected quantiles under Normal distribution.
- If Normality assumption is valid then high correlation is expected between sample quantiles and expected(theoretical quantiles under normal distribution) quantiles.
- The Y axis plots the actual quantiles values based on sample. The X axis plots theoretical values.
- If the data is truly sampled from a Normal distribution, the QQ plot will be linear.

QQ Plot in Python For Variable csi

```
#Import Data
```

```
import pandas as pd
data=pd.read_csv('Normality Testing Data.csv')
```

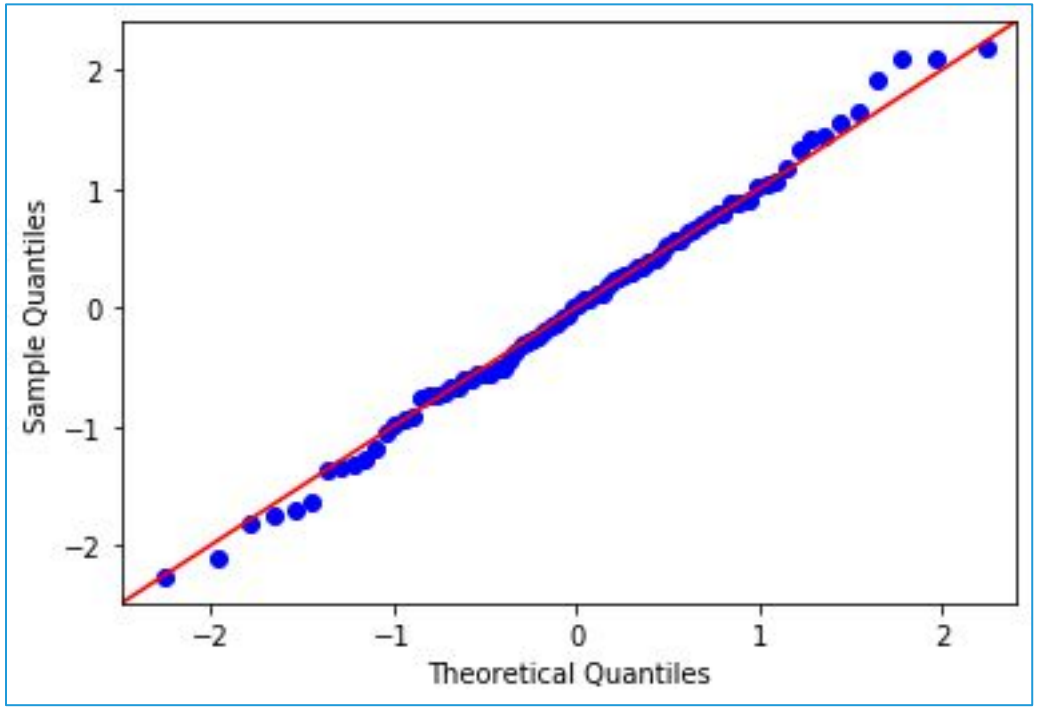
```
#QQ Plot
```

```
import statsmodels.api as sm
sm.graphics.qqplot(data.csi, line='45', fit=True)
```

- **qqplot()** produces a plot with theoretical quantiles on x axis against the sample quantiles on y axis. Column for which normality is being tested is specified in the first argument.
- **line=** is an argument that adds reference line to the qqplot. Here it adds a 45-degree line
- **fit=True** indicates, parameters are fit using the distribution's fit() method

QQ Plot in Python For Variable csi

```
# Output:
```



Interpretation :

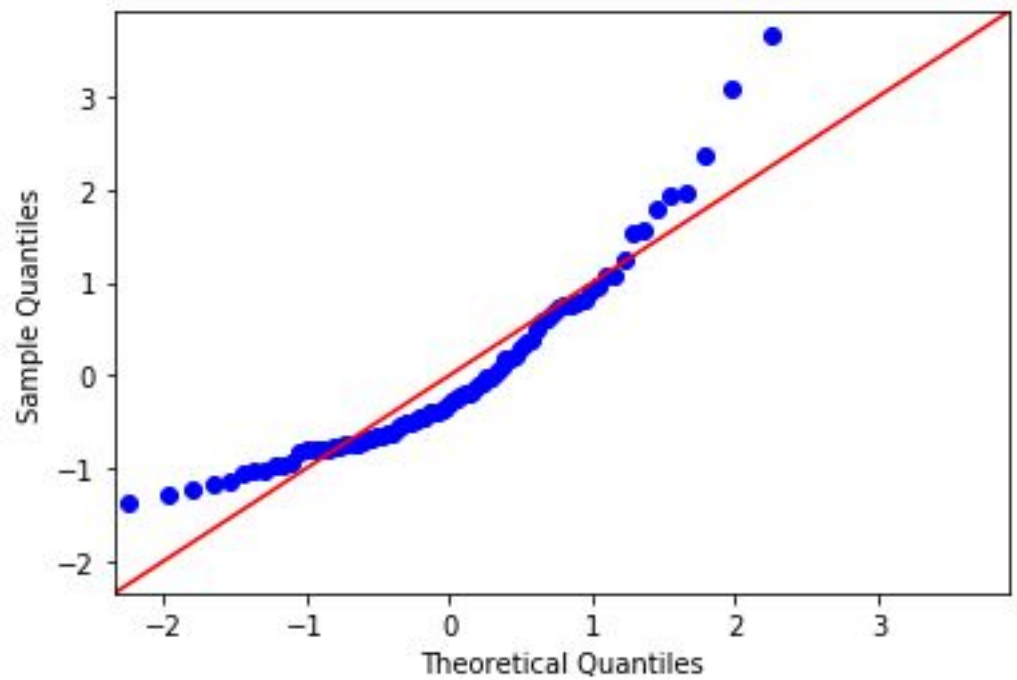
- Q-Q plot is Linear. Distribution of ‘csi’ can be assumed to be normal.

Q-Q plot in Python For Variable billamt

```
# Q-Q plot for the variable billamt
sm.graphics.qqplot(data.billamt, line='45', fit=True)
```

- **data.billamt** is the variable for which normality is to be checked.

```
# Output:
```



Interpretation :

- Q-Q plot is deviated from linearity. Distribution of ‘billamt’ appears to be non-normal.

Shapiro-Wilk test

Shapiro-Wilk test is widely used statistical test for assessing **Normality**.

Objective	To test the **normality** of the data.

Null Hypothesis (H_0): **Sample is drawn from Normal Population**

Alternate Hypothesis (H_1): **Sample is drawn from Non-Normal Population**

The test is performed for the variables, 'csi' and 'billamt' separately.

Test Statistic	It correlates sample ordered values with expected Normal scores. (actual calculation is very complex so we will avoid details)
Decision Criteria	Reject the null hypothesis if p-value < 0.05

* Here we continue to use previous data of CSI and Bill Amount for Shapiro-Wilk test.

Shapiro Wilk Test For Variable csi

```
# Shapiro Wilk Test
import scipy as sp
sp.stats.shapiro(data.csi)

# Output
(0.9919633269309998, 0.9037835597991943)
```

shapiro() from scipy package, returns correlation coefficient w and p-value.

Interpretation :

- Since p-value is >0.05,do not reject H0. Distribution of ‘csi’ can be assumed to be normal.

Shapiro-Wilk test For Variable bilamt

```
# Shapiro Wilk test for the variable billamt

sp.stats.shapiro(data.billamt)
```

- **data.billamt** is the variable for which normality is to be checked.

```
# Output:

(0.8903077244758606, 4.858443844568683e-06)
```

Interpretation :

- Since p-value is <0.05, reject H0. Distribution of 'billamt' appears to be non-normal.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov test is another widely used statistical test for assessing Normality.

Objective	To test the **normality** of the data.

Null Hypothesis (H_0): Sample is drawn from Normal Population

Alternate Hypothesis (H_1): Sample is drawn from Non-Normal Population

The test is performed for the variables, ‘csi’ and ‘billamt’ separately.

Test Statistic	Kolmogorov-Smirnov Test: It compares empirical (sample) cumulative distribution function (CDF) with Normal distribution CDF. The test statistic is maximum difference between CDF's.
Decision Criteria	Reject the null hypothesis if p-value < 0.05

* Here we continue to use previous data of CSI and Bill Amount for Shapiro-Wilk test.

Kolmogorov-Smirnov test in Python

```
# Kolmogorov Smirnov test
sm.stats.diagnostic.lilliefors(data.csi)
```

- Instead **of lilliefors, kstest_normal()** from **statsmodels** can also be used to perform a Lilliefors (KS) Normality Test.
- Both tests returns Kolmogorov-Smirnov test statistic and p-value.
- **data.csi** is the variable for which normality is to

```
# Output:
(0.04238708824708459, 0.9859314950919987)
```

Interpretation :

- Since p-value is >0.05,do not reject H0. Distribution of ‘csi’ can be assumed to be normal.

Kolmogorov-Smirnov test in Python

```
# Kolmogorov Smirnov test for the variable billamt
sm.stats.diagnostic.lilliefors(data.billamt)
```

- **data.billamt** is the variable for which normality is to be checked.

```
# Output:
(0.1424429511673755, 0.00099999999999998899)
```

Interpretation :

- Since p-value is <0.05, reject H0. Distribution of 'billamt' appears to be non-normal.