Statistical Inference

Parametric Tests - I

Contents

Normality test

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

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

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

* A box-whisker plot is used to asses symmetry rather than normality. Hence, only the Q-Q plot method is explained.

Case Study - 1

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 the normality of the data

Sample Size

Sample size: 80
Variables: id, csi, billamt

Data Snapshot

Variables

Normality Testing Data

Observations

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
6	63.04	9.49
7	43.17	19.58
8	35.14	6.15
9	38.33	13.29
10	38.7	9.62
11	31.44	8.51

Column	Description	Type	Measurement	Possible Values
id	Customer ID	Numeric		
csi	Customer Satisfaction Index	Numeric		Positive value
billamt	Total Bill Amount in thousand euros.	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 distributions.
- If the normality assumption is valid then, a high correlation is expected between the sample quantiles and the expected (theoretical quantiles under normal distribution) quantiles.
- The Y axis plots the actual quantile values based on the sample. The X axis plots theoretical values.
- If the data is truly sampled from a normal distribution, the QQ plot will be linear.

Q-Q plot using R

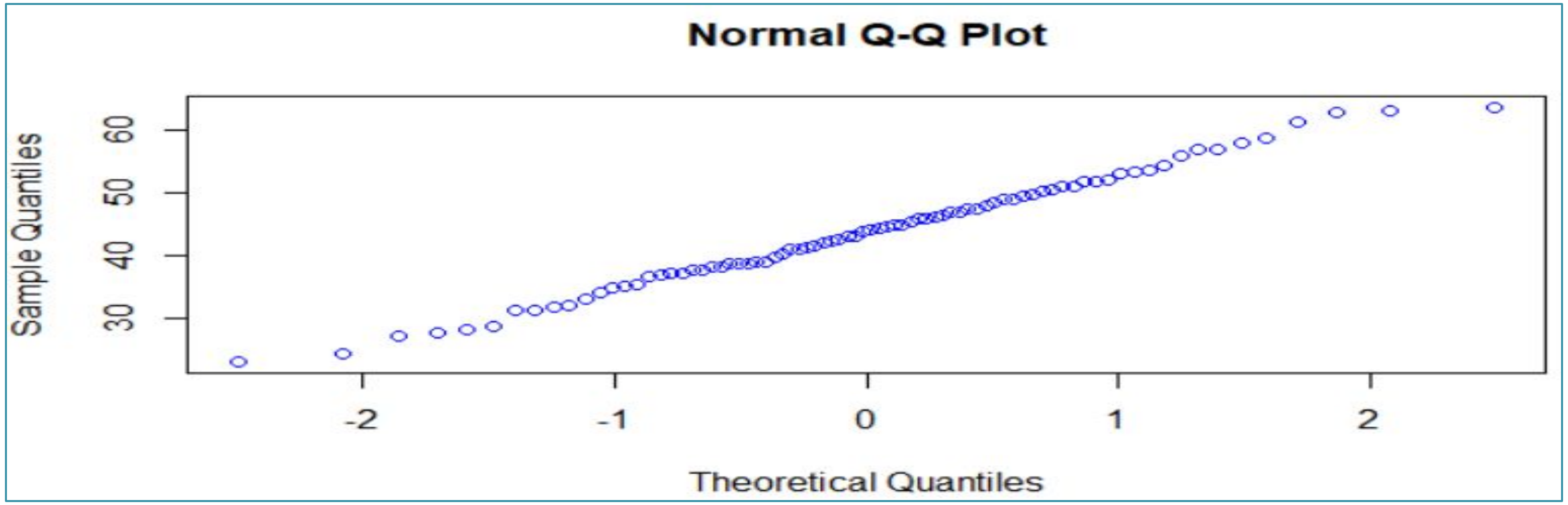
```
# Import data
data<-read.csv("Normality Testing Data.csv", header=TRUE)

# Q-Q plot for the variable csi
qqnorm(data$csi,col="blue")
```

- *data$csi is the variable for which normality is to be checked.*
- *Col=blue specifies the line color on graph.*

Q-Q plot using R

```
# Output:
```



Interpretation :

- *Q-Q plot is Linear. Distribution of 'csi' can be assumed to be normal.*

Q-Q plot using R

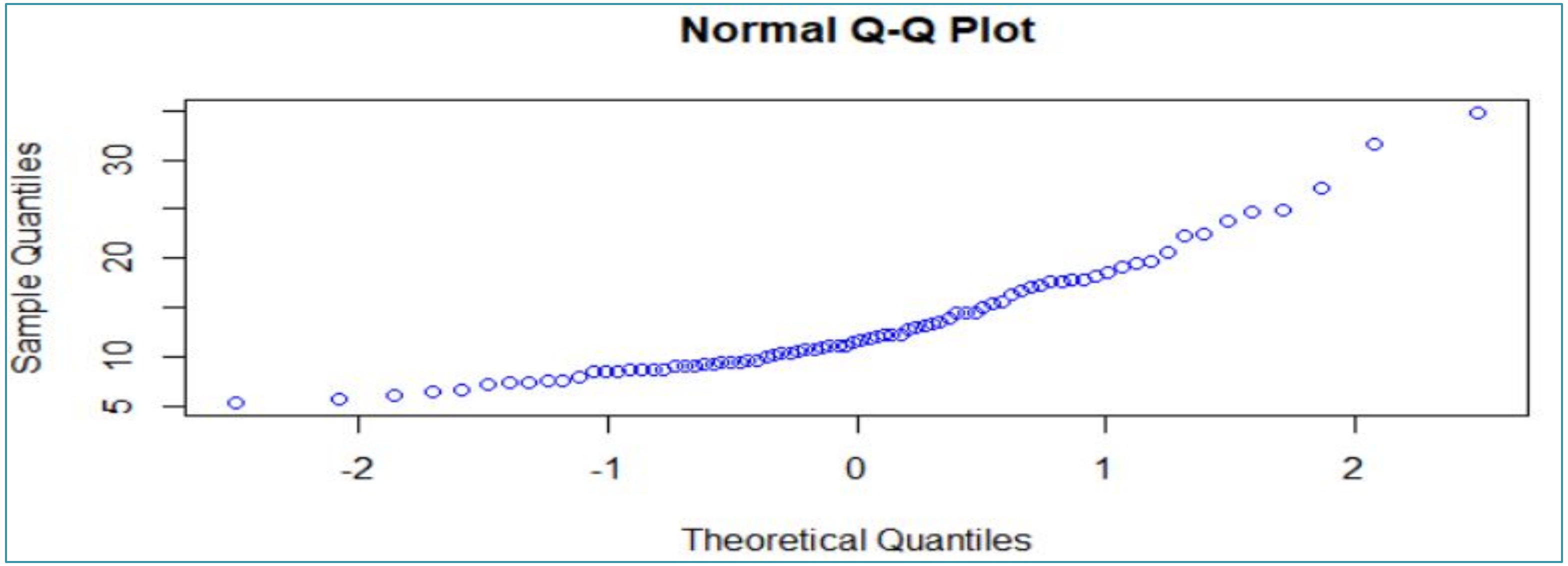
```
# Q-Q plot for the variable billamt
```

```
qqnorm(data$billamt,col="blue")
```

- *data$billamt is the variable for which normality is to be checked.*
- *Col=blue specifies the line color on graph.*

Q-Q plot using R

```
# Output:
```



Interpretation :

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

Shapiro-Wilk test

The 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. (the 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 in R

```
# Shapiro Wilk test for the variable csi

shapiro.test(data$csi)
```

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

```
# Output:

        Shapiro-Wilk normality test

data:  data$csi
W = 0.99196, p-value = 0.9038
```

Interpretation :

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

Shapiro-Wilks test in R

```
# Shapiro Wilks test for the variable billamt
```

```
shapiro.test(data$billamt)
```

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

```
# Output:
        Shapiro-Wilk normality test

data:  data$billamt
W = 0.89031, p-value = 4.858e-06
```

Interpretation :

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

Kolmogorov-Smirnov test

The 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 the 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 R

```
# Install and use package ‘nortest’
```

```
install.packages("nortest")

library(nortest)
```

- *Package nortest contains the Kolmogorov smirnov test.*

```
# Kolmogorov Smirnov test
```

```
lillie.test(data$csi)
```

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

Kolmogorov-Smirnov test in R

```
# Output:
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$csi
D = 0.042387, p-value = 0.9764
```

Interpretation :

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

Kolmogorov-Smirnov test in R

```
# Kolmogorov Smirnov test for the variable billamt
```

```
lillie.test(data$billamt)
```

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

```
# Output:
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$billamt
D = 0.14244, p-value = 0.0003753
```

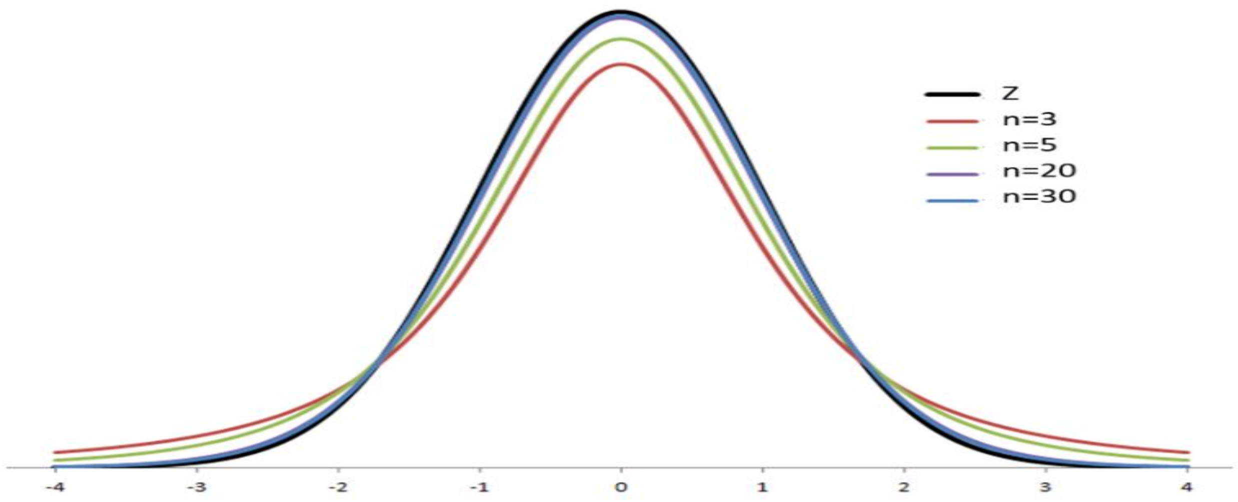
Interpretation :

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

t-distribution

- The t distribution is symmetric and its overall shape resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider.

- As the sample size increases so as the number of degrees of freedom grows, the t-distribution approaches the normal distribution with mean 0 and variance 1.



- In the above graph, z is a normal distribution with mean 0 and variance 1.

A note on Degrees of Freedom (DF)

- Degrees of freedom (df) is defined as the number of independent terms.
- "Sum of the squared deviations about mean of n values" has n-1 degrees of freedom. Knowing n-1 values, we can find last value since sum of deviations about mean is always zero.
- Sampling distributions like t, F and chi square have shapes based on degrees of freedom.
- Example , Give 5 numbers such that sum is 20. You can use 4 numbers freely but the fifth number should be such that sum is 20. Here df =4

One sample t-test

- The one sample t test is used to test a hypothesis about a single population mean.

- We use the one-sample t-test when we collect data on a single sample drawn from a defined population.

- For this design, we have one group of subjects, collect data on these subjects and compare a sample statistic to the hypothesized value of a population parameter.

- Subjects in the study can be patients, customers, retail stores etc.

Case Study - 2

Background

A large company is concerned about time taken by employees to complete the weekly MIS report.

Objective

To check if the average time taken to complete the MIS report is more than 90 minutes

Sample Size

Sample size: 12
Variables: Time

Data Snapshot

ONE SAMPLE t TEST

Variables

Observations

Time
85
95
105
85
90
97
104
95
88
90
94
95

Columns	Description	Type	Measurement	Possible values
Time	Time taken to complete MIS	Numeric	Minutes	Positive Values

Assumptions for one sample t-test

The assumptions of the one-sample t-test are listed below:

- Random sampling from a defined population
  (employees are selected at random from the company)

- The population is normally distributed
  (Time taken to complete MIS report should be normally distributed).

- The variable under study should be continuous.

A normality test can be performed by any of the methods explained earlier.

The validity of the test is not seriously affected by moderate deviations from 'normality' assumption.

One sample t-test

Testing whether mean is equal to a test value.

Objective	To test the average time taken to complete MIS is more than 90 minutes

Null Hypothesis (H_0): $\mu = 90$

Alternate Hypothesis (H_1): $\mu > 90$

Test Statistic	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	12
Mean		93.5833
Standard Deviation	S	6.4731
Standard Error	s/√n	1.8686
Difference		93.5833-90=3.5833
t	$\frac{\bar{x} - \mu_0}{S.E}$	1.9176

One sample t-test in R

```
# Import data
data<-read.csv("ONE SAMPLE t TEST.csv",header=TRUE)

# t-test for one sample
t.test(data$time, alternative="greater", mu=90)
```

- *data$time is the variable under study.*
- *alternative="greater" ,Since under H1, value is tested for greater than 90.*
- *mu=90 is the value to be tested.*

* Before performing t test, normality test is done to ensure time variable is normally distributed.

One sample t-test in R

Output:

```
        One Sample t-test

data:  data$time
t = 1.9176, df = 11, p-value = 0.04074
alternative hypothesis: true mean is greater than 90
95 percent confidence interval:
 90.22748      Inf
sample estimates:
mean of x
 93.58333
```

Interpretation :

- *Since the p-value is <0.05, reject H0. The average time taken to complete the MIS report is more than 90 minutes ’.*

Quick Recap

Normality Test

- Normal data is an underlying assumption in parametric testing.
- Two approaches to test normality:
- Graphical (Box-Whisker plot,Quantile-Quantile plot
- Statistical (Shapiro-Wilks test,Kolmogorov-Smirnov test

One sample t test

- Used to test the hypothesis about a single population mean.
- H0: $\mu = \mu 0$