1.1.9 Frequency Polygon

■ Frequency polygon depicts the shape and trends of data. It can be drawn with or without a histogram.

▶ Joining the upper mid points of a histogram leads to a frequency polygon.

Example 1.13. *Suppose you have a data on no. of peas and frequency from* 198 *pea pods. Using R draw frequency polygon based on the data given below.*

Discrete Data	
No. of Peas	Frequency
1	4
2	33
3	76
4	50
5	26
6	8
7	1

Table 1.6: Pea pods data. [Source: Fundamental Vol. 1]

▶ *R Code*:

```
a=c(0,1,2,3,4,5,6,7,8)
b=c(0,4,33,76,50,26,8,1,0)
plot(a,b,type="o",xlab="Number of Peas",ylab="Frequency")
```

▶ *R Plot*: *See the plot Fig. 1.11.*

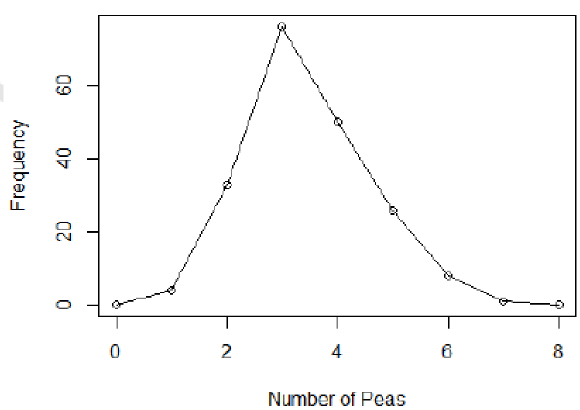

Figure 1.11: Frequency polygon with the data given in Table 1.6.

[Do It Yourself] 1.9. *Suppose you have a data on no. of peas and frequency from* 198 *pea pods. Using R draw frequency polygon based on the data given below.*

Continuous Data	
Score	Frequency
40-48	5
48-56	10
56-64	30
64-72	40
72-80	16
80-88	9
88-96	4

Table 1.7: Score in a test vs. Frequency data

[*Hint* : *Use mid value of Score*]

[Do It Yourself] 1.10. *Create a discrete frequency data of length* 10 *in your own and draw a frequency polygon.*

1.1.10 Pie Chart

Example 1.14. *Suppose you have a data on no. of peas and frequency from* 198 *pea pods. Using R draw frequency polygon based on the data given below.*

Discrete Data	
States	Procurement of Rice
Punjab	5486
Haryana	1248
Tamil Nadu	589
Andhra Pradesh	3987
Uttar Pradesh	1296
Others	1654

Table 1.8: Rice procurement of different States data. [Source: Fundamental Vol. 1]

▶ *R Code*:

```
states=c("Punjab","Haryana","Tamil Nadu","Andhra Pradesh","Uttar Pradesh","Others")
procurement=c(5486,1248,589,3987,1296,1654)
rice.proc=data.frame(states,procurement)
rice.proc=transform(rice.proc,percentage=(rice.proc$procurement/14260)*100)
rice.proc
```

```
attach(rice.proc)
pie(procurement,labels=c("Punjab(38.47%)","Haryana(8.75%)","Tamil Nadu(4.13%)",
"Andhra Pradesh(27.96%)","Uttar Pradesh(9.09%)","Others(11.60%)"),
main="Procurement of Rice('000 tonnes)\nduring Oct. 1993 to Sept. 1994",
col=c("royalblue4","orchid","chocolate4","olivedrab3","orange3","purple1","violetred4"))
detach(rice.proc)
```

▶ R Plot: *See the plot Fig. 1.12.*

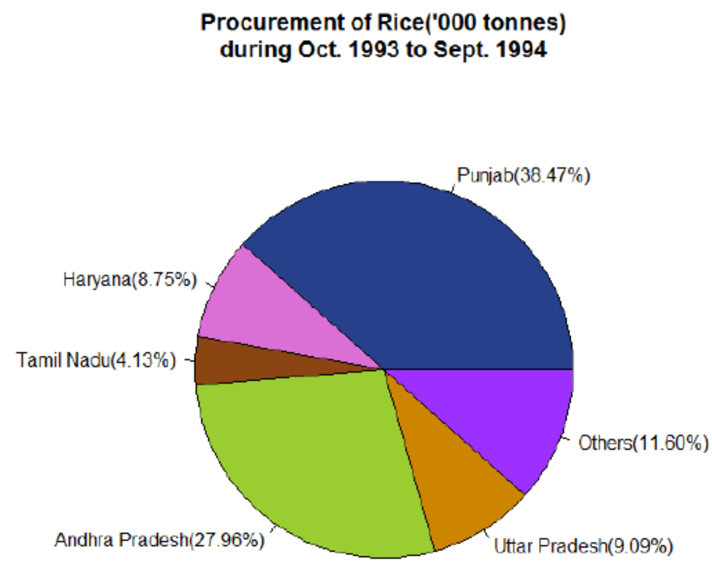

Figure 1.12: Pie Diagram.

[Do It Yourself] 1.11. *Construct a pie chart to visually display the favorite fruits of the students in a class based on the given data: Mango - 45, Orange - 30, Plum - 15, Pineapple - 30, Melon - 30.*

1.1.11 Plot Ogives

Example 1.15. *Suppose you have a data (see Table 1.3) on heights (cm) and frequency of students in a class. Using R draw Ogives (Cummulative frequency diagram).*

▶ R Code:

```
Height=c(149.55,154.55,159.55,164.55,169.55,174.55,179.55,184.55)
a = c(1,3,24,58,60,27,2,2)
# Less than type
b=cumsum(a)
plot(Height,b,type="o",ylab="Cummulative Frequency",col="blue",xlim = c(144,185))
a1=rev(a)
b1=cumsum(a1)
# Greater than type
b2=rev(b1)
lines(Height-5,b2,type="o",col="red")
```

▶ R Plot: *See the plot Fig. 1.13.*

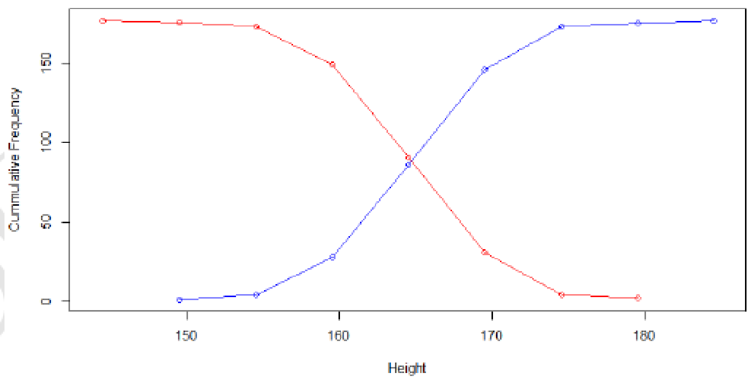

Figure 1.13: Cummulative frequency diagram or Ogive. Here blue and red curve denotes less than and greater than type respectively.

1.1.12 Q-Q Plot

■ In Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line $y = x$.

▶ Suppose we want to check for 10 data points $(y_1, \cdots, y_{10})$ follows normal distribution or not. So first we will standardized and sort the data in increasing order and these will act as data quantiles. Now we will divide the standard normal curve into 11 equal areas with 1/11 sq. unit each. After that we will find 10 x-axis values (i.e. z score) from standard normal distribution. Here $(-\infty < x_1 < x_2 < \cdots < x_{10} < \infty)$ and the area under the curve from $(-\infty, x_1) = \frac{1}{11}$, $(-\infty, x_2) = \frac{2}{11}$ and so on. Using 'qnorm()' function in R, we can find the theoretical quantile values $x_1, \cdots, x_{10}$ e.g. $x_1 = qnorm(1/11)$, $x_2 = qnorm(2/11), \cdots$. Now we can obtain Q-Q plot by drawing theoretical quantiles (x_i) in x-axis and data quantiles (y_i) in y-axis. If they lie on a straight line then the data points follows normal distribution.

▶ Using the above idea we can check for other distributions also.

Example 1.16. *Suppose you have a dataset: 5.81 4.00 5.22 14.02 4.84 12.53 2.25 4.32 5.01 7.04 5.02 6.61 1.96 2.17 0.86 6.51 9.69 4.07 1.72 11.48 2.09 8.40 6.60 8.21 1.68 3.92 1.62 11.52 0.72 6.26 1.84 15.55 3.63 2.98 8.26 3.44 0.62 2.31 2.26 8.65 4.51 6.92 0.69 3.50 0.30 5.10 9.90 1.96 1.10 9.01. Using R check if the data follow i) Standard normal distribution, ii) Chi squared distribution with df* 5.

▶ *R Code*:

```
a=c(5.81,4.00,5.22,14.02,4.84,12.53,2.25,4.32,5.01,7.04,5.02,6.61,1.96,2.17,0.86,
  6.51,9.69,4.07,1.72,11.48,2.09,8.40,6.60,8.21,1.68,3.92,1.62,11.52,0.72,6.26,
  1.84,15.55,3.63,2.98,8.26,3.44,0.62,2.31,2.26,8.65,4.51,6.92,0.69,3.50,0.30,5.10,
  9.90,1.96,1.10,9.01)
library(car)
qqPlot(a,"norm",mean=0,sd=1,main="Normal Q-Q Plot",xlab="Theoretical Quantiles",
ylab="Sample Quantiles",col="dark green",envelope=F)
qqPlot(res,"chisq",df=5,main="Chi-squared Q-Q Plot",xlab="Theoretical Quantiles",
ylab="Sample Quantiles",col="dark green",envelope=F)
```

▶ R Plot: *See the plot Fig. 1.14.*

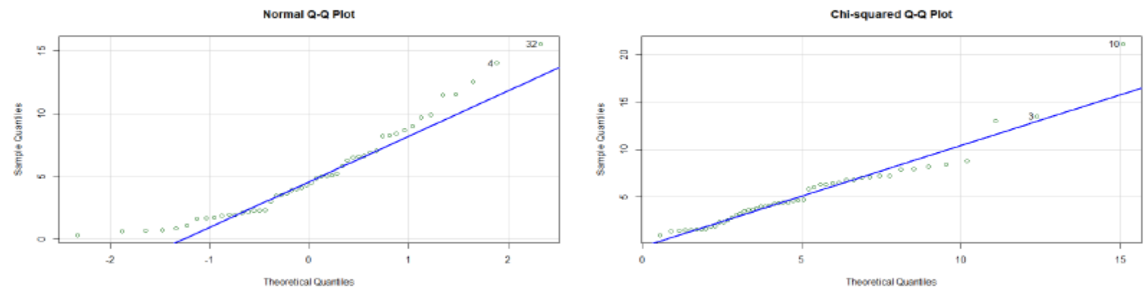

Figure 1.14: QQ plot: a) For standard normal distribution b) For χ^2_5 distribution.

[Do It Yourself] **1.13.** *Suppose you have a dataset: 1.46 2.14 1.70 1.95 2.13 1.03 1.95 0.93 2.43 4.02 2.48 3.15 2.70 0.50 1.13 0.39 1.29 2.56 1.50 2.75 2.18 1.12 2.77 2.22 2.57 -0.27 0.62 4.26 3.53 2.57 1.64 1.54 2.03 1.52 2.47 2.59 2.41 2.41 1.24 1.39 0.55 2.39 2.20 1.43 0.56 0.81 2.50 3.46 3.48 3.83 1.91 1.59 2.84 0.83 1.79 1.39 1.18 0.49 1.60 3.57 1.62 2.80 -0.14 0.66 1.79 0.70 1.61 1.76 2.68 2.10 0.32 0.49 2.00 1.41 0.24 3.78 1.67 2.56 1.55 0.96 2.60 1.94 2.06 1.66 0.61 1.28 2.97 1.26 1.41 1.59 1.57 1.48 2.49 4.22 3.27 1.01 0.68 2.50 1.21 2.46. Using R check if the data follow i) N(0,1), ii) N(2,1) iii)* χ^2_1 *iv)* $t_{(2)}$.

1.1.13 Graphical Summaries of Data

■ Advantages:

▶ Acceptability: Such a report is acceptable to busy persons because it easily highlights the theme of the report. This helps to avoid wastage of time.

▶ Comparative Analysis: Information can be compared in terms of graphical representation. Such comparative analysis helps for quick understanding and attention.

▶ Less cost: Information if descriptive involves huge time to present properly. It involves more money to print the information but the graphical presentation can be made in a short but catchy view to make the report understandable. It obviously involves less cost.

▶ Decision Making: Business executives can view the graphs at a glance and can make a decision very quickly which is hardly possible through descriptive reports.

▶ Helpful for less literate Audience: Less literate or illiterate people can understand graphical representation easily because it does not involve going through line by line of any descriptive report.

▶ Less Effort and Time: To present any table, design, image, or graphs require less effort and time. Furthermore, such a presentation makes a quick understanding of the information.

■ Limitations:

▶ Dimensions: As the graphs are 2 dimensional (sometimes extend to 3 dimension) so for a high dimensional data it is very difficult (or sometimes impossible) to visualize through graphs.

▶ Lack of Exactness: We can visualize the data pattern through graph but it's subjective and not giving the exact measure. For example, from a scatter plot we can observe any correlation pattern but not exactly.

▶ Problems to select a suitable method: Information can be presented through various graphical methods and ways. Which should be the suitable method is very hard to select.