Chapter 2

Descriptive Statistics & Regression

2.1.1 Descriptive Statistics

■ Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population.
■ Descriptive statistics mainly divided into four parts:

1. $\boxed{\text{Measure of Central Tendency}}$: A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. Measures are: $Mean = \frac{\sum x_i f_i}{\sum f_i}$, $Median = \ Middle\ Most\ Value\ (Sort)$, $Mode = \ Highest\ Frequency\ Value$.

2. $\boxed{\text{Measure of Dispersion}}$: A measure of dispersion is a single value that attempts to describe a set of data by identifying the spread of the data. It often describe the spread of data around a central value. Measures are: $Range = \ Max - Min$, $Variance = \frac{1}{n}\sum(x_i - \overline{x})^2$, $Standard\ Deviation = \ \sqrt{Variance}$, $Mean\ deviation = \frac{1}{n}\sum|x_i - A|$, $Inter\ Quartile\ Range = Q_3(75\%\ values\ are\ below\ the\ limit) - Q_1(25\%\ values\ are\ below\ the\ limit)$, $Coefficient\ of\ Variation = \frac{sd}{mean} \times 100\%$.

3. $\boxed{\text{Measure of Skewness}}$: A measure of skewness is a single value that attempts to describe a set of data by identifying the asymmetry of the data. It often describe the degree of departure from the symmetry. Measures are: Moment Coefficient of Skewness $g_1 = \frac{m_3}{m_2^{3/2}}$, $Pearson's\ 1^{st}\ measure = \frac{mean - mode}{sd}$, $Pearson's\ 2^{nd}\ measure = \frac{3(mean - median)}{sd}$, $Bowley's\ measure = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$. Here measures $(0, >, <)$ means data are $(symmetric, +ve\ skew, -ve\ skew)$ respectively. In practice, usually measures lies in $(-0.8, 0.8)$ considered as symmetric.

4. $\boxed{\text{Measure of Kurtosis}}$: A measure of kurtosis is a single value that attempts to describe a set of data by identifying the peakdness of the data. In other words, it defines how heavily the tails of a distribution differ from the tails of a normal distribution. Measure: Moment Coefficient of Kurtosis $g_2 = \frac{m_4}{m_2^2} - 3$. Here g_2 $(0, >, <)$ means data are $(meso\ kurtic,\ lepto\ kurtic,\ platy\ kurtic)$ respectively.

■ Raw Moments: $m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$, Central Moments: $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$.

Example 2.1. *Find the mean, median and mode for the data* (x)*: 45 26 19 49 43 with frequencies* (f)*: 2 4 3 5 4.*

▶ *R Code*:

```
x=c(45,26,19,49,43)
f=c(2,4,3,5,4)
a=rep(x,f)
a
mean(a) # Mean
median(a) # Median
x[which(f==max(f))] # Mode
```

[Do It Yourself] 2.1. *Find the mean, median and mode for the data* (x)*: 14 26 39 86 60 77 36 59 84 39 33 54 70 74 26 with frequencies* (f)*: 9 5 7 2 7 4 4 5 8 3 5 9 5 4 9.*

Example 2.2. *Find the range, variance, sd, mean deviation about 40, mean deviation about mean,* Q_1, Q_3*, Inter Quartile Range and Coefficient of Variation for the data* (x)*: 70 48 84 21 84 33 64 36 61 50 27 87 41 65 33.*

▶ *R Code*:

```
x=c(70,48,84,21,84,33,64,36,61,50,27,87,41,65,33)
max(x)-min(x) # Range
var(x) # Variance
sd(x) # Standard Deviation
sum(abs(x-40))/length(x) # Mean deviation about 50
sum(abs(x-mean(x)))/length(x) # Mean deviation about mean
quantile(x) # Quartiles
IQR=67.5-64.5 # Using Quartile Data
(sd(x)/mean(x))*100 # Coefficient of Variation
```

[Do It Yourself] 2.2. *Find the range, variance, sd, mean deviation about 55, mean deviation about mean,* Q_1, Q_3*, Inter Quartile Range and Coefficient of Variation for the data* (x)*: 16 40 12 46 47 64 50 64 56 37 82 78 27 13 36 50 51 75 70 87.*

Example 2.3. *Find the raw and central moments upto order* 4 *for the data* (x)*: 60 39 75 33 69 47 69 32 59 52 49 63.*

▶ *R Code*:

```
x=c(60,39,75,33,69,47,69,32,59,52,49,63)
r=3 # r=1,2,3,4
sum(x^r)/length(x) # Raw Moments
sum((x-mean(x))^r)/length(x) # Central Moments
```

[Do It Yourself] 2.3. *Find the raw and central moments upto order* 5 *for the data* (x)*: 26 34 38 14 24 37 32 26 11 34 18 38 38 21 20 29 39 13 12 24.*

2.1.2 Correlation

■ The basic idea of correlation between two variables x, y is to check if there is a linear dependencies among them or, not.
▶ The measure of correlation i.e. Pearson's correlation coefficient is denoted by r_{xy} and defined by $\boxed{r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)\ Var(y)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\ \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})}}}$.
▶ Here $-1 \leq r_{xy} \leq 1$, if r_{xy} towards $1(-1)$ then the variables are high positively (negatively) correlated. If $r_{xy} = 0$ then there is no correlation.

▶ The variance-covariance matrix defined as: $\begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{pmatrix}$.

▶ The correlation matrix defined as: $\begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix}$.

▶ The covariance is symmetric i.e. $Cov(x,y) = Cov(y,x)$. Furthermore, $r_{xy} = r_{yx}$.
■ Scatterplot: Suppose there are n pairs of values (x_i, y_i), $i = 1(1)n$. Then draw each point (x_i, y_i) in a $2-dimensional$ plane XY leads to a scatterplot.

Example 2.5. *Suppose we have a bivariate data of* 20 *students, Marks in School* (x)*: 148 134 131 146 135 159 136 161 166 165 160 136 176 145 144 153 158 145 126 170 and Marks in College* (y)*: 423 405 362 369 333 417 301 425 372 438 415 393 349 306 380 338 450 326 381 359. Plot a scatter diagram, correlation coefficient between* x, y *and correlation matrix.*

▶ *R Code*:

```
x=c(126,157,153,156,152,135,145,132,153,143,165,132,180,161,170,176,165,163,157,180)
y=c(292,367,361,366,343,326,331,307,343,322,363,296,412,377,394,384,370,365,349,403)
z=data.frame(x,y)
colnames(z)=c("Marks in School","Marks in College")
plot(z,col="blue") # Scatter Plot
cor(x,y) # Correlation Coefficient
cor(z) # Correlation Matrix
```

2.1.3 Regression

■ If two variables x, y are linearly correlated. Now we take x is independent and y is dependent variable. Then for an unknown data point x_j we can predict the value of y_j.
▶ The regression model is defined as $y = a + bx + \epsilon$. Here a, b are unknown parameters will be estimated from the data and ϵ is error with 0 mean and constant variance.
▶ There are some assumptions in the regression model and the parameters are usually estimated through least square or, maximum likelihood method.
▶ The fitted line is $\boxed{E(y) = \hat{a} + \hat{b}x}$. Here $\boxed{\hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}}$ and $\hat{a} = \overline{y} - \hat{b}\overline{x}$.

Example 2.6. *Fit a regression line based on the data given in Example 2.5. Fit a regression line* Y *on* X *and discuss the summary statistics. Draw the regression line over the scatter plot. Also draw confidence and prediction bands.*

▶ R Code:

```
x=c(126,157,153,156,152,135,145,132,153,143,165,132,180,161,170,176,165,163,157,180)
y=c(292,367,361,366,343,326,331,307,343,322,363,296,412,377,394,384,370,365,349,403)
fit=lm(y~x) # y on x
summary(fit)
plot(x,y,xlab = ’Marks in School’, ylab=’Marks in  College’,col=’blue’,cex=1.2,
pch=16,main=’Regression Fit’)
abline(fit,lwd=2,col=’green’)
# For confidence and Prediction Bands
x1 = data.frame(x=seq(min(x),max(x),0.5))
y1=predict(fit,interval="conf",newdata=x1)
y2=predict(fit,interval="pred",newdata=x1)
matlines(x1$x,y1,lty=c(1,2,2),col=c("green","red","red"),lwd=c(2,1,1))
matlines(x1$x,y2,lty=c(1,3,3),col=c("green","black","black"),lwd=c(2,1,1))
```

Example 2.7. *Fit a 2^{nd} degree polynomial regression curve based on the data: X: 107 110 121 124 129 132 131 153 142 151 158 152 164 152 162 156 175 178 172 177 174 183 172 176; Y: 99 82 102 81 89 87 59 112 74 91 116 91 96 85 131 136 162 121 137 82 165 159 91 193. Discuss the summary statistics and hence find the correlation index of order 2 i.e. r_2. Draw the regression line over the scatter plot. Also draw confidence and prediction bands.*

▶ R Code:

```
x=c(107,110,121,124,129,132,131,153,142,151,158,152,164,152,162,156,175,178,172,
177,174,183,172,176)
y=c(99,82,102,81,89,87,59,112,74,91,116,91,96,85,131,136,162,121,137,82,165,159,
91,193)
fit=lm(y ~ x + I(x^2)) # Polnomial Fit
summary(fit)
plot(x,y,xlab = ’X’, ylab=’Y’,ylim=c(20,200),col=’blue’,main=’Polynomial Fit’,
cex=1.2,pch=16)
x1 = data.frame(x=seq(min(x),max(x),0.5))
lines(x1$x,predict(fit,newdata=x1),col="green",lwd=2)
# For confidence and Prediction Bands
y1=predict(fit,interval="conf",newdata=x1)
y2=predict(fit,interval="pred",newdata=x1)
matlines(x1$x,y1,lty=c(1,2,2),col=c("green","red","red"),lwd=c(2,1,1))
matlines(x1$x,y2,lty=c(1,3,3),col=c("green","black","black"),lwd=c(2,1,1))
```

▶ *Here from the summary statistics we have $r_2^2 = 0.4792$.*

Example 2.8. *Fit a 3^{rd} degree polynomial regression curve based on the data in Example 2.7 and hence find the correlation index of order 3 i.e. r_3. Also find the correlation coefficient r_{xy} and show that $r_{xy}^2 \leq r_2^2 \leq r_3^2$.*

▶ $\boxed{\textit{R Code}}$:

```
x=c(107,110,121,124,129,132,131,153,142,151,158,152,164,152,162,156,175,178,172,
177,174,183,172,176)
y=c(99,82,102,81,89,87,59,112,74,91,116,91,96,85,131,136,162,121,137,82,165,159,
91,193)
rxy=cor(x,y)
fit=lm(y ~ x + I(x^2)) # Quadratic Fit
summary(fit)
fit1=lm(y ~ x + I(x^2) + I(x^3)) # Cubic Fit
summary(fit1)
```

▶ *Here* $r_{xy} = 0.6177 \Rightarrow r_{xy}^2 = 0.3816$ *and from the summary statistics we have* $r_2^2 = 0.4792,\ r_3^2 = 0.4851.$

Example 2.9. *Fit a* 3^{rd} *degree polynomial regression curve based on the data in Example 2.7 and hence find the correlation index of order* 3 *i.e.* r_3*. Also find the correlation coefficient* r_{xy} *and show that* $r_{xy}^2 \leq r_2^2 \leq r_3^2$.

▶ $\boxed{\textit{R Code}}$:

```
x=c(107,110,121,124,129,132,131,153,142,151,158,152,164,152,162,156,175,178,172,
177,174,183,172,176)
y=c(99,82,102,81,89,87,59,112,74,91,116,91,96,85,131,136,162,121,137,82,165,159,
91,193)
rxy=cor(x,y)
fit=lm(y ~ x + I(x^2)) # Quadratic Fit
summary(fit)
fit1=lm(y ~ x + I(x^2) + I(x^3)) # Cubic Fit
summary(fit1)
```

▶ *Here* $r_{xy} = 0.6177 \Rightarrow r_{xy}^2 = 0.3816$ *and from the summary statistics we have* $r_2^2 = 0.4792,\ r_3^2 = 0.4851.$