UNIT 7

IMAGE STATISTICS

Structure

7.1 INTRODUCTION

You are aware that remote sensing is a technique for gathering information about any object or feature from a distance without being making any physical contact with it. This information is recorded in the form of raster data or images by using sensors or cameras mounted on the various remote sensing platforms. The remote sensors detect terrain features by using reflected solar radiation, heat emitted by objects and back-scattered energy in the microwave region of the electromagnetic spectrum. Remotely sensed images give an objective record of the electromagnetic spectrum pertaining to the physical and chemical properties of the terrain or ground. Extracting information from the remotely sensed images requires the knowledge of the image statistics.

You have learnt about basic principles, data formats, data products and applications of various types of remote sensing techniques such as multispectral, thermal, hyperspectral, microwave, LiDAR and UAV-based in the previous block. Images are the final product of remote sensing and they can be microwave, panchromatic, multispectral and hyperspectral.

Contributors: Dr. Manu Mehta and Mr. Akhilesh Kumar

In addition, images can be acquired in single or multiple bands. Image statistics involves the study of remotely sensed images and their statistical properties. These properties are useful in the description, classification, quantification, quality assessment and reduction of images. They also guide us for choosing suitable secondary treatments like filtering, restoring, coding and shape detection for giving to the image. In this unit, we will discuss image statistics and its types such as univariate and multivariate image statistics.

Expected Learning Outcomes

After studying this unit, you should be able to:

- ❖ define image statistics;
- ❖ discuss various types of image statistics;
- ❖ describe univariate descriptive image statistics and its measures; and
- ❖ explain multivariate image statistics and its visualisation.

7.2 IMAGE STATISTICS

As mentioned earlier that image statistics involves the study of remotely sensed images and their statistical properties. It calculates raster cell statistics from one or multiple selected images. It is used to compute the statistical measurements of the images consisting of image histogram, mean, median, standard deviation, dispersion, distribution, covariance and correlation. Image statistical analysis gives us an easily interpretable visual format of an image. It helps us to find typical problems in an image and to visually improve it. It also allows us to determine whether an image is suitably exposed or not. The poorly exposed images result in the loss of the information and this type of error should be recognised at the earlier stage of image processing because it is not possible to recover such loss at the later stages of image processing (Burger and Burge, 2022).

It is important to note that the volume of data available from remote sensors is astounding both in spatial and temporal domains. For instance, while Terra and Aqua satellites onboard the Moderate Resolution Imaging Spectroradiometer (MODIS) provide satellite images at high temporal resolution of 1-2 days, others from satellites like WorldView – 3 provide data at high spatial resolution of almost 31 cm. Thus, making meaningful sense of the huge amount of remotely sensed data requires knowledge of not just data handling but also its basic statistical properties. Prior to discussing image statistics, it is important to understand the basics of statistics especially different types of data and levels of measurement for selecting appropriate statistical techniques.

7.3 BASICS OF STATISTICS

Statistics is a branch of applied mathematics that deals with the collection, organisation, presentation, analysis and interpretation of data. Let us discuss main concepts of statistics applicable to the statistical analysis of remotely sensed images.

Contributors: Dr. Manu Mehta and Mr. Akhilesh Kumar

7.3.1 Data Types

In order to decide which statistics is well suited for a dataset, it is important to first understand the type of data we are dealing with. Broadly speaking, data can be categorised into two types such as qualitative and quantitative.

i) **Qualitative Data**

It mainly provides insights related to subjective opinions, preferences or qualitative aspects of the phenomenon. The statistics pertaining to qualitative data primarily revolve around generating descriptive statistics or summaries, aiming to unravel patterns and trends or explore associations. Some of the commonly used statistics for qualitative data includes frequencies, proportions and modes.

ii) **Quantitative Data**

It represents numerical data and the statistics addressing the quantitative data can be descriptive, providing information regarding the measures of central tendency (e.g., mean and median), measures of dispersion or variability (e.g., range and standard deviation) or measures of distribution (e.g., skewness and kurtosis). Quantitative data can also be used to generate inferential statistics, using advanced statistical techniques like hypothesis testing, regression analysis and analysis of variance (ANOVA), among others to draw inferences or predictions based on the sample data.

Both qualitative and quantitative data provide broad understanding of the nature of the information being measured or conveyed, level of measurements, order of hierarchy or categories and the significance of zero points.

7.3.2 Levels of Measurement

The levels of measurement play an important role in selecting type of statistical analysis for a particular image (Stommel and Katherine, 2014). There are some statistical analyses which require more precise levels of measurement as compared to the others. There are four levels of measurement, which are described below:

i) **Nominal Data**

It represents labels or categories that do not have a numerical value or an order assigned to it. Examples of such data could include land use classes (urban, agricultural or vegetation) or colours (red, green, blue). The statistical techniques pertaining to nominal data generally include frequency analysis, where the focus is on counting and summarising the occurrences of different categories or labels. Frequency, proportions and modes are some of the commonly used measures for describing nominal data.

ii) **Ordinal Data**

It also represents labels or categories. However, these categories have a meaningful order or hierarchy attached to them. Though these categories are not numerically quantifiable, their relative position or rank has a meaning. Common examples could include education levels (high school, graduate, post-graduate) or literacy (illiterate, semi-literate, literate). Central tendency and dispersion of ordinal data is generally summarised using median and quartiles.

iii) **Interval Data**

It is the third level of measurements and extends the characteristics of ordinal data by exhibiting equal intervals between values. Examples include temperature measured in Celsius or Fahrenheit scales. Statistical techniques for interval data include measures of central tendency using mean or median and measures of dispersion using range or standard deviation.

iv) **Ratio Data**

It is the highest level of measurement. In addition to possessing all the characteristics of interval data, ratio data also have a meaningful zero point and thus, it is possible to make statements about ratios and proportions. Examples include age, height, income or weight. The statistical techniques that can be applied to ratio data range from as basic as descriptive statistics to hypothesis testing, regression, and other advanced modelling techniques.

It may be noted that the numerical analysis is possible for interval and ratio data. However, it is not possible for nominal or ordinal data. Similarly, while assessment of hierarchy is possible for ordinal or ratio data, it is not possible for nominal or interval data.

7.3.3 Population and Sample

In statistics, population refers to the entire group of events for which we have the information and sample is used to refer to a small subset of the population which can be used as a smaller subset or representation of the entire population. The importance of the concept of sample lies in the fact that often at times, it is neither feasible nor appropriate to survey or collect information for the entire population, either due to scale, resource constraints or some other ethical reason preventing the testing or application to the entire population. In such situation, sample can be used effectively to test the hypothesis and then generalise the findings to the entire population.

Using a subset of the population to make generalisation about the entire population often leads to uncertainty, commonly referred to as sampling error and is generally measured by a confidence interval. For example, if a survey states with 90% confidence that the average height in the age group 18-30 is 160 cm, it means that if a random person is picked up in this age group, 9 out of 10 times, his/her height will be around 160 cm.

While there are many formulae that can be used to determine the sample size at a particular confidence threshold for a given population, one of them is the *Slovin's formula*, which states that–

$$n = \frac{N}{1 + Ne^2} ----(1)$$

where n is the sample size, N is the population size and e is the error tolerance.

Let us calculate sample size (n) by taking a population size of 10,000 and 90% confidence level (i.e., 10% or 0.10 tolerance level) and equation (1) as given below:

$$n = \frac{10{,}000}{1 + 10{,}000 * 0.10^2} = 99$$

7.3.4 Types of Statistical Analysis

Generally, statistical analysis is categorised into three main types – univariate, which deals with a single variable, bivariate which deals with two variables and multivariate statistics that deals with a number of variables.

Univariate Analysis: It focuses on one variable at a time and can be used to examine the central tendency, variability or distribution of the data using various measures like mean, median, mode, variance and standard deviation. It is commonly used to summarise data, identify the outliers and gain insights about the behaviour of a specific variable.

Bivariate Analysis: It seeks to study the relationship between two variables and examines how the change in one variable impacts the other variables. It can be done by using techniques like correlation, scatter plots or simple linear regression to explore the cause-and-effect relationships, measure the direction or strength of association and to make predictions based on the association.

Multivariate analysis: It extends the capabilities of bivariate analysis to three or more variables and is focussed on understanding the complex relationships and associations existing among them. Multivariate analysis relies on statistical techniques like principal component analysis (PCA), multiple regression analysis, and cluster analysis.

7.3.5 Significance

Understanding type of data and appropriate analysis technique is essential in order to draw meaningful inferences and for making informed decisions. For instance, while frequency and proportion are more suited to nominal data, ordinal data can be better explained using measures of central tendency and dispersion. Interval and ratio data, though, require more advanced statistical techniques like correlation and regression.

The choice and outcome of statistical analysis is also dependent on number of variables. As discussed earlier, where univariate analysis provides information about individual variables, bivariate variables explore associations between two variables. Multivariate analysis, though, can help unravel even more complex dependencies and associations by taking all the variables into account.

7.4 TYPES OF IMAGE STATISTICS

In digital images, statistics mainly deals with the probability distribution function of the brightness levels of a particular image and density function of the brightness levels of the same image. It may be noted that majority of image processing systems perform two types of statistical analyses such as univariate and multivariate analysis on single and multiple band remotely sensed data (Jensen, 2018). Let us discuss these two types of image statistics.

7.4.1 Univariate Descriptive Image Statistics

It describes and summarises individual image variable separately. It does not work on cause-and-effect relationship. Its main aim is to describe image and infer patterns that exist within the image. The most common univariate descriptive statistics in remotely sensed data are described below.

i) Measure of Central Tendency in Remotely Sensed Data

A measure of central tendency of a given dataset describes the characteristics of a dataset using a single value, typically the one that occupies the central position within the dataset.

a) Mean

It is the arithmetic average of all the values for a variable. In terms of image statistics, mean can be defined as the average value of all the pixel brightness.

If N is the number of pixels and $x_{i,j}$ is the value of the pixel at $(i,j)th$ position, then mean μ is given by –

$$\mu = \frac{\sum_{i=1,j=1}^{N} x_{i,j}}{N} ----(2a)$$

If sample size n is considered, instead of the entire population, then mean ($\bar{x}$) is given by –

$$\bar{x} = \frac{\sum_{i=1,j=1}^{n} x_{i,j}}{n} ----(2b)$$

For example, the pixel values in an image of 3x3 dimension $= \begin{bmatrix} 1 & 2 & 7 \\ 3 & 4 & 2 \\ 5 & 2 & 3 \end{bmatrix}$, then the mean pixel value will be calculated as –

$$= \frac{(1+2+7+3+4+2+5+2+3)}{9} = 3.22$$

While the simple mean gives equal weightage to all the observations, there may be situations where there are multiple datasets but number of observations in each of them is not same. In such cases, **weighted mean** can be calculated as:

$$\bar{x} = \frac{\sum_{i=1,j=1}^{N} w_{i,j} x_{i,j}}{\sum_{i,j}^{N} w_{i,j}} ----(2c)$$

where $w_{i,j}$ is the corresponding weight and other terms have the same meaning as in Equation 2a.

Let us consider the following example. Suppose, we have an image array A with the array of weights as W.

A = $\begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix}$, W = $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

Now, the mean of A = $\frac{2+1+3+2}{4} = 2$ and the weighted mean of

A = $\frac{2*1+1*2+3*2+2*1}{6} = 2$

b) Median

Median represents the middle value of the dataset arranged in an order – either ascending or descending and unlike mean, is unaffected by the extreme values or outliers. To calculate the median, first arrange the observations in a dataset in an order and then pick the middle value.

In order to calculate the median of a given image array below (2c), we first need to arrange it into an order (2d) –

$$\text{Image array} = \begin{bmatrix} 1 & 2 & 7 \\ 3 & 4 & 2 \\ 5 & 2 & 3 \end{bmatrix} ----(2c)$$

$$\text{Arrangedarray} = [1, 2, 2, 2, 3, 3, 4, 5, 7] ----(2d)$$

Since the number at the middle of the arranged order is 3, the median pixel value of the image is 3. In case the number of data points is even, then median value is taken as the average of the two middle values.

Let us take another example of an image array given at 2e for understanding importance of median value over mean.

Image array = $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 100 \end{bmatrix} ----(2e)$

The mean value $= \frac{1+2+3+4+5+100}{6} = 19.17$

But the median value $= \frac{3+4}{2} = 3.5$

For a dataset where there is huge difference between the values or where there is possibility of having outliers, median values are recommended measures for central tendency instead of mean.

c) **Mode**

Mode is defined as the value occurring most often in a dataset. It is a useful statistic for determining the values that dominate the dataset.

Take an example of an image array matrix given below:

$$\text{Image array} = \begin{bmatrix} 1 & 2 & 7 \\ 3 & 4 & 2 \\ 5 & 2 & 3 \end{bmatrix}$$

The pixel value 2 appears maximum number of times (3 in total), hence its mode is 2.

It is worth mentioning here that a dataset can have more than one mode value. In that case, the data is called as bimodal (2 modes), trimodal (3 modes) or multimodal (more than 3 nodes). For instance, in the following case –

$$\begin{bmatrix} 1 & 4 & 4 & 2 \\ 2 & 5 & 6 & 8 \\ 7 & 7 & 9 & 3 \end{bmatrix}$$

The mode values are = 4, 2, 7 and the dataset can be considered as trimodal.

ii) **Measures of dispersion**

Measures of the dispersion give us extent of variability, which helps us to understand the distribution of the image data. In other words, it gives us information on how stretched or squeezed the values are around their central values. Let us discuss different measures of dispersion.

a) **Range**

It is the simplest measure of dispersion. The range of a band of imagery is calculated as the difference between the maximum and minimum values.

Let us take the following image array –

$$A = \begin{bmatrix} 10 & 15 & 20 \\ 25 & 30 & 10 \end{bmatrix}$$

Here the maximum valve = 30 and minimum value = 10, hence, the range = 30 – 10 = 20

But just like the mean, range is impacted by the presence of outliers in the dataset. Consider the following dataset –

$$A = \begin{bmatrix} 10 & 15 & 20 \\ 25 & 30 & 150 \end{bmatrix}$$

Now, the maximum value = 150 and minimum value = 10, hence, the range = 150 – 10 = 140.

But from the above dataset, it is evident that most values lie around 20±10 and 150 is the outlier. So, range is not a very reliable measure in case outliers are present in the dataset.

b) **Variance**

Variance is described as the average of the squared distance of data points from their mean. Just like the range, variance too is affected by the presence of outliers. Variance for a band of imagery can be calculated as –

$$\sigma^2 = \frac{\sum_{i=1,j=1}^{N}\left(x_{i,j} - \mu\right)^2}{N} ----(3a)$$

For sample, variance is calculated as –

$$\sigma^2 = \frac{\sum_{i=1,j=1}^{N}\left(x_{i,j} - \bar{x}\right)^2}{n-1} ----(3b)$$

Where, σ^2 is the variance and σ is the standard deviation. The other terms have the same meaning as in Equations 2a and 2b.

Let us calculate the variance by considering the following dataset

$$A = \begin{bmatrix} 4 & 8 & 6 \\ 2 & 10 & 6 \end{bmatrix}$$

In order to calculate variance, we first need to calculate the mean of the dataset.

Mean $= \frac{4+8+6+2+10+6}{6} = 6$, then variance =

$$\frac{(4-6)^2+(8-6)^2+(6-6)^2+(2-6)^2+(10-6)^2+(6-6)}{6} = 6.67$$

To understand the importance of variance, consider the following as an example. There are two Images – A and B of size 3x3. With all pixel values equal to 255, A represents a mostly white image, while in B the first two rows are black and the third row is white.

$$A = \begin{bmatrix} 50 & 50 & 50 \\ 50 & 50 & 50 \\ 50 & 50 & 50 \end{bmatrix}, \text{ and } B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 255 & 255 & 255 \end{bmatrix}$$

Now, Mean of A $= \frac{50+50+50+50+50+50+50+50+50}{9}$ = 50, and

Mean of B = $\frac{0+0+0+0+0+0+255+255+255}{9}$ = 85.Similarly, variance of A = $\frac{9*(50-50)^2}{9}$ = 0, and

Variance of B = $\frac{6*(85-0)^2+3*(85-255)^2}{9}$ = 3211.11

So, while the means of both images are quite close, the variance indicates that the variability in image B is much more than that in image A. As such, mean alone could not provide the conclusive information about an image.

c) Standard Deviation

Standard deviation is simply the square root of the variance. Since standard deviation has the same units, unlike variance, as the variable, a direct comparison between the two is easily possible.

The standard deviation of the brightness value of pixels in band imagery is given by –

$$\sigma = \sqrt[2]{\frac{\sum_{i=1,j=1}^{N}\left(x_{i,j}-\mu\right)^2}{N}} ----(4a)$$

If sample is considered instead of population, then standard deviation –

$$\sigma = \sqrt[2]{\frac{\sum_{i=1,j=1}^{N}\left(x_{i,j}-\bar{x}\right)^2}{n-1}} ----(4b)$$

Where, the terms have same meanings as in Equations 2a, 2b, 3a, and 3b.

Considering the dataset from the previous section, the standard deviation = $\sqrt{variance} = \sqrt{6.67} = 2.58$. Hence, the average distance of the data points from the mean is 2.58.

The standard deviation is routinely used in digital image processing for performing linear contrast enhancement, spatial filtering, parallelepiped image classification and for calculating accuracy assessment of a classified image (Jenson, 2018). However, both mean and standard deviations are to determine the degree of similarity or dissimilarity among neighbouring pixels in a digital remote sensing image. If the standard deviation is closer to 1, then the pixels are homogenous. Standard deviation farther away from 1 shows heterogeneity in the pixels.

In most cases, the image statistics is commonly represented with the help of its mean and standard deviation. The variance, though useful, is less widely used.

iii) **Measures of Distribution (Histogram)**

Measures of distribution provide information on the shape, symmetry and concentration of individual data points in a dataset/histogram. Two of the most used measures of distribution are –

a) **Skewness**

Skewness is the measure of asymmetry of a dataset or histogram. It is used to determine the extent and direction of the deviation from the asymmetry distribution. Skewness can be calculated using the following equation –

$$\text{Skewness} = \frac{\sqrt[3]{n} * \sum_{i=1,j=1}^{n}(x_{i,j} - \bar{x})^3}{\left[\sum_{i=1,j=1}^{n}(x_{i,j} - \bar{x})^2\right]^{\frac{3}{2}}} - - - - (5)$$

Here, n is the number of observations, $\bar{x}$ is the mean.

A skewness with value of 0 indicates a perfect symmetry, a positive skewness indicates a longer tail on the right side of the distribution with more values concentrated on the left, and negative skewness indicates a longer left tail with values concentrated more on the right side of the distribution.

Let us calculate the value of skewness by considering the following dataset:

$$\begin{bmatrix} 2 & 4 & 6 \\ 6 & 8 & 10 \end{bmatrix}$$

Its mean = $\frac{2+4+6+6+8+10}{6} = 6$, the skewness can be calculated as –

$$= \frac{\sqrt[3]{6} * [(2-6)^3 + (4-6)^3 + (6-6)^3 + (6-6)^3 + (8-6)^3 + (10-6)^3]}{[(2-6)^2 + (4-6)^2 + (6-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2]^{\frac{3}{2}}}$$
$$= 1.39$$

The skewness value of 1.39 indicates that the data points are positively skewed towards the right side of the distribution.

b) **Kurtosis**

Kurtosis measures the sharpness of the peak and strength of the tails. Kurtosis can be calculated as –

$$\text{Kurtosis} = \frac{n * \sum_{i=1,j=1}^{n}(x_{i,j} - \bar{x})^4}{\left[\sum_{i=1,j=1}^{n}(x_{i,j} - \bar{x})^2\right]^2} - 3 - - - - (6)$$

A kurtosis with value of 0 indicates a normal distribution, a positive kurtosis (also called as leptokurtic distribution) indicates heavier tails, thereby indicating presence of more extreme values, and a negative kurtosis (also called as platykurtic distribution) is an indicative of lighter tails and a fewer extreme value.

Let us calculate the value of kurtosis by considering the following dataset:

$$\begin{bmatrix} 2 & 4 & 6 \\ 6 & 8 & 10 \end{bmatrix}$$

Its mean = 6. So, kurtosis

$$= \frac{6 * [(2-6)^4 + (4-6)^4 + (6-6)^4 + (6-6)^4 + (8-6)^4 + (10-6)^4]}{[(2-6)^2 + (4-6)^2 + (6-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2]^2}$$
$$= -0.96$$

The negative kurtosis value indicates that while the distribution has a relatively flatter peak, the tails are lighter and the dataset does not appear to have extreme values.

iv) **Representation and Visualisation of Univariate Statistics**

Univariate statistics are important for an exploratory analysis of the image or dataset and some of the most common graphical ways of representing them include the following:

- Histogram, which displays the frequency distribution of the variable, with each bin representing the number of occurrences within that interval.
- Boxplot graphically summarises the distribution of the dataset using the minimum, first quartile, median (second quartile), third quartile and maximum range of the dataset and is very effective at visualising the outliers in the dataset.
- Probability distribution function (PDF) can also be used to visualise shape of the distribution.

The univariate statistics discussed above can also be calculated using python or other programming languages. Fig. 7.1 shows the code snippets to create a sample dataset of 1000 random points from a normal distribution and calculate the common statistics. The outputs have also been displayed. The sample dataset was also visualised using different graphical methods shown in Fig. 7.2.

```python
# Installing the necessary libraries
# pip install numpy matplotlib seaborn scipy

# Importing the necessary libraries
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# Set a random seed for reproducibility
np.random.seed(42)

# Generate a dummy dataset with 1000 samples from a normal distribution
data = np.random.normal(loc = 50, scale = 10, size = 1000)

# Calculate and print the univariate statistics
print("Mean:", np.mean(data))
print("Median:", np.median(data))
print("Mode:", float(stats.mode(data)[0]))
print("Standard Deviation:", np.std(data))
print("Skewness:", float(stats.skew(data)))
print("Kurtosis:", float(stats.kurtosis(data)))

# ------------------------------------------------------------------------------
# Output:
# Mean: 50.193320558223256
# Median: 50.25300612234888
# Mode: 17.58732659930927
# Standard Deviation: 9.787262077473542
# Skewness: 0.11680083110533478
# Kurtosis: 0.06620589292148482
```

Fig. 7.1: Calculation of common univariate statistics using python. Each band in remote sensing images is mostly a 2-D array and the commands used here can also be used on images, using suitable geospatial libraries like *rasterio* or *GDAL.*

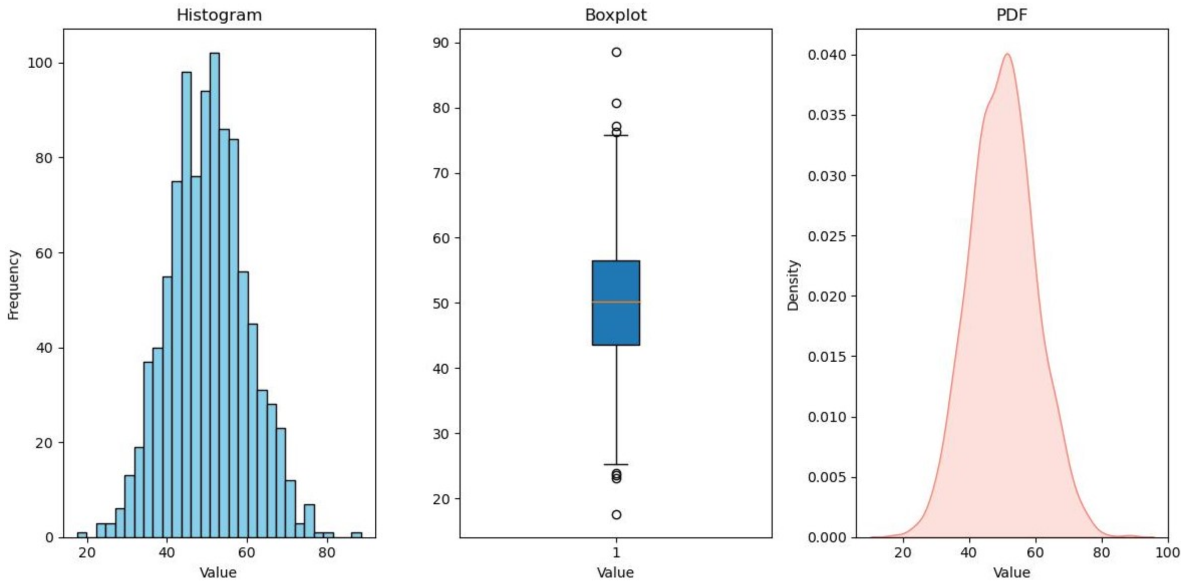

Fig. 7.2: Graphical representation of the sample dataset shown in Fig. 7.1 using different visualisation techniques.

SAQ I

a) Define mode.

b) What does a data with negative skewness suggest?

7.4.2 Multivariate Image Statistics

In remote sensing, it is also important to measure of how much radiant energy is emitted or reflected back from a feature/object in more than one band (Jensen, 2018). The measures of multivariate image statistics among the bands allow us to know how the measurements covary. In this section, we look at different multivariate statistical techniques like covariance, correlation, regression and principal component analysis.

a) Covariance

In many digital image processing techniques, understanding the relationship between any two bands is an important aspect of image classification and information extraction. This measure is known as covariance and it simply refers to a measure of the joint variance between two variables. Covariance can be positive or negative depending on the variability of one variable with respect to other. If the change in both variables is in the same direction, covariance is positive, else it is negative.

Just like variance, covariance is the average product of the differences in the values of both variables from their respective means. The unit, too, is squared. The covariance is generally expressed as a covariance matrix of dimension nxn with n being the number of variables, or bands. Covariance can be calculated as –

$$Cov(x, y) = \frac{\sum_{i=1,j=1}^{n}(x_{i,j} - \bar{x}) * (y_{i,j} - \bar{y})}{N - 1} ----(7)$$

While $\bar{x}, and\ \bar{y}$ are mean values, $x_{i,j}$ and $y_{i,j}$ are pixel values at $(i,j)th$ location and N is the number of observations.

Let us calculate the value of covariance of two image array A and B of size 2x2 as given below –

A = $\begin{bmatrix} 5 & 4 \\ 1 & 1 \end{bmatrix}$, and B = $\begin{bmatrix} 2 & 3 \\ 5 & 1 \end{bmatrix}$

The covariance matrix would look like –

$$\begin{bmatrix} cov(A,A) & cov(A,B) \\ cov(B,A) & cov(B,B) \end{bmatrix}$$

Since the diagonal elements in the covariance matrix represent the covariance between the same variables (x or y), it can also be said that the diagonal elements of a covariance matrix are simply the variances of x or y and the covariance matrix can also be written as –

$$\begin{bmatrix} var(A,A) & cov(A,B) \\ cov(B,A) & var(B,B) \end{bmatrix}$$

Since μ_A andμ_B= 2.75, the covariance can be calculated as shown in Table 7.1.
Table 7.1: Calculation of terms used for determining covariance matrix

Pixel	$A_{i,j}$	$B_{i,j}$	$A_{i,j} - \mu_A$	$B_{i,j} - \mu_B$	$(A_{i,j} - \mu_A) * (B_{i,j} - \mu_B)$	$(A_{i,j} - \mu_A)^2$	$(B_{i,j} - \mu_B)^2$
(1,1)	5	2	2.25	-0.75	-1.69	5.06	0.56
(1,2)	4	3	1.25	0.25	0.31	1.56	0.06
(2,1)	1	5	-1.75	2.25	-3.94	3.06	5.06
(2,2)	1	1	-1.75	-1.75	3.06	3.06	3.06

Then, $cov(A,A)$ or $var(A,A)$ = $\frac{5.06+1.56+3.06+3.06}{3}$ = 4.24,

$cov(B,B)$ or $var(B,B)$ = $\frac{0.56+0.06+5.06+3.06}{3}$ = 2.91, and

$cov(A,B) = cov(B,A)$ = $\frac{-1.69+0.31-3.94+3.06}{3}$ = - 0.75

The covariance matrix can then be filled as $= \begin{bmatrix} 4.25 & -0.75 \\ -0.75 & 2.92 \end{bmatrix}$

Covariance matrix suggests that A and B share a negative relationship.

b) Correlation coefficient

You are aware that covariance depends on the variance, but correlation coefficient removes this dependency. It is a standardised measure of the relationship of one variable with the other. Unlike covariance, correlation coefficient provides insights on both the strength and direction of the relationship between the variables and can be positive, indicating a change in the same direction, or negative, if the change is in opposite direction.

Furthermore, correlation coefficient, being a scaled measure of covariance, is a dimensionless quantity with values ranging between -1 and 1. The strength of the dependency is indicated by how close the value is to either 1 or -1. There are two common methods of calculating correlation coefficient as listed below:

- **Pearson correlation coefficient:** It works with continuous variables and only evaluates the linear relationship between variables. It can be calculated as –

$$r = \frac{\sum_{i,j}^{n} x_{i,j} * y_{i,j} - \frac{\sum_{i,j}^{n} x_{i,j} * \sum_{i,j}^{n} y_{i,j}}{n}}{\sqrt{\left(\sum_{i=1,j=1}^{n} x_{i,j}^2 - \frac{\left(\sum_{i=1,j=1}^{n} x_{i,j}\right)^2}{n}\right) * \left(\sum_{i=1,j=1}^{n} y_{i,j}^2 - \frac{\left(\sum_{i=1,j=1}^{n} y_{i,j}\right)^2}{n}\right)}} ----(8)$$

- **Spearman correlation coefficient**: It can be used for both continuous and ordinal variables. It is used to evaluate any monotonic relationships. Since Spearman correlation is based on rank order, it can be calculated using the following steps –
  - Rank the values for both X and Y in some order. Generally, this is based on the supposed hierarchy of one value over the other.
  - Calculate the difference between the assigned rank of X and Y, and square it.
  - Use the equation below to calculate the Spearman correlation coefficient

$$r_S = 1 - \frac{6 * \sum_{i=1,j=1}^{n} \left(x_{i,j} - y_{i,j}\right)^2}{n * (n^2 - 1)} ----(9)$$

While the sign of correlation coefficient indicates the nature of association, the value indicates its strength.

Let us consider the following image arrays for calculating then Pearson correlation –

A = $\begin{bmatrix} 7 & 6 & 8 \\ 5 & 6 & 9 \end{bmatrix}$ and B = $\begin{bmatrix} 12 & 8 & 12 \\ 10 & 11 & 13 \end{bmatrix}$, then Pearson correlation coefficient can be calculated by calculating various terms as shown in Table 7.2.

Table 7.2: Calculation of terms used for determining Pearson correlation coefficient.

Pixel number	$x_{i,j}$	$y_{i,j}$	$\left(x_{i,j}\right)^2$	$\left(y_{i,j}\right)^2$	$x_{i,j} * y_{i,j}$
(1,1)	7	12	49	144	84
(1,2)	6	8	36	64	48
(1,3)	8	12	64	144	96
(2,1)	5	10	25	100	50
(2,2)	6	11	36	121	66
(2,3)	9	13	81	169	117

So, $r = \frac{[84+48+96+50+66+117] - \frac{[7+6+8+5+6+9]*[12+8+12+10+11+13]}{6}}{\sqrt{\left([49+36+64+25+36+81] - \frac{[7+6+8+5+6+9]^2}{6}\right) * \left([144+64+144+100+121+169] - \frac{[12+8+12+10+11+13]^2}{6}\right)}}$

$$= \frac{461 - 451}{\sqrt{(291 - 280.17) * (742 - 726)}} = 0.7596$$

Hence, both images show a strong positive correlation of 0.76.

Let us consider two thematic images (or arrays), in which one represents education levels (an ordinal data) and another gives income levels (a ratio data) for calculating Spearman correlation coefficient.

If the educational level (X) = $\begin{bmatrix} Preparatory & Primary & University \\ Secondary & Secondary & Illiterate \end{bmatrix}$, and

income (Y) = $\begin{bmatrix} 25 & 10 & 8 \\ 10 & 15 & 50 \end{bmatrix}$, then Spearman correlation coefficient (Table 7.3) can be calculated as –

Table 7.3: Calculation of Spearman correlation coefficient between two images representing education and income levels.

Pixels	$X_{i,j}$	Original Rank $(X_{i,j,old})$	Assigned Rank $(X_{i,j,new})$	$(Y_{i,j})$	Original Rank $(Y_{i,j,old})$	Assigned Rank $(Y_{i,j,new})$	$(X_{i,j,new} - Y_{i,j,new})^2$
(1,1)	Preparatory	5	5.0	25	2	2.0	9.00
(1,2)	Primary	4	4.0	10	4	4.5	0.25
(1,3)	University	1	1.0	8	6	6.0	25.00
(2,1)	Secondary	2	2.5	10	5	4.5	4.00
(2,2)	Secondary	3	2.5	15	3	3.0	0.25
(2,3)	Illiterate	6	6.0	50	1	1.0	25.00

So, $r_S - 1 - \frac{6*[9+0.25+25+4+0.25+25]}{6*(36-1)} - -0.014$

Hence, there appears to be a strong negative correlation between the level of education and the income levels.

In order to calculate covariance and correlation between different bands of a remotely sensed data, we can use image processing and GIS software like ENVI, ArcMap or we can use programming languages like R and Python. While the example shown below (Fig. 7.3) generates 1000 random pixels for 4 hypothetical bands (Fig. 7.4), they can easily be replaced by actual bands once they have been converted to arrays using suitable geospatial libraries like *rasterio*.

```
1  import numpy as np, pandas as pd, seaborn as sns, matplotlib.pyplot as plt
2
3  # Set a random seed for reproducibility
4  np.random.seed(42)
5
6  # Generate a dummy multivariate dataset
7  n_samples = 1000
8
9  # Generate random data for 1000 pixels from normal distribution for each band
10 data = pd.DataFrame({
11     'Band 1': np.random.normal(loc=50, scale=10, size=n_samples),
12     'Band 2': np.random.normal(loc=20, scale=5, size=n_samples),
13     'Band 3': np.random.normal(loc=30, scale=20, size=n_samples),
14     'Band 4': np.random.normal(loc=10, scale=10, size=n_samples),
15 })
16
17 # Calculate covariance and correlation
18 covariance_matrix = data.cov()
19 correlation_matrix = data.corr()
20
21 # Print covariance matrix
22 print("Covariance Matrix: \n", covariance_matrix, "\n")
23
24 # Print correlation matrix
25 print("Correlation Matrix: \n", correlation_matrix, "\n")
26
27 # Plot histograms of each band
28 plt.figure(figsize=(10, 10))
29 for i, column in enumerate(data.columns):
30     plt.subplot(3, 2, i+1)
31     sns.histplot(data[column], kde=True)
32     plt.title(column)
33     plt.xlabel('Values')
34
35 # Plot covariance matrix
36 plt.subplot(3, 2, 5)
37 sns.heatmap(covariance_matrix, annot=True, cmap='coolwarm', fmt='.2f', square=True)
38 plt.title('Covariance Matrix')
39
40 # Plot correlation matrix
41 plt.subplot(3, 2, 6)
42 sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1, fmt='.2f', square=True)
43 plt.title('Correlation Matrix')
44
45 plt.tight_layout()
46 plt.savefig(r'E:\UGC Book\bands.jpeg')
47 plt.show()
48
49 # --------------------------------------------------------------------------------
50 # Output:
51 # Covariance Matrix:
52 #             Band 1     Band 2      Band 3      Band 4
53 # Band 1  95.886385  -1.972975    4.262082   -1.339816
54 # Band 2  -1.972975  24.872881   -1.098604   -2.801943
55 # Band 3   4.262082  -1.098604  386.872929    4.361007
56 # Band 4  -1.339816  -2.801943    4.361007  105.500122
57
58 # Correlation Matrix:
59 #            Band 1    Band 2    Band 3    Band 4
60 # Band 1  1.000000 -0.040400  0.022129 -0.013321
61 # Band 2 -0.040400  1.000000 -0.011199 -0.054698
62 # Band 3  0.022129 -0.011199  1.000000  0.021586
63 # Band 4 -0.013321 -0.054698  0.021586  1.000000
```

Fig. 7.3: Calculating covariance and correlation for multiple bands of a remotely sensed image in python. To use actual bands, convert them to arrays and substitute them in *line 10-15*.

The correlation and covariance matrices can be visualised using histograms and heatmaps (Fig. 7.4).The histogram shows that while band 1 has similar frequency distribution as band 3, the distribution of band 2 is similar to band 4

(Fig. 7.4a-d). In the covariance heatmap, the diagonal elements represent variances of each band and the off-diagonal elements represent the covariance between pairs of bands. From Fig. 7.4e, it appears that while some bands share a positive or direct relationship (e.g., band 1 and 3, 3 and 4, etc.), others share a negative relationship (e.g., band 1 and 2, 2 and 4,
et

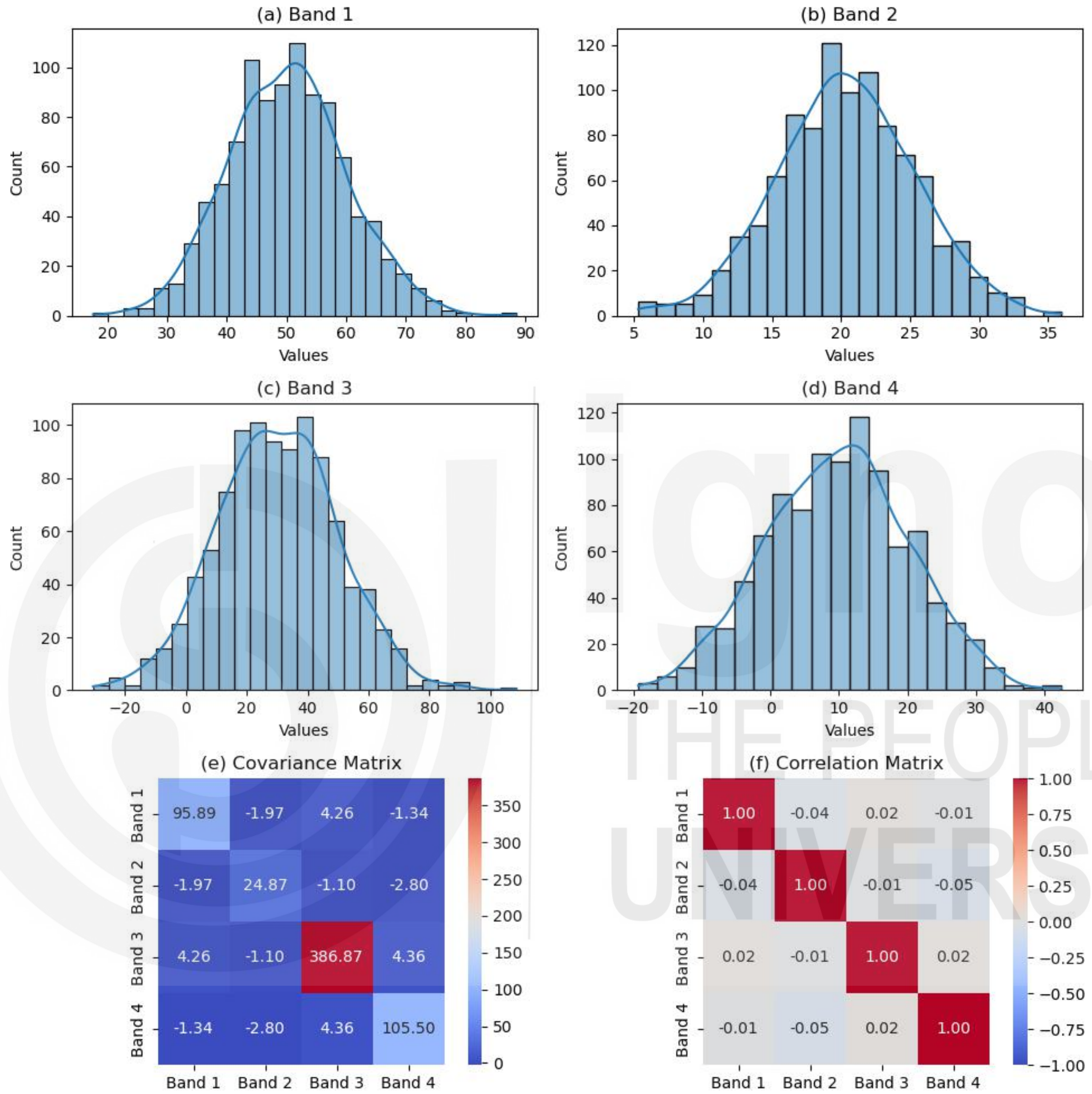

Fig. 7.4: Plot showing histogram of each sample band (a-d) and the heatmaps of (e) covariance and (f) correlation matrix.

The correlation heatmap (Fig. 7.4f), on the other hand, indicates a mostly weak or no correlation between bands, at least in a linear sense. This is expected since the pixel values in the example has been generated randomly

c) Regression

Regression is a statistical measure that attempts to determine the relationship between a dependent variable (generally referred to as Y) and one or more independent variables (generally referred to as X). Regression can be used to predict the dependent variables if the corresponding value of the independent variable is known.

By using the least squares method, regression finds the best fitting line and the corresponding regression equation. Slope and intercept are the two most important measures that are obtained through regression. While the slope indicates the rate of change of dependent variable with respect to the independent variable, intercept provides information regarding the initial offset value for the dependent variable. A regression equation is given as–

$$\hat{Y} = a + bX - - - - (10a)$$

Here, a is the y-intercept, $\hat{Y} = \bar{y} + b(x - \bar{x})$, and b is the slope given by the equation (Chatterjee and Hadi, 2006) –

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} - - - - (10b)$$

Let us consider an array X representing the independent variables

$= \begin{bmatrix} 7 & 6 & 8 \\ 5 & 6 & 9 \\ 4 & 3 & 10 \\ 6 & 6.4 & 6.4 \end{bmatrix}$, and an array Y representing the dependent variables

$= \begin{bmatrix} 12 & 8 & 12 \\ 10 & 11 & 13 \\ 9 & 5 & 15 \\ 12 & 10.7 & 10.7 \end{bmatrix}$,

As shown in Table 7.4 for finding the regression equation, first, we need to estimate the slope and y-intercept.

Table 7.4: Calculation of squared and product terms of each independent and dependent variables used in the regression.

Pixels	X	Y	X^2	$X * Y$
(1, 1)	7	12	49	84
(1, 2)	6	8	36	48
(1, 3)	8	12	64	96
(2, 1)	5	10	25	50
(2, 2)	6	11	36	66
(2, 3)	9	13	81	117
(3, 1)	4	9	16	36
(3, 2)	3	5	9	15
(3, 3)	10	15	100	150
(4, 1)	6	12	36	72
(4, 2)	6.4	10.7	40.96	68.48
(4, 3)	6.4	10.7	40.96	68.48

From the above data, $\bar{x} = 6.4$, $\bar{y} = 10.7$, so –

$$b = \frac{870.96 - \frac{76.8*128.4}{10}}{533.92 - \frac{76.8*76.8}{10}} = 1.1604,$$

And, taking x = 7 (i.e., first observation), $\hat{Y} = 10.7 + 1.1604 * (7 - 6.4)$, or $\hat{Y} = 11.396$

So, $a = 11.396 - 1.1604 * 7 = 3.2736$

Hence, the regression equation is y = 3.273 + 1.1604x. Using this equation, let us try to predict the value y at x = 8.5.

i.e., $y_{8.5} = 3.2736 + 1.1604 * 8.5 = 13.137.$

The regression line along with the data points have been represented as a scatterplot shown in fig. 7.5. The predicted value at x = 8.5 has also been marked in Fig. 7.5. R is the Pearson correlation coefficient between X and Y.

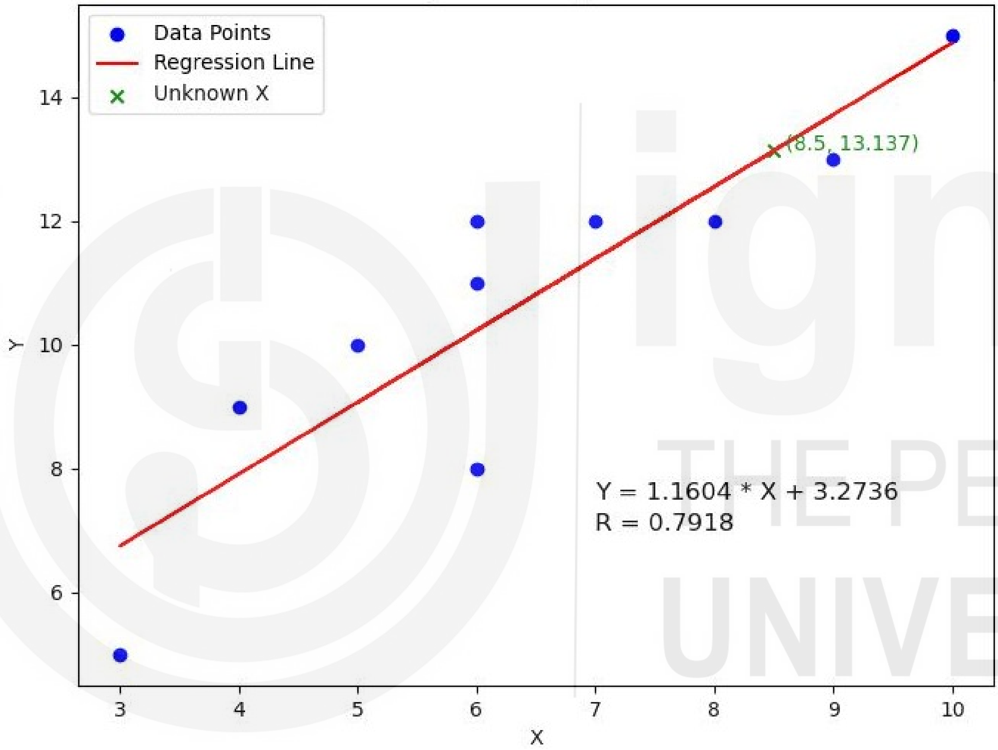

Fig. 7.5: Scatterplot showing the distribution of data and the regression line. The regression equation, correlation coefficient and the predicted value at x = 8.5 has also been marked.

d) Visualisation of Multivariate Image Statistics

Some of the most common ways of representing the multivariate image statistics have already been discussed. To summarise, here are some of the possible choices of visualisations –

- **Heatmap**: It is a tiled plot which is extensively used to represent covariances or correlations between two or more variables. Fig. 7.4 shows two such heatmaps, one for covariance and another for correlation.
- **Scatterplot**: It is a useful tool to represent the pair-wise comparison of two variables and is the preferred choice of visualisation for correlation or regression. Fig. 7.5 is an example of scatterplot showing the linear regression between two variables.

SAQ II

a) If y=-3x; how are y and x correlated?

b) What is the covariance between two independent random variables?

7.5 ACTIVITY

Read the first four bands of Landsat 8 using python and calculate the following statistics:

a) Univariate statistics – mean, variance and standard deviation

b) Find the correlation between green and red band as well as between red and NIR band. Analyse the results.

Hint: Use `rasterio` library (ref.: https://rasterio.readthedocs.io/en/stable/) to read and convert the data into `numpy` arrays and then use the code snippet given in Fig. 7.1 and Fig. 7.3.

7.6 SUMMARY

Let us summarise what we have learnt in this unit:

- Statistics is a branch of applied mathematics that deals with the collection, organisation, presentation, analysis and interpretation of data.
- In digital images, statistics mainly deals with the probability distribution function of the brightness levels of a particular image and density function of the brightness levels of the same image.
- Image statistics involves the study of remotely sensed images and their statistical properties. It calculates raster cell statistics from one or multiple selected images.
- Data can be categorised into two types such as qualitative and quantitative.
- There are four levels of measurement namely, nominal, ordinal, interval and ratio.
- Univariate descriptive image statistics describes and summarises individual image variable separately
- Though common univariate statistics like mean, variance and standard deviation are affected by the presence of outliers; median and IQR are not.
- These methods provide a measure of the central tendency and dispersion of the data. The distribution of a dataset is best captured through skewness and kurtosis
- The measures of multivariate image statistics among the bands allow us to know how the measurements covary.
- Multivariate statistics, on the other hand, use methods like covariance and correlation to measure the strength and direction to explore associations, patterns, and dependencies between more than one variable.
- The choice of an appropriate statistical technique is dependent not just on data type or number of variables in the question, but also on the intended application

7.8 TERMINAL QUESTIONS

1. Define image statistics. Discuss why it is important in digital image processing.
2. Describe various types of statistical analysis.
3. Discuss various measures of central tendency in remotely sensed data.
4. Write notes on covariance and correlation.

7.9 REFERENCES

- Burger W and Burge MJ (2022) Histograms and Image Statistics. In: Digital Image Processing. Texts in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-031-05744-1_2
- Chatterjee S and Handi AS (2006) Regression analysis by example, John Wiley & Sons.
- Jensen JR (2018) Digital Image Processing: A Remote Sensing Perspective, 4th edition, Pearson India Education Services Pvt Ltd, Uttar Pradesh, India.
- Stommel M and Katherine JD (2014) Properties of Variables: Levels of Measurement. Springer Publishing Company,New York.

7.9 FURTHER/SUGGESTED READINGS

- Bhatta B (2008) Remote Sensing and GIS, Oxford University Press, USA.
- Datar R and Garg H (2019) Hands-On Exploratory Data Analysis with R: Become an expert in exploratory data analysis using R packages, Packt Publishing Ltd.
- Everitt B and Hothorn T (2011) An introduction to applied multivariate analysis with R, Springer Science & Business Media.
- Mukhiya SK and Ahmed U (2020) Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data, Packt Publishing Ltd.

7.10 ANSWERS

SAQ I

a) Mode is defined as the value occurring at the most in a dataset.

b) Negative skewness indicates a longer left tail with values concentrated more on the right side of the distribution.

SAQ II

a) y and x are negatively correlated.

b) The covariance between 2 independent random variables is zero. At the same time, a zero covariance between 2 random variables does not imply that they are independent.

Terminal Questions

1. Please refer to section 7.2.
2. Please refer to subsection 7.3.4.
3. Please refer to subsection 7.4.1.
4. Please refer to subsection 7.4.2.