

Modern Surveying Techniques

Prof. S. K. Ghosh

Department of Civil Engineering

Indian Institute of Technology, Roorkee

Lecture - 11

Statistical Evaluation of RS Data

Previous session, I had introduced the concept of classification of remote sensing data. In that 2 approaches; one is visual and the other digital were introduced. In the digital image processing, the various processes involved in classifying a digital image were introduced and a brief discussion of the same was provided.

Now, in this session, I am going to take one by one each of the processes as discussed earlier.

(Refer Slide Time: 1:07)

Image Quality Assessment and Statistical Evaluation

- Many remote sensing datasets contain high-quality, accurate data.
- Unfortunately, sometimes error (or noise) is introduced into the remote sensor data by:
 - the *environment* (e.g., atmospheric scattering),
 - *random or systematic malfunction* of the remote sensing system (e.g., an uncalibrated detector creates striping), or
 - *improper airborne or ground processing* of the remote sensor data prior to actual data analysis (e.g., inaccurate analog-to-digital conversion).

The first task or process to be undertaken is the statistical evaluation of remote sensing datasets. As we know, many remote sensing datasets contain high quality and accurate data. Unfortunately, sometimes error or noise is introduced into the remote sensing data by the environment that is atmospheric scattering, random or systematic malfunctioning of the remote sensing system that is an uncalibrated detector create striping or improper

airborne or ground processing of the remote sensing data prior to actual data analysis that is in accurate analog to digital conversion.

(Refer Slide Time: 1:56)

- Therefore, the analyst should first assess its quality and statistical characteristics.
- This is normally accomplished by:
 - looking at the frequency of occurrence of individual brightness values in the image displayed in a *histogram*
 - viewing on a computer monitor *individual pixel brightness values* at specific locations or within a geographic area,
 - computing *univariate descriptive statistics* to determine if there are unusual anomalies in the image data, and
 - computing *multivariate statistics* to determine the amount of between-band correlation (e.g., to identify redundancy).

Therefore, the analyst should first assess its quality and its statistical characteristics. This is normally accomplished by looking at the frequency of occurrences of individual brightness values in the image displayed in a histogram or by viewing on a computer monitor, individual pixels brightness value at specific locations or within a geographic area, by computing univariate descriptive statistics to determine if there are unusual anomalies in the image data and computing multivariate statistics to determine the amount of between band correlation that is to identify redundancy.

First of all, let us have a look at the remote sensing sampling theory. Some of the terms need to be understood before we can go in for this statistical analysis.

(Refer Slide Time: 3:05)

Remote Sensing Sampling Theory

- A *population* is an infinite or finite set of elements. An infinite population could be all possible images that might be acquired of the Earth in 2007.
- A *sample* is a subset of the elements taken from a population used to make inferences about certain characteristics of the population.
- If observations with certain characteristics are systematically excluded from the sample either deliberately or inadvertently (such as selecting images obtained only in the spring of the year), it is a *biased* sample.
- *Sampling error* is the difference between the true value of a population characteristic and the value of that characteristic inferred from a sample.

First is population. It is an infinite or finite set of elements and infinite population could be all possible images that might be acquired of the earth in a particular year. A sample is a subset of the elements taken from a population used to make inferences about certain characteristics of the population.

If observations within certain characteristics are systematically excluded from the sample, either deliberately or inadvertently such as selecting images obtained only in the spring of the year, it is a biased sample.

Sampling error is the difference between the true value of a population characteristic and the value of that characteristic inferred from a sample.

(Refer Slide Time: 4:04)

Remote Sensing Sampling Theory

- Large samples drawn randomly from natural populations usually produce a *symmetrical frequency distribution*.
- Most values are clustered around some central value, and the frequency of occurrence declines away from this central point.
- A graph of the distribution appears bell shaped and is called a *normal distribution*.
- Many statistical tests used in the analysis of remotely sensed data assume that the brightness values recorded in a scene are normally distributed.
- Unfortunately, remotely sensed data may *not* be normally distributed and the analyst must be careful to identify such conditions.
- In such instances, *nonparametric* statistical theory may be preferred.

Generally, large sample drawn randomly from natural populations, usually produce a symmetrical frequency distribution. Most values are clustered around some central value and the frequency of occurrence declines away from this central point.

A graph of this distribution appears bell shaped and is called a normal distribution. Many statistical test used in the analysis of remotely sense data assume that the brightness values recorded in a scene are normally distributed. Unfortunately, remote sensing data may not be normally distributed and the analyst must be careful to identify such conditions. In such circumstances, non parametric statistical theory may be preferred.

(Refer Slide Time: 5:04)

Histograms of Symmetric and Skewed Distributions

The diagram illustrates five types of frequency distributions:

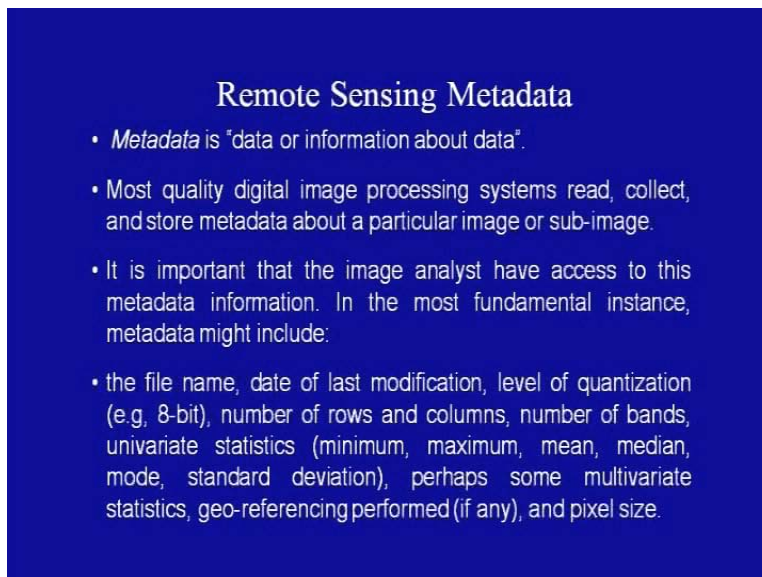
- a. Normal Distribution (Symmetric):** A bell-shaped curve where Mean, Median, and Mode are all at the same central point.
- d. Negatively Skewed Distribution (Skewed):** A curve that is skewed to the left, where the Mode is on the left, followed by the Median, and the Mean is on the right.
- b. Multimodal Distribution (Symmetric):** A curve with two peaks, where the Mean and Median are located between the two peaks.
- e. Positively Skewed Distribution (Skewed):** A curve that is skewed to the right, where the Mode is on the left, followed by the Median, and the Mean is on the right.
- c. Uniform Distribution:** A flat rectangular distribution where the Mean and Median are at the center, but no mode exists.

Common Symmetric and Skewed Distributions in Remotely Sensed Data

Here we can see, on the basis of the information which is available, histograms can be created and these histograms depict that around a central value which could be the mean median or mode. In a normal distribution case, this is symmetric. However, if the mean, median or mode are different, then we may get skewed distribution. If the skewness is towards the left, then we call this as negative skewed distribution. If this skewness is towards the right, then we call this as positive distribution.

Many a times, there may be two peaks or two bell shaped and this is what we call it as a multi model distribution. When we look at the remote sensing data, it is very important to know about the data itself and this is what we call it as metadata that is data or information about the data.

(Refer Slide Time: 6:08)



Remote Sensing Metadata

- *Metadata* is "data or information about data".
- Most quality digital image processing systems read, collect, and store metadata about a particular image or sub-image.
- It is important that the image analyst have access to this metadata information. In the most fundamental instance, metadata might include:
 - the file name, date of last modification, level of quantization (e.g, 8-bit), number of rows and columns, number of bands, univariate statistics (minimum, maximum, mean, median, mode, standard deviation), perhaps some multivariate statistics, geo-referencing performed (if any), and pixel size.

Most quality digital image processing system read, collect and store metadata about a particular image or sub image. It is important that the image analyst have access to this metadata information. In the most fundamental instance, metadata may include the file name, the date of last modification, level of quantization that is whether it is a 8 bit image or a seven bit image, the number of rows and columns in the image, number of bands that the data has, univariate statistics such as the min, maximum, mean, median, mode, standard deviation. Perhaps, some multivariate statistics; geo referencing performed, if any and the pixel size.

Now, let us look at how we go about with the assessment part. The first is viewing individual pixels.

(Refer Slide Time: 7:21)

Viewing Individual Pixels

- *Viewing individual pixel brightness values* in a remotely sensed image is one of the most useful methods for assessing the quality and information content of the data.
- Virtually all digital image processing systems allow the analyst to:
 - use a mouse-controlled *cursor* (cross-hair) to identify a geographic location in the image (at a particular row and column or geographic x,y coordinate) and display its brightness value in *n* bands,
 - display the individual brightness values of an individual band in a matrix (raster) format.

By viewing individual pixel brightness values in a remotely sense data is one of the most useful methods of assessing the quality and the information content of the data. Virtually, all digital image processing system allow the analyst to use a mouse controlled cursor to identify a geographic location in the image and display its brightness value in all datasets that are present that is in n bands in which it has been collected. Then we can display the individual brightness values of an individual band in the form of a matrix that is in raster form.

(Refer Slide Time: 8:08)

Cursor and Raster Display of Brightness Values

The screenshot shows a software interface with a main window displaying a grayscale thermal infrared image of the Savannah River. A white crosshair cursor is positioned over a bright area of the river. To the right of the image is a control panel with various icons and a 'Cursor' label. Below the image, a table displays the brightness values for a 12x12 geographic area.

a. Nighttime thermal infrared image of the Savannah River.

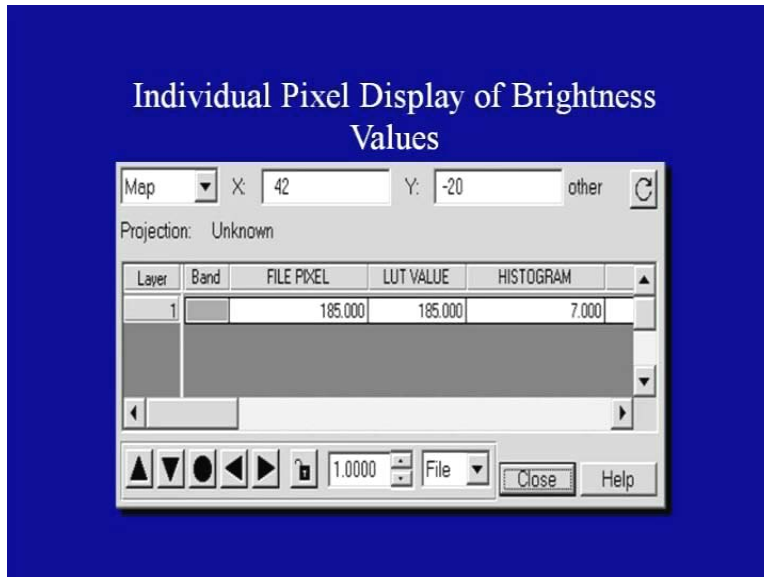
b. Individual brightness value extracted using the cursor.

Row	0	1	2	3	4	5	6	7	8	9	10	11
0	101	113	101	124	105	121	104	104	101	101	101	101
1	101	121	121	109	121	105	101	101	101	101	101	101
2	101	121	121	109	121	105	101	101	101	101	101	101
3	101	121	121	109	121	105	101	101	101	101	101	101
4	101	121	121	109	121	105	101	101	101	101	101	101
5	101	121	121	109	121	105	101	101	101	101	101	101
6	101	121	121	109	121	105	101	101	101	101	101	101
7	101	121	121	109	121	105	101	101	101	101	101	101
8	101	121	121	109	121	105	101	101	101	101	101	101
9	101	121	121	109	121	105	101	101	101	101	101	101
10	101	121	121	109	121	105	101	101	101	101	101	101
11	101	121	121	109	121	105	101	101	101	101	101	101
12	101	121	121	109	121	105	101	101	101	101	101	101
13	101	121	121	109	121	105	101	101	101	101	101	101
14	101	121	121	109	121	105	101	101	101	101	101	101
15	101	121	121	109	121	105	101	101	101	101	101	101
16	101	121	121	109	121	105	101	101	101	101	101	101
17	101	121	121	109	121	105	101	101	101	101	101	101
18	101	121	121	109	121	105	101	101	101	101	101	101
19	101	121	121	109	121	105	101	101	101	101	101	101

c. Brightness values in a geographic area for an individual band.

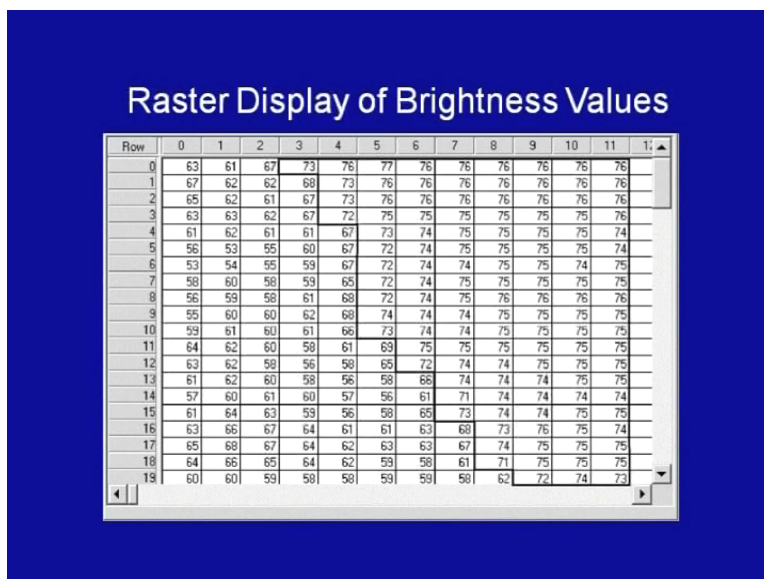
This particular slide shows that a cursor has been placed at a particular location in the image and corresponding to this, the next window shows what could be the location of the pixel that is at the top where x is equal to 42.

(Refer Slide Time: 8:29)



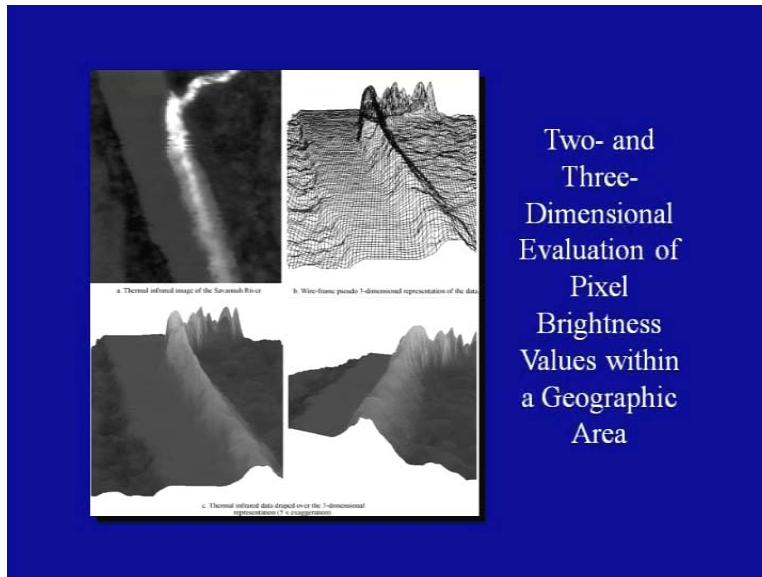
It means that this is the number of column that we are looking at and why where the value is minus 20, it tells us that this is the number of row. And below that, the value of this particular point in terms of brightness is indicated and such number of points in the histogram or that is the number of occurrences of this particular value in the whole image is also provided.

(Refer Slide Time: 9:01)



We may look it into the raster form wherein, the data would be now viewed as if it is a plethora of numbers in a graded manner.

(Refer Slide Time: 9:16)

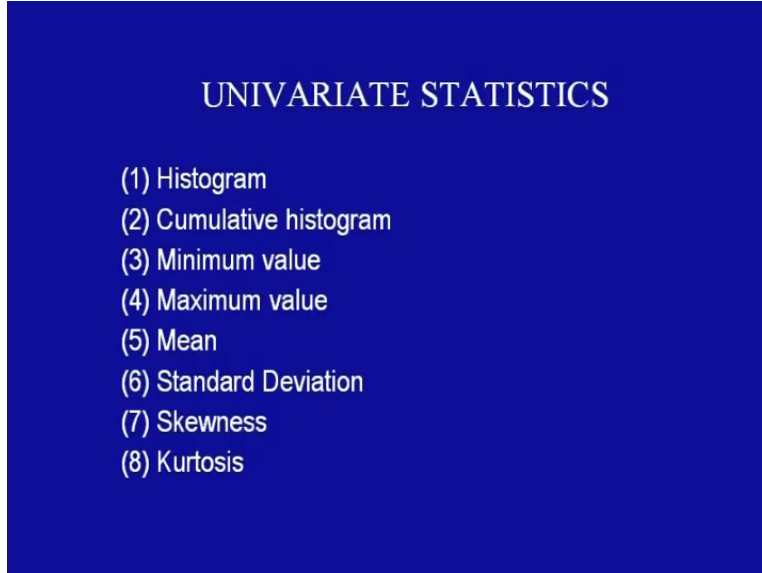


We can also view the image in the form of 2 or 3D dimensional evaluation wherein, based on the brightness values, the elevation of the information is provided. So, all those regions which are defected in bright white would have the highest elevation in the 3 dimensional representation, whereas, the dark object will be having a value of 0 and may represent a land or a surface which is at the datum level. This 2D or 3D evaluation also allows the analyst to view the image by rotating it in such a manner that the desired area can be viewed effectively.

Now, let us look at the various statistical tests that may be subjected to a remote sensing dataset. There are two types of statistics which may be computed. First is the univariate statistics and the second is multivariate statistics.

So, now let us look at the various univariate statistics which are there. It could consist of computing the histogram, the cumulative histogram, finding out the minimum value, the maximum value, mean, standard deviation, skewness and kurtosis.

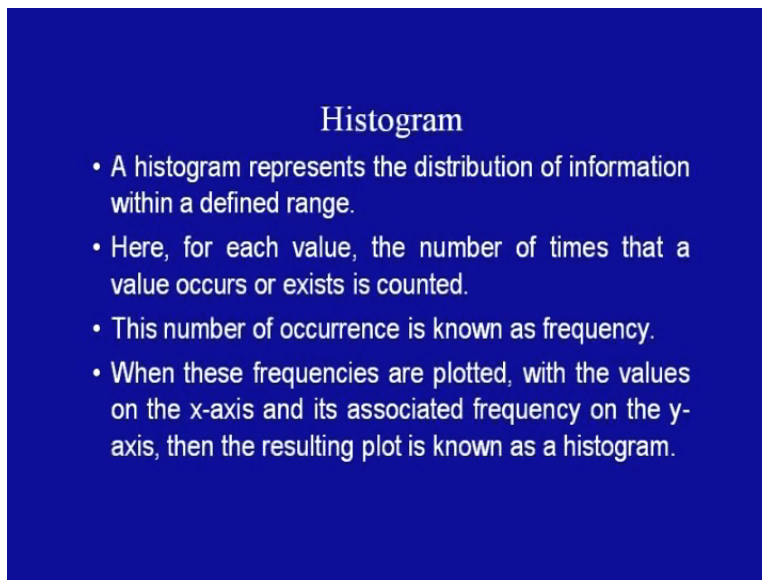
(Refer Slide Time: 10:52)



UNIVARIATE STATISTICS

- (1) Histogram
- (2) Cumulative histogram
- (3) Minimum value
- (4) Maximum value
- (5) Mean
- (6) Standard Deviation
- (7) Skewness
- (8) Kurtosis

(Refer Slide Time: 10:56)



Histogram

- A histogram represents the distribution of information within a defined range.
- Here, for each value, the number of times that a value occurs or exists is counted.
- This number of occurrence is known as frequency.
- When these frequencies are plotted, with the values on the x-axis and its associated frequency on the y-axis, then the resulting plot is known as a histogram.

First of all, let us look at what actually is a histogram. A histogram represents the distribution of information within a defined range. Here, for each value, the number of times that a value occurs or existed is counted. This number of occurrence is known as frequency. When these frequencies are plotted with the values on the x-axis and is associated frequency on the y-axis, then the resulting plot is known as a histogram.

(Refer Slide Time: 11:34)

- If each frequency is represented by a dot, its resultant appearance is in form of a curve.
- If the frequency is represented by a line, then the resultant appearance is in the form of a series of skyscraper.

If each frequency is represented by a dot, its resultant appearance is in the form of a curve. If the frequency is represented by a line, then the resultant appearance is in the form of a series of skyscrapers.

(Refer Slide Time: 11:53)

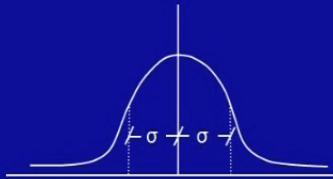
- In image processing, histogram is useful information, and is used at various steps of image processing.
- Since the image data is an independent, distributed set of samples from a random process, it is often convenient mathematically to assume that the histogram is Gaussian in nature.
- Here the distribution is centred about the mean value and its width is proportional its standard deviation.

In image processing, histogram is a useful information and is used at various steps of image processing. Since the image data is an independent distributed set of samples from a random process, it is often convenient mathematically to assume that the histogram is Gaussian in nature. Here, the distribution is centred about the mean value and its width is proportional to its standard deviation.

(Refer Slide Time: 12:24)

- The distribution can be expressed as

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\left[\frac{(x - \mu)^2}{2\sigma}\right]}$$



This distribution can be expressed by this relationship that is $f(x)$ that is the frequency of a particular value x can be expressed as 1 divided by sigma multiplied by under root 2 pi exponential raise to the power minus x minus μ whole square divided by 2 sigma; where sigma is the standard deviation and μ is the mean value. These parameters will be discussed later.

One thing which can be seen regarding the histogram is that it is symmetrical about its centered axis and that it has a direct correlation to the standard deviation.

(Refer Slide Time: 13:19)

- The shape of the histogram provides information regarding the number of features within an image.
- Generally, histogram of image is typically unimodal in nature i.e. it has a single peak, with extended tail at the both ends.
- However, in the presence of multiple features in an image, the histogram may be multi-modal i.e., there are many peaks, where each peak represents a feature.
- However, it may not be possible to clearly pinpoint which peak represents a feature.

The shape of the histogram provides information regarding the number of features within an image. Generally, histogram of an image is typically unimodal in nature that is it has a single peak with extended tails at the both ends. However, in the presence of multiple features in an image, the histogram may be multi-modal that is there are many peaks where each peak represents a feature. However, it may not be possible to clearly identify or pinpoint which peak represents a particular feature.

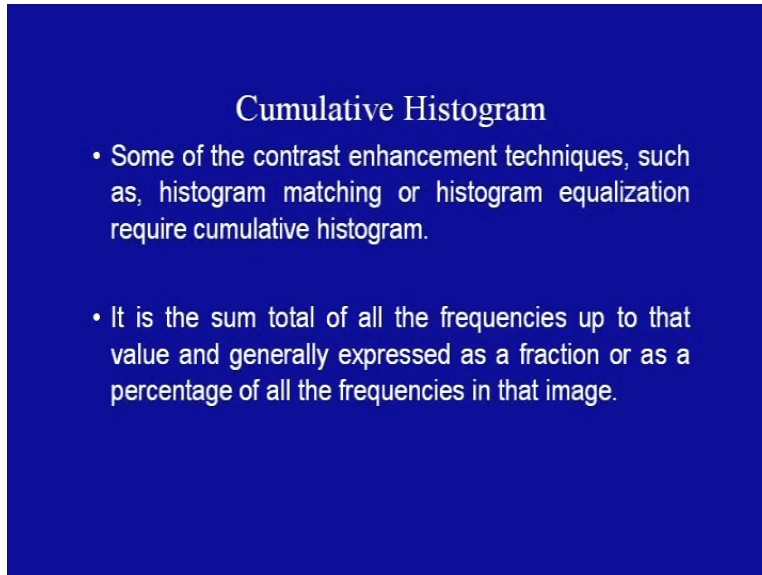
(Refer Slide Time: 14:00)

- Further; it may noted that an image histogram only provides the frequency at each level, but not the spatial correlation.
- Image histogram is a very useful graphical representation of the image information content.
- Based on this information, the contrast of an image can be improved.

Further, it may be noted that an image histogram only provides the frequency at each level but not the spatial correlation. Image histogram is a very useful geographical

representation of the image information content. Based on this information, the contrast of an image can be improved that is the appearance on a display device can be improved by having the input information regarding the distribution of information represented by a histogram.

(Refer Slide Time: 14:40)

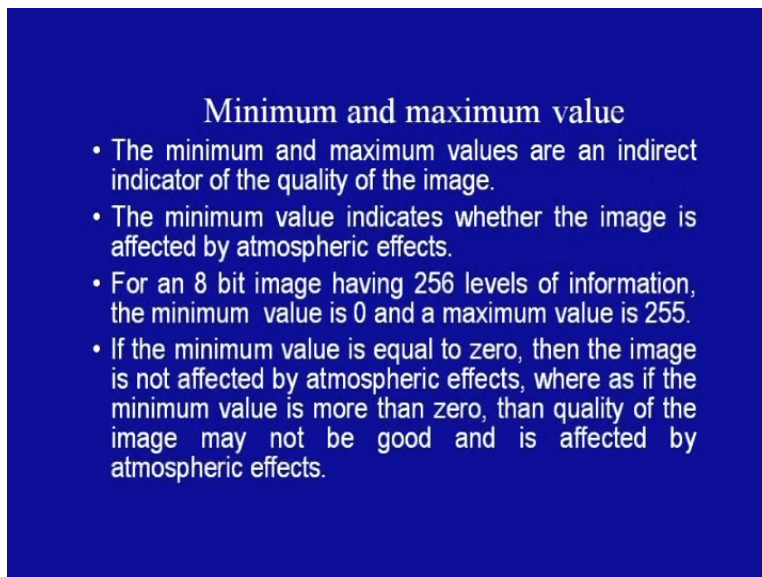


Cumulative Histogram

- Some of the contrast enhancement techniques, such as, histogram matching or histogram equalization require cumulative histogram.
- It is the sum total of all the frequencies up to that value and generally expressed as a fraction or as a percentage of all the frequencies in that image.

The next parameter is cumulative histogram. Some of the contrast enhancement techniques such as histogram matching or histogram equalization require cumulative histogram. It is the sum total of all the frequencies upto that value and generally expressed as a fraction or as a percentage of all the frequencies in that particular image.

(Refer Slide Time: 15:10)

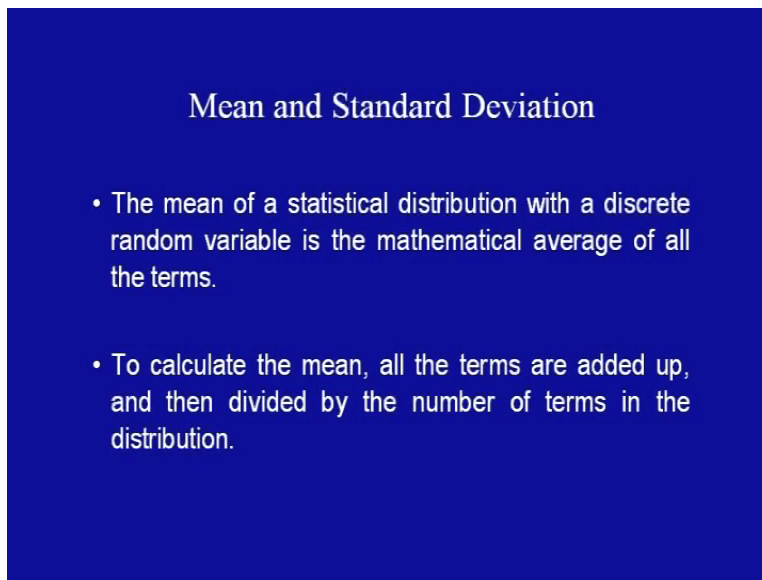


Minimum and maximum value

- The minimum and maximum values are an indirect indicator of the quality of the image.
- The minimum value indicates whether the image is affected by atmospheric effects.
- For an 8 bit image having 256 levels of information, the minimum value is 0 and a maximum value is 255.
- If the minimum value is equal to zero, then the image is not affected by atmospheric effects, where as if the minimum value is more than zero, than quality of the image may not be good and is affected by atmospheric effects.

Next is determination of the minimum and maximum value. The minimum and maximum values are an indirect indicator of the quality of the image. The minimum value indicates whether the image is affected by atmospheric effects. For an 8 bit image having 256 levels of information, the minimum value is 0 and the maximum value is 256. If the minimum value is equal to 0, then the image is not affected by atmospheric effects, whereas, if the minimum value is more than 0, the quality of the image may be affected and the quality of the image may not be good and it is and it is affected by atmospheric effects.

(Refer Slide Time: 16:11)



Mean and Standard Deviation

- The mean of a statistical distribution with a discrete random variable is the mathematical average of all the terms.
- To calculate the mean, all the terms are added up, and then divided by the number of terms in the distribution.

The next statistical parameter to be determined are the mean and the standard deviation. The mean of a statistical distribution with a discrete random variable is the mathematical average of all the terms or values in an image. To calculate the mean, all the terms are added up and then divided by the number of terms in the distribution.

(Refer Slide Time: 16:38)

- This type of mean is also called the arithmetic mean (or more commonly known as, the "average").
- The mean of a statistical distribution with a continuous random variable is the value of that random variable, denoted by the lowercase Greek letter μ (μ).

$$\mu = \frac{1}{N} \sum_{i=1}^N BV_i$$

This type of mean is also called the arithmetic mean or more commonly known as average. The mean of a statistical distributor with continuous random variable is the value of that random variable denoted by the lower Greek letter mu. So, mean can be expressed as 1 divided by N summation of all the points within the given dataset.

(Refer Slide Time: 17:12)

STANDARD DEVIATION

- The standard deviation is the root mean square (RMS) deviation of the values from their arithmetic mean.
- Standard deviation is the most common measure of statistical dispersion, measuring how the values in a data set are distributed.
- If the data points are all close to the mean, then the standard deviation is close to zero.

Next is standard deviation. Standard deviation is the root mean square deviation of the values from their arithmetic mean. Standard deviation is the most common measure of statistical dispersion, measuring how the values in a dataset is distributed. If the dataset points are all close to the mean, then the standard deviation is close to 0.

(Refer Slide Time: 17:41)

- If many data points are far from the mean, then the standard deviation is far from zero.
- If all the data values are equal, then the standard deviation is zero.
- A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (BV_i - \mu)^2}$$

If many data points are far from the mean, then the standard deviation is far from 0. If all the data values are equal, then also the standard deviation is equal to 0. A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

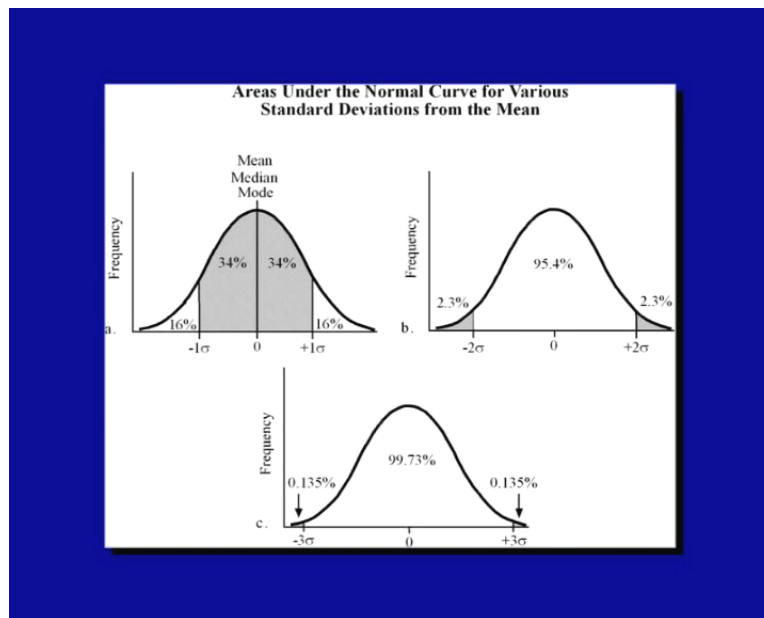
Well, standard deviation can be expressed by the following relationship that is by taking the square root of 1 by N minus 1 and taking the summation of the difference between the input values minus the mean and taking the square of this particular difference.

(Refer Slide Time: 18:35)

- The mean and variance are sufficient to specify a normal, or Gaussian, distribution and the area under the probability distribution function curve is symmetrical about its mean (μ).
- If the histogram is unimodal and symmetrical, a Gaussian distribution may be an appropriate model for representing the actual data.
- However, histograms of an image tend to be asymmetric, and also multimodal, hence information of mean and variance is useful.
- Further, standard deviation can be used as a measure of image contrast, since it is a measure of the histogram width, i.e. the spread in brightness values.

The mean and variance are sufficient to specify a normal or Gaussian distribution and the area under the probability distribution function curve is symmetrical about its mean that is μ . If the histogram is unimodal and symmetrical; a Gaussian distribution may be an appropriate model for representing the actual data. However, histograms of an image tend to be asymmetric and also multimodal in nature and hence information about mean and the standard deviation is useful to the analyst. Further, standard deviation can be used as a measure of image contrast, since it is a measure of the histogram width that is the spread in the brightness values.

(Refer Slide Time: 19:33)



This particular slide shows the area under the normal curves for various standard deviations from the mean. The first graph shows the area under the normal distribution curve with the range of ± 1 plus minus 1 sigma. It is observed that in a normal distributed dataset, this would be 34% about the mean on both the sides that is the total area coverage under 1 sigma with respect to μ is 68%. When this is increased to 2 sigma, the area under the curve increases to 95.4% and if the 3 sigma condition is taken, then we find that the area under the normal distribution curve is 99.73%.

Basically, one has to understand what does this represents. It just tells us the compactness with which we would like to adhere to or dataset and also it gives an idea as to the rejection criteria. For a 1 sigma, spread about the mean, about 32% of the data is considered to be a part of the error system. However, when it is increased to 2 sigma, then this is only 44.3% and for 3 sigma, this is very much less of the order of about 0.27% that means **only point** nearly 0.3% of the data is to be considered as outlier or what we can call it as noise in the data.

(Refer Slide Time: 21:40)

SKEWNESS

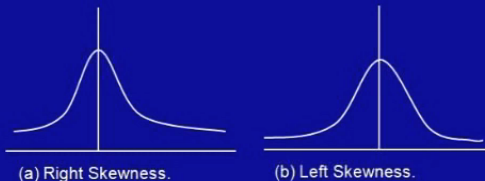
- It is a measure of the asymmetry of the probability distribution of a real-valued random variable.
- A distribution has positive skew (right-skewed), if the right (higher value) tail is longer and negative skew (left-skewed) if the left (lower value) tail is longer.
- The skewness can be calculated using formula

$$\frac{1}{N} \sum_{p=1}^N \left(\frac{BV_p - \mu}{\sigma} \right)^3$$

The next parameter to be examined is the skewness. It is a measure of the asymmetry of the probability distribution of a real-valued random variable. A distribution has positive skew or right-skewed if the (not audible) and negative skew or left skew if the left tail is longer. The skewness can be calculated using the following formula; that is it 1 divided by N, taking the summation of all the points in the dataset of the difference between the input value minus its mu divided by sigma and taking the cube of the same.

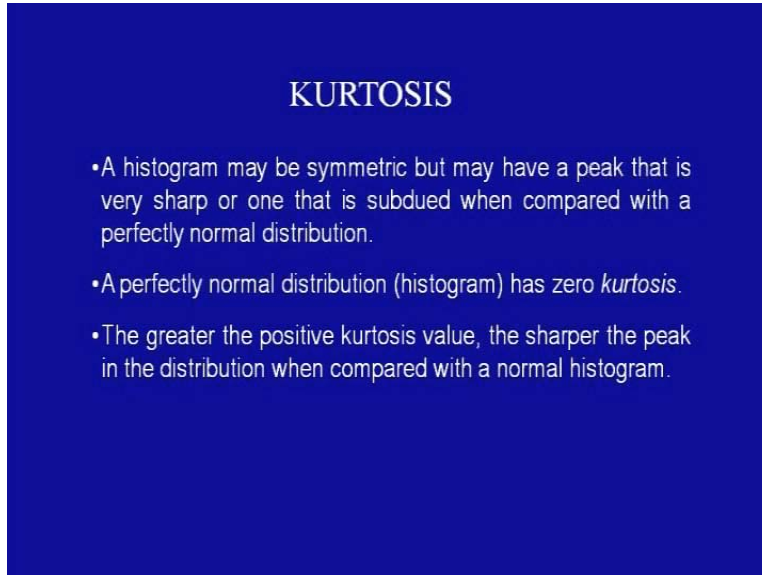
(Refer Slide Time: 22:28)

- Skewness is zero for any symmetric histogram.
- A histogram with a long tail towards larger BV has a positive skewness, and this is typical of remote-sensing images



Skewness is 0 for any symmetric histogram. A histogram with a long tail towards larger brightness values has a positive skewness as shown in the graph below and this is very typical of the remote sensing datasets.

(Refer Slide Time: 22:50)



KURTOSIS

- A histogram may be symmetric but may have a peak that is very sharp or one that is subdued when compared with a perfectly normal distribution.
- A perfectly normal distribution (histogram) has zero *kurtosis*.
- The greater the positive kurtosis value, the sharper the peak in the distribution when compared with a normal histogram.

Next is kurtosis. A histogram may be symmetric but it may have a peak that is very sharp or that is it is subdued when compared with a perfectly normal distribution. A perfectly normal distribution histogram has 0 kurtosis. The greater the positive kurtosis value, the sharper the peak in the distribution when compared to a normal histogram. Conversely, a negative kurtosis value suggests that the peak in the histogram is less sharp than that of a normal distribution.

(Refer Slide Time: 23:34)

Kurtosis

- So, it is a measure of the “peakedness” of the probability distribution of a real-valued random variable.
- Higher kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to moderately sized deviations.

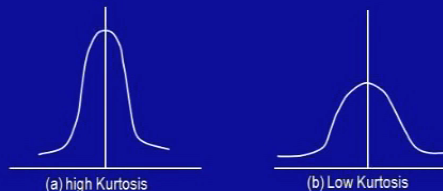
$$\left[\frac{1}{N} \sum_{p=1}^N \left(\frac{BV_p - \mu}{\sigma} \right)^4 \right] - 3$$

So, it can be said that it is a measure of the peakedness of the probability distribution of a real-valued random variable. Higher kurtosis means more of the variance is due to infrequent extreme deviations as opposed to moderately size deviations.

Kurtosis can be expressed by the following relationship; that is 1 divided by N taking the summation of all the N points of the dataset of the difference between the input values minus the mu divided by sigma raised to the power 4 and this expression is subtracted by a value of 3.

(Refer Slide Time: 24:24)

- A high kurtosis distribution has a sharper “peak” and fatter “tails”, while a low kurtosis distribution has a more round peak with wider “shoulders”.

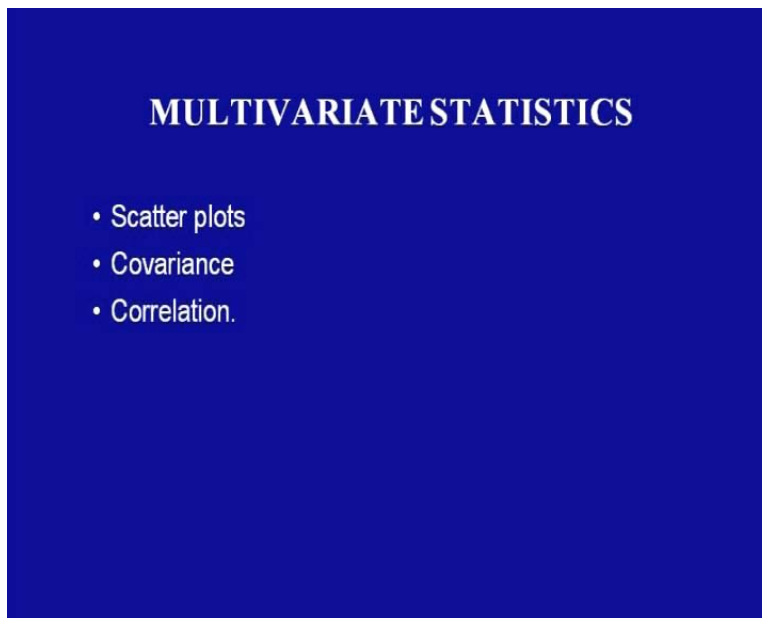


Types of Kurtosis

The higher kurtosis distribution has a sharper peak and the fatter tails; while, a low kurtosis distribution has a more round peak with wider shoulders. The same has been shown in the 2 graphs depicted below. A higher kurtosis has a higher peak along the central value, whereas, lower kurtosis is more bell shaped and the shoulders are much more rounded.

Having had a look at the univariate statistics, now let us look at the multivariate statistics which are there.

(Refer Slide Time: 25:08)



Here, we are now going to take, instead of 1 dataset; we are going to compare it with another dataset. **This can be** this examination can be performed by plotting scatter plots or by computing covariance or correlation. First of all, let us look at what is a scatter plot.

(Refer Slide Time: 25:36)

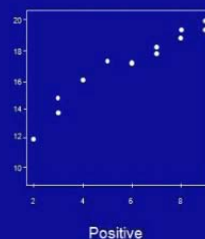
SCATTERPLOT

- A scatterplot is a useful summary for a set of bi-variate data, usually drawn before working out a linear correlation or fitting a regression line.
- It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.
- Each unit contributes one point to the scatterplot, on which points are plotted but not joined.
- The resulting pattern indicates the type and strength of the relationship between the two variables.

A scatter plot is a useful summary of a set of bi-variate data, usually drawn before working out the linear correlation or fitting a regression line. It gives a good visual picture of the relationship between the 2 variables and aids the interpretation of the correlation coefficient or the regression model. Each unit contributes one point to the scatter plots on which the points are plotted but not joined. The resulting pattern indicates the type and the strength of the relationship between the 2 variables.

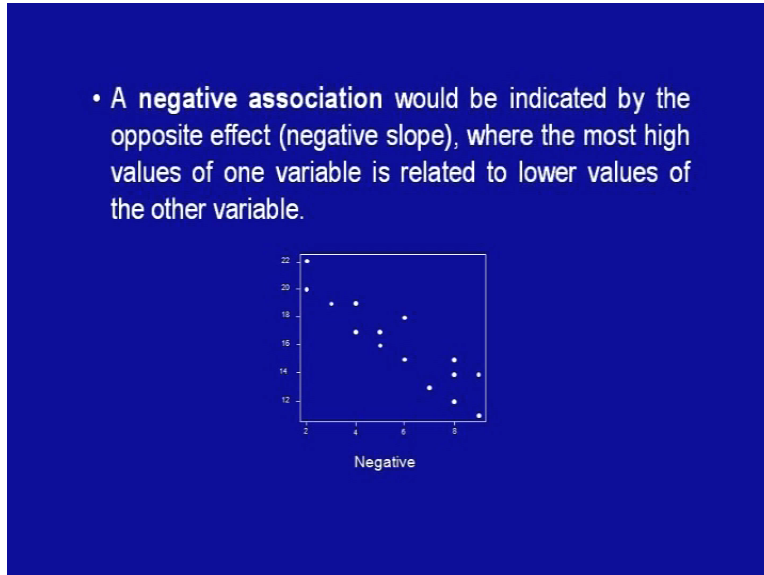
(Refer Slide Time: 26:20)

- A **positive association** between two variables would be indicated on a scatterplot by an upward trend (positive slope where small values of one variable is related to small value of the other variable).



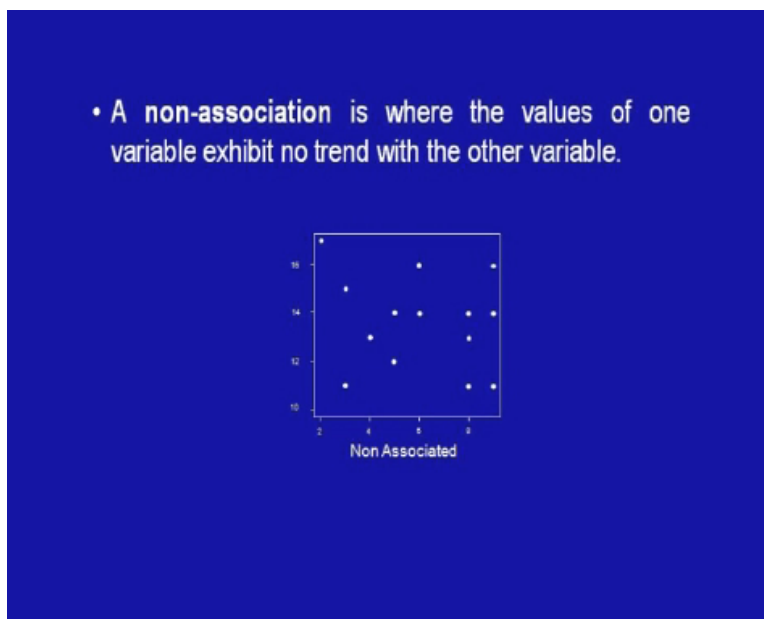
The scatter plot, based on its trend can be categorized as positive association between 2 variables; wherein, we would find an upward trend that is a positive slope where small value of one variable is related to the small value of the other variable.

(Refer Slide Time: 26:44)



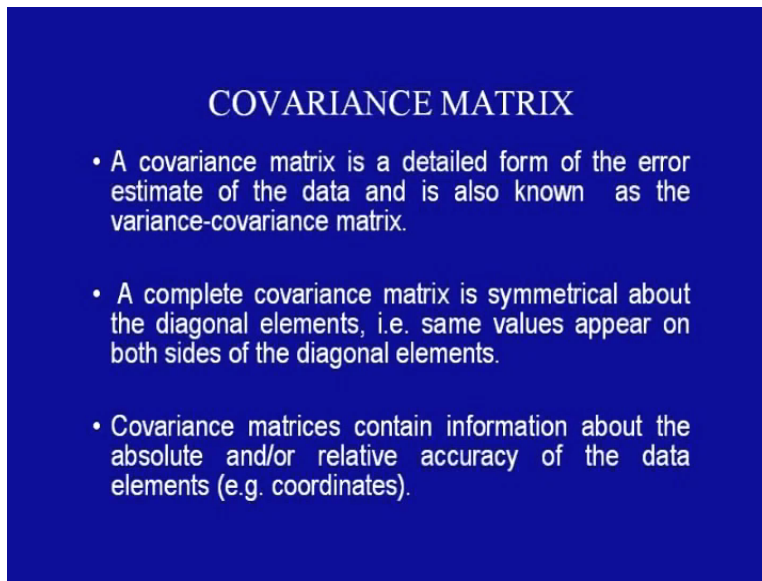
A negative association would be indicated by the opposite effect that is a negative slope where most high values of one variable is related to the lower value of the other variable.

(Refer Slide Time: 27:01 min)



Non-association is where the values of one variable exhibit no trend with the other variable. The next multivariate statistic parameter to be estimated is the covariance and this is in the form of a matrix in case where we have n number of data sets contained within a particular image. In remote sensing, we normally have more than 1 band of information.

(Refer Slide Time: 27:36)



COVARIANCE MATRIX

- A covariance matrix is a detailed form of the error estimate of the data and is also known as the variance-covariance matrix.
- A complete covariance matrix is symmetrical about the diagonal elements, i.e. same values appear on both sides of the diagonal elements.
- Covariance matrices contain information about the absolute and/or relative accuracy of the data elements (e.g. coordinates).

So, a covariance matrix is a matrix which provides us a detailed form of the error estimate of the data and is also known as the variance-covariance matrix. A complete covariance matrix is symmetrical about the diagonal elements that is same values appear on both sides of the diagonal elements. Covariance matrices contain information about the absolute and or relative accuracy of the data elements that is coordinates.

(Refer Slide Time: 28:15)

- The absolute accuracy information is contained in the diagonal matrix elements.
- Relative accuracy is a function of multiple diagonal and off-diagonal elements.
- A complete covariance matrix for N specific points in 3D space would contain $3N$ rows by $3N$ columns.

The absolute accuracy information is contained in the diagonal matrix elements. The relative accuracy is a function of the multiple diagonal and the off-diagonal terms. A complete covariance matrix of N specific points in 3D space would contain N number of rows, $3N$ rows by $3N$ columns.

(Refer Slide Time: 28:42)

- For example, for three coordinates, a covariance matrix is a 3 by 3 matrix, with the matrix rows and columns each corresponding to the three coordinates.
- For just two horizontal coordinates, a covariance matrix is a 2 by 2 matrix, with the matrix rows and columns each corresponding to the two horizontal coordinates.
- Similarly, for two image coordinates, a covariance matrix is a 2 by 2 matrix, with the matrix rows and columns each corresponding to the two image coordinates.

For example; for a 3 coordinate, a covariance matrix is a 3 by 3 matrix with the matrix rows and the columns each corresponding to the 3 coordinates. For just 2 horizontal coordinates, a covariance matrix is 2 by 2 matrix with the matrix rows and the columns each corresponding to the 2 horizontal coordinates. Similarly, for 2 image coordinates, a

covariance matrix is a 2 by 2 matrix with the matrix rows and columns each corresponding to the image coordinates.

(Refer Slide Time: 29:19)

	Band 1	Band 2	Band 3	Band 4
Band 1	σ_1^2	Cov_{12}	Cov_{13}	Cov_{14}
Band 2	Cov_{21}	σ_2^2	Cov_{23}	Cov_{24}
Band 3	Cov_{31}	Cov_{32}	σ_3^2	Cov_{34}
Band 4	Cov_{41}	Cov_{42}	Cov_{43}	σ_4^2

σ_i^2 = the square of standard deviation or variance, and

where

$$\text{Cov}_{ij} = \frac{SP_{ij}}{n-1} = \frac{n \sum_{k=1}^n (x_{ki} \times x_{kj}) - \sum_{k=1}^n x_{ki} \sum_{k=1}^n x_{kj}}{n(n-1)}$$

n = the total number of pixels in the image, and

x_{ki} = the brightness value of k^{th} pixel in band i .

x_{kj} = the brightness value of k^{th} pixel in band j .

Here, we see a typical representation of a covariance matrix for a 4 band remote sensing datasets consisting of band 1, 2 and 3. The diagonal elements are represented by the square of the standard deviation which was also known as the variance; whereas, the off-diagonal terms are represented as the interrelationship between variable 1 and 2 or between 1 and 3 or 1 and 4 and similarly, we find that the variation for the second band with respect to the other can be represented as covariance 2, 1 or 2, 3 or 2, 4.

So, what does this covariance c_{21} actually mean? This can be represented in the general form that is between 2 variables i and j , the covariance between them can be expressed as the sum of the products between the 2 variables i and j divided by n minus 1 which can be further expanded and simplified as n summation of the cross products **of the** of a single value in both the variables i and j minus the sum of the variables in band in variable 1 and the summation of variable 2 whole divided by n , n minus 1 where n is the total number of pixels in the image and x_{ki} is nothing but the brightness value of the k^{th} pixel in band i and x_{kj} is the brightness value of the k^{th} pixel in band j . Having computed the **correlation**, covariance matrix, we then compute the correlation matrix.

(Refer Slide Time: 31:26)

CORRELATION

- It is also known as **correlation coefficient**, and it indicates the strength and direction of a linear relationship between two random variables.
- In statistics, correlation refers to the departure of two variables from independence.
- There are several coefficients, measuring the degree of correlation, adapted to the nature of data.
- The best known is the Pearson Product Moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations.

Correlation matrix is also known as the correlation coefficient and it indicates the strength and direction of a linear relationship between 2 random variables. In statistics, correlation refers to the departure of 2 variables from independence. There are several coefficients measuring the degree of correlation adapted to the nature of data. The best known is the Pearson Product Moment correlation coefficient which is obtained by dividing the covariance of the 2 variables by the product of their standard deviation.

(Refer Slide Time: 32:11)

- The correlation coefficient between two variables i and j can be expressed as

$$r_{ij} = \frac{Cov_{ij}}{\sigma_i \sigma_j}$$

- where σ_i & σ_j are standard deviation of variable i and j .
- Correlation coefficient is a ratio, hence it is a dimensionless parameter and it ranges between -1 to $+1$.

So, for 2 variables i and j , the coefficient, correlation coefficient can be expressed as r_{ij} is equal to covariance_{ij} divided by the standard deviation of variable i and standard

deviation of variable j . The correlation coefficient is a ratio and hence, it is a dimensional parameter and it ranges between minus 1 to plus 1.

(Refer Slide Time: 32:46)

- The correlation is defined only if both variables have finite standard deviation and both of them are nonzero.
- It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed more than 1 in absolute value.
- The correlation is 1 in the case of an increasing linear relationship, while it is -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables.

The correlation is defined only if 2 variables have a definite standard deviation and both of them are non-zero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed more than 1 in absolute value. The correlation is 1 in the case of a increasing linear relationship; while, it is minus 1 in case of a decreasing linear relationship and some value in between in all other cases, indicating the degree of linear dependence between the 2 variables.

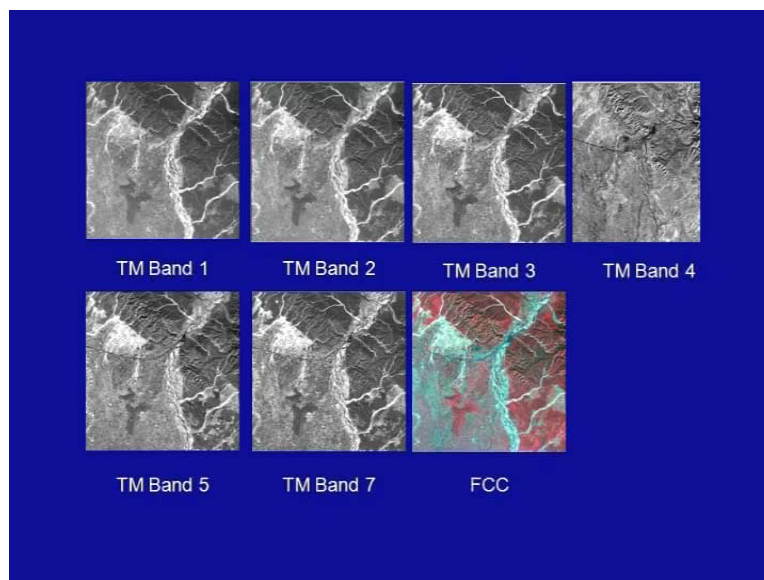
(Refer Slide Time: 33:31)

- Closer the coefficient is to either -1 or 1, stronger is the correlation between the variables.
- If the variables are independent, then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.
- In image processing, Variance-covariance and correlation matrices provide an insight to data redundancy, and are used in principal component analysis, feature selection, and classification.

Closer the coefficient is to either minus 1 or plus 1, stronger is the correlation between the variables. If the variables are independent, then the correlation is zero. But the converse is not true because the correlation coefficient detects only linear dependencies between 2 variables.

In image processing, variance - covariance and correlation matrices provide an insight to data redundancy and are used in principal component analysis, feature selection and classification. Having had a theoretical description of the various parameters which help us in identifying, **what are**, what is the characteristic of an image in terms of its statistical nature. Now, let us understand all this by undertaking a small illustrative example.

(Refer Slide Time: 34:36)



Here, a TM dataset for a particular area has been taken and the 6 band which is the thermal data has not been taken into consideration. The rest, other; that is TM band 1, 2, 3, 4, 5 and 7 have been taken for analysis. The seven image which we see in the color is the false color composite wherein, 3 bands have been taken where the brightness values of each of the bands have been projected in a particular color to create a color image. We call this false color because it is not the absolute measure of the information. But it is the reflected energy of the information that we have.

So, we will be now undertaking the analysis of the 6 bands of the TM dataset. In order to do this, it does imagine software has been used to compute the various variables.

(Refer Slide Time: 35:53)

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Minimum	44	25	13	6	1	1
Maximum	147	139	183	164	227	255
Mean	66.1	51.84	49.73	70.78	65.08	44.53
Standard Deviation	10.01	12.01	18.13	12.93	21.56	22.51
Median	65	50	45	71	62	38
Mode	67	53	36	72	59	29

This particular table shows the various statistical parameters which have been discussed earlier. These are the univariate statistics for all this 6 bands; 1, 2, 3, 4, 5 and 7 and we find that the minimum value in band 1 is 44, in band 2, it is 25, band 3 is the 13, 6 in band 4, one each in band 5 and 7.

Similarly, the maximum values are 147, 139, 183, 164, 227 and 255 respectively. The mean has been computed for each of these bands and it is 66.1, 51.84, 49.73, 70.78, 65.08 and 44.53 respectively for the 6 bands with a standard deviation of 10.01, 12.01, 18.13, 12.93, 21.56 and 22.51 respectively. For the 6 bands the median is 65, 50, 45, 71, 62, 38; while, mode is 67, 53, 36, 72, 59, 29 respectively. Having obtained these values through the software, now we have to interpret what actually these values provide us in terms of a qualitative measure of the quality of the data.

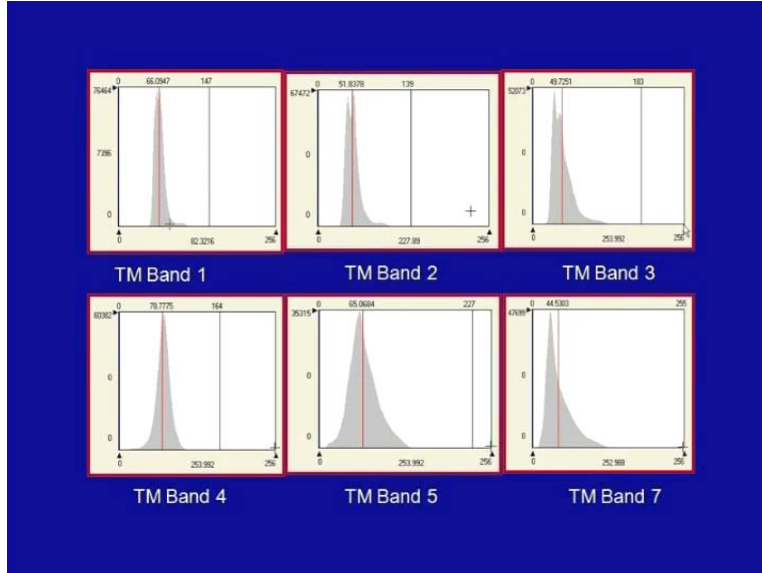
(Refer Slide Time: 37:41)

- TM Bands 1, 2 and 3 have high minimum value, an indication of atmospheric scattering, resulting in an offset to the brightness value to the darkest object in the image.
- This is not evident for TM Bands 4, 5 and 7 because atmospheric scattering is negligible in Infra-red bands.
- Hence, removal of atmospheric effects has to be undertaken for TM bands 1, 2 and 3.
- The maximum values for TM Band 1, 2, 3, 4 and 5 reveals that the full brightness scale is not being utilized and hence the image may have a poor background and may require contrast enhancement.

We observe that TM bands 1, 2 and 3 have high minimum value and indication of atmospheric scattering, resulting in an offset to the brightness values to the darkest object in the image. This is not evident in TM band 4, 5 and 7 because atmospheric scattering is negligible in infra-red bands. Hence, removal of atmospheric effects has to be undertaken for TM band 1, 2 and 3.

Well, this is a very conclusive decision or information that one has now derived. The maximum value of TM band 1, 2, 3, and 5 reveals that is the full brightness scale is not being utilized and hence the image may have a poor background and may require contrast enhancement.

(Refer Slide Time: 38:42)



Now, we look at the histogram of all the 6 bands that we have.

(Refer Slide Time: 38:50)

- The mean values for all the bands are skewed from the central value of the brightness scale.
- A critical examination of the histogram shows that especially in TM Band 1, 2 and 3, the information is concentrated within a small range of brightness values and has multi-modal distribution,
- For TM Band 4, 5 and 7, the histogram has a wider base and unimodal in nature.

The mean values of all the bands are skewed from the central value of the brightness scale. A critical examination of the histogram shows that especially in TM band 1, 2 and 3, the information is concentrated within a small range of the brightness value and has multi-modal distribution. For TM band 4, 5 and 7, the histogram has a wider base and unimodal in nature.

(Refer Slide Time: 39:24)

- This implies that in case of Bands 1, 2 and 3, a contrast enhancement if undertaken will enhance features on the brighter side of the scale, since the maximum values are close to the central value of the histogram.
- In case of the other bands, little or no improvement in the image quality may be there, if contrast enhancement is undertaken.

This implies that in case of band 1, 2 and 3; a contrast enhancement if undertaken will enhance the features on the brighter side of the scale, since the maximum values are closure to the central value of the histogram. In case of the other bands, little or no improvement in the image quality may be there if contrast enhancement is undertaken. So, another process in image processing has now been understood by examining the initial statistics of the dataset.

(Refer Slide Time: 40:05)

COVARIANCE MATRIX

Band	1	2	3	4	5	7
1	100.26					
2	116.50	144.32				
3	171.10	211.67	328.52			
4	32.58	49.15	60.65	167.15		
5	167.45	215.05	338.59	134.86	464.93	
7	192.27	242.01	383.70	79.98	457.03	506.79

Now, we examine the covariance matrix. Here, we find that the diagonal terms are providing are nothing but the square of the standard deviation or what we call it as

variance. Variance is nothing but it is like entropy as in thermodynamics which tells us the amount of heat energy.

Similarly, in remote sensing, this tells us the amount of information content available in each of the datasets. So, if we really look at these diagonal terms; band 1, probably has the lowest value that means it has the lowest information content in comparison to band 5 and 7 where the values are very high.

The off diagonal terms which indicate the deviation from or the deviation from the central value and they represent a form of dependency or independence from each variable. However, the analyst may not be able to derive any possible information by only examining the covariance matrix because **there is** this is an unscaled information. In order to have a scaled information, this covariance matrix is subsequently converted into a correlation matrix.

(Refer Slide Time: 41:51)

CORRELATION MATRIX

Band	1	2	3	4	5	7
1	1.00					
2	0.97	1.00				
3	0.94	0.97	1.00			
4	0.25	0.32	0.26	1.00		
5	0.78	0.83	0.87	0.48	1.00	
7	0.85	0.89	0.94	0.27	0.94	1.00

When we look at the correlation matrix; probably, all the information has now been scaled between the range of minus 1 to plus1 as already indicated earlier and what we find that all the diagonal terms are having a value of 1 because we are comparing the same variable to itself.

However, when we look at the comparison between or the correlation between band 1 and band 2 or between band 1 and band 3; what we find is that the value is close to 1. In case of band 1 and 2, the correlation is of the order of 0.97 or in case of band 1 and 3, it is of the order of 0.94. Similarly, between band 2 and 3, we find it to be of the order of 0.97.

Well, what does this mean? Well, it is a clear indication that the information contained between these variables that is the TM band 1, 2 and 3 are similar; more or less similar and thus, there is no additional or extra information which is present in this datasets

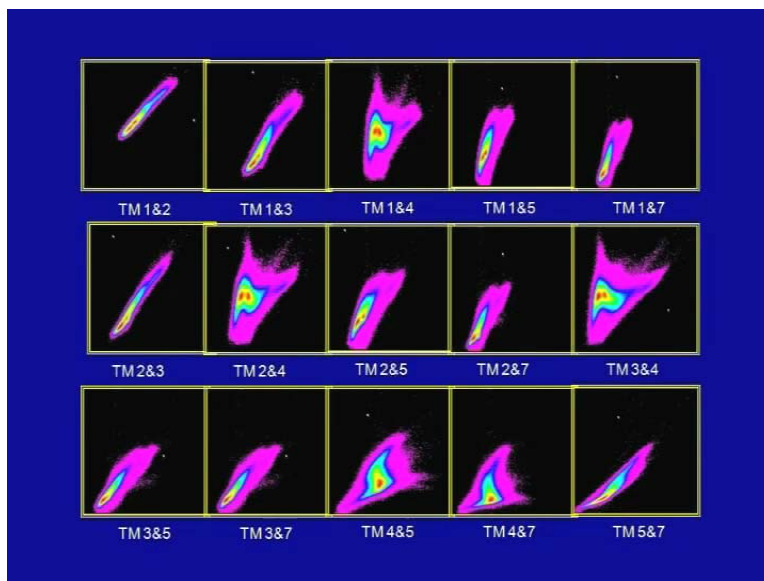
which gives the analyst another indication that out of these 3 datasets; one, anyone of them can be used subsequently for classification of the image.

A very unique trend may be observed with TM band 4 and it reveals that the correlation amongst the other 5 datasets which is there is probably of the lowest order of the order of 0.25 and so on; which means that band 4, when it is compared to all other bands, it contains all together a very different information. And, this is a very positive conclusion which can be drawn and that is that when we are proceeding for classification of the image; wherein, we will be using a multispectral datasets that is large number of datasets would be taken. A minimum of 3 is required; so it means band 4 is a strong candidate in the whole analysis procedure.

However, when we look at the correlation between band 5 and 7, what we find; it is of the order of 0.94 that means these 2 data are strongly correlated and hence are redundant. And thus, now what we can understand is that out of band 1, 2 and 3, either of them that is either 1 of them can be selected. Amongst 5 and 7, one of them can be selected and band 4 being one which is totally distinct in nature.

Now, we look at the scatter plots. Here, the scatter plots between all the datasets have now been plotted.

(Refer Slide Time: 45:19)



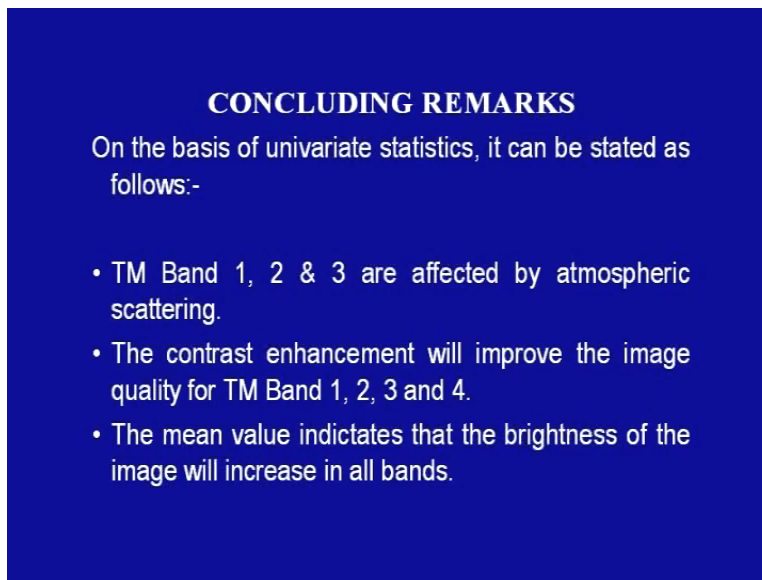
One can see; in some of the cases, the spread of the information is like a very fine line and in some cases, it spreads out into a triangular shape. Since, we have already derived one information that is the correlation between band 1 and 2 or band 1 and 3 or between 2 and 3 is very high, the same trend is also depicted in the scatter plots.

So, if we look at the scatter plots between TM 1 and 2 or 1 and 3 or between 2 and 3, we find the spread is along a line, along an elongated line but at an inclination from both the

axes. This is primarily because of the offsets that may have been produced by the atmospheric effects which has been reflected in the terms of minimum value.

However, if you look at the scatter plot between TM band 4 and the other way other TM bands 1, 2, 3, 5 and 7; we find that the spread of the information is in a triangular shape that means there is a very plausible information which is available and the spread of the information is very well distinct and can be utilized subsequently by the analyst to identify the plausible information classes which are present there.

(Refer Slide Time: 47:07)



CONCLUDING REMARKS

On the basis of univariate statistics, it can be stated as follows:-

- TM Band 1, 2 & 3 are affected by atmospheric scattering.
- The contrast enhancement will improve the image quality for TM Band 1, 2, 3 and 4.
- The mean value indicates that the brightness of the image will increase in all bands.

So, based on the observations that we have had up till now, in this particular example; we can make some concluding remarks. That is on the basis of univariate statistics, it can be stated as follows; TM band 1, 2 and 3 are affected by atmospheric scattering, the contrast enhancement will improve the image quality of TM band 1, 2, 3 and 4. The mean value indicates that the brightness of the image will increase in all the bands.

(Refer Slide Time: 47:42)

The multivariate statistic shows that

- TM Band 4, 5 and 7 have more information content compared to TM 1, 2, 3
- TM 4 Band has low correlation with all bands, while TM Band 1, 2, 3 are highly correlated to each other and thus either of them can be used.
- Scatter plots show that TM 4 with all other bands has a good spread of information.

The multivariate statistic shows that TM band 4, 5 and 7 have more information content compared to TM band 1, 2 and 3. TM band 4 has low correlation with all bands while TM band 1, 2 and 3 are highly correlated to each other and thus either of them can be used. Scatter plots show that TM band 4 with all other bands has a good spread of information.

With this, we have now made a assessment of the quality of the data. In my next session, I would be looking at some of the pre processing which are required in order to provide a strong special correlation of the dataset to the actual earth of which the image has been taken.

Thank you.