**STATISTICS FOR DATA SCIENCEData Visualization and Interpretation** 

**Prof. Uma D**

**Prof. Silviya Nancy J**

**Prof. Suganthi S**

Department of Computer Science and Engineering

**STATISTICS FOR DATA SCIENCE**

**Data Visualization and Interpretation – Histogram**

**Prof. Uma D**

**Prof. Silviya Nancy J**

**Prof. Suganthi S**

****

**STATISTICS FOR DATA SCIENCE** 

**Data Visualization**

Data visualization is a **BIG buzz word** these days, but what does it actually mean ????

At a basic level, data is just information — **facts, figures, words, percentages, measurements, and observations**, but it’s just computerized information.

In order for you to make it useful, you need to **find creative ways** to **make it user friendly** for your audience.

This is where the art of **data visualization** comes in! Source: nutsandboltspeedtraining.com

**STATISTICS FOR DATA SCIENCE** 

**Data Visualization**

**Data visualization** is the **secret art** of **turning data** into **visual graphics** that people can understand (**graphs, charts, info graphics**, etc.).

Words may be mightier than the sword, but in a battle for our brains, **visual images win every time**. - Colin Ware.

Source: nutsandboltspeedtraining.com

**STATISTICS FOR DATA SCIENCETypes of Visualizations** 

• Histogram

• Box plot

• Scatter plot

• Bar chart

• Heat map

**STATISTICS FOR DATA SCIENCE** 

**Data Distribution**

• To understand some of the fundamental concepts of statistical analysis, it is important to **appreciate** the **importance** of the **distribution of data points** in the sample.

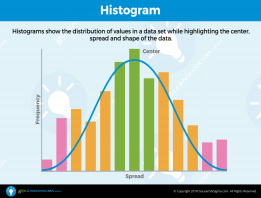
• **Data type** and the **distribution pattern** of their values **influence** the **choice of appropriate statistical tests**. • Emphasis will be placed on the **normal, or Gaussian, distribution.**

• This is an important distribution to understand because the **assumption of this distribution** underlies the **use of many** common **statistical tests**.

Source: Sciencedirect.com

**STATISTICS FOR DATA SCIENCE** 

**Histogram**

• **Histograms** show the **distribution of data values** in a data set while highlighting the **center, spread and shape** of the data. 

• Histograms, also known as **Frequency Plots**, are a visual displays of **how much variation exists** in a process. • They highlight the **center of the data** measured as the **mean, median and mode.**

• They highlight the **distribution of the data** measured as the **range and standard deviation.**

• The **shape of a Histogram** indicates whether the distribution is **normal, bi-modal, or skewed.**

Source: goLEANSIXSIGMA.com **Text Book – Chapter 1.3 – Pg. No - 27**

**STATISTICS FOR DATA SCIENCE** 

**Histogram**

A histogram is used to **summarize discrete or continuous data.**

In other words, it provides a visual interpretation of numerical data by showing the **number of data points** that fall within a specified range of values (called “**bins**”).

It is similar to a vertical bar graph.

However, a histogram, unlike a vertical bar graph, shows **no gaps between the bars.**

**STATISTICS FOR DATA SCIENCE** 

**Importance of Histogram**

A histogram is a **graphical display of data** using bars of different heights.

In a histogram, each bar groups **numbers into ranges**. **Taller bars** show that **more data falls** in that **range.**

A histogram displays the **shape and spread** of **continuous** sample data.

A histogram is used to **summarize discrete or continuous data.**

**STATISTICS FOR DATA SCIENCE** 

**Example – Construction of Histogram for PM emissions of 62 Vehicles**

Class Interval Frequency Relative Frequency Density 1 - 3 12 0.1935 0.0968 3 – 5 11 0.1774 0.0887 5 – 7 18 0.2903 0.1452 7 – 9 9 0.1452 0.0726

9 – 11 5 0.0806 0.0403 11 – 13 1 0.0161 0.0081 13 – 15 2 0.0323 0.0161 15 – 17 0 0.0000 0.0000 17 – 19 2 0.0323 0.0161 19 – 21 1 0.0161 0.0081 21 – 23 0 0.0000 0.0000 23 - 25 1 0.0161 0.0081

**Text Book – Chapter 1.3 – Pg. No - 27**

**STATISTICS FOR DATA SCIENCE** 

**Example – Class Intervals**

The intervals in the left-hand column are called **class intervals.**

**They divide the** sample into groups.

For most histograms, the class intervals all have the same width. In the example all classes have width 2.

The notation 1–< 3, 3–< 5, and so on, indicates that a point on the boundary will go into the class on its right.

For example, a sample value equal to 3 will go into the class 3–< 5, not 1–< 3.

**Text Book – Chapter 1.3 – Pg. No - 27**

**STATISTICS FOR DATA SCIENCE** 

**Example – Width of Class Intervals**

There is no hard-and-fast rule as to how to choose the endpoints of the class intervals.

It is good to have more intervals rather than fewer, but it is also good to have large numbers of sample points in the intervals.

When the number of observations n is large (several hundred or more), some have suggested that reasonable starting points for the number of classes may be log2n or 2n1/3.

When the number of observations is smaller, more classes than these are often needed.

**Text Book – Chapter 1.3 – Pg. No - 27**

**STATISTICS FOR DATA SCIENCE** 

**Example – Frequency and Relative Frequency**

The column labeled “Frequency” in Table presents the numbers of data points that fall into each of the class intervals.

The column labeled “Relative Frequency” presents the frequencies divided by the total number of data points, which for these data is 62.

The relative frequency of a class interval is the proportion of data points that fall into the interval.

Note that since every data point is in exactly one class interval, the relative frequencies must sum to 1.

**Text Book – Chapter 1.3 – Pg. No - 28**

**STATISTICS FOR DATA SCIENCE** 

**Example - Density**

Finally, the column labeled “Density” presents the relative frequency divided by the class width.

In this case all classes have width 2, so the densities are found by dividing the relative frequencies by 2.

Note that when the classes are of equal width, the frequencies, relative frequencies, and densities

are proportional to one another.

**Text Book – Chapter 1.3 – Pg. No - 28**

**STATISTICS FOR DATA SCIENCE** 

**Example – Interpretation of Histogram**

The units on the horizontal axis are the units of the data, in this case grams per gallon. 

Each class interval is represented by a rectangle.

When the class intervals are of equal width, the heights of the rectangles may be set equal to the frequencies, the relative frequencies, or the densities.

Since these three quantities are proportional, the shape of the histogram will be the same in each case.

**Text Book – Chapter 1.3 – Pg. No - 28**

**STATISTICS FOR DATA SCIENCE** 

**Bins in Histogram**

Histograms are based on area, not height of bars.

In a histogram, it is the area of the bar that indicates the frequency of occurrences for each bin.

In statistics, the Freedman – Diaconis rule can be used to select the size of the bins to be used in a histogram



**STATISTICS FOR DATA SCIENCE** 

**Equal Bins / Class Widths**

Suppose we are looking at the history grades of students in 10th grade and have the classes corresponding to letter grades: A, B, C, D, F. The number of each of these grades gives us a frequency for each class:



**Frequency**

**Relative Frequency**

Source: www.thoughtco.com

**STATISTICS FOR DATA SCIENCE** 

**Unequal Bin / Class Widths**

Histograms are drawn with class intervals of differing widths rarely.

When the **class intervals** are of **unequal widths**, the **heights** of the **rectangles or bars** must be set equal to the densities.

**Compute** the **density** for each class, according to the formula,



The **areas of the rectangles** will then be the **relative frequencies**.

**Text Book – Chapter 1.3 – Pg. No - 28**

**STATISTICS FOR DATA SCIENCE** 

**Parts of Histogram**

**The title:** The title **describes the information** included in the histogram. 

**X-axis:** The X-axis are intervals that show the **scale of values** which the measurements fall under.

**Y-axis:** The Y-axis shows the **number of times** that the values occurred within the intervals set by the X-axis.

**The bars:** The **height of the bar** shows the **number of times** that the values occurred within the interval, while the **width** of the bar shows the **interval that is covered**.

For a histogram with **equal bins**, the **width** should be the **same** across all bars.

Source: Mathispower4u

**STATISTICS FOR DATA SCIENCE** 

**Distributions of a Histogram – Normal Distribution**

**A normal distribution:**

In a normal distribution, points on one side of the average are as likely to occur as on the other side of the average.



Source: corporatefinanceinstitute.com

**STATISTICS FOR DATA SCIENCE** 

**Distributions of a Histogram – Bimodal Distribution**

**A bimodal distribution:** In a bimodal distribution, there are two peaks.

In a bimodal distribution, the data should be separated and analyzed as separate normal distributions.



Source: corporatefinanceinstitute.com

**STATISTICS FOR DATA SCIENCE** 

**Distributions of a Histogram – Right-Skewed Distribution**

**A right-skewed distribution:** A right-skewed distribution is also called a positively skewed distribution. 

In a right-skewed distribution, a large number of data values occur on the left side with a fewer number of data values on the right side.

A right-skewed distribution usually occurs when the data has a range boundary on the left-hand side of the histogram. For example, a boundary of 0.

Source: corporatefinanceinstitute.com

**STATISTICS FOR DATA SCIENCE** 

**Distributions of a Histogram – Left-Skewed Distribution**

**A left-skewed distribution:** A left-skewed distribution is also called a negatively skewed distribution. 

In a left-skewed distribution, a large number of data values occur on the right side with a fewer number of data values on the left side.

A right-skewed distribution usually occurs when the data has a range boundary on the right-hand side of the histogram. For example, a boundary such as 100.

Source: corporatefinanceinstitute.com

**STATISTICS FOR DATA SCIENCE** 

**Distributions of a Histogram – Random Distribution**

**A random distribution:**

A random distribution lacks an apparent pattern and has several peaks. 

In a random distribution histogram, it can be the case that different data properties were combined.

Therefore, the data should be separated and analyzed separately.

Source: corporatefinanceinstitute.com

**STATISTICS FOR DATA SCIENCE** 

**Summary**

Choose boundary points for the class intervals.

Compute the frequency and relative frequency for each

class.

Compute the density for each class, according to the formula



Draw a rectangle for each class. If the classes all have the

same width, the heights of the rectangles may be set equal

to the frequencies, the relative frequencies, or the densities.

If the classes do not all have the same width, the heights of

the rectangles must be set equal to the densities.**Text Book – Chapter 1.3 – Pg. No - 30**

**STATISTICS FOR DATA SCIENCE** 

**Symmetry**

A distribution is said to be symmetrical if it can be divided into two equal sizes of the same shape.



Source: Simple Learning Pro **Text Book – Chapter 1.3 – Pg. No - 31**

**STATISTICS FOR DATA SCIENCE** 

**Distribution - Symmetry**

When a histogram is roughly **symmetric**, the mean and the median are approximately equal.



Source: Simple Learning Pro **Text Book – Chapter 1.3 – Pg. No - 31**

**STATISTICS FOR DATA SCIENCE** 

**Distribution – Left Skewed**

When a histogram is **left-skewed**, the mean is less than the median.

Source: Simple Learning Pro **Text Book – Chapter 1.3 – Pg. No - 31**

**STATISTICS FOR DATA SCIENCE** 

**Distribution – Right Skewed**

When a histogram is **right-skewed**, the mean is greater than the median.

Source: Simple Learning Pro **Text Book – Chapter 1.3 – Pg. No - 31**

**STATISTICS FOR DATA SCIENCE Unimodal and Bimodal Histogram** 

A histogram with only one peak is what we call **unimodal**. 

If a histogram has two peaks 

then we say that it is **bimodal**. **Text Book – Chapter 1.3 – Pg. No - 31**

**STATISTICS FOR DATA SCIENCE** 

**Outlier**

An outlier falls far from the rest of the data.

**THANK YOU** 

**Prof. Uma D**

**Prof. Silviya Nancy J**

**Prof. Suganthi S**

Department of Computer Science and Engineering