# Dimensionality Reduction

Data encoding or transformation methods are applied – to obtain either a reduced or compressed representation of the original data

Lossless methods
Lossy methods

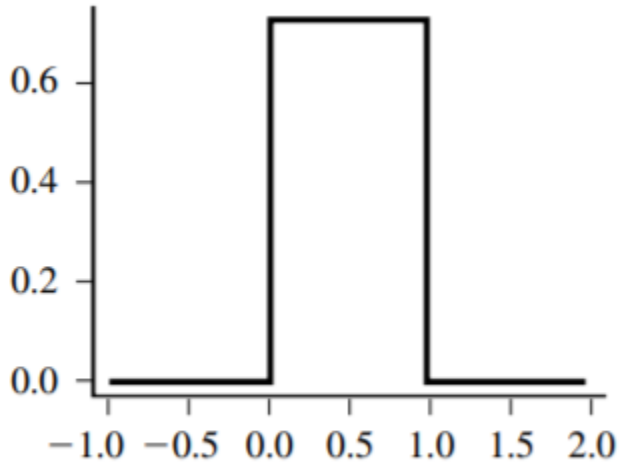
# Effective methods for Lossy Dimensionality Reduction


Wavelet Transformation

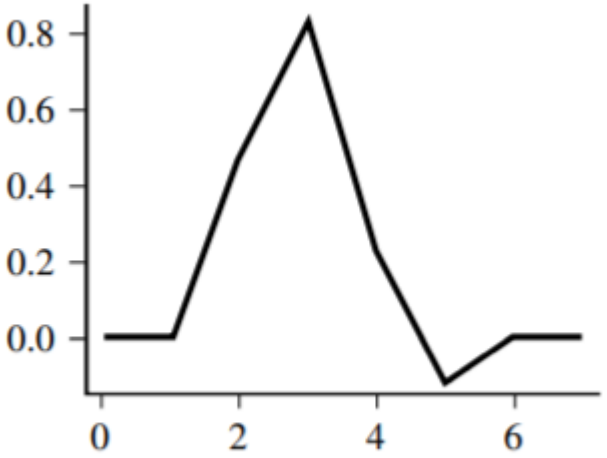
Principal Components Analysis

# Wavelet Transforms

**The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector X, transforms it to a numerically different vector, X0 , of wavelet coefficients**



(a) Haar-2



(b) Daubechies-4

# Principal Components Analysis

**The original data are thus projected onto a much smaller space, resulting in dimensionality reduction.**

**Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes**

**searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$.**

# Principal Components Analysis

**The basic procedure is as follows:**

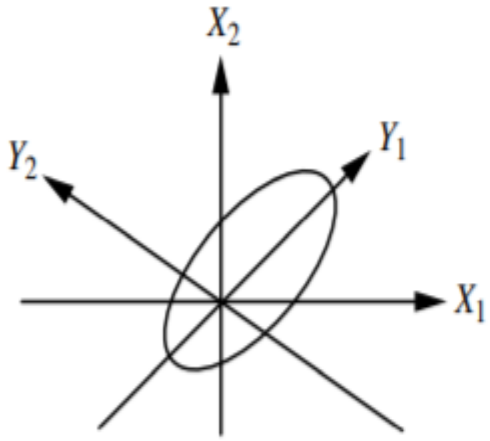
1. **The input data are normalized, so that each attribute falls within the same range**

2. **PCA computes k orthonormal vectors that provide a basis for the normalized input data.**

3. **The principal components are sorted in order of decreasing "significance" or strength.**

4. **Because the components are sorted according to decreasing order of "significance," the size of the data can be reduced by eliminating the weaker components**

# Principal Components Analysis




Principal components analysis. $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are the first two principal components for the given data.

# Numerosity Reduction

**Original data replaced by alternative, smaller data representations.**

**Parametric methods**

**Non-parametric methods**

# Parametric Methods

**Store the data parameters instead of actual data**

**Regression**

**Log-Linear models**

# Regression

## Linear Regression

**Data are modelled to fit a straight line**

$$y=wx+b$$

**w and b are regression coefficients**
**x- random variable**
**y-response variable**

## Multi-Linear Regression

**Which allows a response variable, y, to be modelled as a linear function of two or more predictor variable**

## Log-Linear Regression

**Log-linear models can be used to estimate the probability of each point in a multidimensional space for a set of discretized attributes**

This allows a higher-dimensional data space to be constructed from lower dimensional spaces

# Non-Parametric Methods

**Histograms**

**Clustering**

**Sampling**

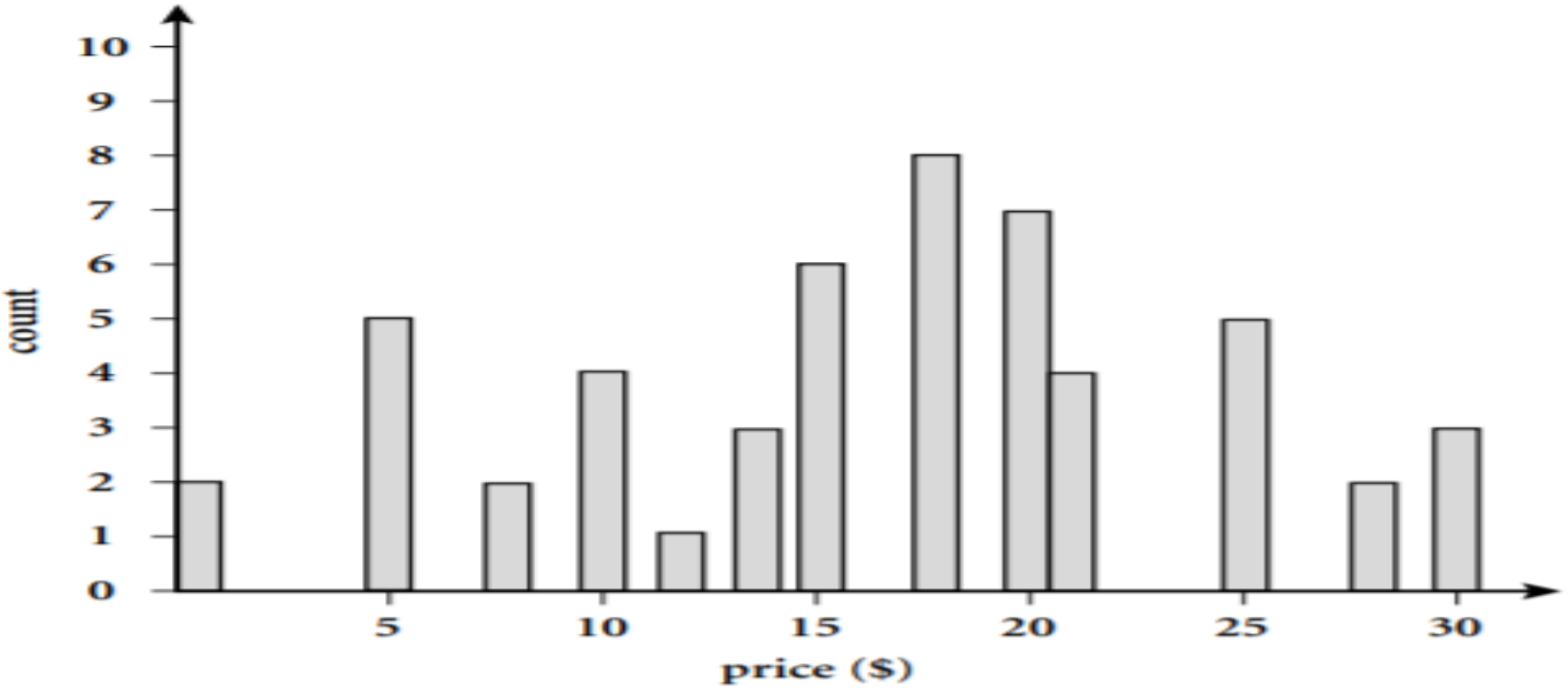
# Non-Parametric Methods

## Histograms

A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, or buckets. If each bucket represents only a single attribute-value/frequency pair, the buckets are called singleton buckets.

# Histogram : Example

The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30



# Histogram (Partitioning Rules)

**Equal-Width**

**Equal-Frequency**

**V-optimal**

**MaxDiff**

# Histogram

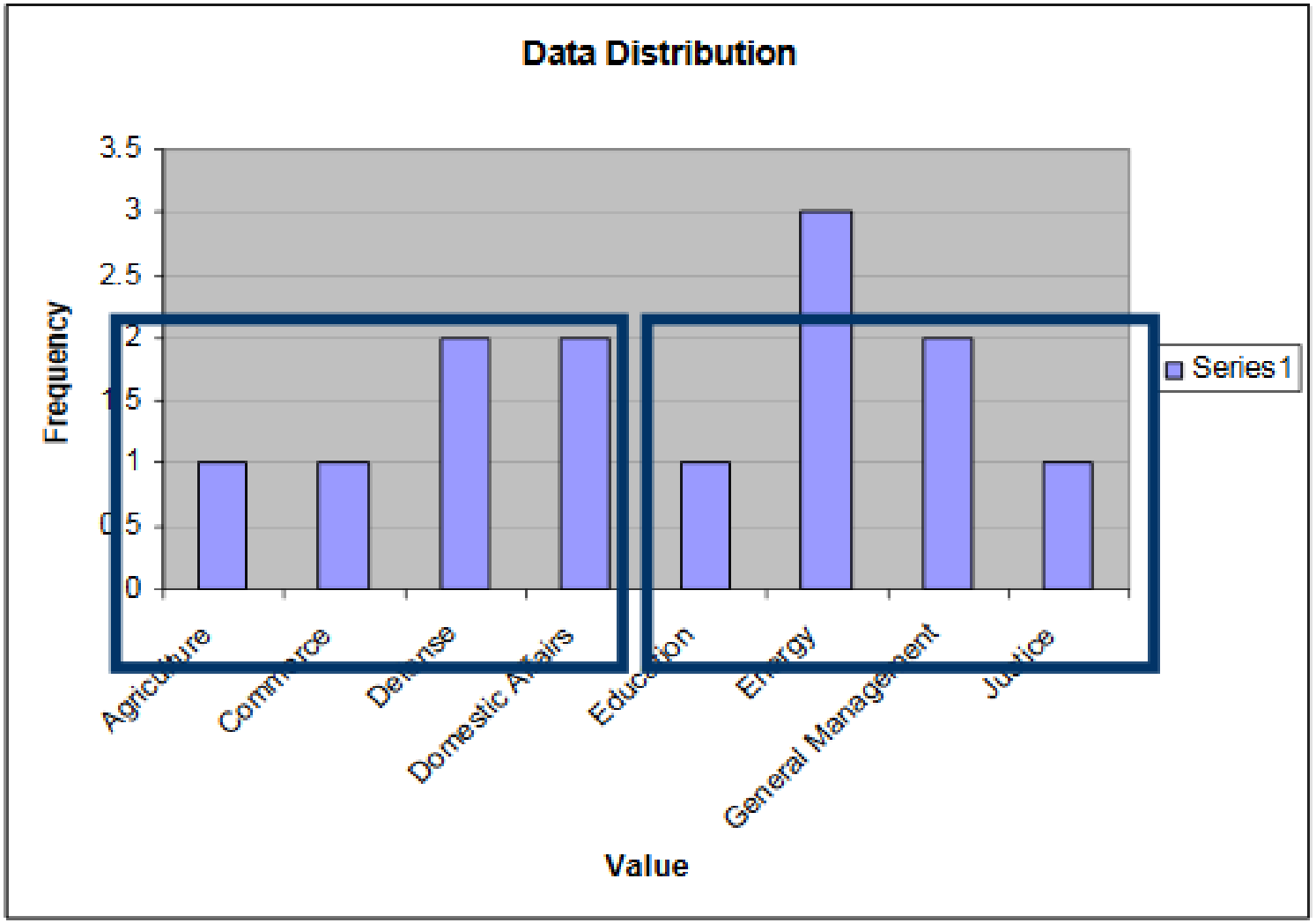
Histograms as approximations of data distribution

Data distribution is a set of (attribute value, frequency)pairs

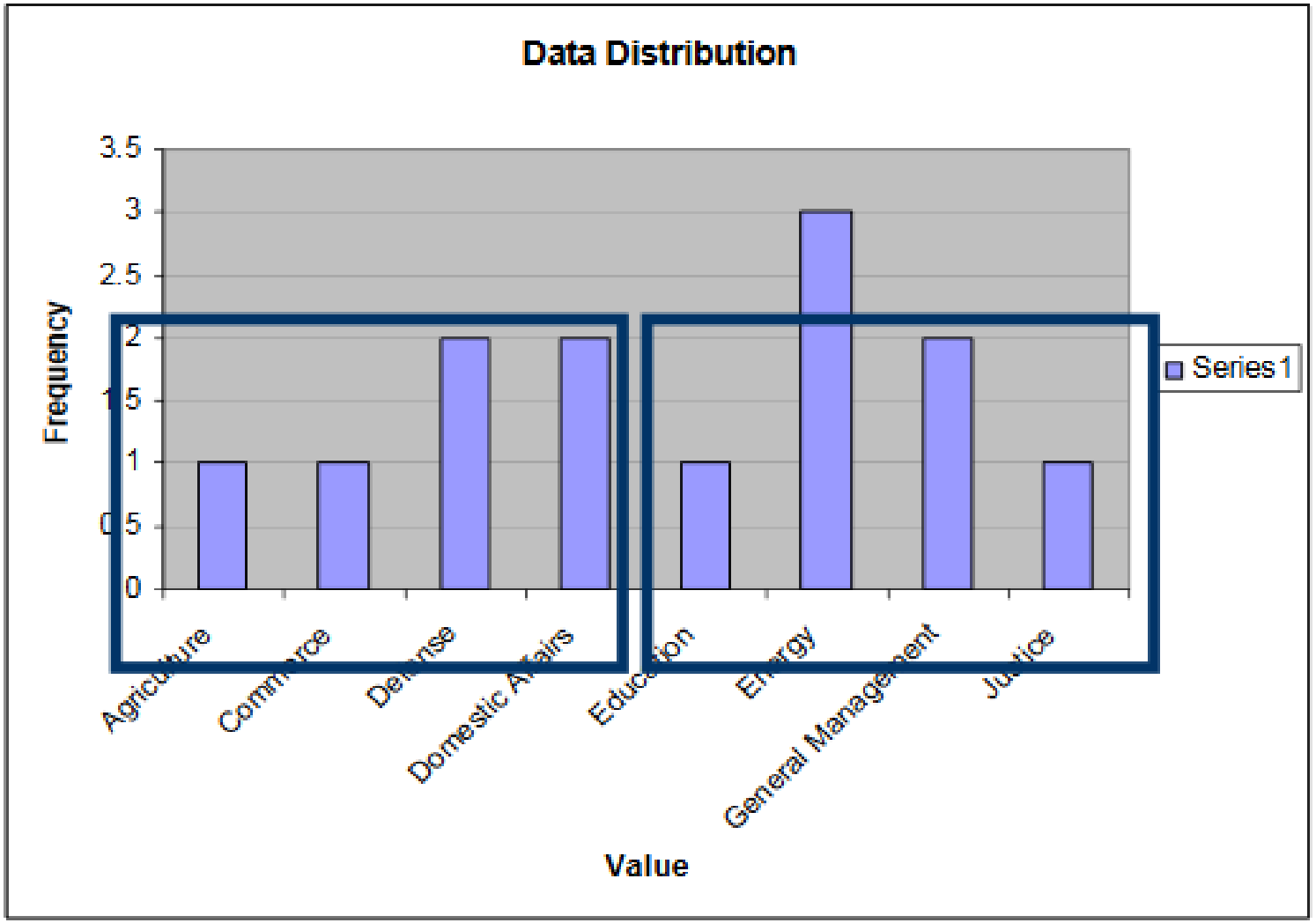
| Name | Salary | Department |
|---|---|---|
| Zeus | 100K | General Management |
| Poseidon | 80K | Defense |
| Pluto | 80K | Justice |
| Aris | 50K | Defense |
| Ermis | 60K | Commerce |
| Apollo | 60K | Energy |
| Hefestus | 50K | Energy |
| Hera | 90K | General Management |
| Athena | 70K | Education |
| Aphrodite | 60K | Domestic Affairs |
| Demeter | 60K | Agriculture |
| Hestia | 50K | Domestic Affairs |
| Artemis | 60K | Energy |

| Department | Frequency |
|---|---|
| General Management | 2 |
| Defense | 2 |
| Education | 1 |
| Domestic Affairs | 2 |
| Agriculture | 1 |
| Commerce | 1 |
| Justice | 1 |
| Energy | 3 |

# Histogram : Example

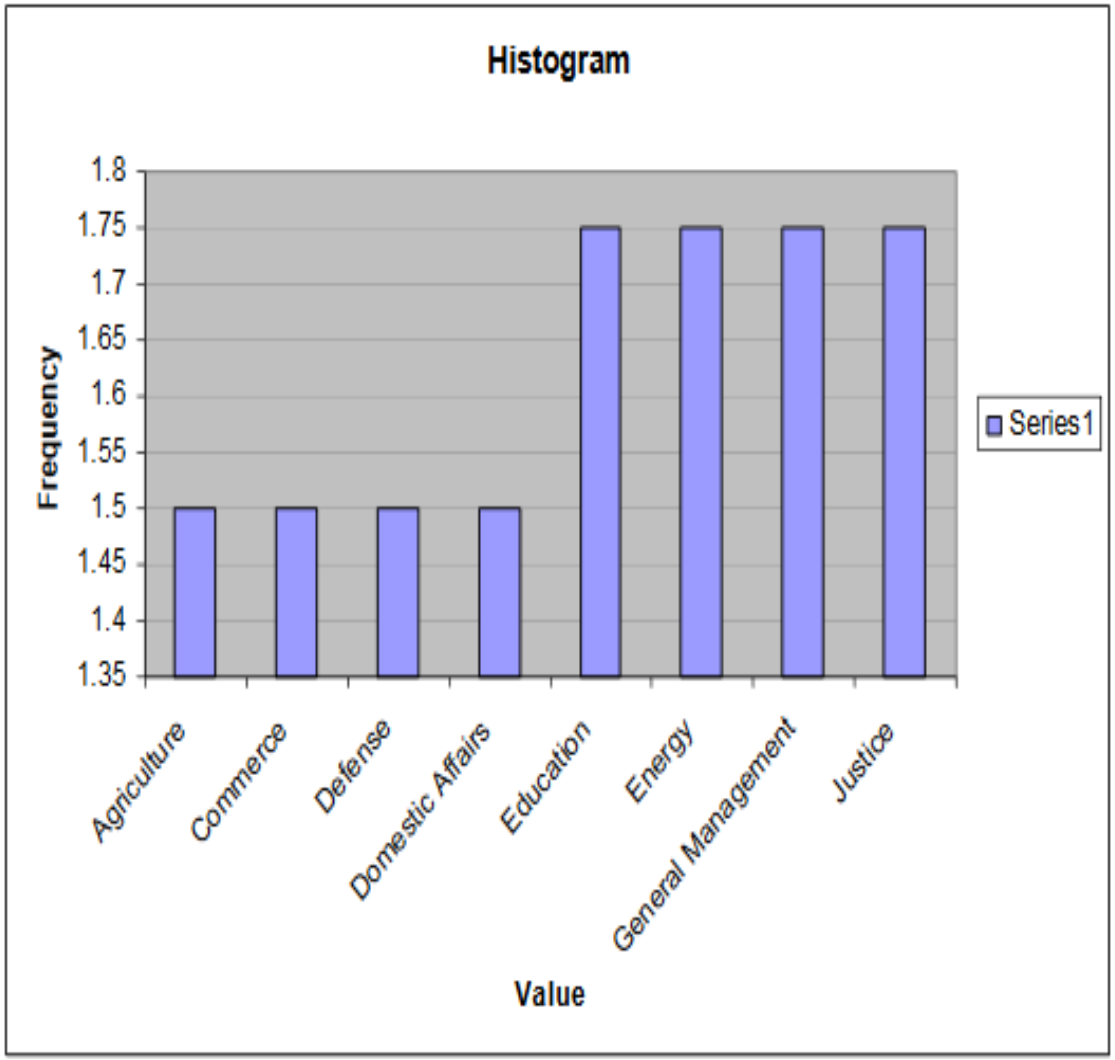


# Histogram : Example



# Histogram : Example

| Department | Histogram H1 | |
|---|---|---|
| | Frequency in Bucket | Approximate Frequency |
| Agriculture | 1 | 1.5 |
| Commerce | 1 | 1.5 |
| Defense | 2 | 1.5 |
| Domestic Affairs | 2 | 1.5 |
| Education | 1 | 1.75 |
| Energy | 3 | 1.75 |
| General Management | 2 | 1.75 |
| Justice | 1 | 1.75 |

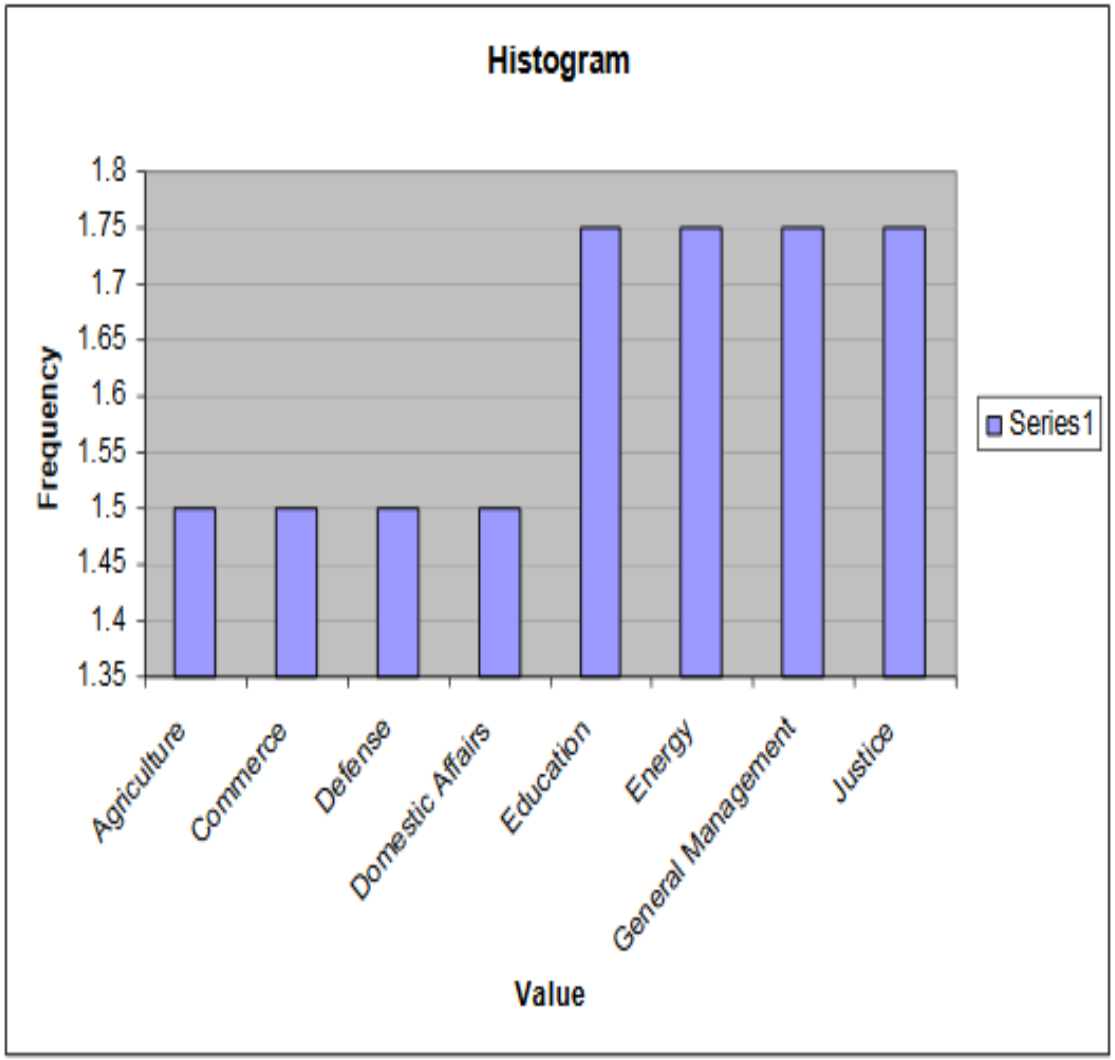


# Histogram : Example

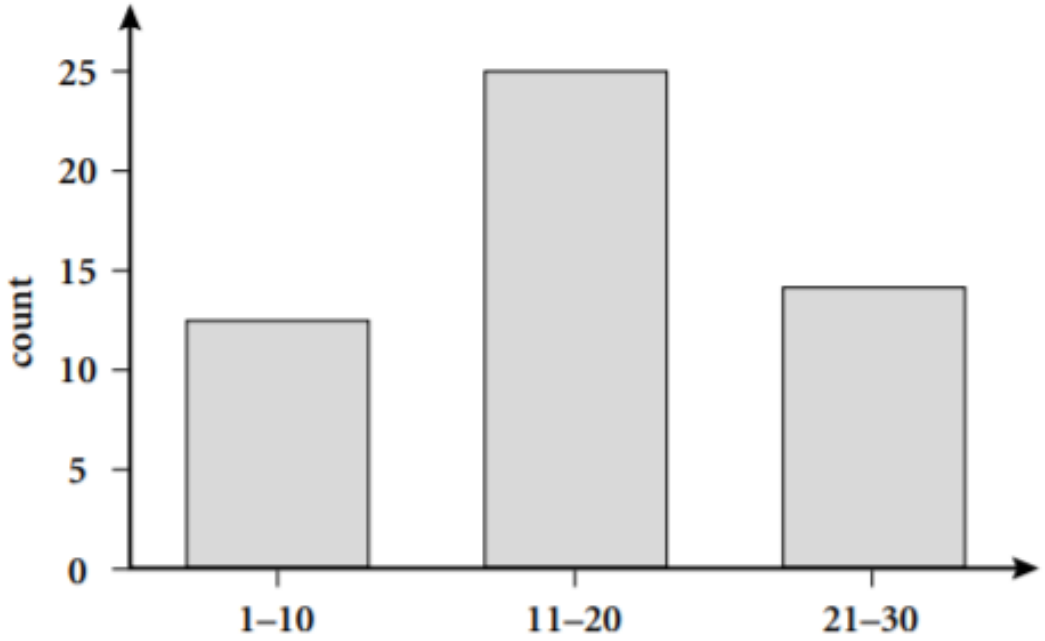
| Department | Histogram H1 | |
|---|---|---|
| | Frequency in Bucket | Approximate Frequency |
| Agriculture | 1 | 1.5 |
| Commerce | 1 | 1.5 |
| Defense | 2 | 1.5 |
| Domestic Affairs | 2 | 1.5 |
| Education | 1 | 1.75 |
| Energy | 3 | 1.75 |
| General Management | 2 | 1.75 |
| Justice | 1 | 1.75 |

# Histogram : Example

| Department | Histogram H1 | |
|---|---|---|
| | Frequency in Bucket | Approximate Frequency |
| Agriculture | 1 | 1.5 |
| Commerce | 1 | 1.5 |
| Defense | 2 | 1.5 |
| Domestic Affairs | 2 | 1.5 |
| Education | 1 | 1.75 |
| Energy | 3 | 1.75 |
| General Management | 2 | 1.75 |
| Justice | 1 | 1.75 |



# Examples



# Histogram : Example –V-optimal

Take a simple set of data, for example, a list of integers:
1, 3, 4, 7, 2, 8, 3, 6, 3, 6, 8, 2, 1, 6, 3, 5, 3, 4, 7, 2, 6, 7, 2

Compute the value and frequency pairs

(1, 2), (2, 4), (3, 5), (4, 2), (5, 1), (6, 4), (7, 3), (8, 2)

"V-optimality rule states that the cumulative weighted variance of the buckets must be minimized"

# Histogram : Example –V-optimal

Option 1: Bucket 1 contains values 1 through 4.
Bucket 2 contains values 5 through 8.

Bucket 1:
Average frequency 3.25
Weighted variance 2.28

Bucket 2:
Average frequency 2.5
Weighted variance 2.19

Sum of Weighted Variance 4.47

# Histogram : Example –V-optimal

Option 2: Bucket 1 contains values 1 through 2.
Bucket 3 contains values 5 through 8.

Bucket 1:
Average frequency 3
Weighted variance 1.41

Bucket 2:
Average frequency 2.88
Weighted variance 3.29

Sum of Weighted Variance 4.70

Option1 : 4.47, Option 2: 4.70
Hence, Option 1 is selected as per V-optimal rule