Non-Parametric Methods

Histograms

Clustering

Sampling

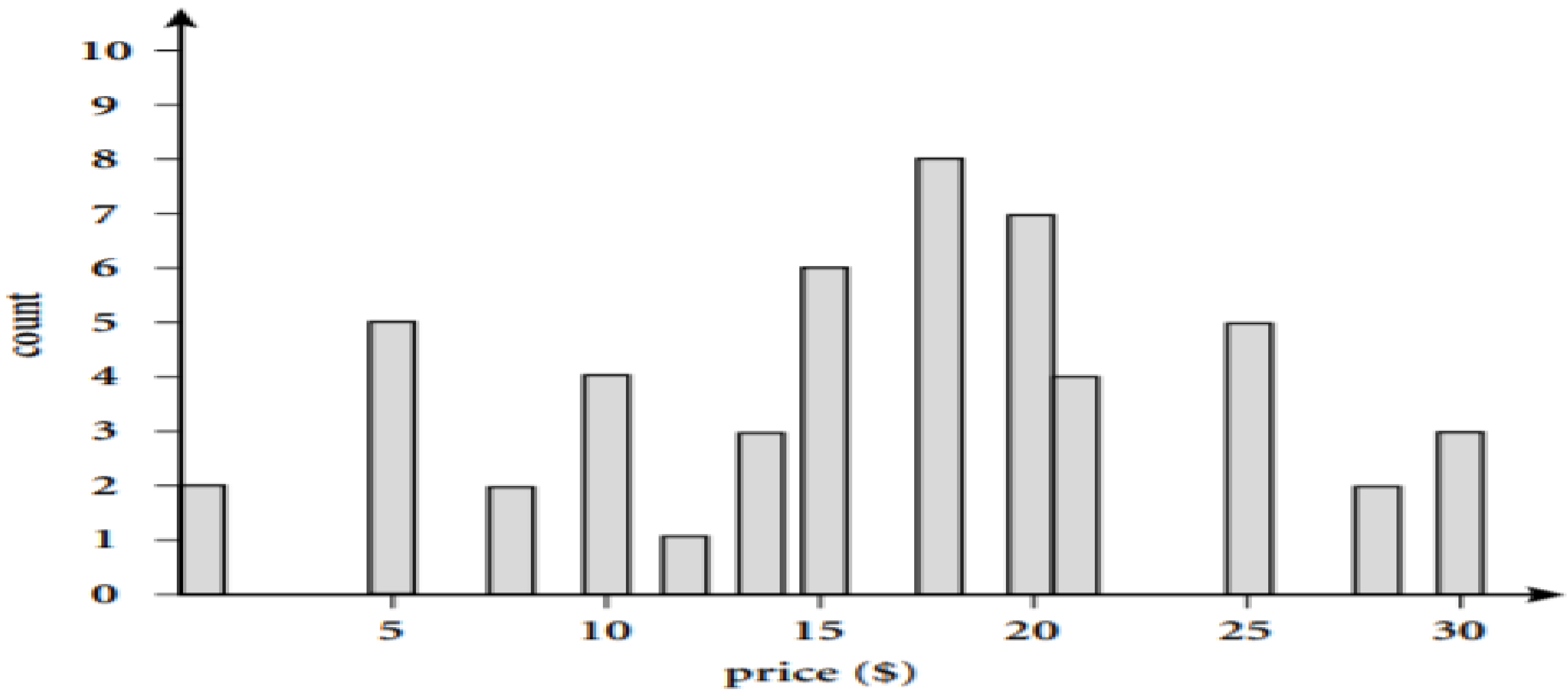
Non-Parametric Methods

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, 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

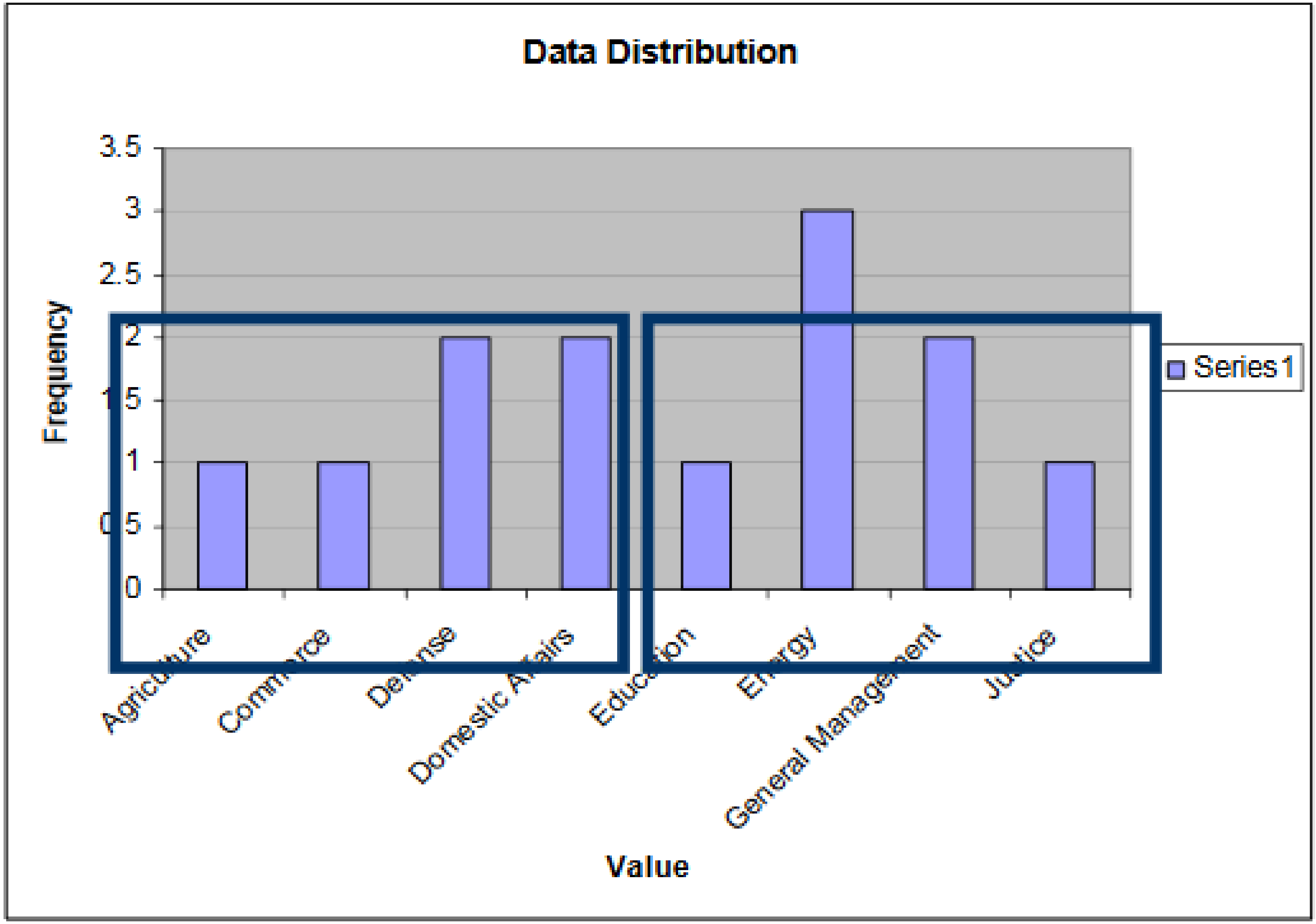
Histograms as approximations of data distribution

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

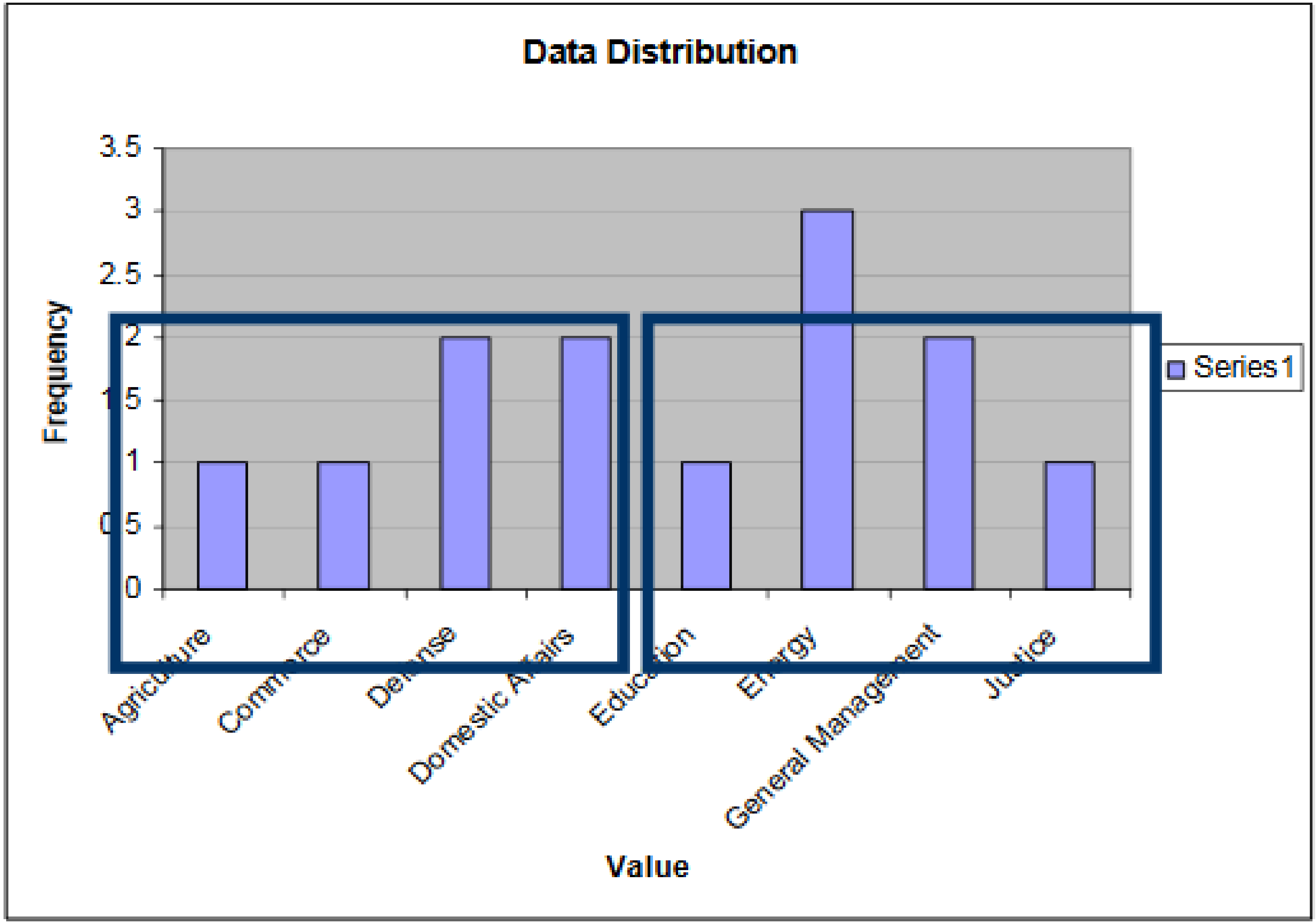
Name	Salary	Department
Zeus	100K	General Management
Poseidon	80K	Defense
Pluto	80K	Justice
Aris	50K	Defense
Ermis	60K	Commerce
Apollo	60K	Energy
Hefestus	50K	Energy
Hera	90K	General Management
Athena	70K	Education
Aphrodite	60K	Domestic Affairs
Demeter	60K	Agriculture
Hestia	50K	Domestic Affairs
Artemis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

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

Histogram : Example

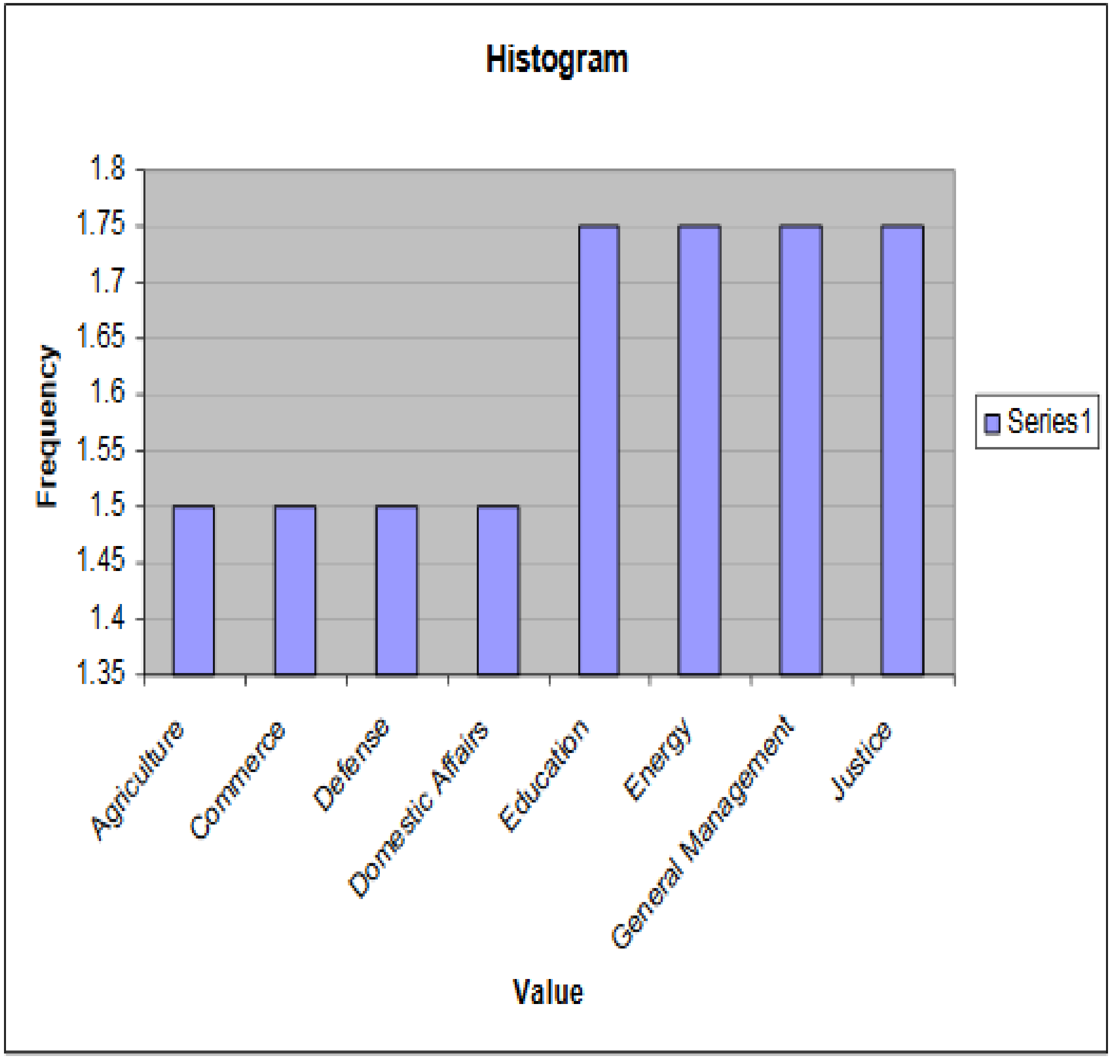


Histogram : Example



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

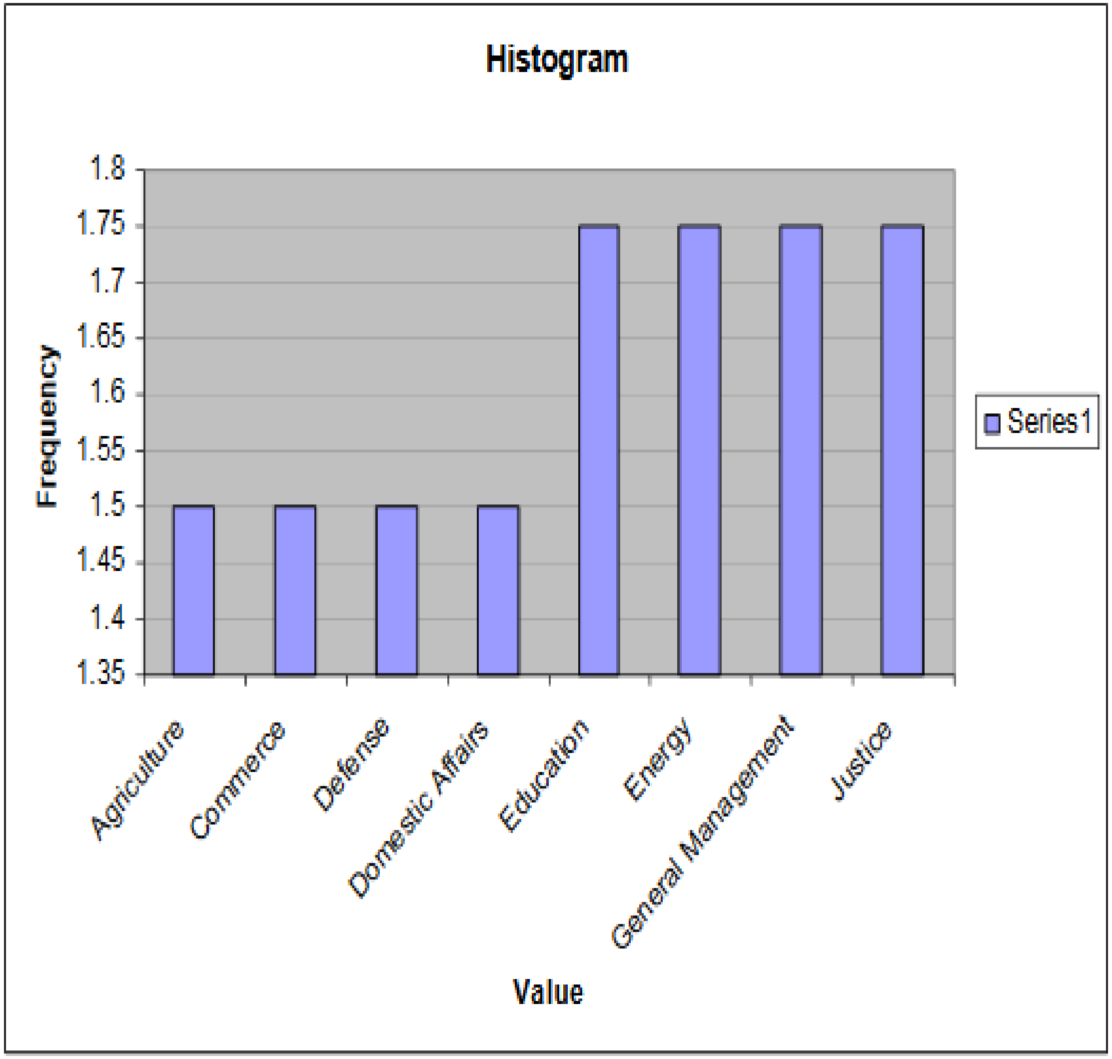


Histogram : Example

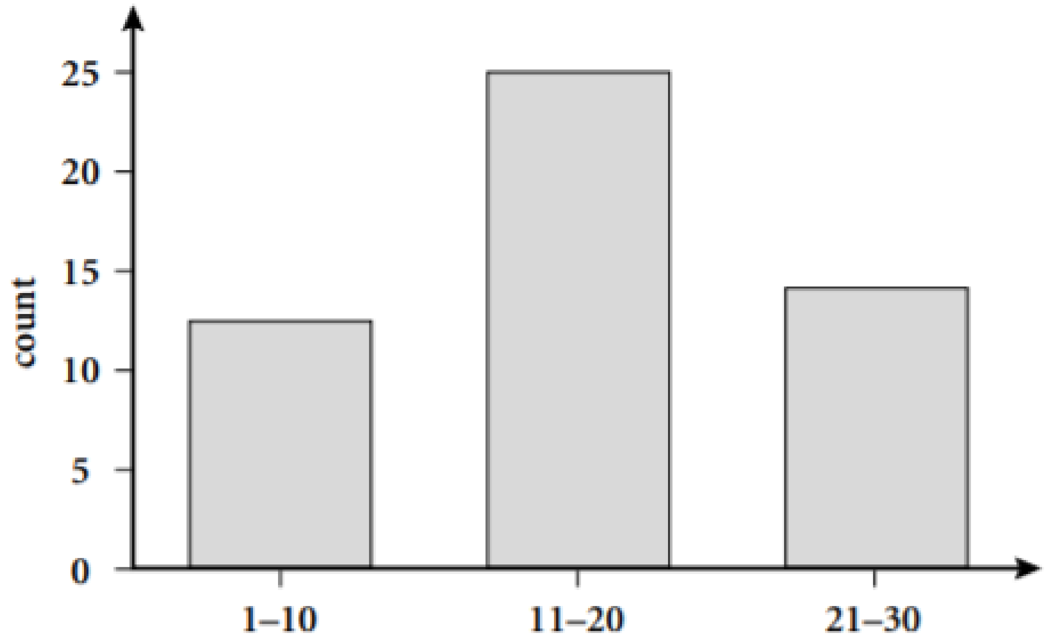
Department	Histogram H1	
	Frequency in Bucket	Approximate Frequency
Agriculture	$\boxed{1}$	1.5
Commerce	$\boxed{1}$	1.5
Defense	$\boxed{2}$	1.5
Domestic Affairs	$\boxed{2}$	1.5
Education	1	1.75
Energy	3	1.75
General Management	2	1.75
Justice	1	1.75

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

$$[(W_1)(D_1-D_m)^2 + (W_2)(D_2 -D_m)^2 + (W_3)(D_3 -D_m)^2] / (W_1+ W_2+ W_3)$$

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

Histogram : MaxDiff

MaxDiff:

“There is a bucket boundary between the adjacent values which have the maximum difference. “

We compute the difference between
f (vi+1)*Si+1 and f(vi)*Si

Si is the spread of attribute value vi, Si = vi+1-vi
f(vi)* Si is the area of v
f(vi):frequency of vi

Histogram : MaxDiff-Example

V_i	f_i
180	2
250	1
260	1
270	2
320	1
345	1
380	1
410	1
450	3
490	1
550	1

TABLE II. COMPUTING THE SPREAD, AREA AND Δ AREA

Value	180	250	260	270	320	345	380	410	450	490	550
Frequency	2	1	1	2	1	1	1	1	3	1	1
Spread	70	10	10	50	25	35	35	40	40	60	-
Area	140	10	10	100	25	35	35	40	120	60	-
Δ Area	130	0	90	75	10	0	5	80	60	-	-

Clustering

Clustering techniques consider the data tuples as data objects.

Partitions the objects into clusters (how the objects are close in the space)

Quality of cluster-Measures

- Diameter
- Centroid distance

Sampling

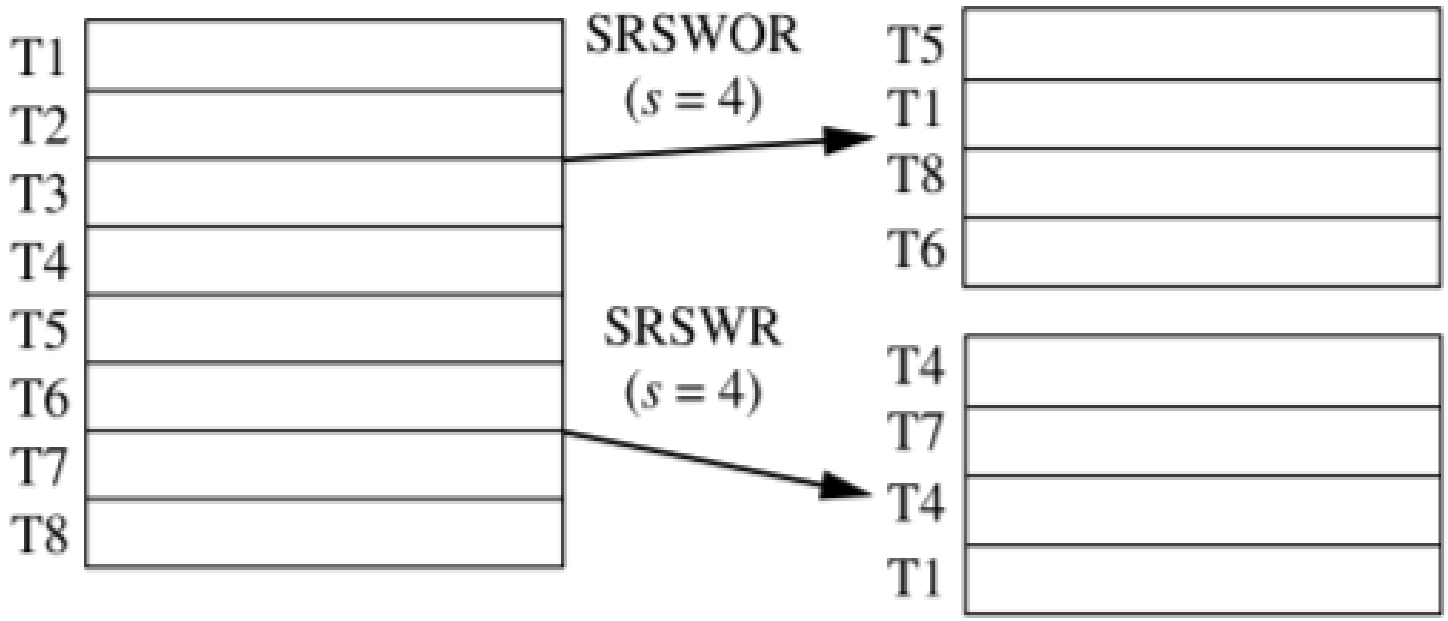
Large data that can be represented by a much smaller random sample

Methodologies:

1. Simple random sample without replacement (SRSWOR) of size s.
Drawing s of the N tuples from D (s < N), where the probability of drawing any tuple in D is 1/N, that is, all tuples are equally likely to be sampled

2. Simple random sample with replacement (SRSWR) of size s.
This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again

Sampling



Sampling

3. Cluster Sample

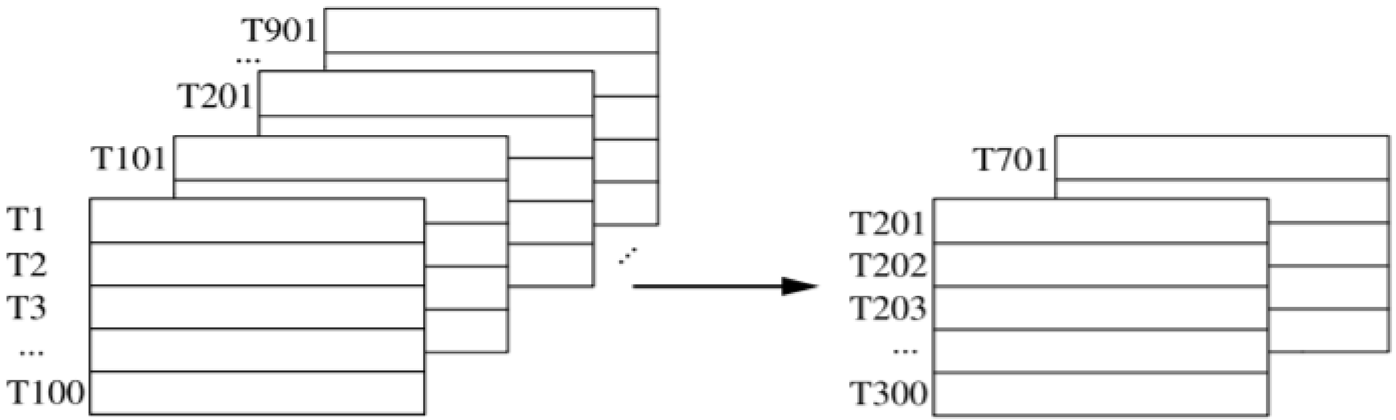
If the tuples in D are grouped into M mutually disjoint "clusters,"

4. Stratified Sample

If D is divided into mutually disjoint parts called strata

Sampling

Cluster sample
($s = 2$)



Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Data Discretization and Concept Hierarchy

"Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals."

Interval labels can then be used to replace actual data values

Supervised discretization
Unsupervised discretization

Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data

Data Discretization and Concept Hierarchy

"Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals."

Interval labels can then be used to replace actual data values

Supervised discretization
Unsupervised discretization

Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data

Data Discretization and Concept Hierarchy

Top-Down Discretization or Splitting

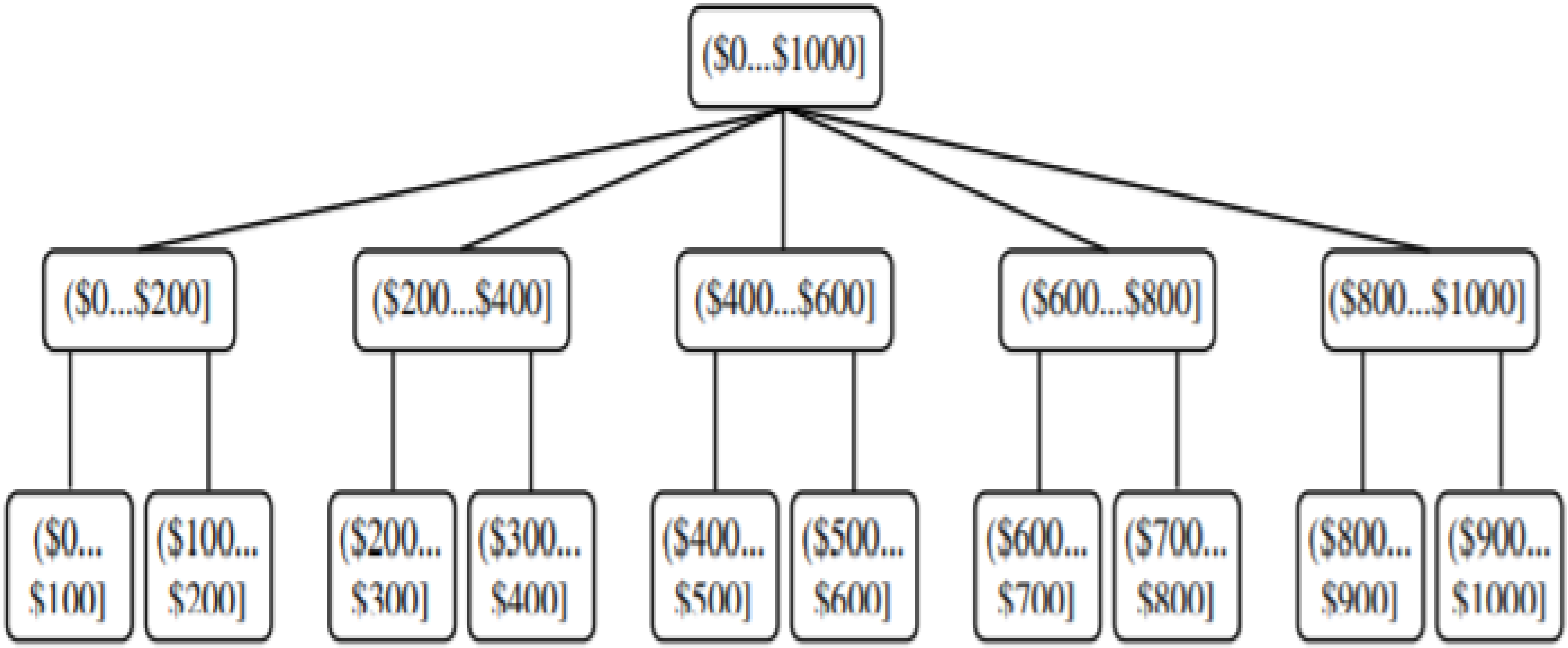
Bottom up discretization

(ESTD-1995)

Top-Down Discretization or Splitting

Bottom up discretization

Data Discretization and Concept Hierarchy



Discretization and Concept Hierarchy for Numerical Data

Binning

Histogram Analysis

Entropy-Based Discretization

Interval Merging by χ2 Analysis

Cluster Analysis