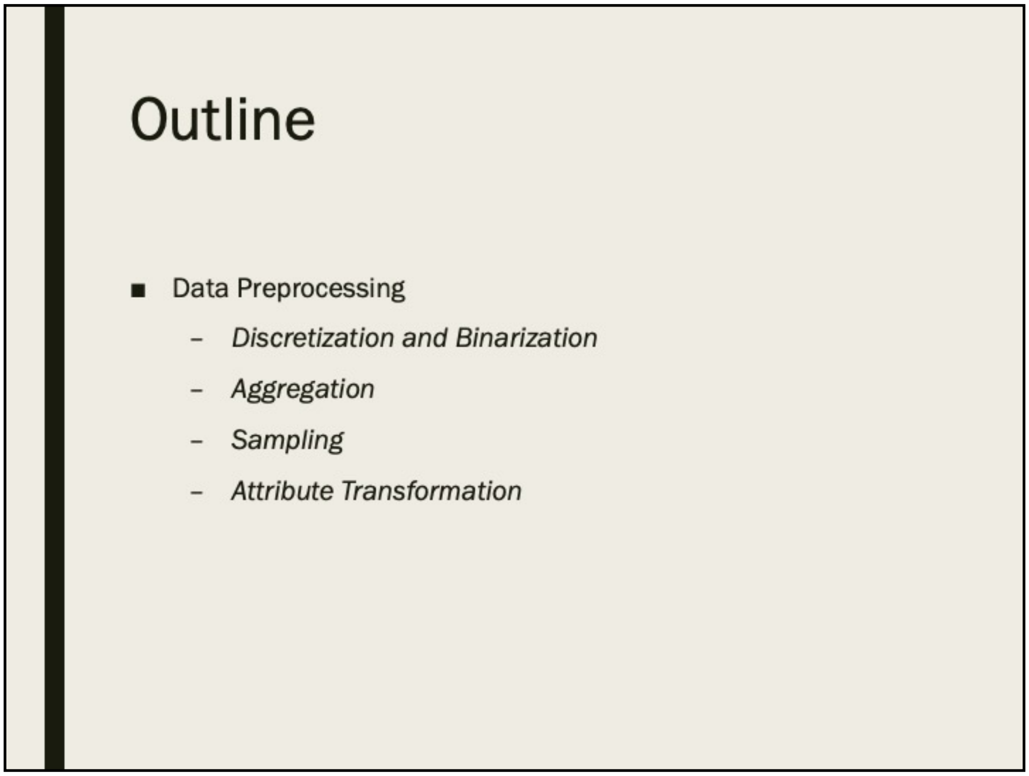
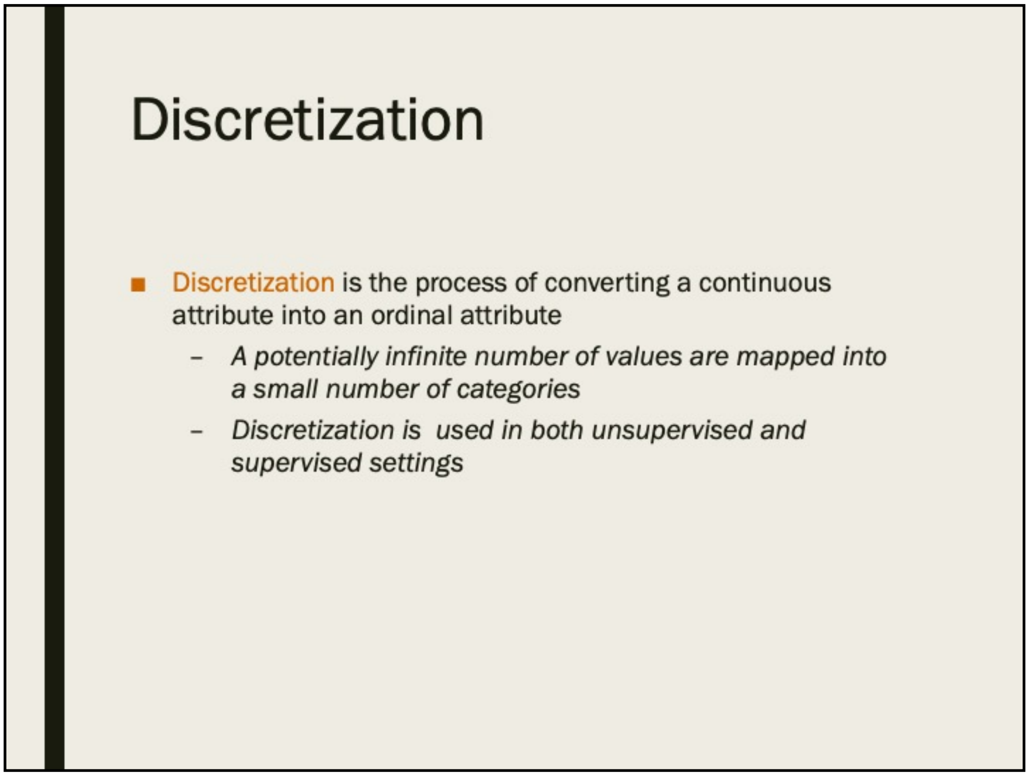


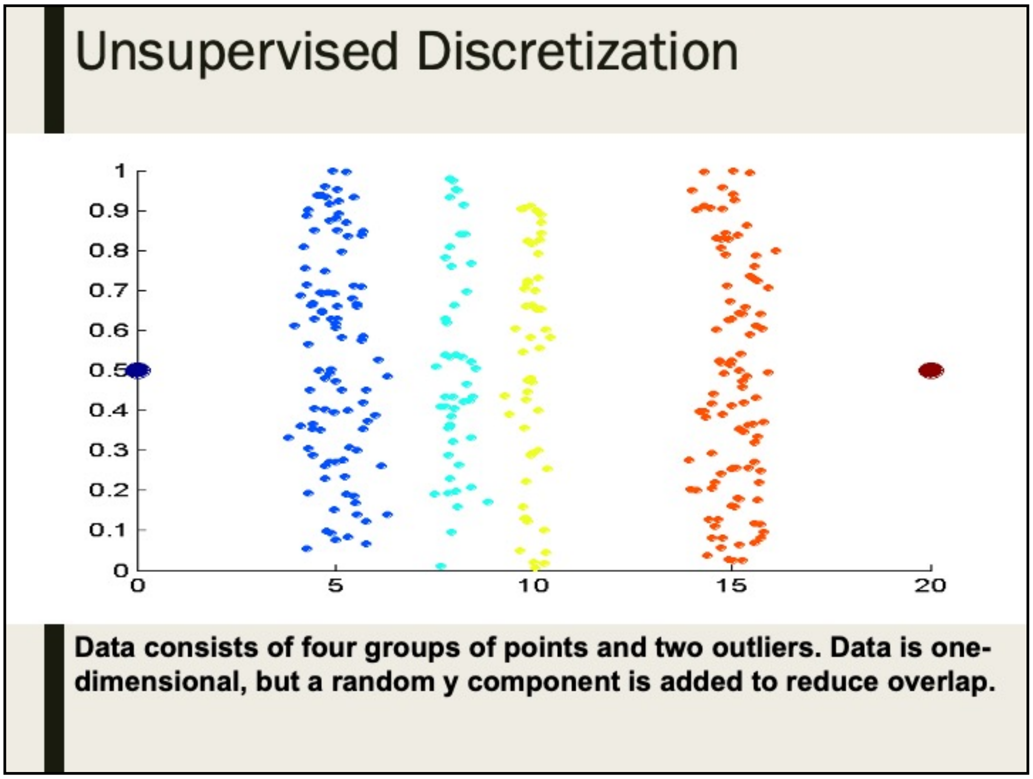
Outline

- Data Preprocessing
  - *Discretization and Binarization*
  - *Aggregation*
  - *Sampling*
  - *Attribute Transformation*



Discretization

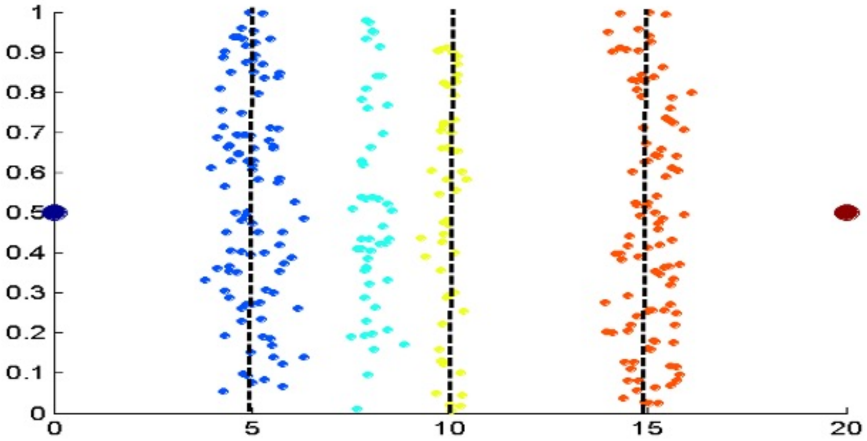
- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - *A potentially infinite number of values are mapped into a small number of categories*
  - *Discretization is used in both unsupervised and supervised settings*



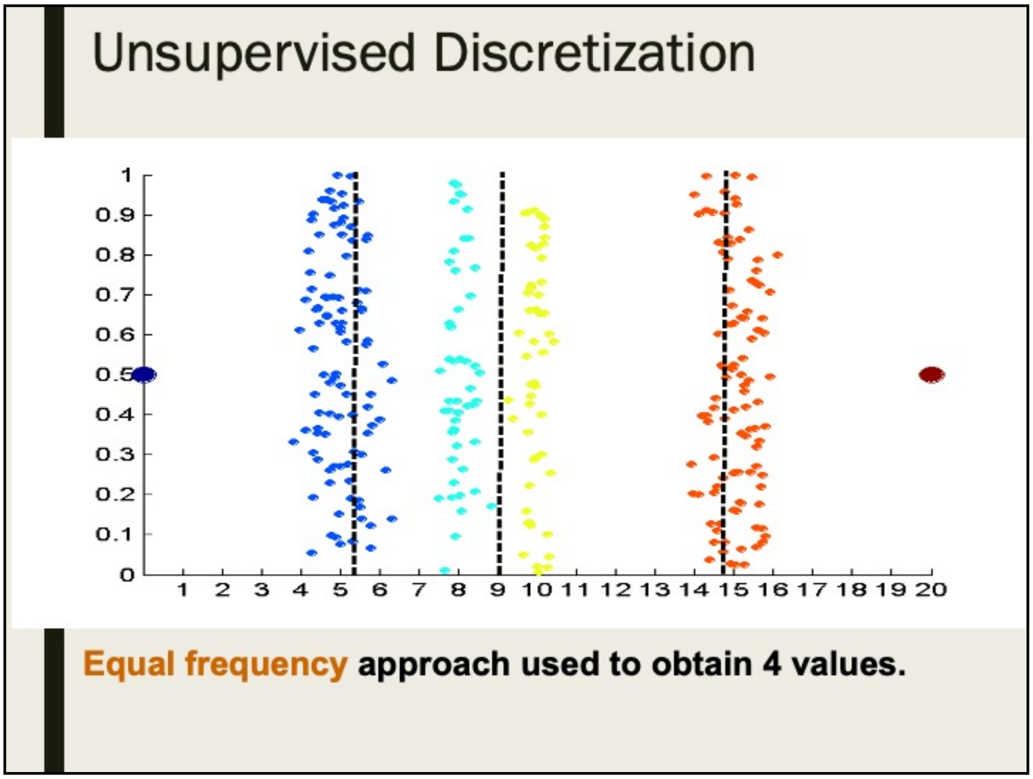
Unsupervised Discretization

Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Unsupervised Discretization



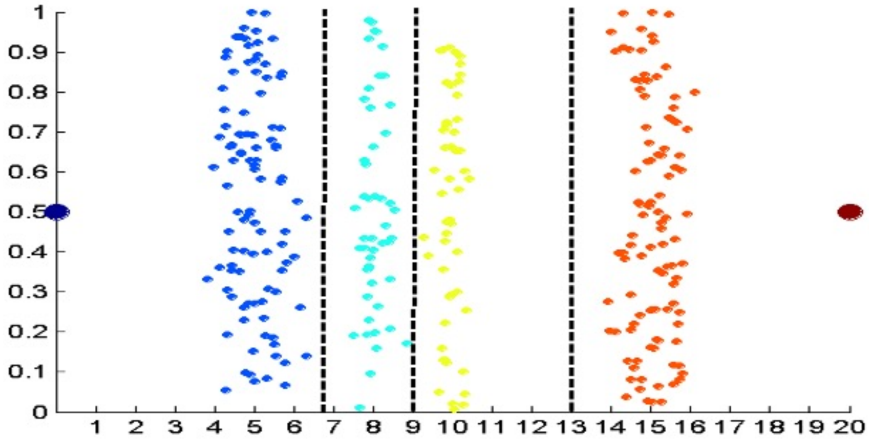
Equal interval width approach used to obtain 4 values.



Unsupervised Discretization

Equal frequency approach used to obtain 4 values.

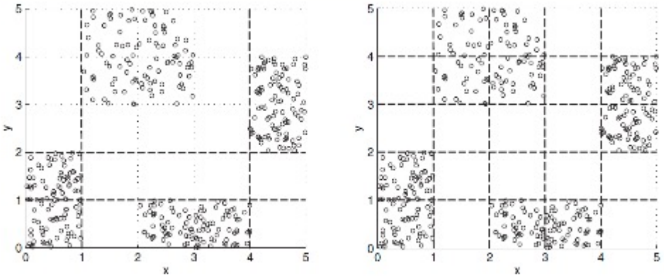
Unsupervised Discretization



K-means approach to obtain 4 values.

Discretization in Supervised Settings

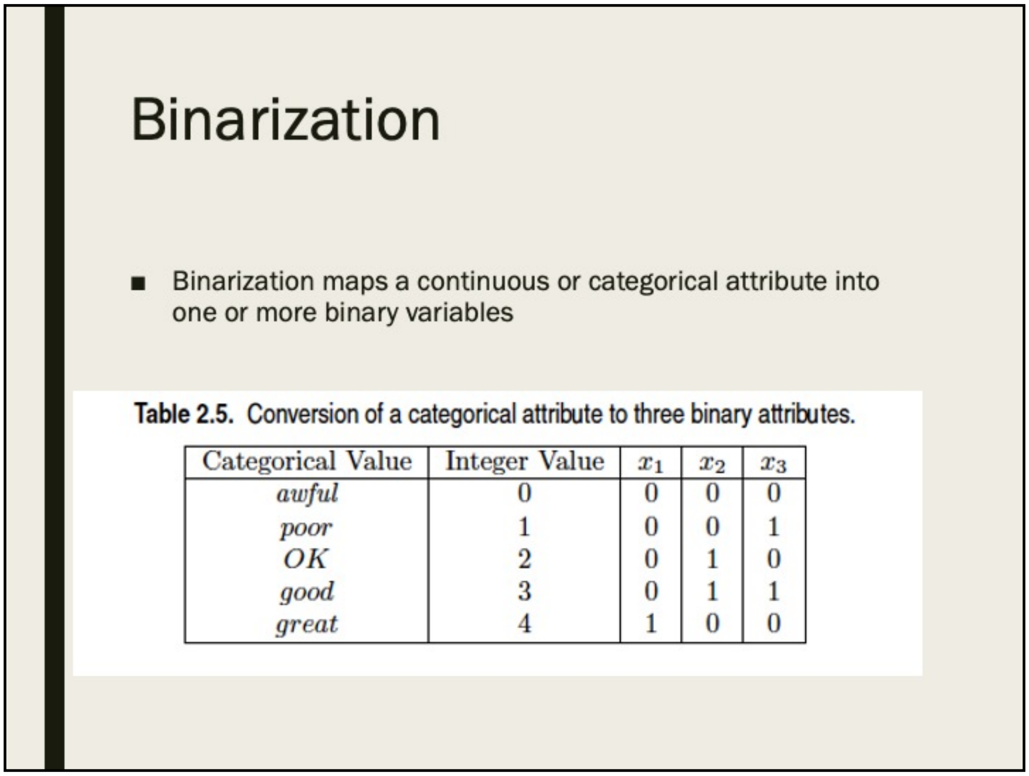
- *Many classification algorithms work best if both the independent and dependent variables have only a few values*
- *We give an illustration of the usefulness of discretization using the following example.*



(a) Three intervals

(b) Five intervals

Figure 2.14. Discretizing x and y attributes for four groups (classes) of points.



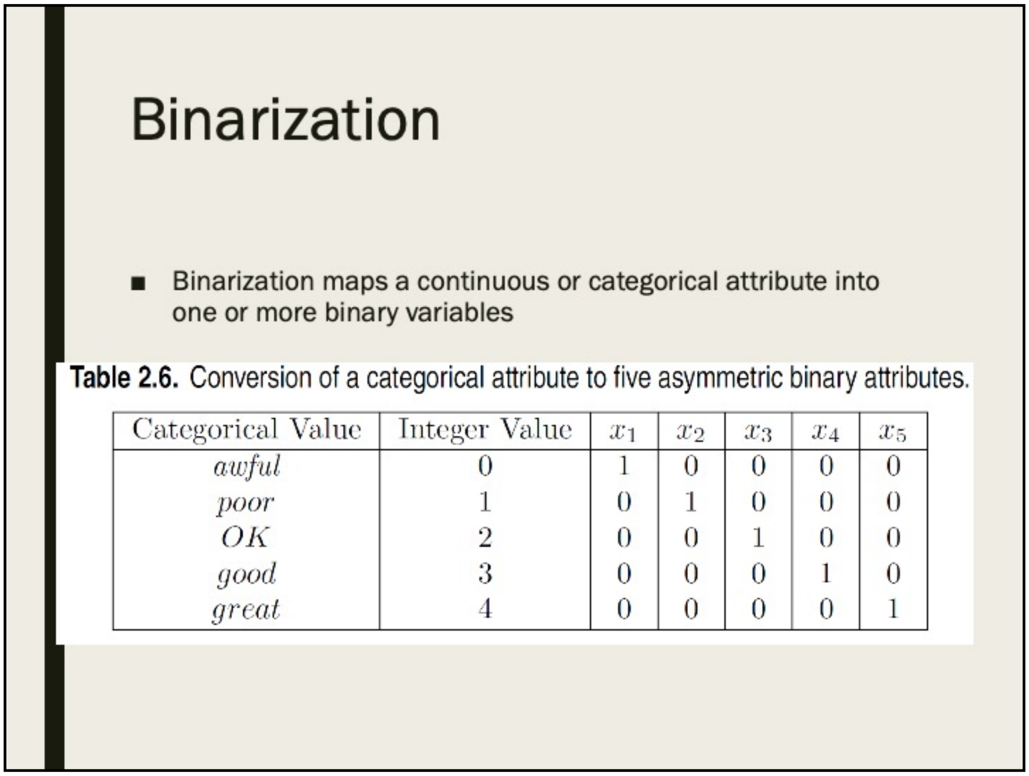
Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

If there are m categorical values, then uniquely assign each original value to an integer n the interval [0,m − 1]. If the attribute is ordinal, then order must be maintained by the assignment. (Note that even if the attribute is originally represented using integers, this process is necessary if the integers are not in the interval [0,m−1].) Next, convert each of these m integers to a binary number. Since n = log2(m) binary digits are required to represent these integers, represent these binary numbers using n binary attributes. To illustrate, a categorical variable with 5 values {awful, poor, OK, good, great} would require three binary variables x1, x2, and x3.



Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Previous slide transformation can cause complications, such as creating unintended relationships among the transformed attributes. For example, in Table 2.5, attributes x2 and x3 are correlated because information about the good value is encoded using both attributes. Furthermore, association analysis requires asymmetric binary attributes, where only the presence of the attribute (value = 1) is important.

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - *Data reduction - reduce the number of attributes or objects*
  - *Change of scale*
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - *More "stable" data - aggregated data tends to have less variability*

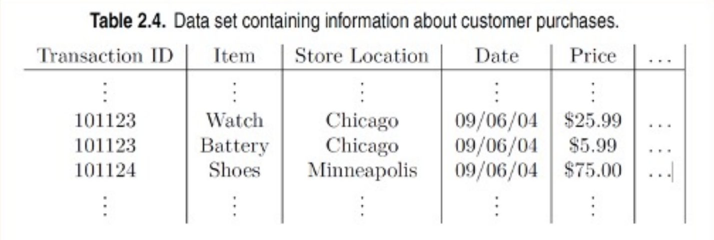
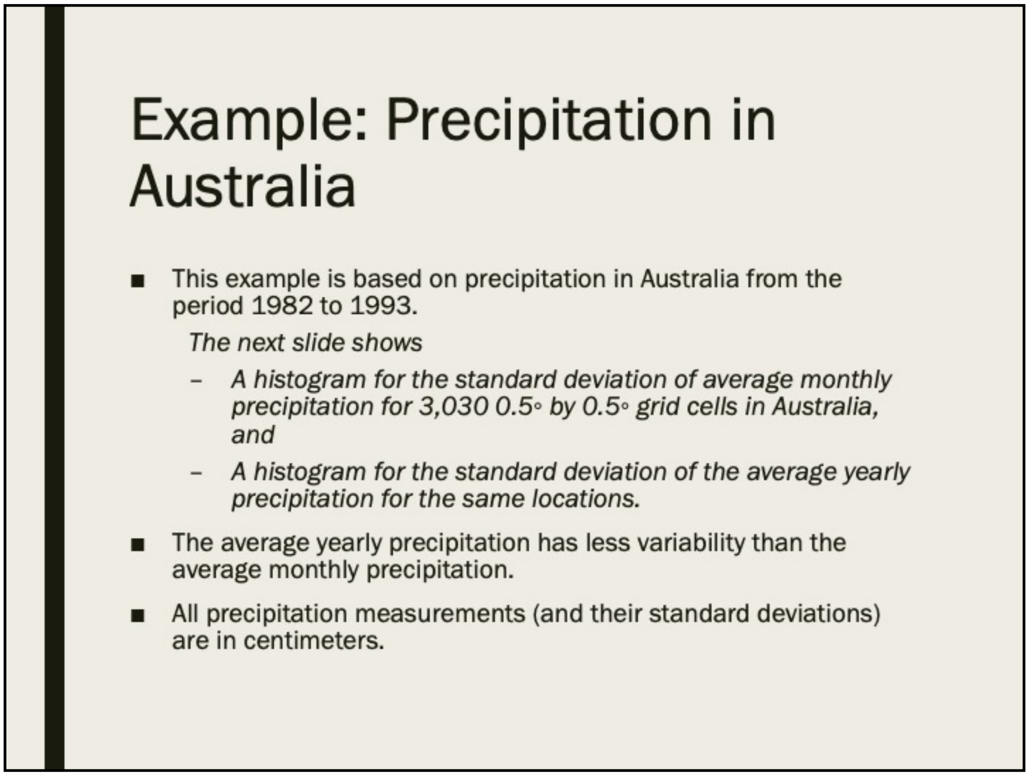


Table 2.4. Data set containing information about customer purchases.

Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	



Example: Precipitation in Australia

- This example is based on precipitation in Australia from the period 1982 to 1993.

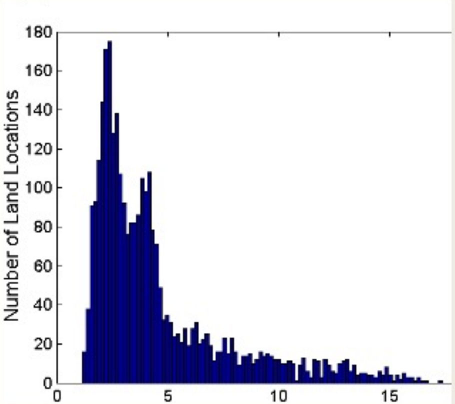
  The next slide shows

  - *A histogram for the standard deviation of average monthly precipitation for 3,030 0.5° by 0.5° grid cells in Australia, and*
  - *A histogram for the standard deviation of the average yearly precipitation for the same locations.*

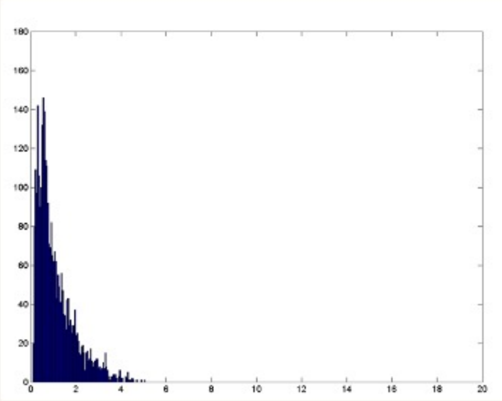
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia …

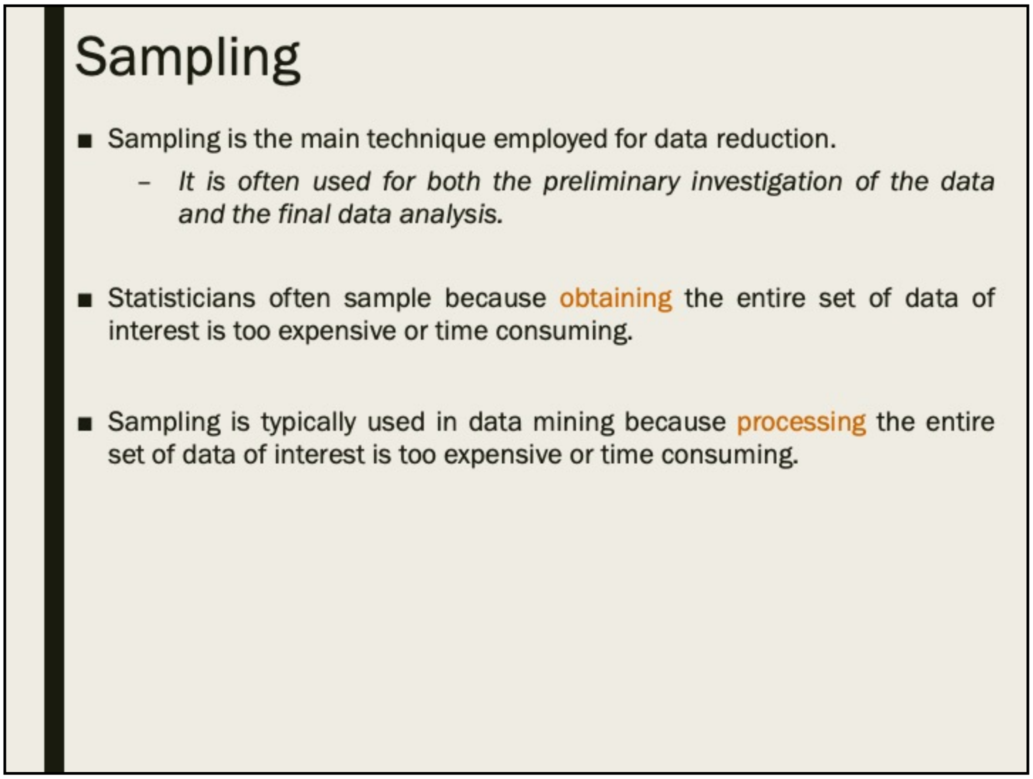
Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation

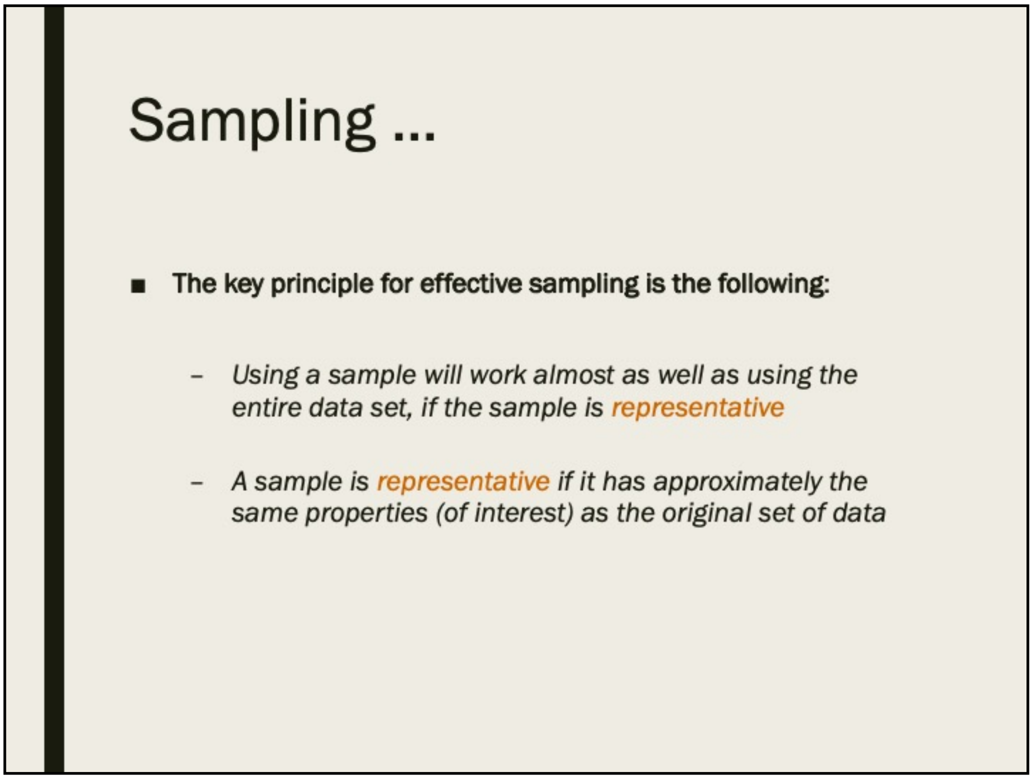


Standard Deviation of Average Yearly Precipitation



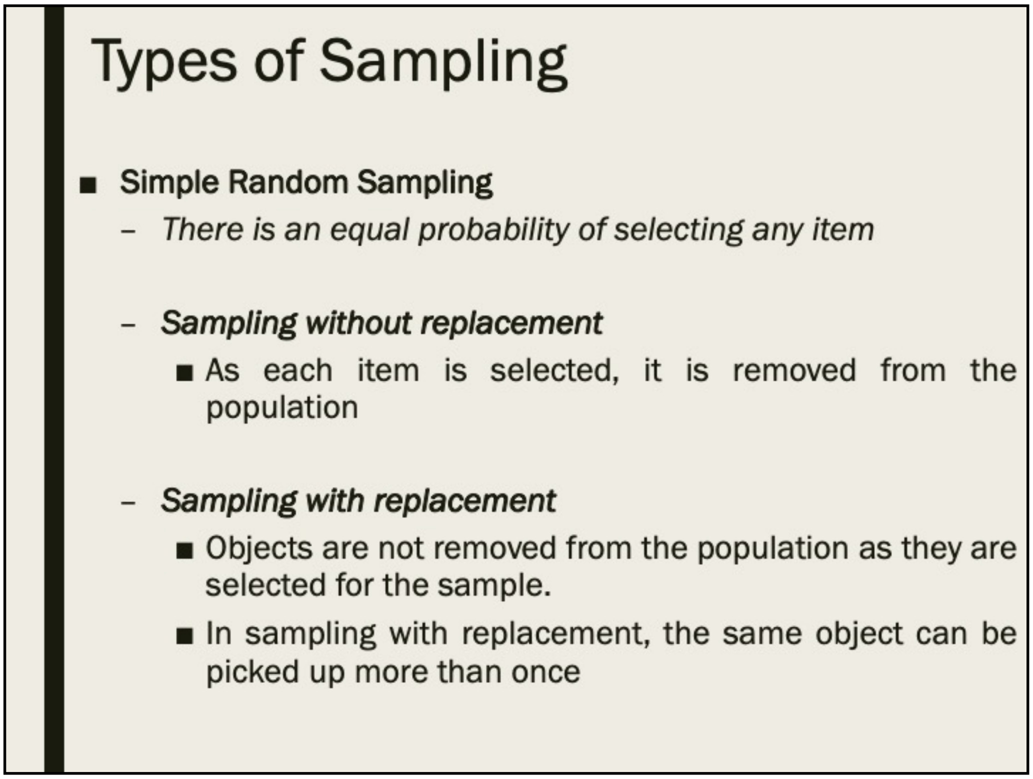
Sampling

- Sampling is the main technique employed for data reduction.
  - *It is often used for both the preliminary investigation of the data and the final data analysis.*
- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.



Sampling …

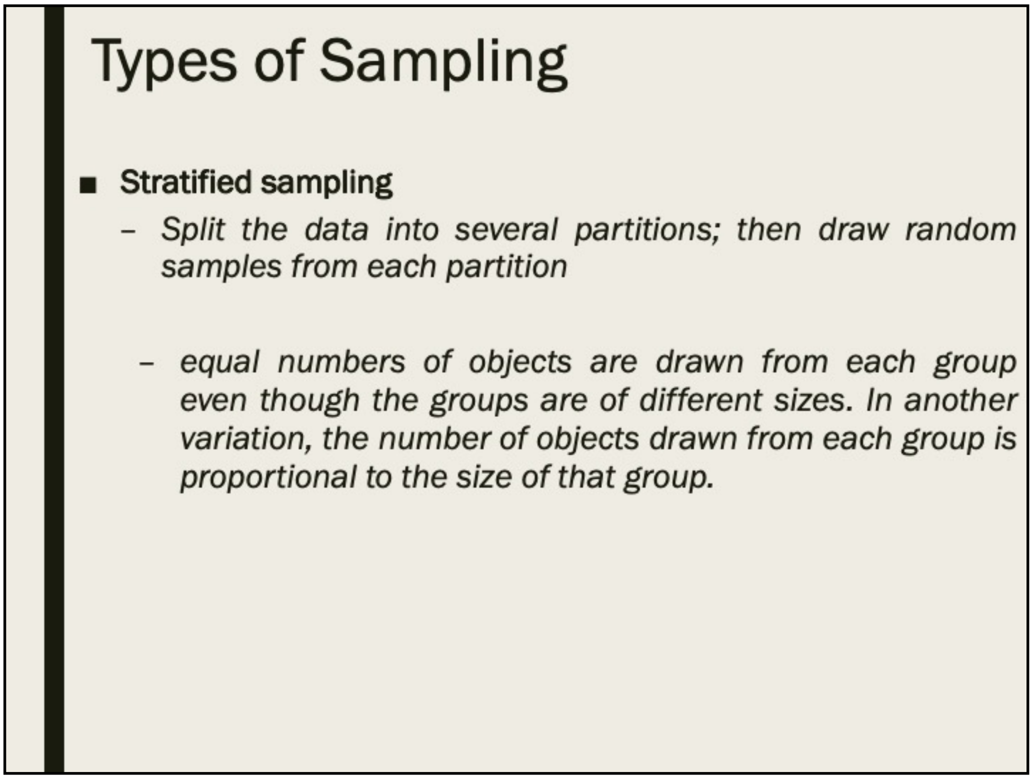
- **The key principle for effective sampling is the following:**
  - *Using a sample will work almost as well as using the entire data set, if the sample is representative*
  - *A sample is representative if it has approximately the same properties (of interest) as the original set of data*



Types of Sampling

- Simple Random Sampling
  - *There is an equal probability of selecting any item*
  - ***Sampling without replacement***
    - As each item is selected, it is removed from the population
  - ***Sampling with replacement***
    - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once

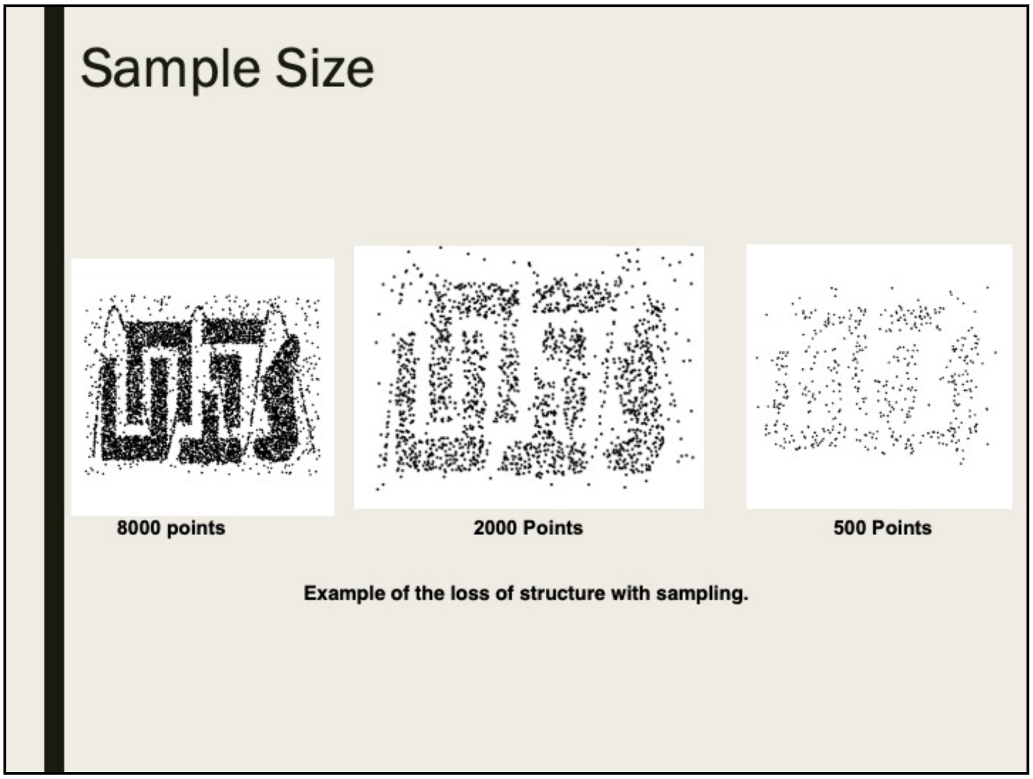
When the population consists of different types of objects, with widely different numbers of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent. This can cause problems when the analysis requires proper representation of all object types. For example, when building classification models for rare classes, it is critical that the rare classes be adequately represented in the sample. Hence, a sampling scheme that can accommodate differing frequencies for the items of interest is needed.



Types of Sampling

- **Stratified sampling**
  - *Split the data into several partitions; then draw random samples from each partition*
  - *equal numbers of objects are drawn from each group even though the groups are of different sizes. In another variation, the number of objects drawn from each group is proportional to the size of that group.*

When the population consists of different types of objects, with widely different numbers of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent. This can cause problems when the analysis requires proper representation of all object types. For example, when building classification models for rare classes, it is critical that the rare classes be adequately represented in the sample. Hence, a sampling scheme that can accommodate differing frequencies for the items of interest is needed.



Sample Size

8000 points

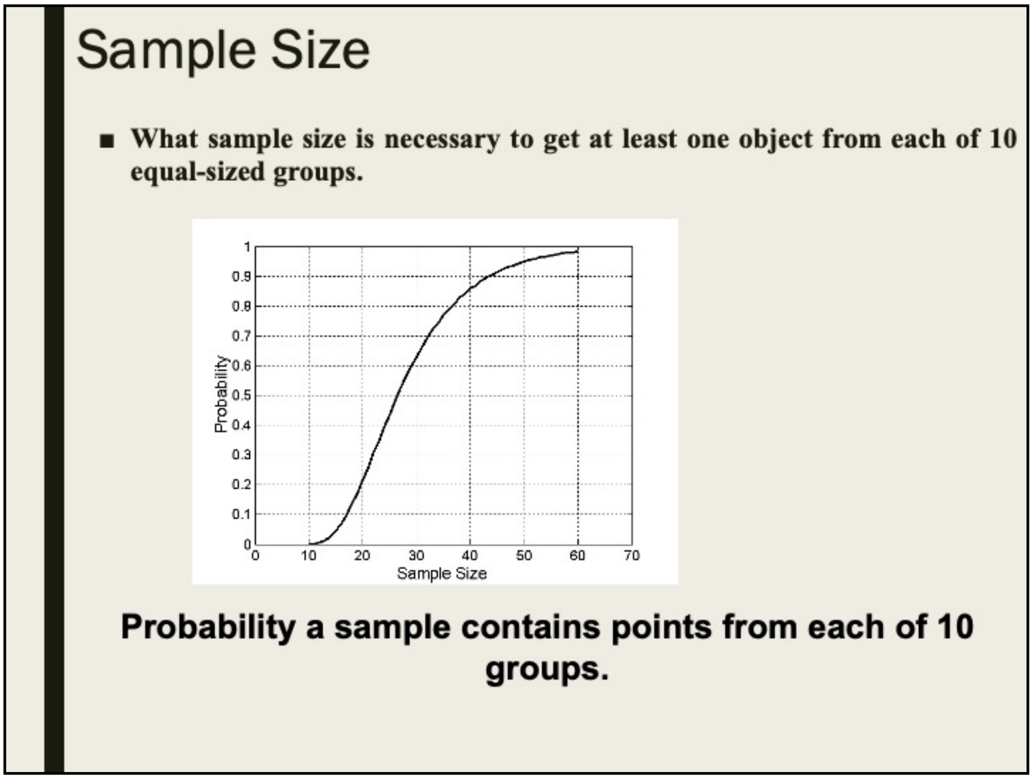
2000 Points

500 Points

Example of the loss of structure with sampling.

(Sampling and Loss of Information). Once a sampling technique has been selected, it is still necessary to choose the sample size.

Larger sample sizes increase the probability that a sample will be representative, but they also eliminate much of the advantage of sampling. Conversely, with smaller sample sizes, patterns may be missed or erroneous patterns can be detected.



Sample Size

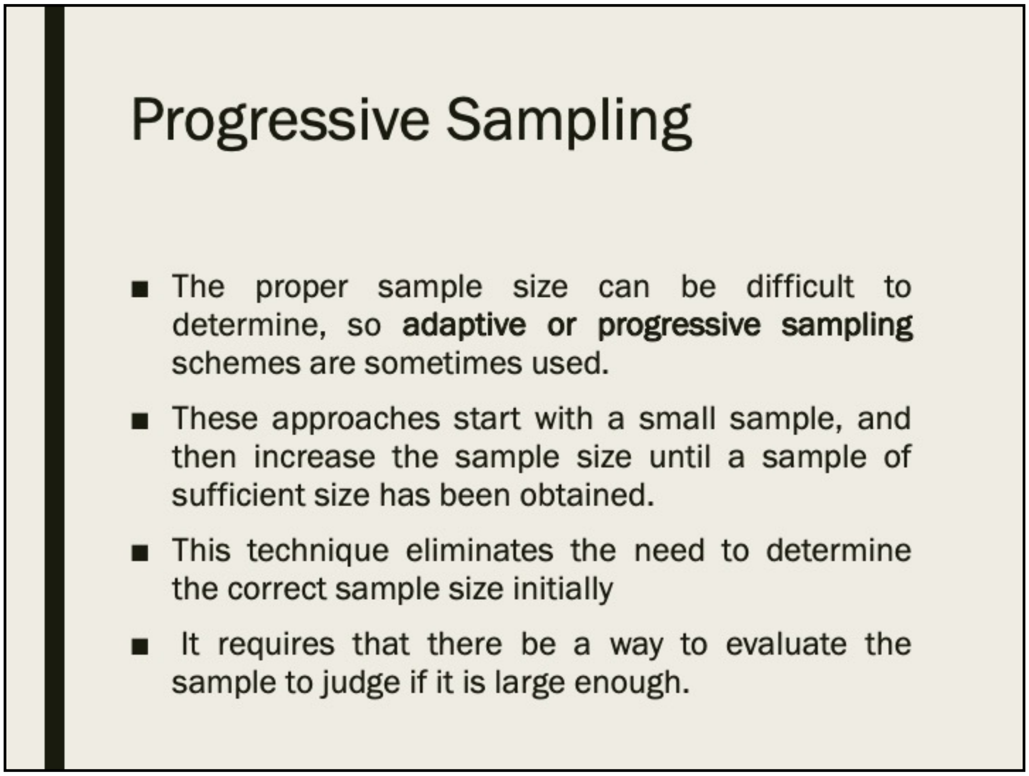
- **What sample size is necessary to get at least one object from each of 10 equal-sized groups.**

Probability a sample contains points from each of 10 groups.

Determining the Proper Sample Size:. To illustrate that determining the proper sample size requires a methodical approach, consider the following task.

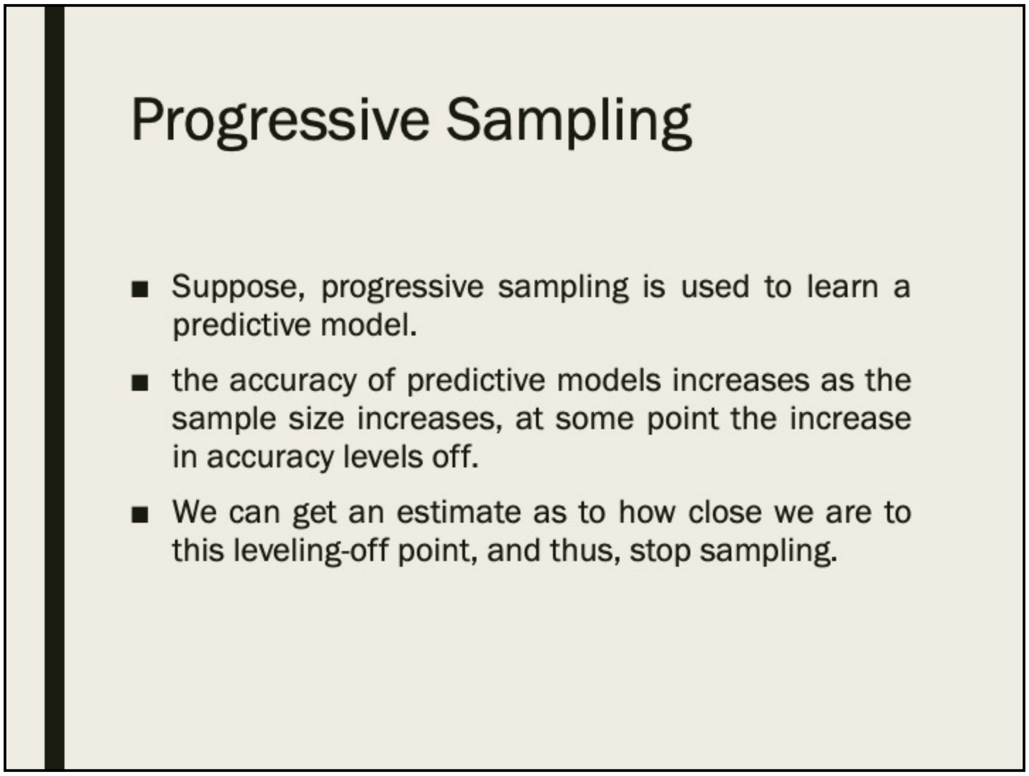
Given a set of data that consists of a small number of almost equalsized groups, find at least one representative point for each of the groups. Assume that the objects in each group are highly similar

to each other, but not very similar to objects in different groups. Also assume that there are a relatively small number of groups, e.g., 10. Figure shows an idealized set of clusters (groups) from which these points might be drawn.



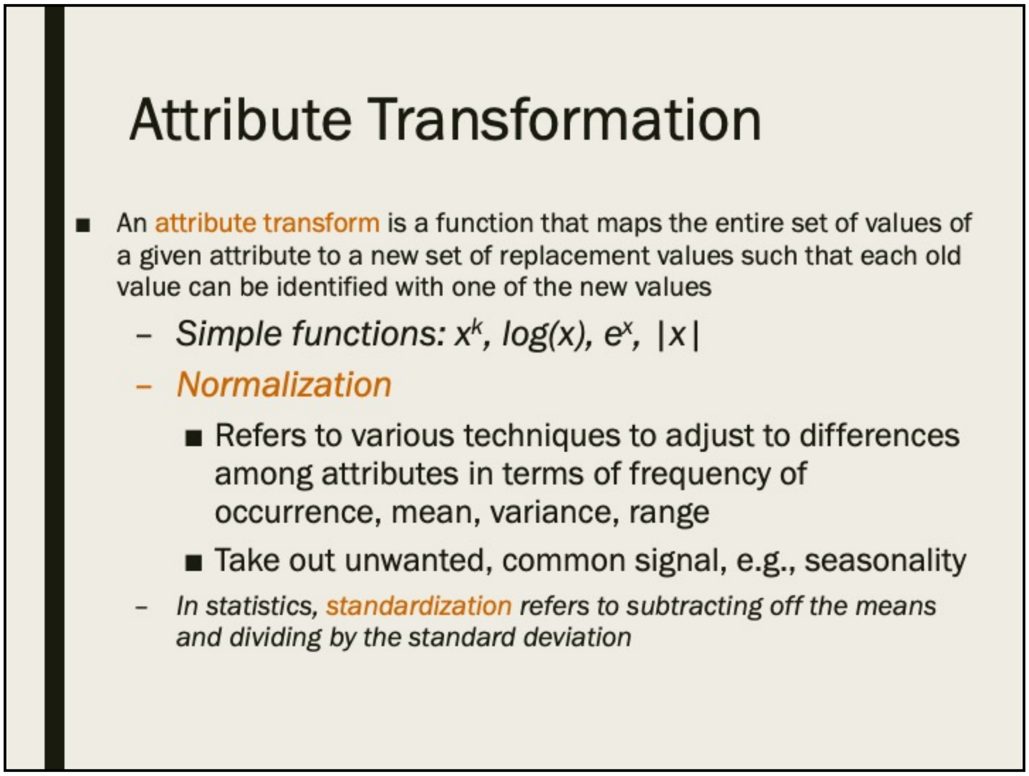
Progressive Sampling

- The proper sample size can be difficult to determine, so **adaptive or progressive sampling** schemes are sometimes used.
- These approaches start with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.
- This technique eliminates the need to determine the correct sample size initially
- It requires that there be a way to evaluate the sample to judge if it is large enough.



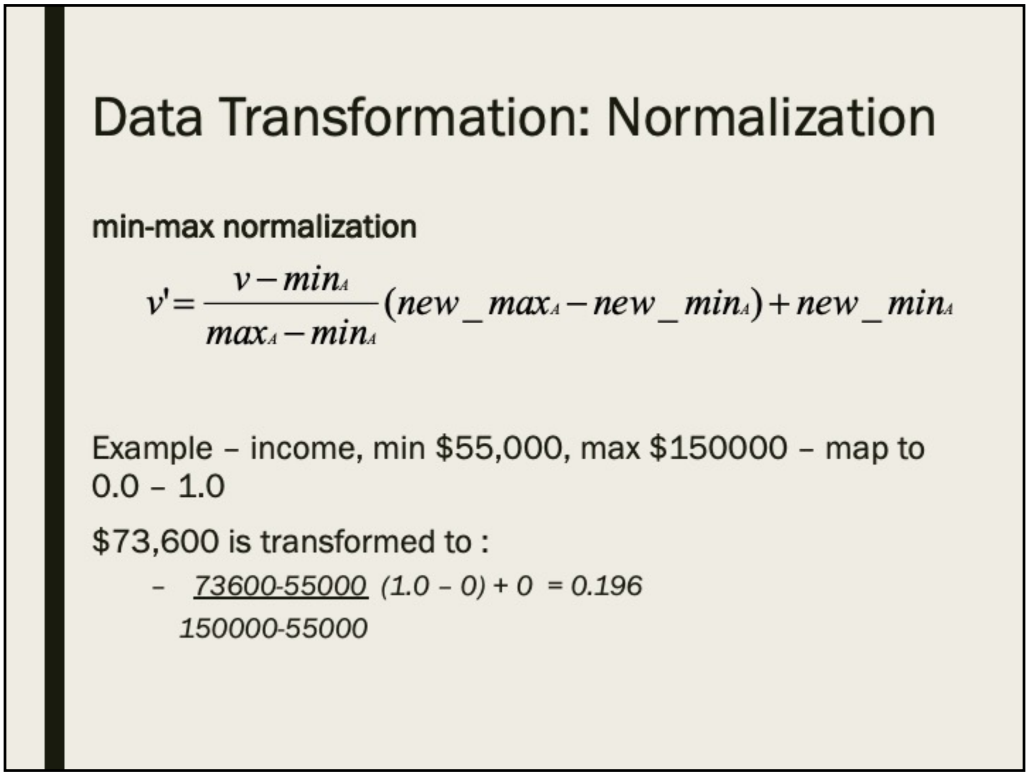
Progressive Sampling

- Suppose, progressive sampling is used to learn a predictive model.
- the accuracy of predictive models increases as the sample size increases, at some point the increase in accuracy levels off.
- We can get an estimate as to how close we are to this leveling-off point, and thus, stop sampling.



Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - *Simple functions: x^k, log(x), e^x, $|x|$*
  - *Normalization*
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - *In statistics, standardization refers to subtracting off the means and dividing by the standard deviation*



Data Transformation: Normalization

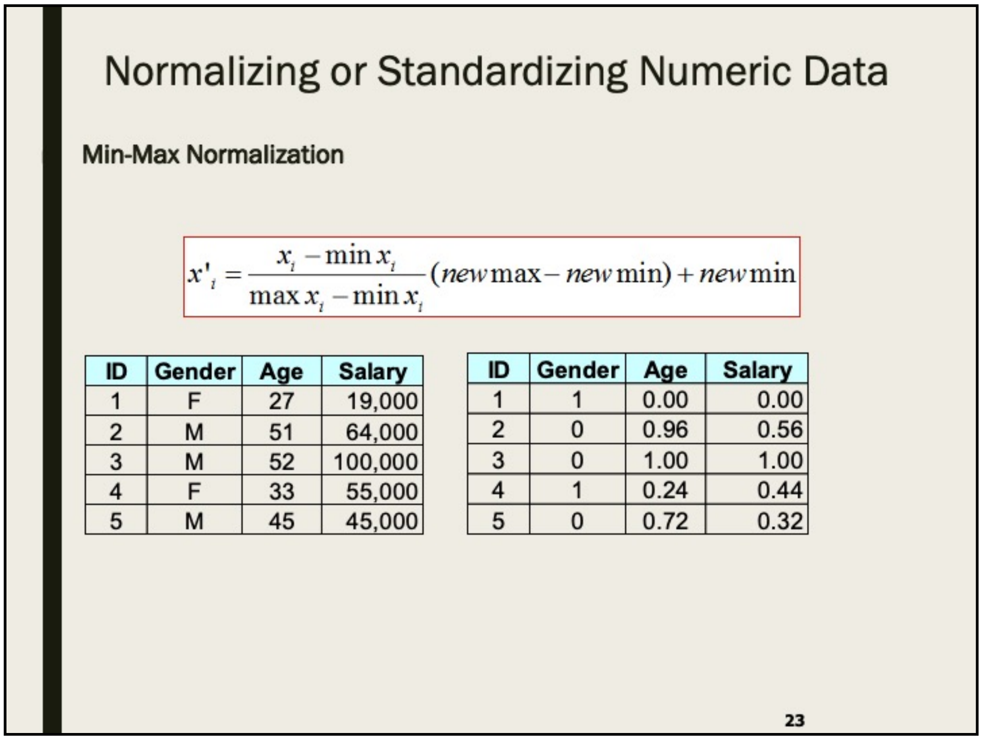
min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Example – income, min $55,000, max $150000 – map to 0.0 – 1.0

$73,600 is transformed to :

- $\frac{73600-55000}{150000-55000}(1.0 - 0) + 0 = 0.196$



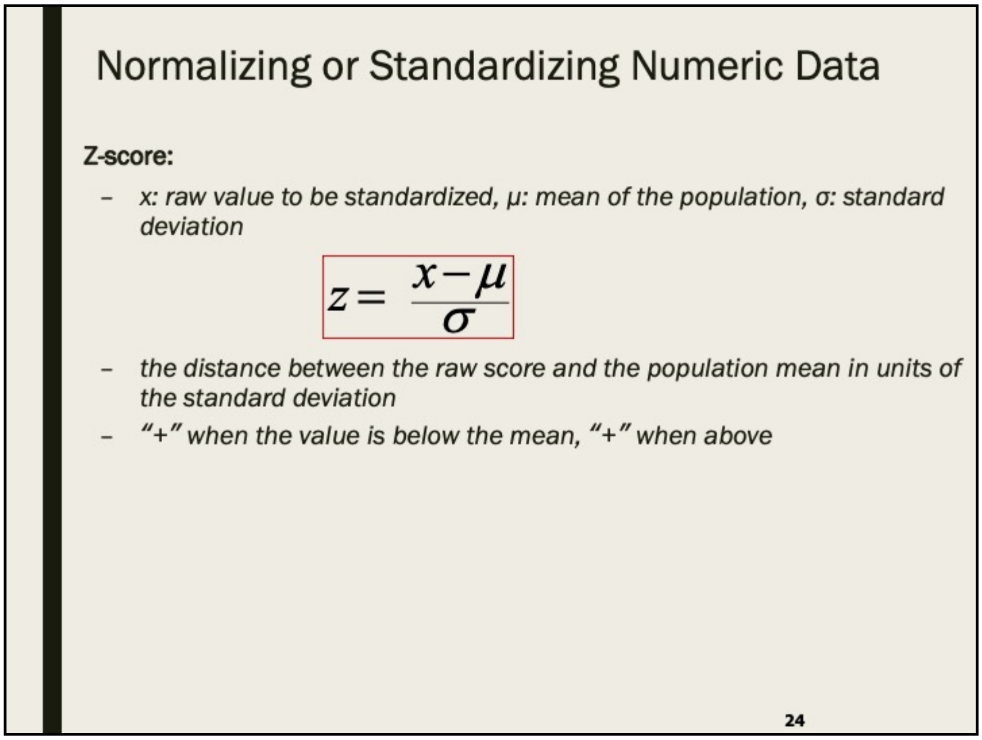
Normalizing or Standardizing Numeric Data

Min-Max Normalization

$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}(new\max - new\min) + new\min$$

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

ID	Gender	Age	Salary
1	1	0.00	0.00
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32



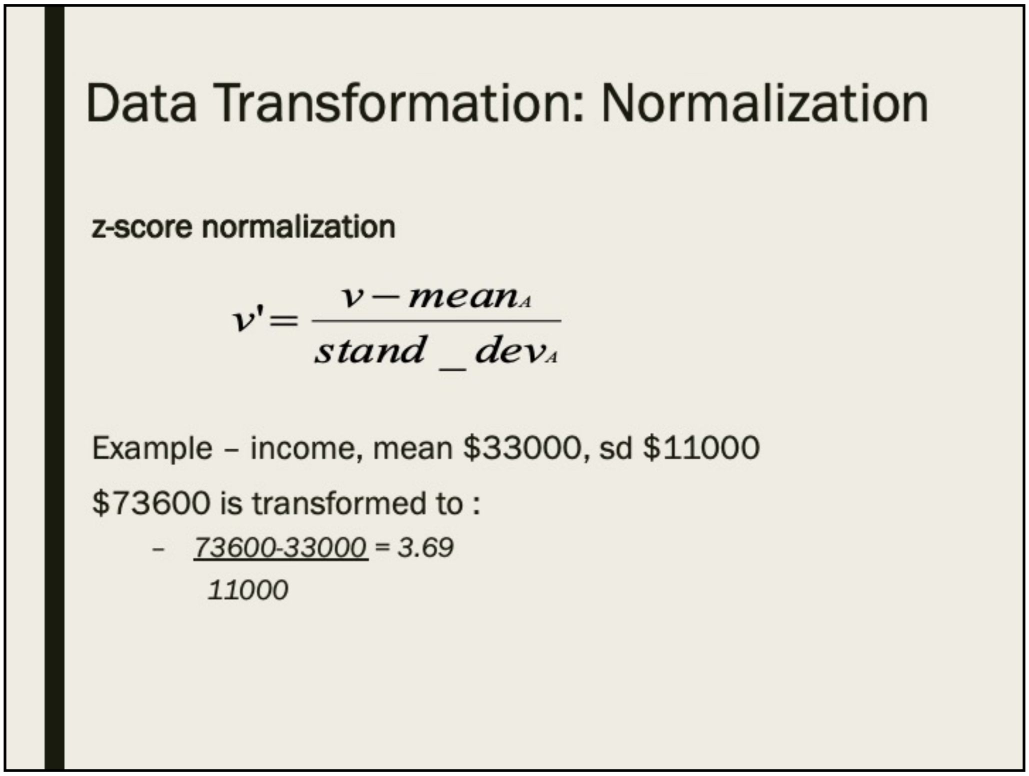
Normalizing or Standardizing Numeric Data

Z-score:

- *x: raw value to be standardized, μ: mean of the population, σ: standard deviation*

$$\boxed{z = \frac{x-\mu}{\sigma}}$$

- *the distance between the raw score and the population mean in units of the standard deviation*
- *"+" when the value is below the mean, "+" when above*



Data Transformation: Normalization

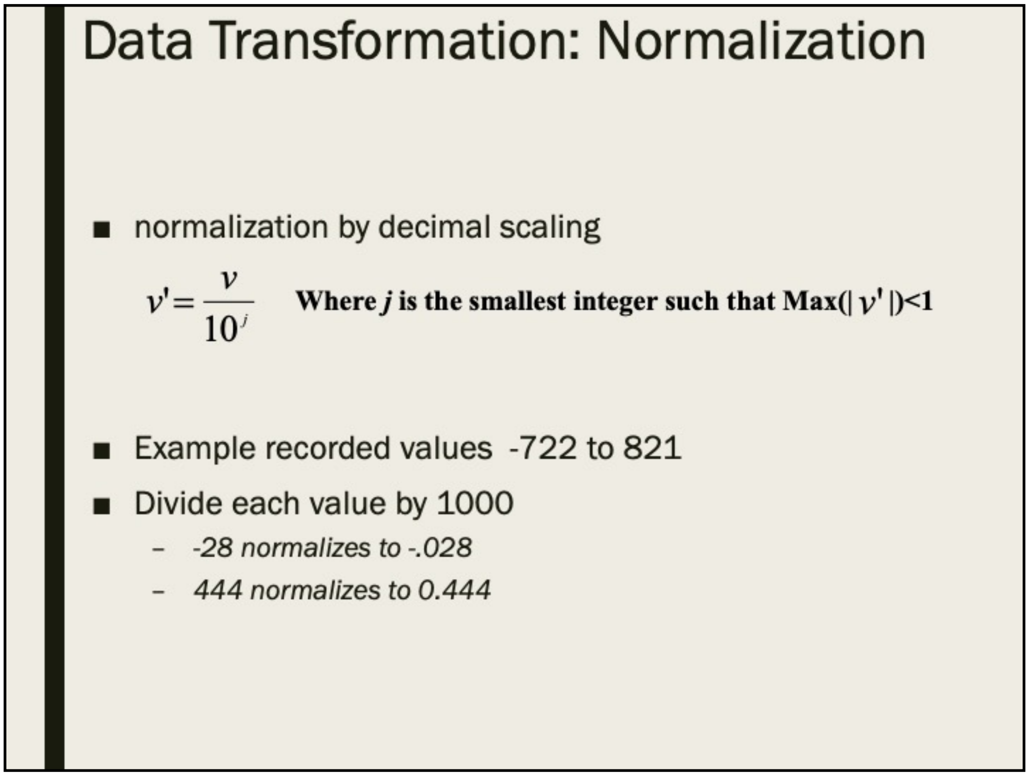
z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Example – income, mean $33000, sd $11000

$73600 is transformed to :

- $\frac{73600-33000}{11000} = 3.69$

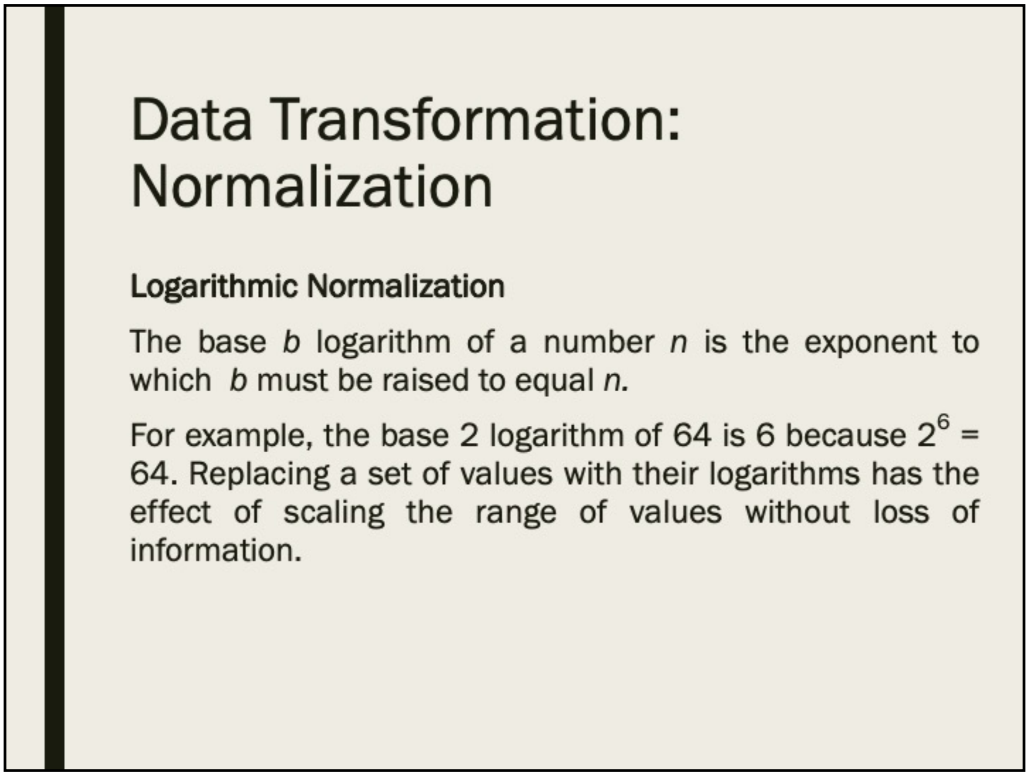


Data Transformation: Normalization

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ **Where *j* is the smallest integer such that Max(| v' |)<1**

- Example recorded values -722 to 821
- Divide each value by 1000
  - *-28 normalizes to -.028*
  - *444 normalizes to 0.444*

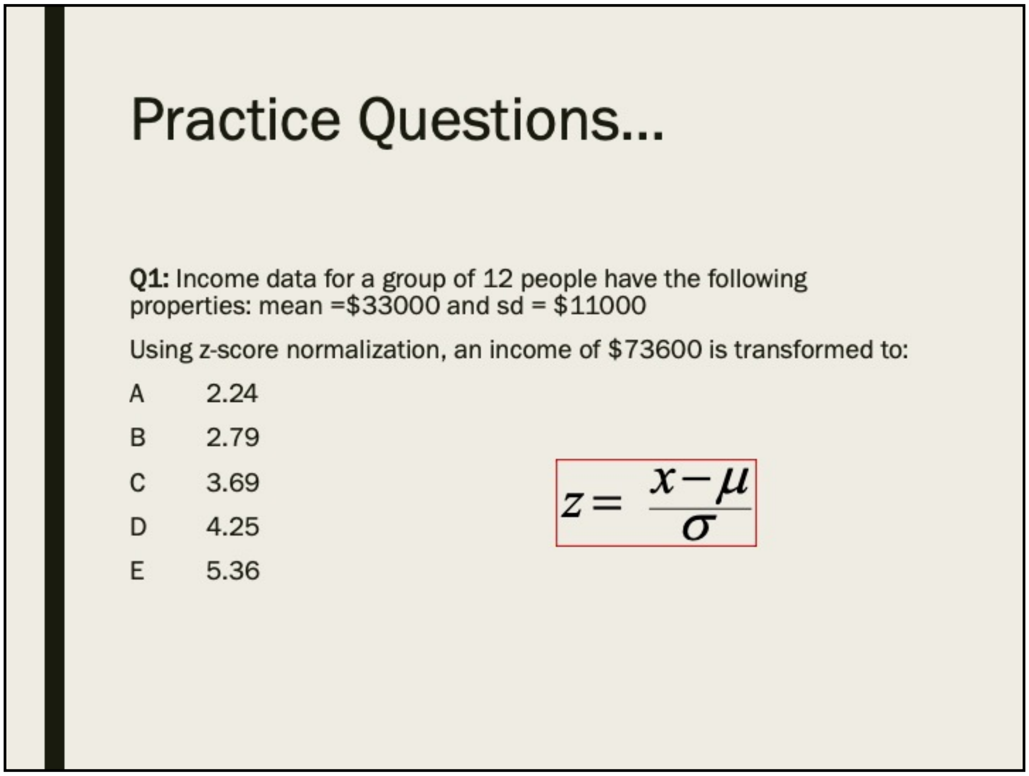


Data Transformation: Normalization

Logarithmic Normalization

The base *b* logarithm of a number *n* is the exponent to which *b* must be raised to equal *n.*

For example, the base 2 logarithm of 64 is 6 because 2^6 = 64. Replacing a set of values with their logarithms has the effect of scaling the range of values without loss of information.



Practice Questions…

Q1: Income data for a group of 12 people have the following properties: mean =$33000 and sd = $11000

Using z-score normalization, an income of $73600 is transformed to:

A 2.24

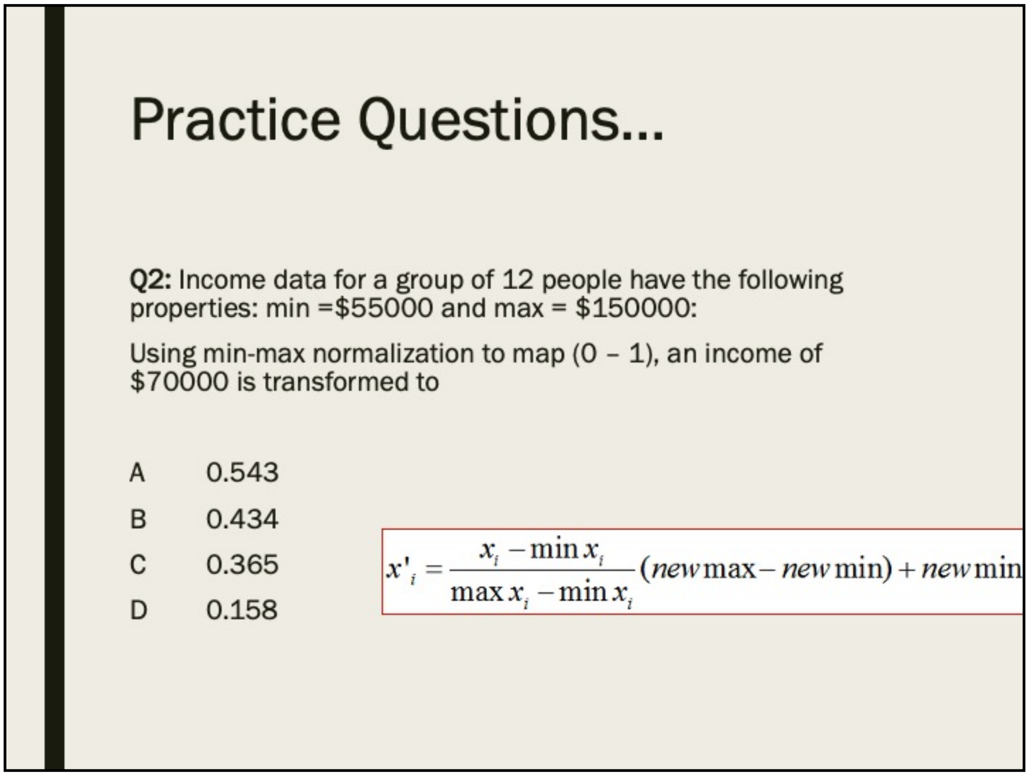
B 2.79

C 3.69

D 4.25

E 5.36

$$z = \frac{x-\mu}{\sigma}$$



Practice Questions…

Q2: Income data for a group of 12 people have the following properties: min =$55000 and max = $150000:

Using min-max normalization to map (0 – 1), an income of $70000 is transformed to

A 0.543

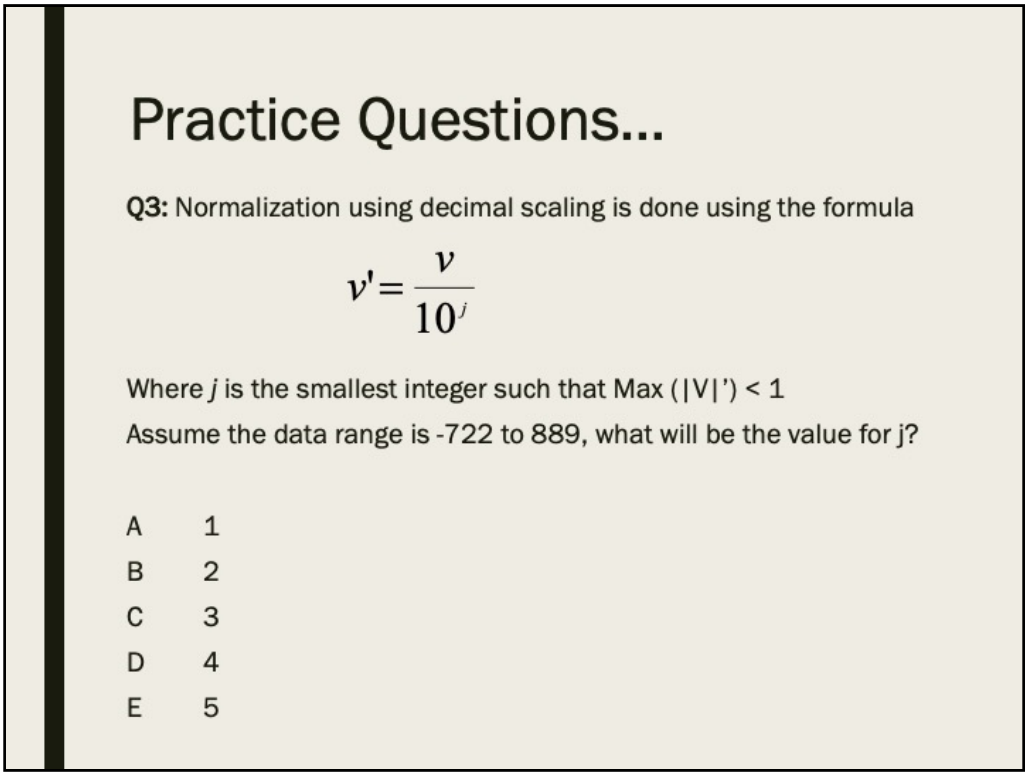
B 0.434

C 0.365

D 0.158

$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}(new\max - new\min) + new\min$$

70000-55000/95000



Practice Questions…

Q3: Normalization using decimal scaling is done using the formula

$$v' = \frac{v}{10^j}$$

Where *j* is the smallest integer such that Max (|V|') < 1

Assume the data range is -722 to 889, what will be the value for j?

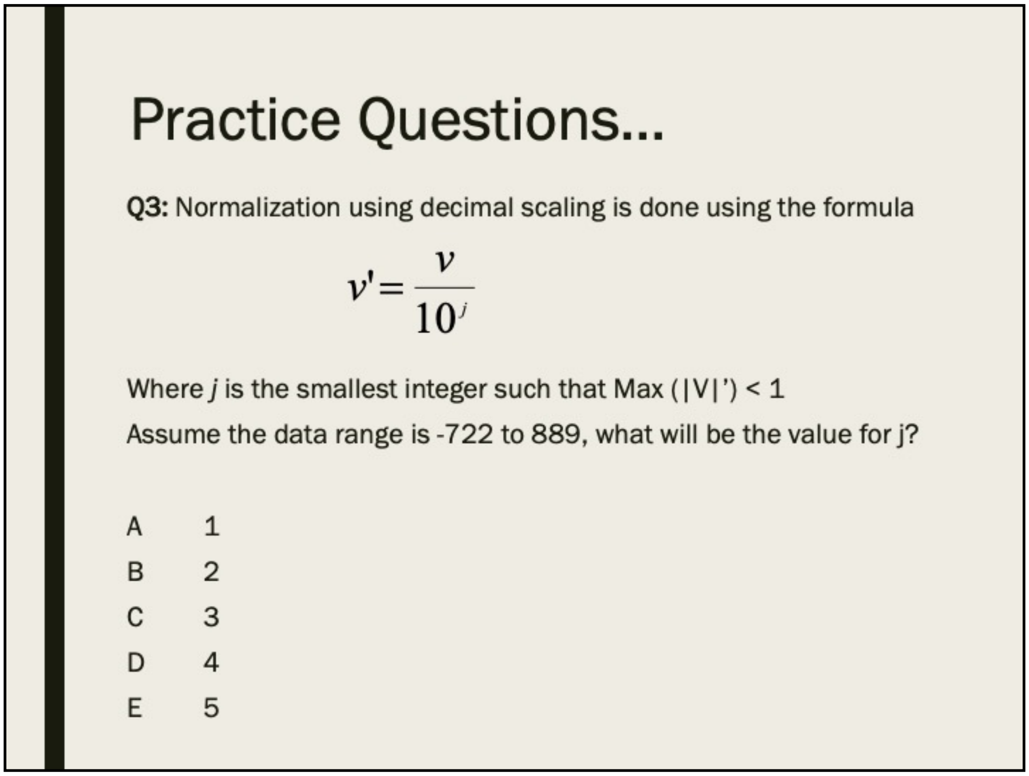
A 1

B 2

C 3

D 4

E 5



Practice Questions…

Q3: Normalization using decimal scaling is done using the formula

$$v' = \frac{v}{10^j}$$

Where *j* is the smallest integer such that Max (|V|') < 1

Assume the data range is -722 to 889, what will be the value for j?

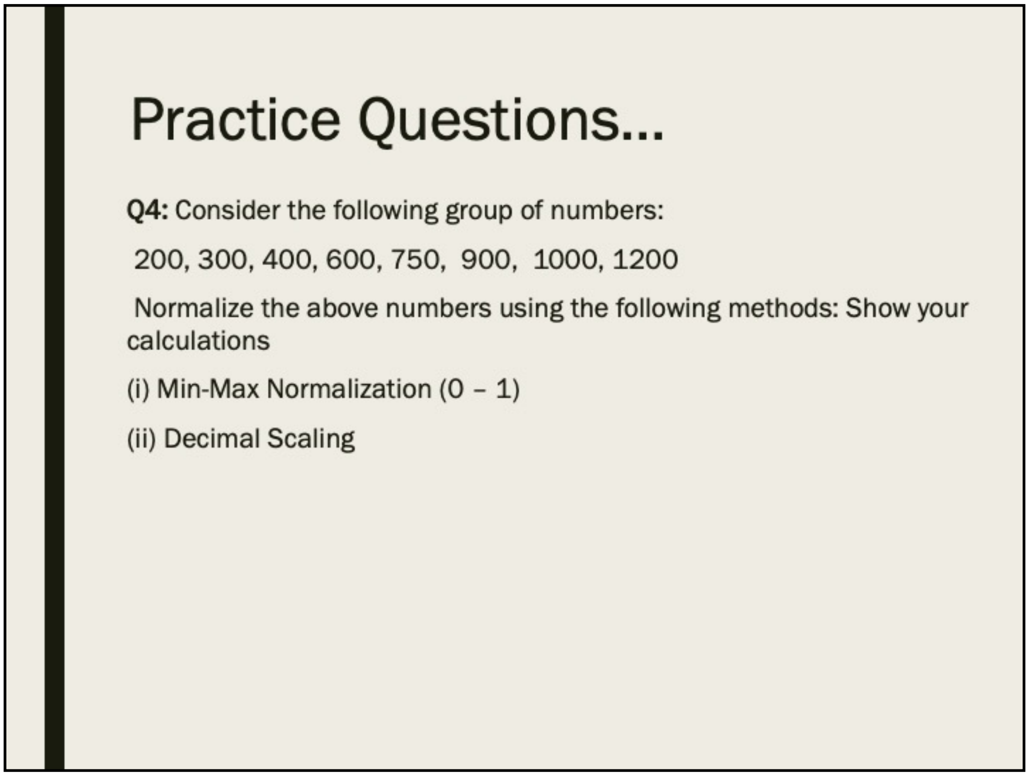
A 1

B 2

C 3

D 4

E 5



Practice Questions…

Q4: Consider the following group of numbers:

200, 300, 400, 600, 750, 900, 1000, 1200

Normalize the above numbers using the following methods: Show your calculations

(i) Min-Max Normalization (0 – 1)

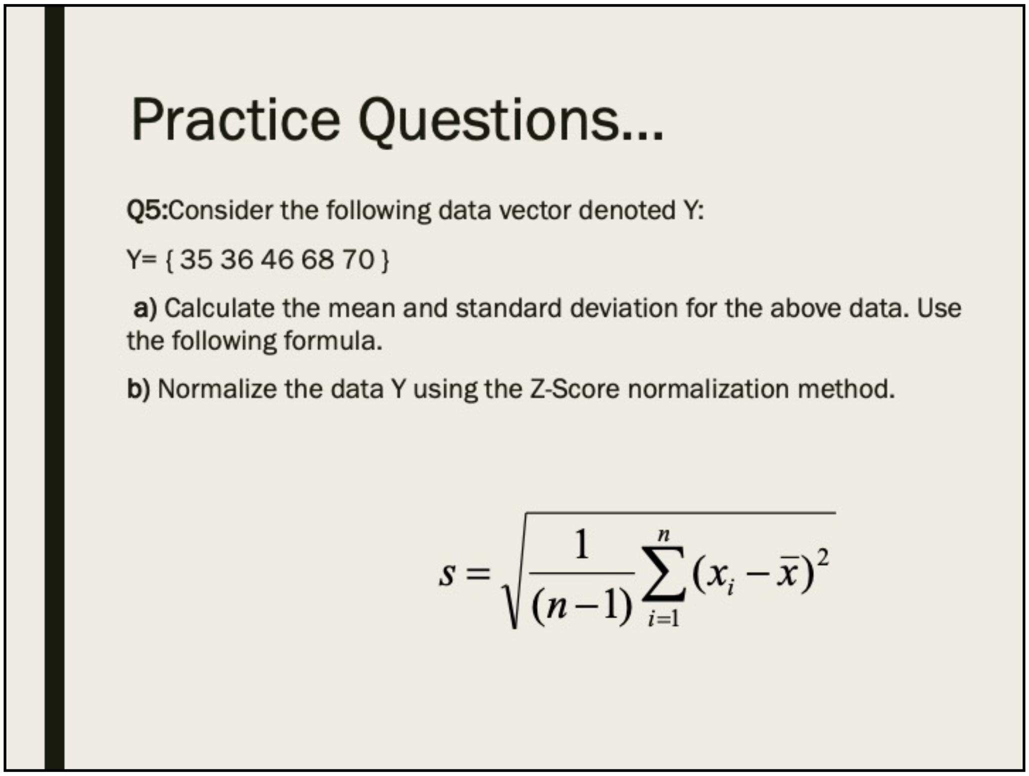
(ii) Decimal Scaling

(i) Min-Max Normalisation (0 – 1)

{0, 0.1, 0.2, 0.4, 0.55, 0.7, 08, 1.0}

(ii) Decimal Scaling

{0.02, 0.03, 0.04, 0.06, 0.075, 0.09, 0.1, 0.12}



Practice Questions...

Q5: Consider the following data vector denoted Y:

Y= { 35 36 46 68 70 }

a) Calculate the mean and standard deviation for the above data. Use the following formula.

b) Normalize the data Y using the Z-Score normalization method.

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standard Deviation Formula

Mean = (35 + 36 + 46 + 68 + 70) / 5 = 51

(35-51)2 + (36-51)2 + (46-51)2 + (68-51)2 + (70-51)2 = 1156

StdDev = SQRT (1156 / 4) = SQRT (289) = 17

(b) Normalise the data Y using the Z-Score normalization method.

New_Y = (Old_y – Mean) / StdDev

NY1 = (35-51 / 17 = - 0.941

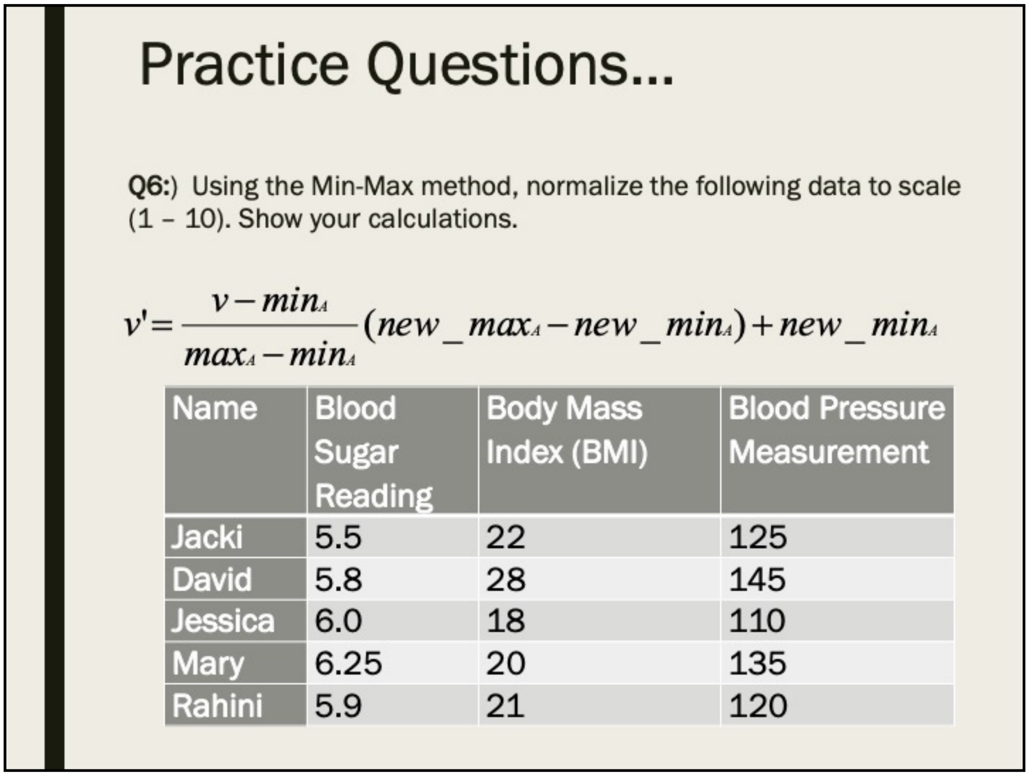
NY2 = (36-51 / 17 = - 0.882

NY3 = (46-51 / 17 = - 0.294

NY4 = (68-51 / 17 = 1

NY5 = (70-51 / 17 = 1.117

Therefore, the normalised vector is {- 0.941, - 0.882, -0.294, 1, 1.117}

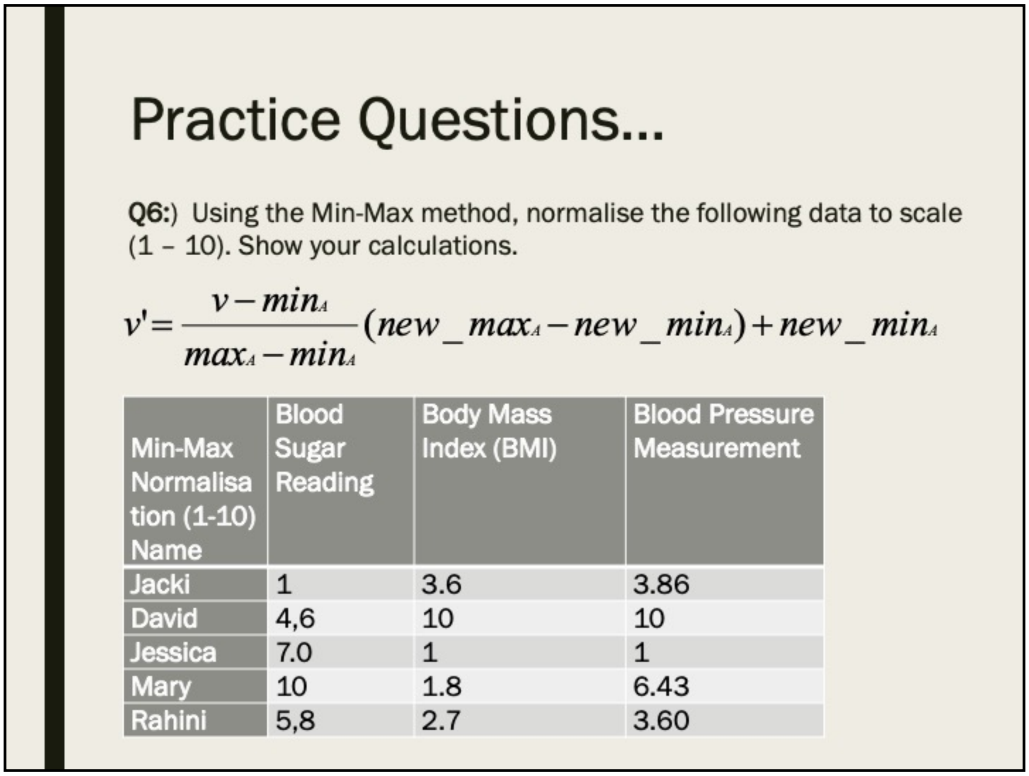


Practice Questions…

Q6:) Using the Min-Max method, normalize the following data to scale (1 – 10). Show your calculations.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Name	Blood Sugar Reading	Body Mass Index (BMI)	Blood Pressure Measurement
Jacki	5.5	22	125
David	5.8	28	145
Jessica	6.0	18	110
Mary	6.25	20	135
Rahini	5.9	21	120



Practice Questions…

Q6:) Using the Min-Max method, normalise the following data to scale (1 – 10). Show your calculations.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Min-Max Normalisation (1-10) Name	Blood Sugar Reading	Body Mass Index (BMI)	Blood Pressure Measurement
Jacki	1	3.6	3.86
David	4,6	10	10
Jessica	7.0	1	1
Mary	10	1.8	6.43
Rahini	5,8	2.7	3.60