BITS Pilani
Pilani Campus

Introduction to Data Science

(BA ZG523 / CSI ZG523)

Tirtharaj Dash
Dept. of Comp. Sc. and APP Center for A.I. Research
BITS Pilani, K.K. Birla Goa Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Introduction to Data Science
Lecture No. 05 (C.H.: 9, 10)

Previous Lecture:

- Data
- Types of data sources
  - Transactions data
  - Unstructured data
  - Qualitative or expert data
  - Publicly available data
- Sampling
- Types of data elements
  - Continuous
  - Categorical (nominal, ordinal, binary)
- Visualizing data
  - Exploration (plotting) and analysis (basic statistical measurements)
- Missing values
  - Replace, delete, keep

Sampling

- **Population:** The entire group that you want to draw conclusions about.
- **Sample:** A subset of population.
  - A well-chosen sample will contain most of the information about the population.
  - The relation between sample and population should be such that the inferences made from the sample should apply to the population as well.
- **Variable:** It is a characteristic that describes the member of a sample.
  - We will refer to it as “attribute” or “feature”.

Sampling methods

We will study two types:

1. Probabilistic sampling
2. Non probabilistic sampling

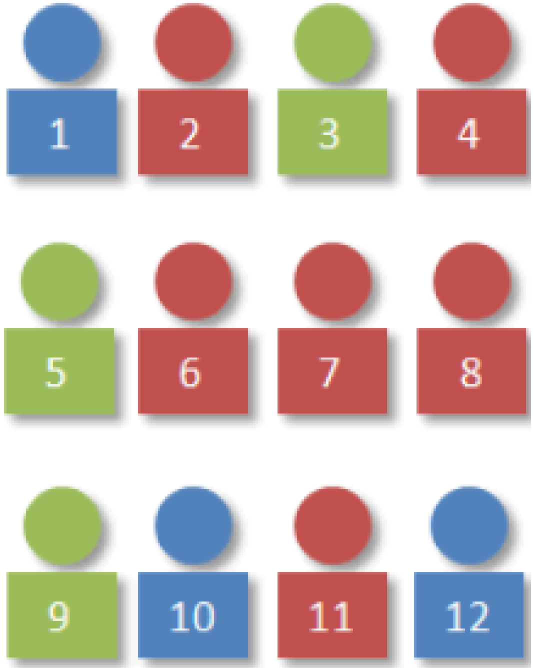
Probabilistic Sampling

- Select the members of population that have equal or non-zero probability.
- Types:
  - Simple sampling
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling

Simple sampling

- A sample of size *n* from a population of size *N* is obtained through **simple random sampling** if every possible sample of size *n* has an equally likely chance of occurring.

- Select 4 members from this group.

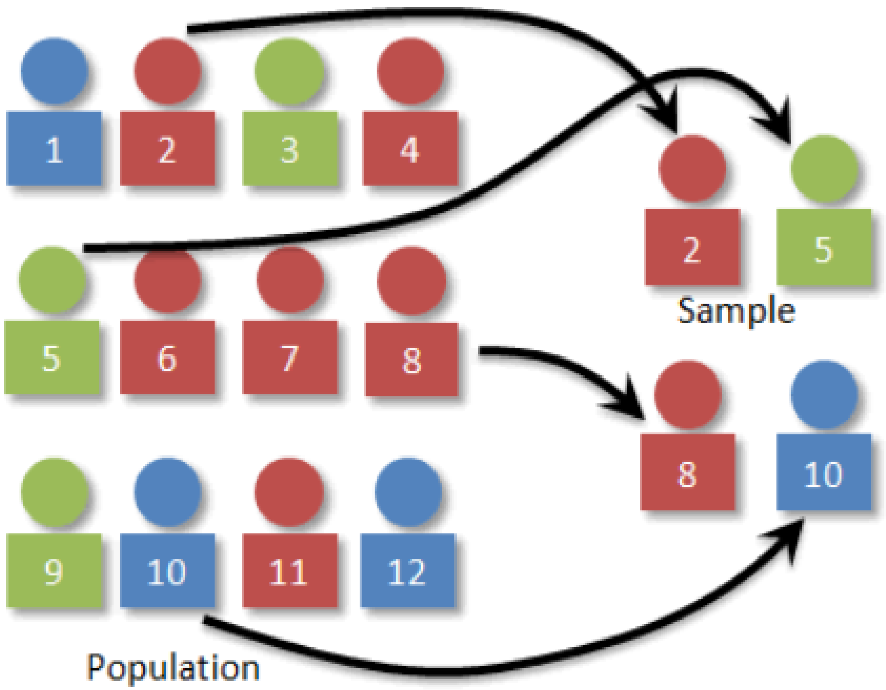


Population [1]

[1] https://faculty.elgin.edu/dkernler/statistics/ch01/1-3.html

Cont.

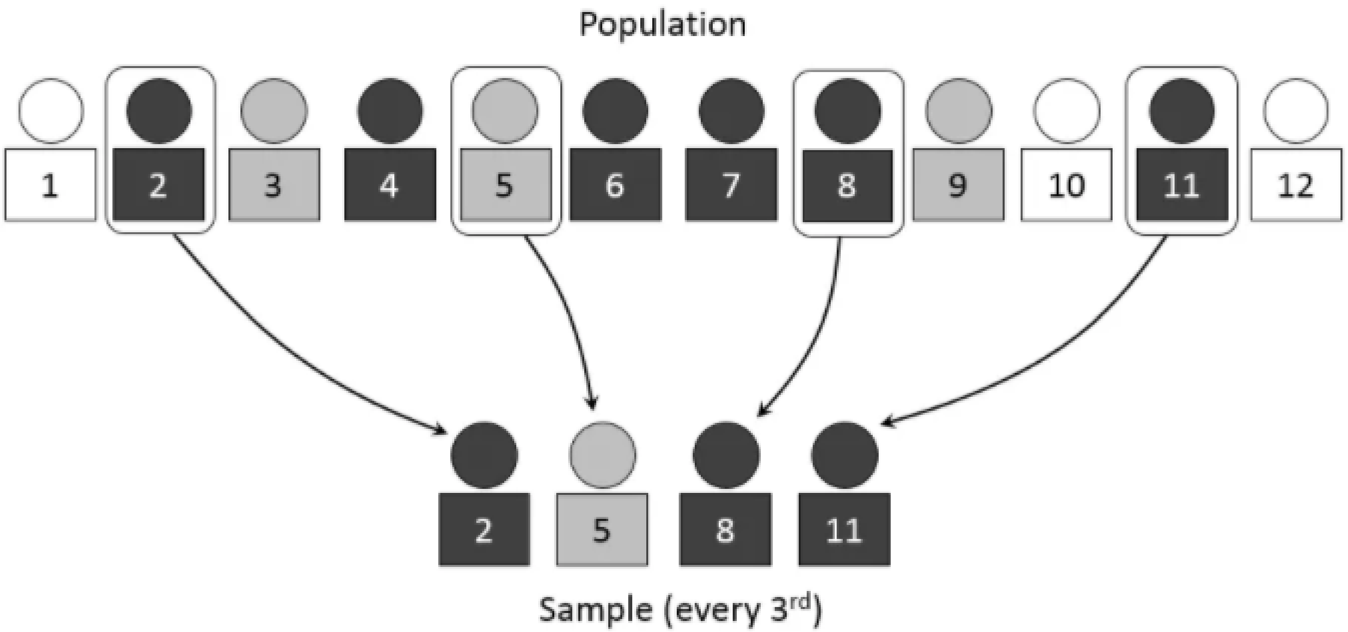
A possible sample from the give population:



Sampled 4 members [1]

Systematic sampling

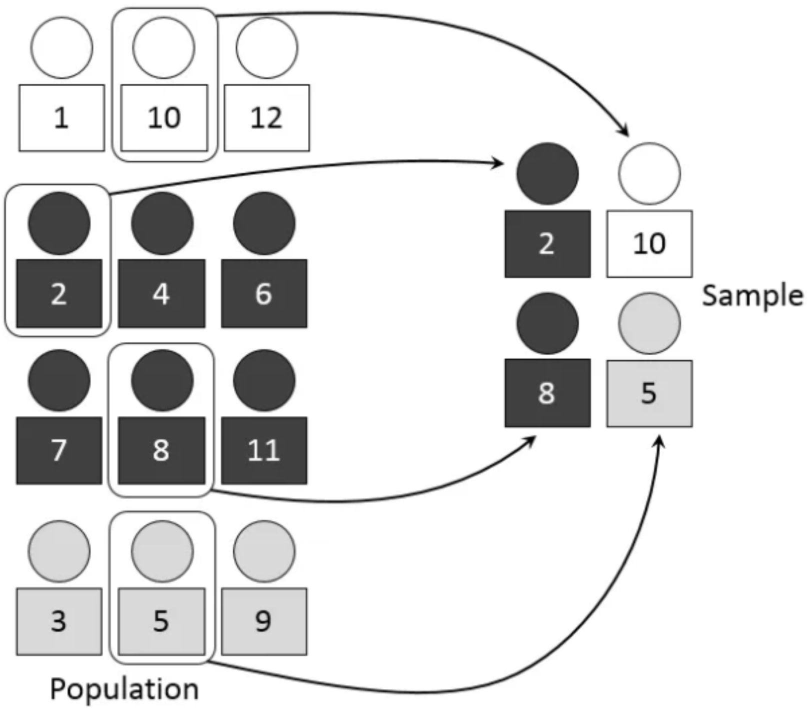
- Pick up the members from the population through a well-defined system to make a Sample.
- Sampling is done based on some given condition.



Source: Wikipedia

Stratified sampling

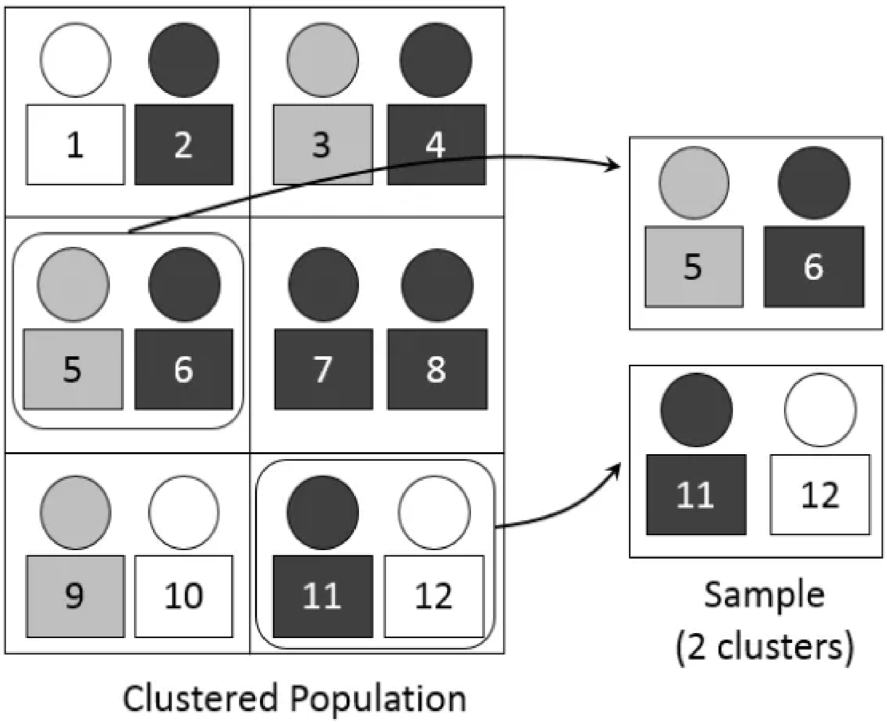
- First, stratify to make an ordered or categorized samples from the population called as **strata**.
- Then, choose members from each stratum for making a sample.



Source: Wikipedia

Cluster sampling

- Divide the population into groups call as clusters.
- Use simple random sampling to select the cluster to form a sample. (*mostly common in geography*)



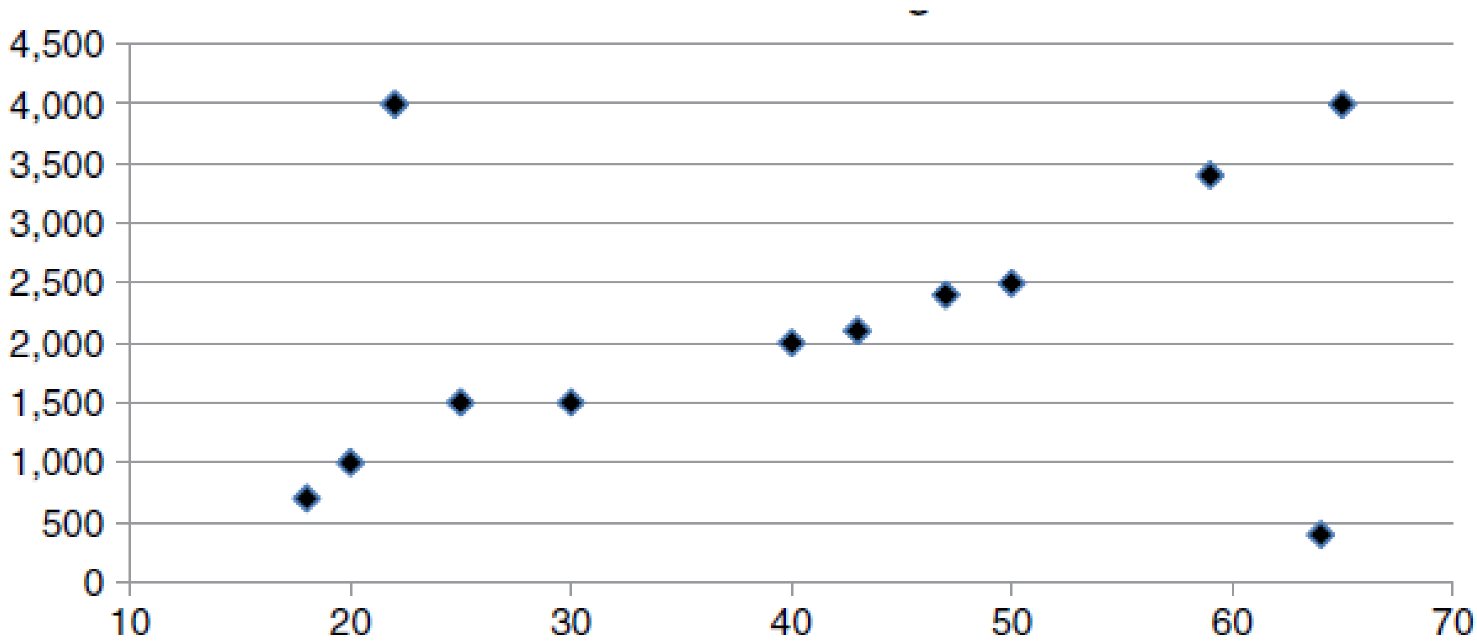
Source: Wikipedia

Outliers

- Outliers are extreme observations that are very dissimilar to the rest of the population.
- Two types of outliers:
  1. Valid observations (e.g., salary of boss is $1 million)
  2. Invalid observations (e.g., age is 300 years)
- Both (1) and (2) are univariate (outlying on 1 dimension).
- Multivariate outliers are observations that are outlying in multiple dimensions.

Cont.

- Multivariate outliers



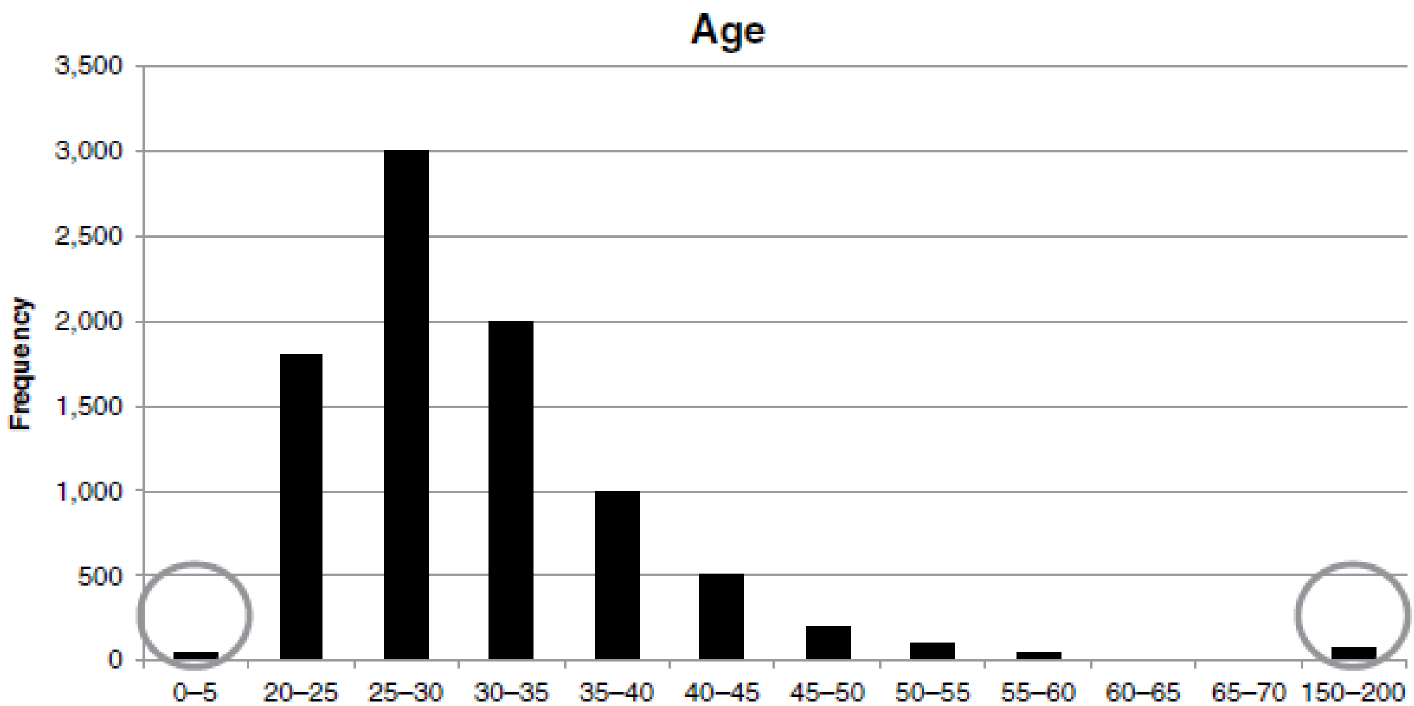
Income and age (Source: T4, Ch.2)

Detecting univariate outliers

- Two standard ways:
  - Visualizing data
  - Statistical measurements
- Visualization based methods:
  - Histogram
  - Box-plot
- Statistical measurements:
  - z-score

Cont.

- Using histogram for outlier detection:

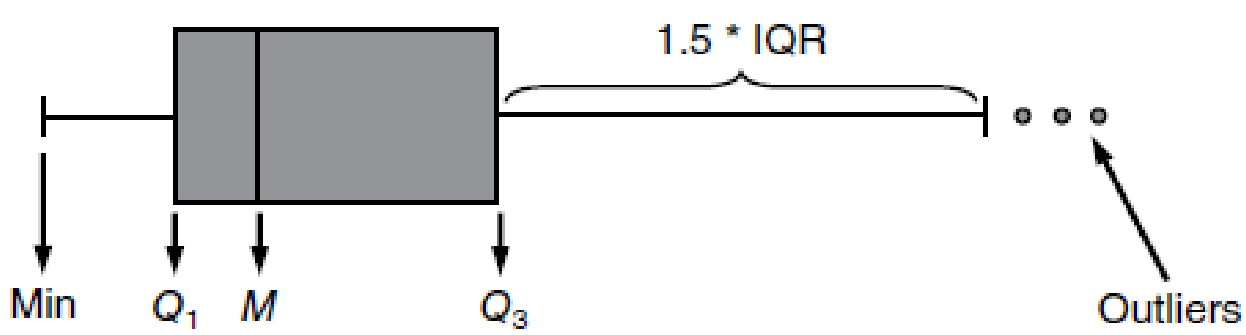


Example: Histogram of age (Source: T4, Ch.2)

Contd.

- Using box-plot for outlier detection:
  - It represents 3 key quartiles of the data: the first quartile (25% of the observations have a lower value), the median (50% of the observations have a lower value), and the third quartile (75% of the observations have a lower value).
  - All three quartiles are represented as a box.
  - The minimum and maximum values are then also added unless they are too far away from the edges of the box.
  - Too far away is then quantified as more than 1.5 * Interquartile Range:

$$IQR = Q3 - Q1$$



Box-plot (Source: T4, Ch.2)

Cont.

- z-score method for outlier detection:
  - Measures how many standard deviation an observation lies away from the mean.

$$z_i = \frac{x_i - \mu}{\sigma}$$

  - Here, μ is the mean of X; x_i is an observation that belongs to X; the denominator is the standard deviation.
  - A practical rule of thumb: Outliers when the absolute value of the *z*-score |z| is greater than 3.
  - Note: z-score relies on the normal distribution.

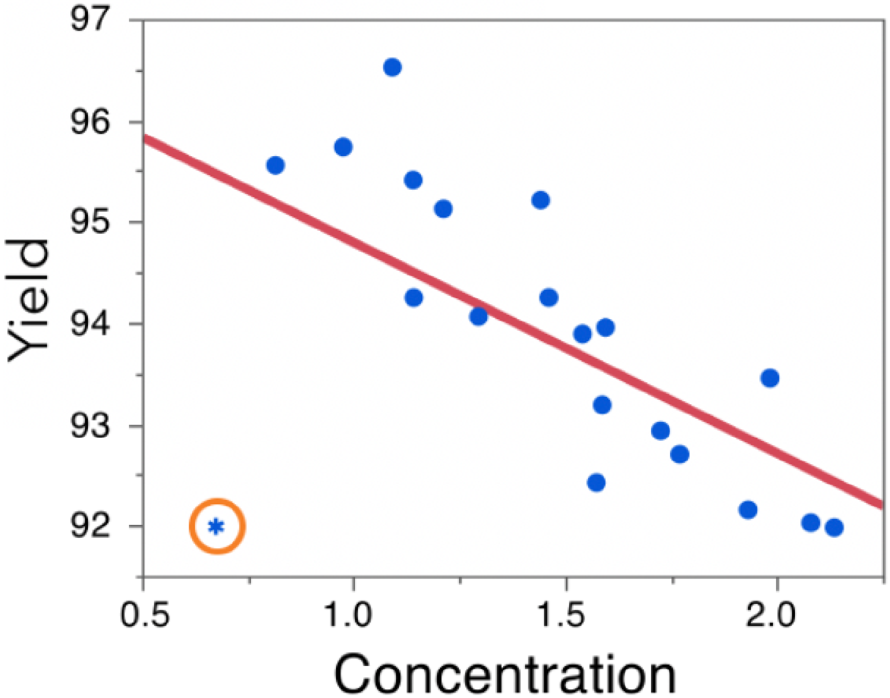
Cont.

Age	Z-Score
30	$(30 - 40)/10 = -1$
50	$(50 - 40)/10 = +1$
10	$(10 - 40)/10 = -3$
40	$(40 - 40)/10 = 0$
60	$(60 - 40)/10 = +2$
80	$(80 - 40)/10 = +4$
...	...
$\underline{\mu = 40}$ $\underline{\sigma = 10}$	$\underline{\mu = 0}$ $\underline{\sigma = 1}$

z-score based outlier detection for Age (Source: T4, Ch.2)

Detecting multivariate outliers

- Multivariate outliers can be detected by fitting regression lines and inspecting the observations with large errors (using, for example, a residual plot).



Source: [2]

[2] https://www.jmp.com/

Cont.

- Alternative methods are clustering or calculating the Mahalanobis distance.
- Note: Although potentially useful, multivariate outlier detection is typically not considered in many modeling exercises due to the typical marginal impact on model performance.

Standardizing Data

- It is a data preprocessing activity targeted at scaling variables to a similar range.
- Why should you standardize data?

Standardizing Data

- It is a data preprocessing activity targeted at scaling variables to a similar range.
- Why should you standardize data?
  - Consider, for example, two variables: gender (coded as 0/1) and income (ranging between $0 and $1 million).
  - When building logistic regression models using both information elements, the coefficient for income might become very small.
  - Hence, it could make sense to bring them back to a similar scale.

Cont.

- Methods:
  - Min/Max standardization
  - z-score standardization
  - Decimal scaling

Contd.

- Min/Max standardization

$$X_{new} = \frac{X_{old} - \min(X_{old})}{\max(X_{old}) - \min(X_{old})}(newmax - newmin) + newmin,$$

  - *newmax* and *newmin* are the newly imposed maximum and minimum (e.g., 1 and 0).

Contd.

- z-score standardization

$$z_i = \frac{x_i - \mu}{\sigma}$$

  - This becomes the standardized value for an observation x_i.

- Decimal scaling
  - Divide by a power of 10 as follows:

$$X_{new} = \frac{X_{old}}{10^n}$$

  - Here, *n* is the number of digits of the maximum absolute value.

Data Quality

- 4 categories:
  - Intrinsic
  - Contextual
  - Representational
  - Accessibility

Cont.

- Intrinsic category
  - **Accuracy**: Data are regarded as correct
  - **Believability**: Data are accepted or regarded as true, real, and credible
  - **Objectivity**: Data are unbiased and impartial
  - **Reputation**: Data are trusted or highly regarded in terms of their source and content

- Contextual category
  - **Value-added:** Data are beneficial and provide advantages for their use
  - **Completeness**: Data values are present
  - **Relevancy**: Data are applicable and useful for the task at hand
  - **Appropriate amount of data**: The quantity or volume of available data is appropriate

Cont.

- Representational category
  - **Interpretability**: Data are in appropriate language and unit and the data definitions are clear
  - **Ease of understanding**: Data are clear without ambiguity and easily comprehended

Cont.

- Accessibility category
  - **Accessibility**: Data are available or easily and quickly retrieved
  - **Security**: Access to data can be restricted and hence kept secure.

High-dimensional Data

- Dimensionality refers to the number of features (attributes) that defines a data point or a data instance.
- Example: In one of our experiments, we saw that the an Iris flower is described by 4 features:
  - Sepal length
  - Sepal width
  - Petal length
  - Petal width

- For Iris dataset, dimensionality is 4.

Cont.

There are some common misunderstandings and wrong definitions:

- Large number of features
  - My question: How much large is “large”? Is it 1000, 10000, 100000000?
- When our computer cannot handle the given data dimensions
  - My question: What if someone else’s computer can?
- There are many such definitions available all over the internet. Do not read such non-scientific definitions.

Cont.

Definition: Let a dataset contain *N* data instances. Each data instance can be described by *d* features. The dataset is called "high-dimensional" data, iff ***d* > *N***.

Example:

- *N*=5, *d* = 10: high-dimensional data
- *N*=100, *d*=101: high-dimensional data
- *N*=1000000, *d*=10000: not high-dimensional data

Question:

- A dataset of 1M images, of size 128×128×3. Will you call this dataset high-dimensional?

Cont.

Definition: Let a dataset contain N data instances. Each data instance can be described by d features. The dataset is called “high-dimensional” data, iff $\boldsymbol{d > N}$.

Example:

- N=5, d = 10: high-dimensional data
- N=100, d=101: high-dimensional data
- N=1000000, d=10000: not high-dimensional data

Question:

- A dataset of 1M images, of size 128×128×3. Will you call this dataset high-dimensional? Answer: **No**

Cont.

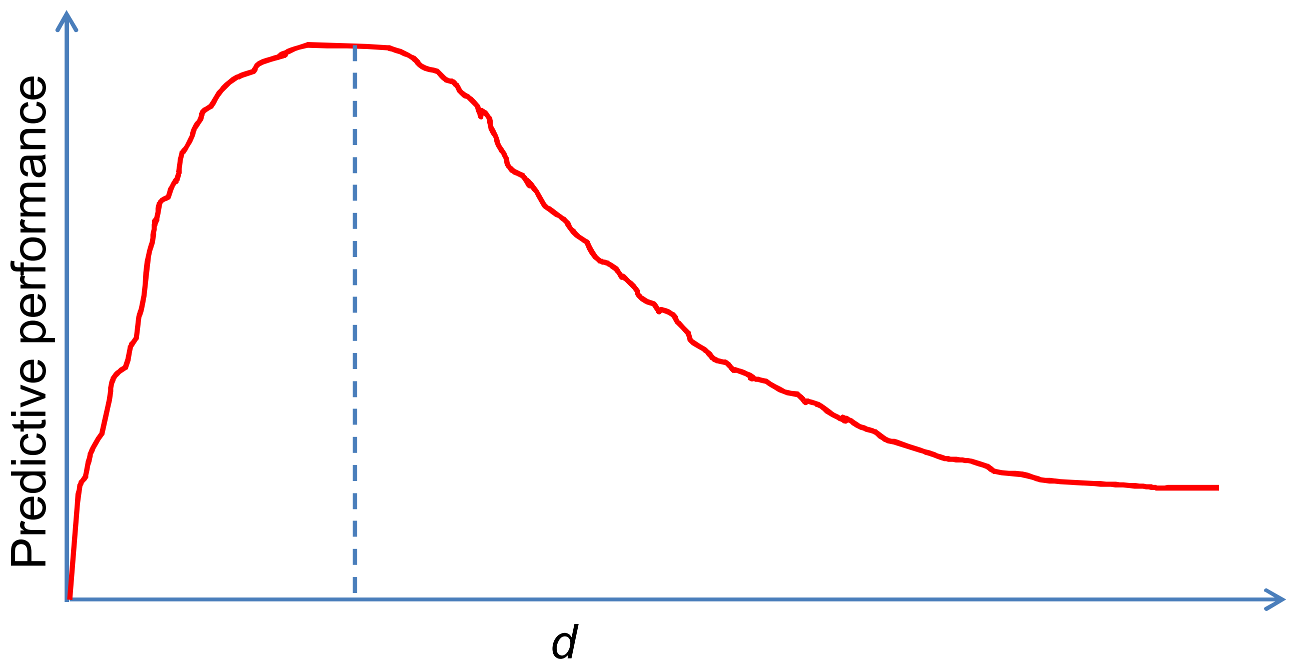
Is high-dimensionality a problem?

- Curse of dimensionality
- Nearest-neighbor problem

Cont.

Curse of dimensionality

- With a fixed number of training samples, the predictive power of a classifier or regressor first increases as d increases, and then decreases with *d*.



Cont.

Nearest-neighbor problem

- Let's look at the following example of determining nearest neighbor (used in *k*-NN, geoinformatics, etc.)

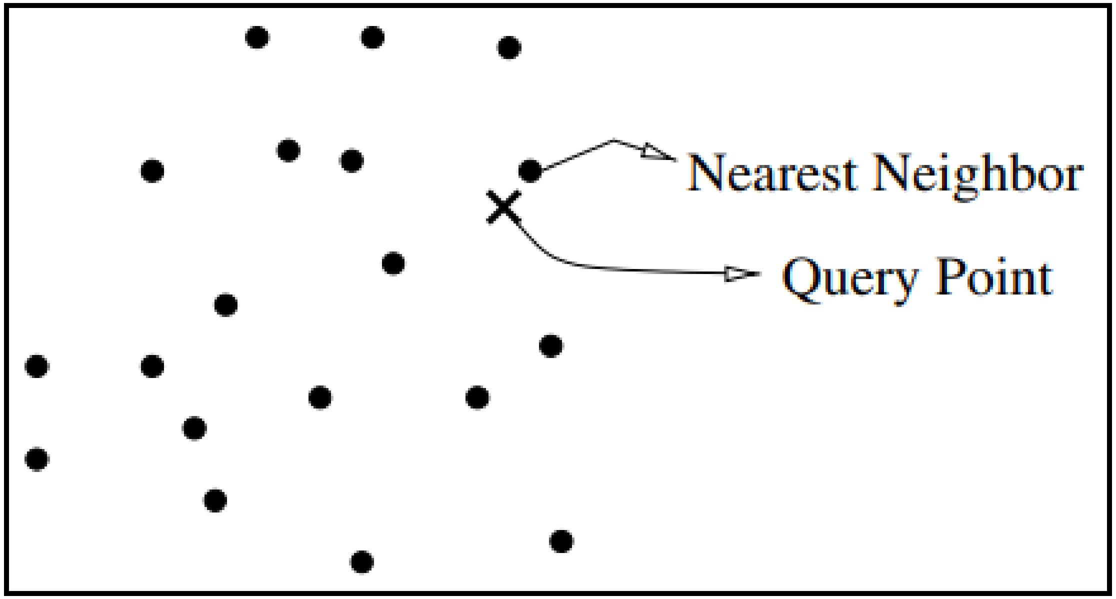


Fig 1: Query point (*Q*), and its neighbor [3]

[3] Beyer *et al.:* https://doi.org/10.1007/3-540-49257-7_15

Cont.

Nearest-neighbor problem

- But, what happens now?

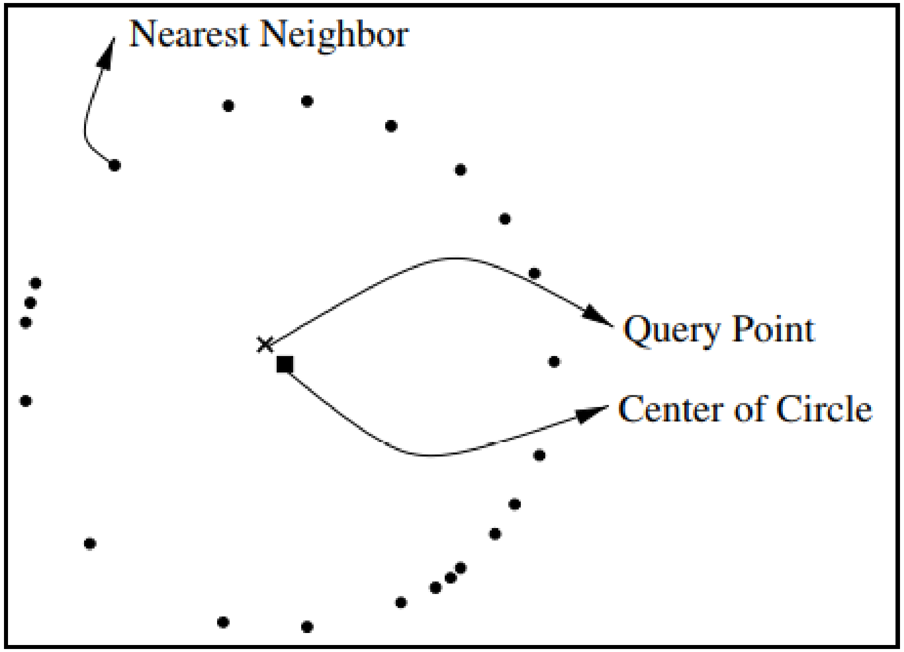


Fig 2: Query point (Q), and its neighbor [3]

[3] Beyer *et al.:* https://doi.org/10.1007/3-540-49257-7_15

Cont.

Nearest-neighbor problem

- So: the distance between the nearest neighbor (point at minimum distance) and any other point in the space is small.
- Specifically, given a query point *Q*, the minimum and maximum distance between *Q* and any list of *n* random points in the space follows:

$$\lim_{d\to\infty} \frac{dist_{max}(d) - dist_{min}(d)}{dist_{min}(d)} \to 0$$

- As *d* increases, distance function loses its usefulness.

Data Models

- We will mean a data model as a statistical model.
- It provides:
  - Quantitative summary of the data
  - Impose a specific structure on the population
- A trivial model is no model at all.
  - Example: Let say, you are building a new product. You did a market survey on a set of people on ‘how much would they be willing to pay for the new product?’ And you got the following price data:

10, 20, 13, 7, 8, …, 11, 13, 27, …, 10, 7, 4

  - A trivial model could be: you draw the histogram of prices i.e. X-axis will be each discrete price, and Y-axis will be frequency.

Data Models

- We will mean a data model as a statistical model.
- It provides:
  - Quantitative summary of the data
  - Impose a specific structure on the population
- A trivial model is no model at all.
  - Example: Let say, you are building a new product. You did a market survey on a set of people on ‘how much would they be willing to pay for the new product?’ And you got the following price data:

    10, 20, 13, 7, 8, …, 11, 13, 27, …, 10, 7, 4

  - A trivial model could be: you draw the histogram of prices i.e. X-axis will be each discrete price, and Y-axis will be frequency.
  - Difficulty: You can’t reliably question like “Will anyone buy my product if I sell it for 30?”

Cont.

Models as expectations

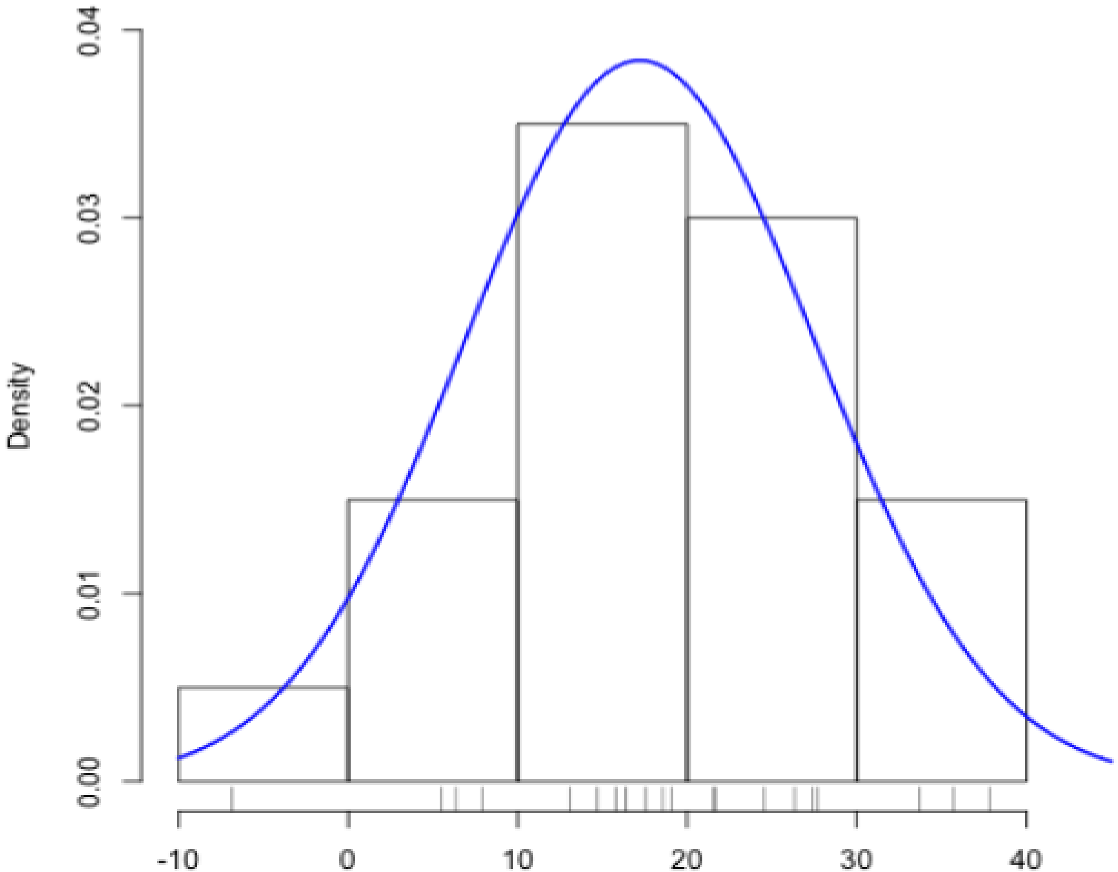
- Imposes a structure on the population
- We will call this structure as a 'distribution'.
- The most common distribution is normal distribution. This is called a normal model, denoted as $\mathcal{N}(\mu, \sigma^2)$
- The probability density function (PDF) is:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \qquad \mu : \text{mean}, \sigma^2 : \text{variance}$$

- We can assume that this distribution generated the obtained data.

Cont.

Models as expectations

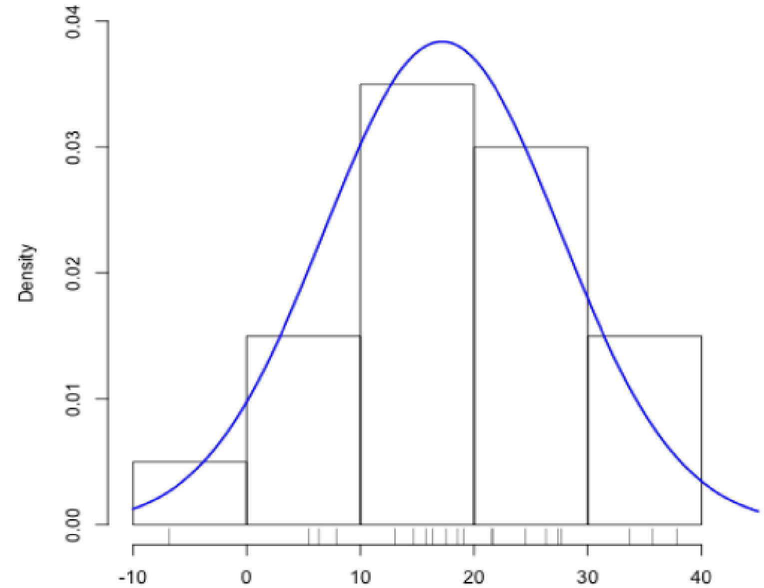


X: price, Y: density (Source: T3, Ch.5)

Cont.

Models as expectations

- Now we can answer questions.



X: price, Y: density (Source: T3, Ch.5)

- What is the probability that the product can be sold for 30? (*Homework!*)

Cont.

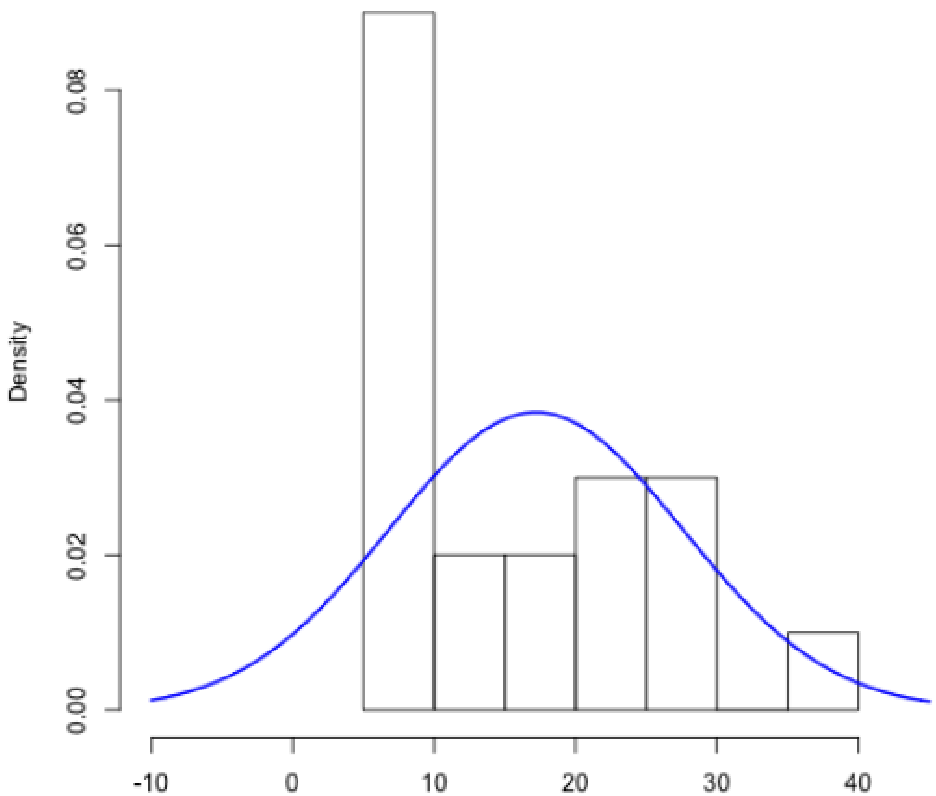
Here are the steps:

- You have the mathematical function for the density.
- Plug in the value of 30 for *x* in the PDF and obtain the answer.
- The mean and variance are the mean and variance of the data sampled during the market survey.

Cont.

Issues with normal model

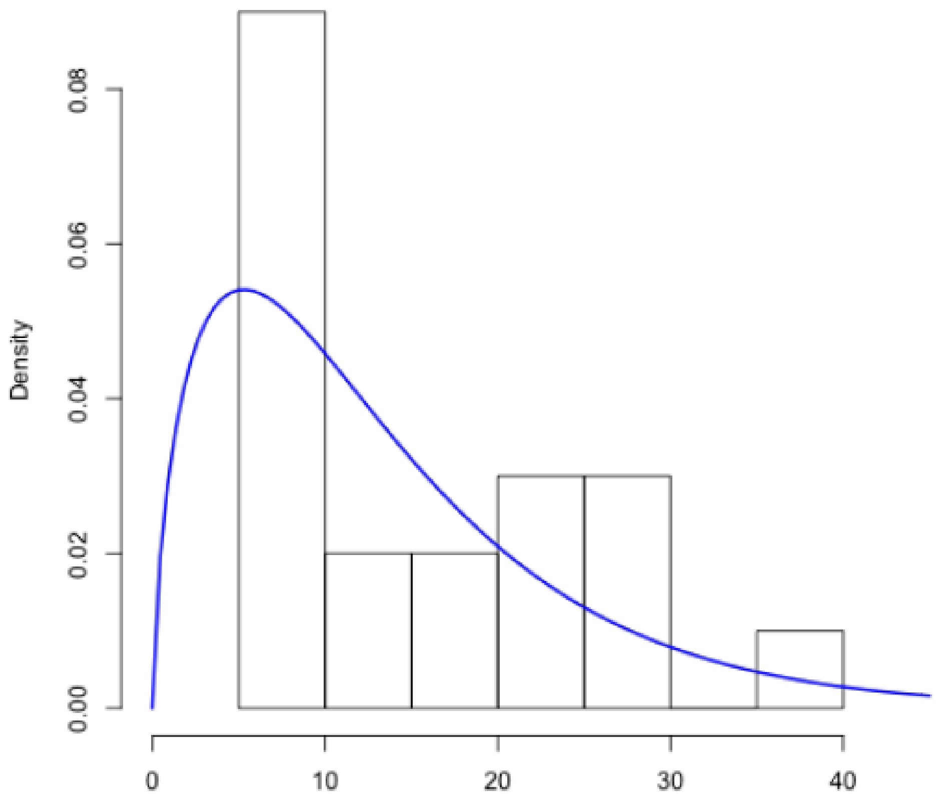
- Assumptions: Normal model will fit the data. What if it does not.



X: price, Y: density (Source: T3, Ch.5)

Cont.

Solution 1: Get a different model (*e.g.* gamma distribution)



X: price, Y: density (Source: T3, Ch.5)

Cont.

Solution 2: Get different or more data: perform the survey again to collect more data or replace the existing data. Getting more data is better.