Unimodal Empirical Formula: mean – mode = 3(mean - median)

Sample Variance:

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

Population Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2$$

Chi-square test:

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

How to derive 90?
450/1500 * 300 = 90

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

We can reject the null hypothesis of independence at a confidence level of 0.001

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

z-score normalization

- Z-score: $z = \frac{x-\mu}{\sigma}$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation

min-max normalization

Min-max normalization: to [new_min_A, new_max_A]

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

- $p = 1$: (L_1 norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{il} - x_{jl}|$
- $p = 2$: (L_2 norm) Euclidean distance

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

- $p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p\to\infty}\sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l}|x_{if} - x_{jf}|$$

$$(a_1, a_2, a_3, a_4, a_5, a_6, a_7) : 1, (a_1, b_2, a_3, b_4, a_5, b_6, a_7) : 1$$

where $a_i \neq b_i$ for $i = 2, 4, 6$. Assume each dimension contains no concept hierarchy.

(1) (6 points) Please list all the (nonempty) closed cells in this data cube.
Solution: The closed cells in this data cube include:
$(a_1, a_2, a_3, a_4, a_5, a_6, a_7) : 1, (a_1, b_2, a_3, b_4, a_5, b_6, a_7) : 1,$
$(a_1, *, a_3, *, a_5, *, a_7) : 2.$

(2) (4 points) How many (nonempty) aggregate cells are there in this data cube?
Solution: $2 * 2^7 - 2 - 2^4 = 238.$

(3) (4 points) How many (nonempty) aggregate closed cells are there in this data cube? Please list them.
Solution: One. $(a_1, *, a_3, *, a_5, *, a_7) : 2.$

(4) (4 points) If we set minimum support = 2, how many (nonempty) aggregate cells are there in the corresponding iceberg cube?
Solution: 2^4.

χ^2:
Answer: $[0, +\infty)$
Pearson correlation coefficient:
Answer: $[-1, +1]$

*Drill down on Department from * to College-level*
*Drill down on Time dimension from * to year-level*
Dice on (i.e., select) college = "Engineering" and Year = "2007"
Drill down on Time to season and slice on season = "Spring"
Drill down on Department to the department-level
Drill down on Student dimension to student name (or ID)
Select top 10 GPA values, and print the corresponding student names

(i) algebraic: average, variance
(ii) holistic: median, Q_1, rank

BoxPlot
A five-number summary is a detailed descriptive statistics of a variable.
1.minimum value 2.Quartile 1 3.Median(quartile 2) 4.Quartile 3 5.Maximum value

Boxplot is the diagramatical visual presentation of the five number summary of the data.
The boxplot can also give us the information regarding:
1.OUTLIERS 2.SYMMTERY of the data(left skewed,symmetric,right skewed)

Two distributions can have same boxplots:
if they have same five number summary i.e
1.minimum value 2.Quartile 1 3.Median(quartile 2) 4.Quartile 3 5.Maximum value

Describe what quantile plots are.
Describe what quantile-quantile plots are.
How is a quantile-quantile plot di erent from a quantile plot? Clearly explain.

Quantile Plot
The quantile is the value such that some fixed proprotion of a distribution is less than equal to that.
quantile -quantile plots are the graphical method for determining whether the two samples came from same population or not.

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
Graphs the quantiles of one univariate distribution against the corresponding quantiles of another

How many cuboids are needed to preprocess all?

$$T = \prod_{i=1}^{n} (L_i + 1)$$