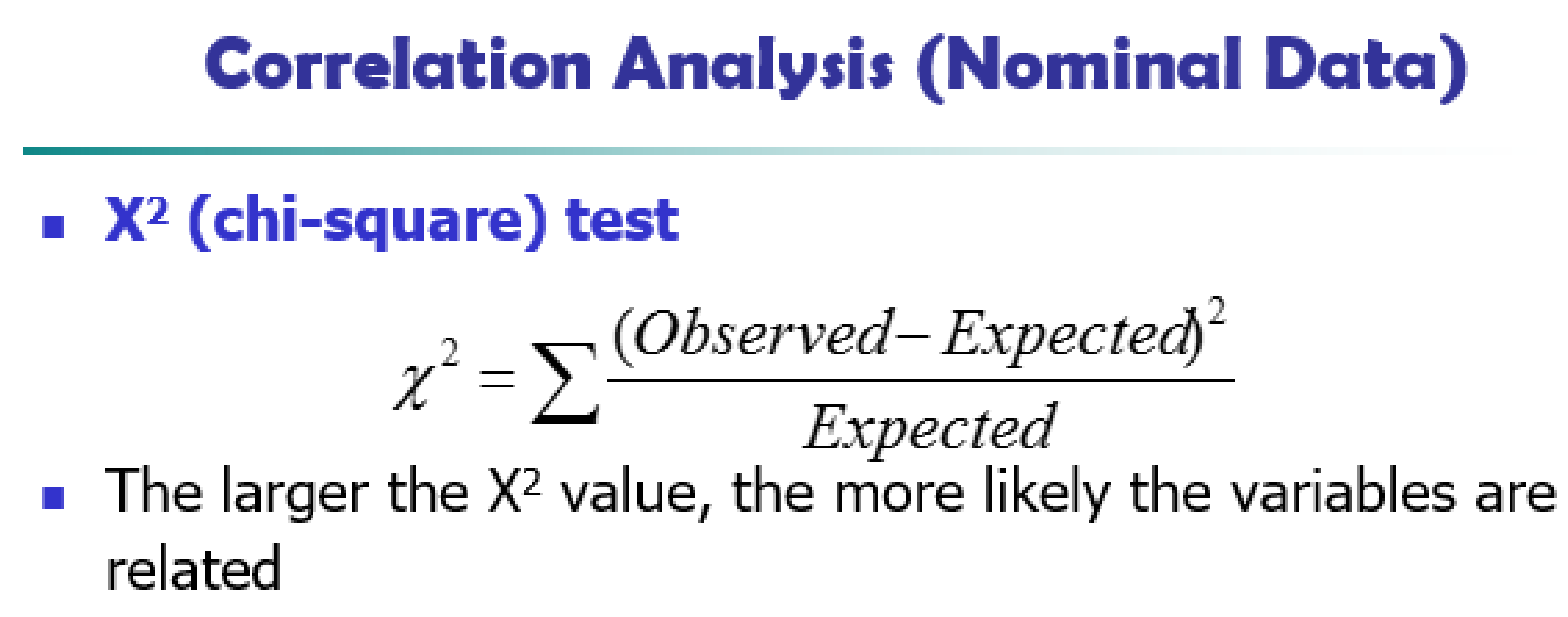
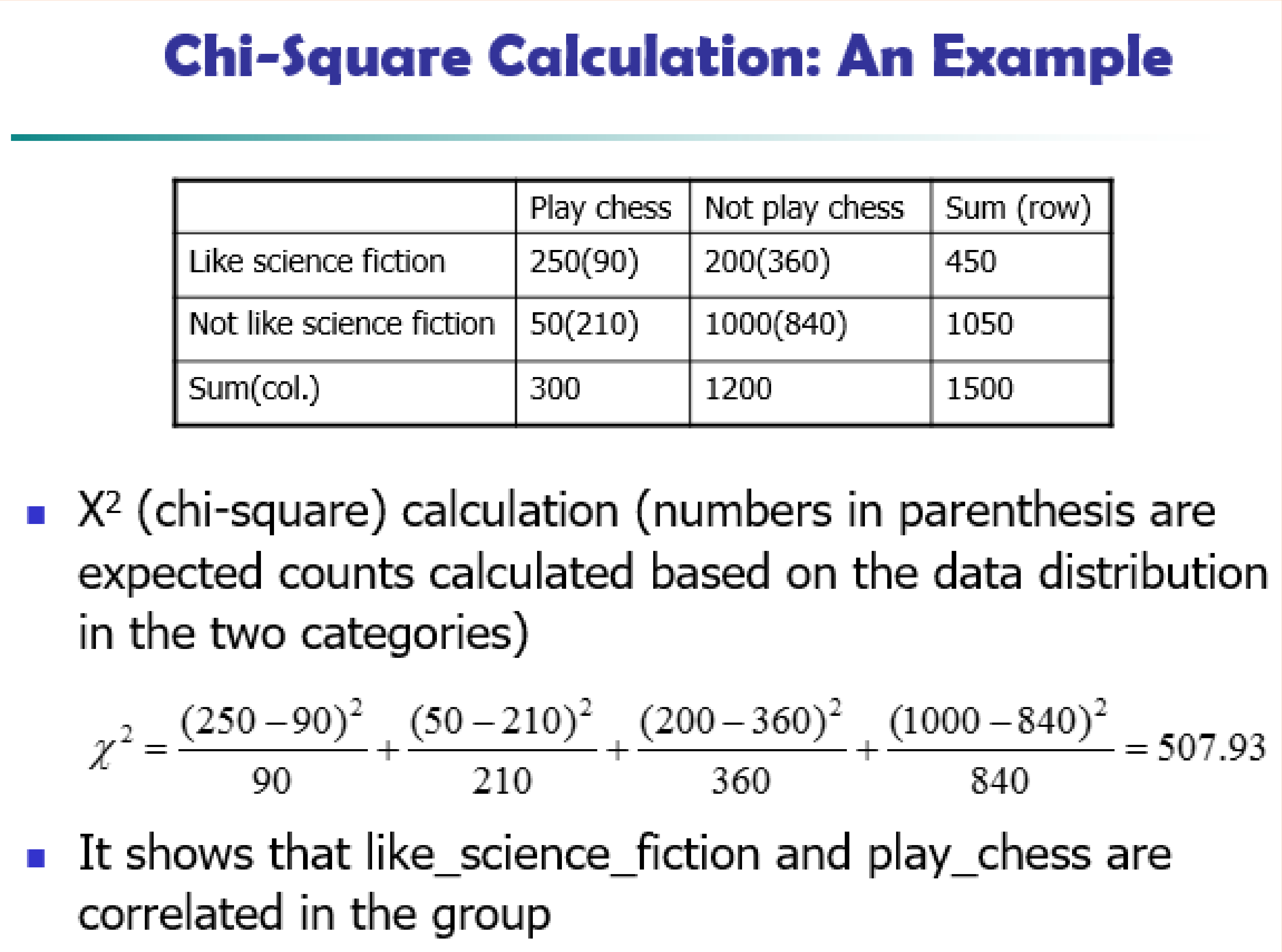


Correlation Analysis (Nominal Data)

- **X² (chi-square) test**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X² value, the more likely the variables are related



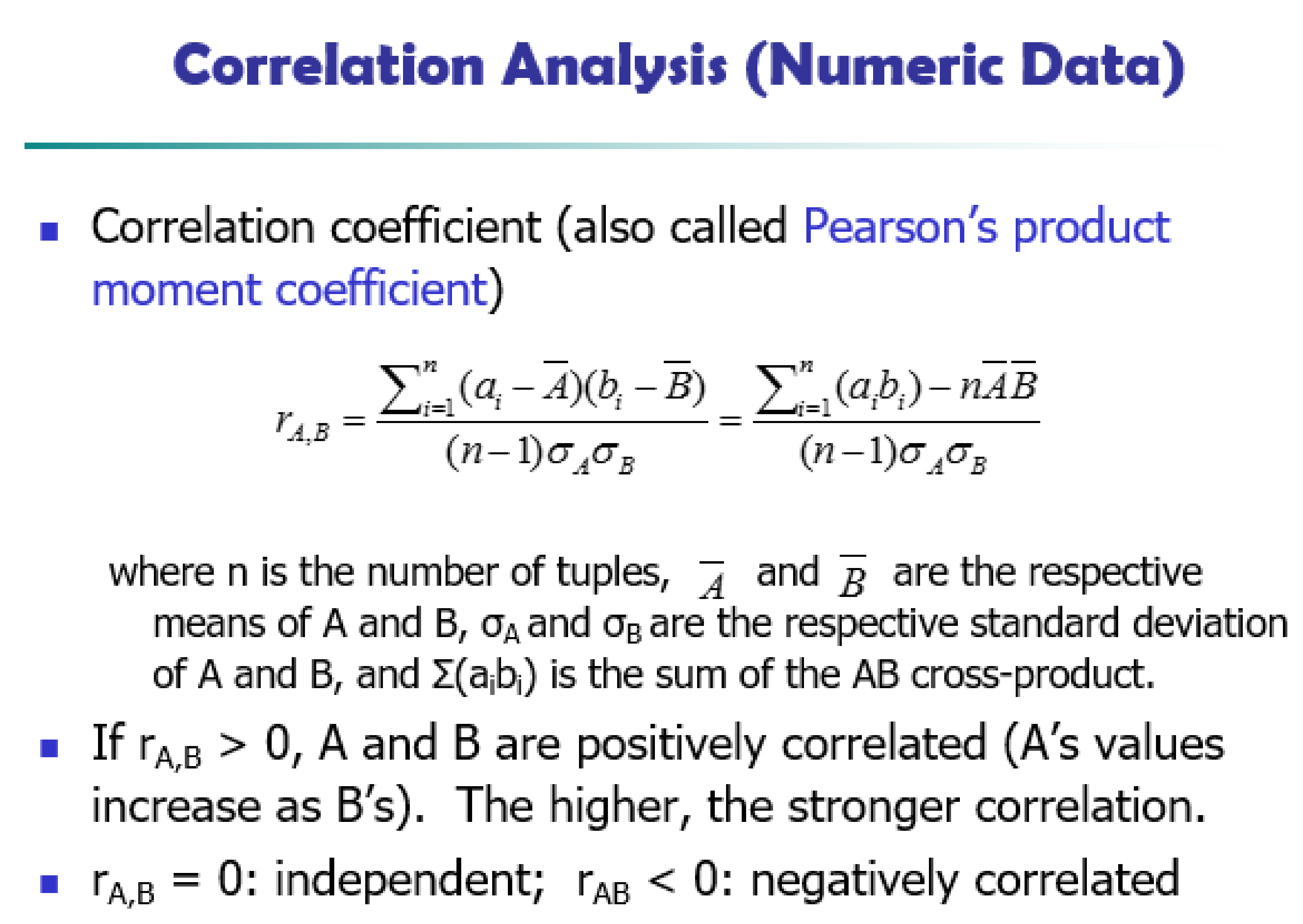
Chi-Square Calculation: An Example

	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

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

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

- It shows that like_science_fiction and play_chess are correlated in the group



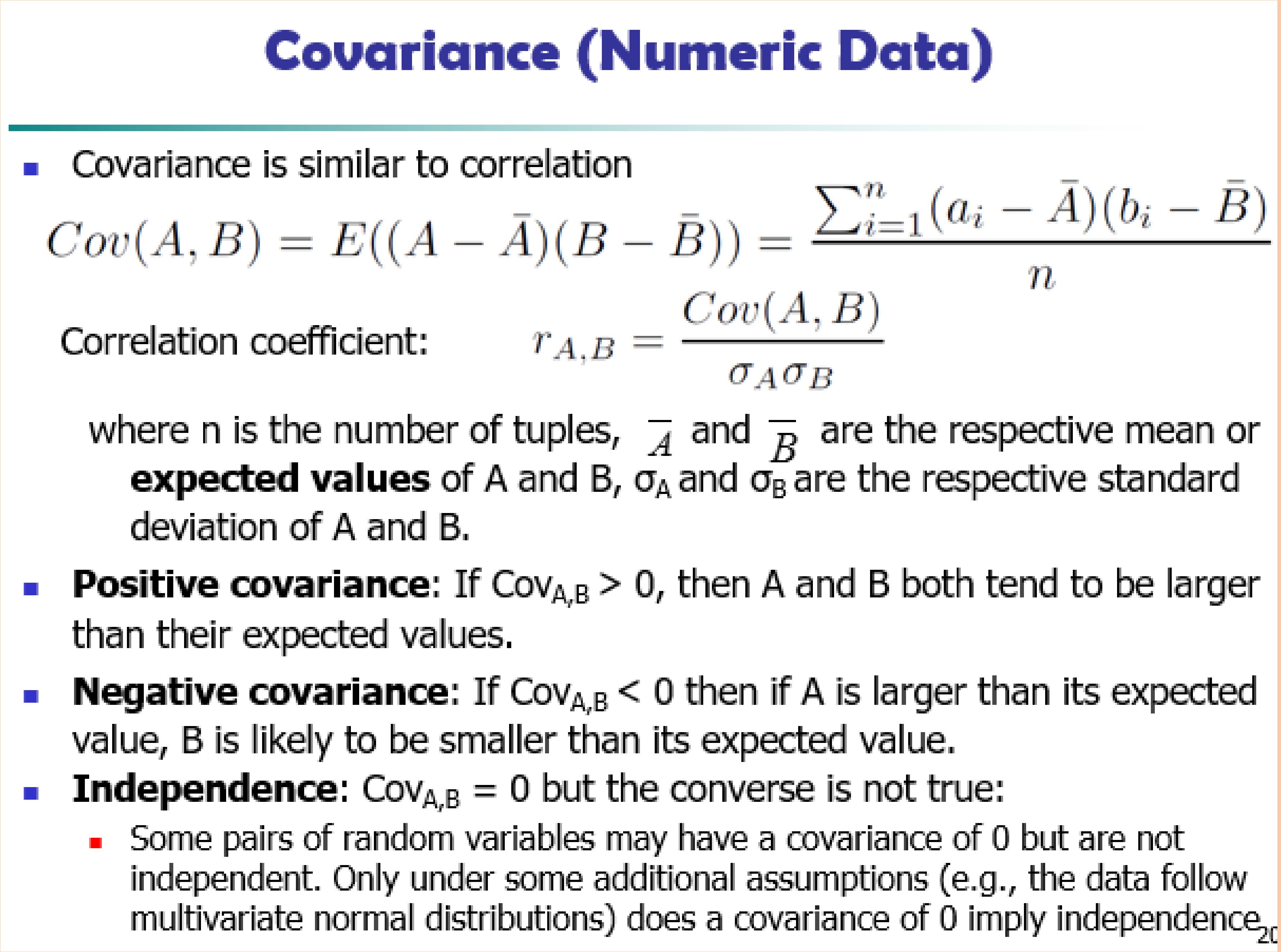
Correlation Analysis (Numeric Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated



Covariance (Numeric Data)

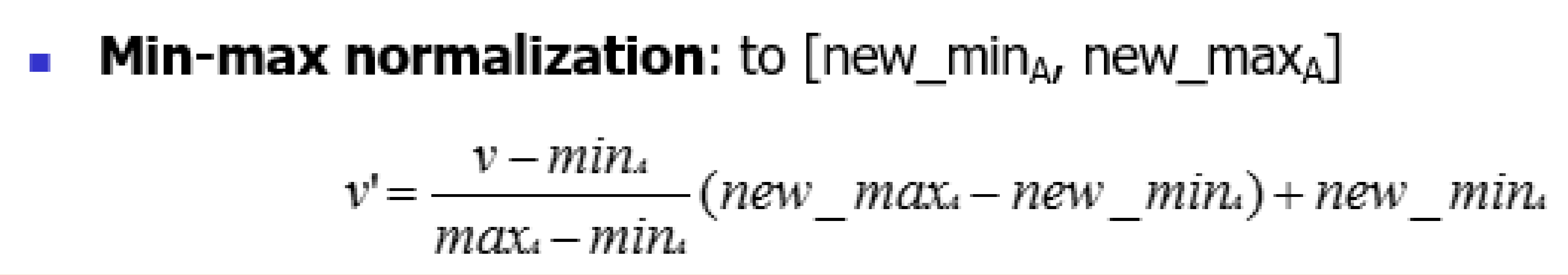
- Covariance is similar to correlation

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient: $$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

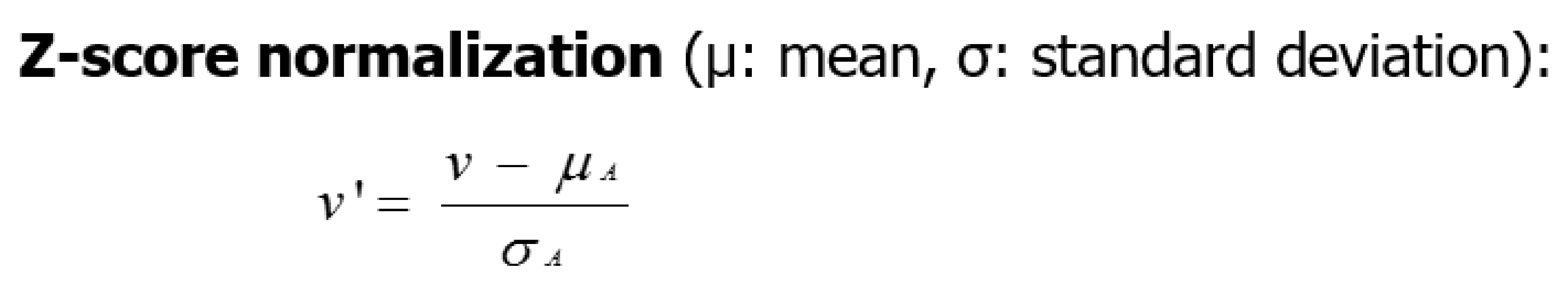
where n is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence



- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

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



Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

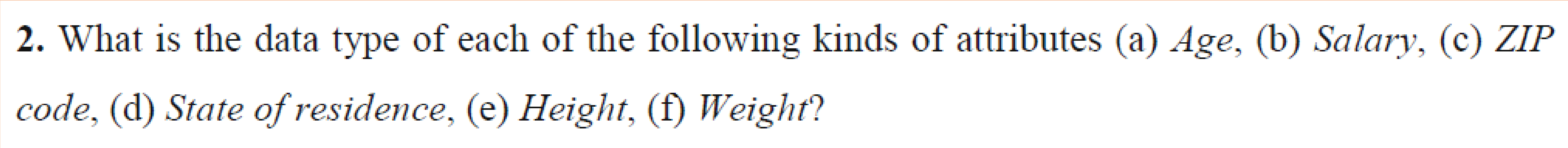
1. An analyst collects surveys from different participants about their likes and dislikes. Subsequently, the analyst uploads the data to a database, corrects erroneous or missing entries, and designs a recommendation algorithm on this basis.

Which of the following actions represent data collection, data preprocessing, and data analysis?

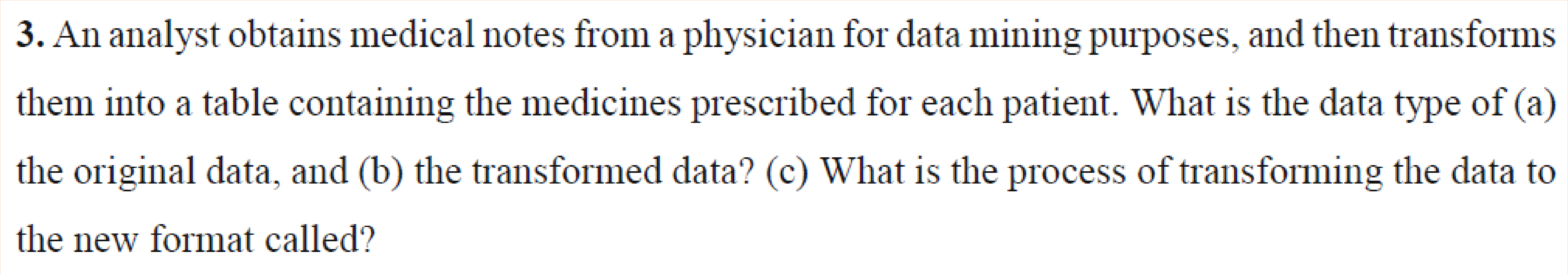
(a)Conducting surveys and uploading to database,

(b) correcting missing entries,

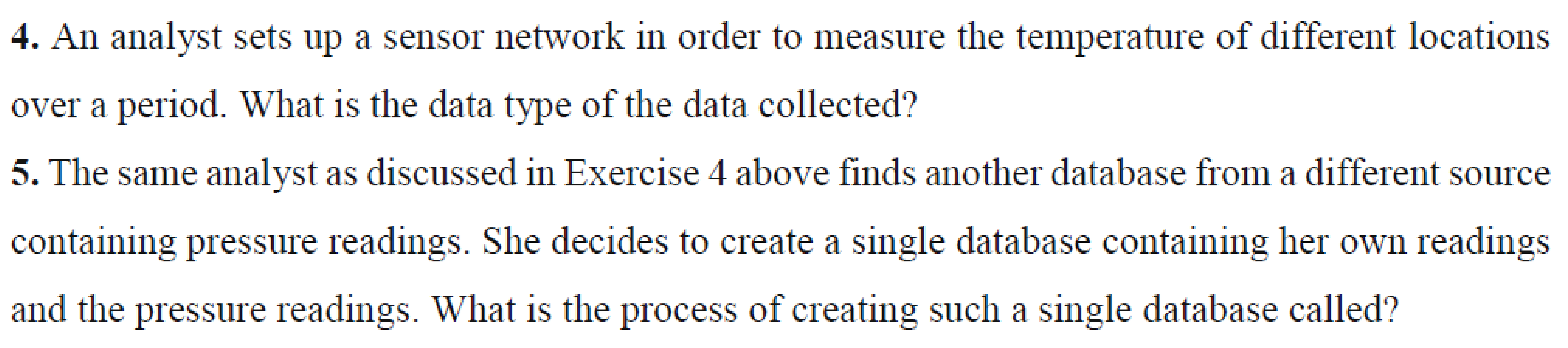
(c) designing a recommendation algorithm.



2. What is the data type of each of the following kinds of attributes (a) *Age*, (b) *Salary*, (c) *ZIP code*, (d) *State of residence*, (e) *Height*, (f) *Weight*?

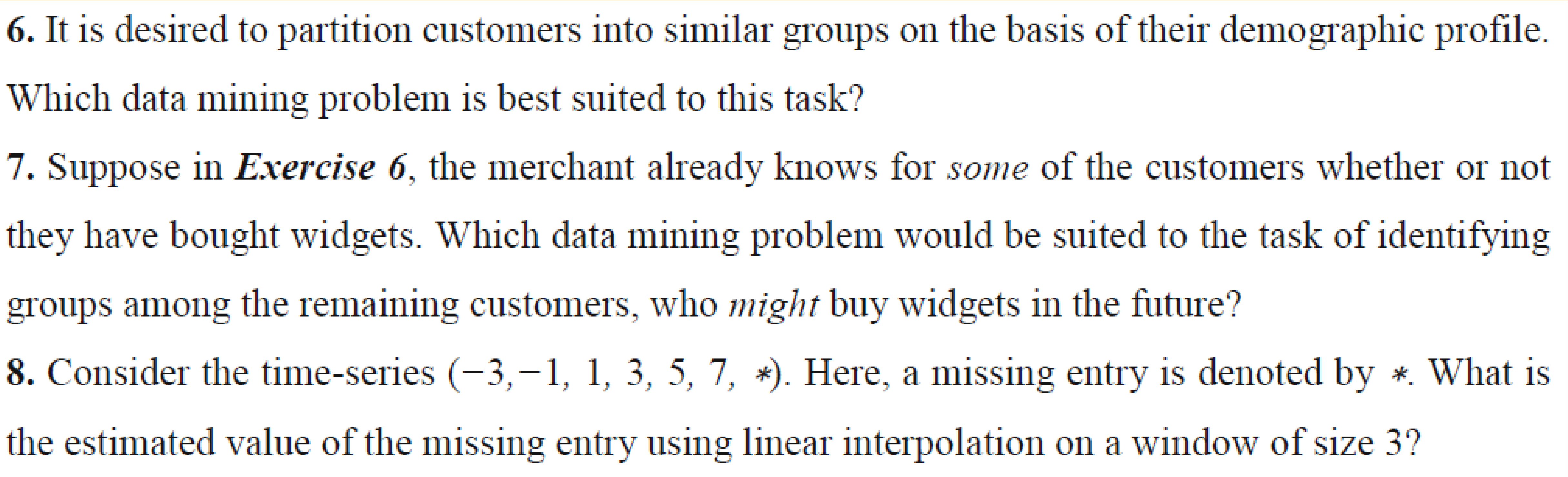


3. An analyst obtains medical notes from a physician for data mining purposes, and then transforms them into a table containing the medicines prescribed for each patient. What is the data type of (a) the original data, and (b) the transformed data? (c) What is the process of transforming the data to the new format called?



4. An analyst sets up a sensor network in order to measure the temperature of different locations over a period. What is the data type of the data collected?

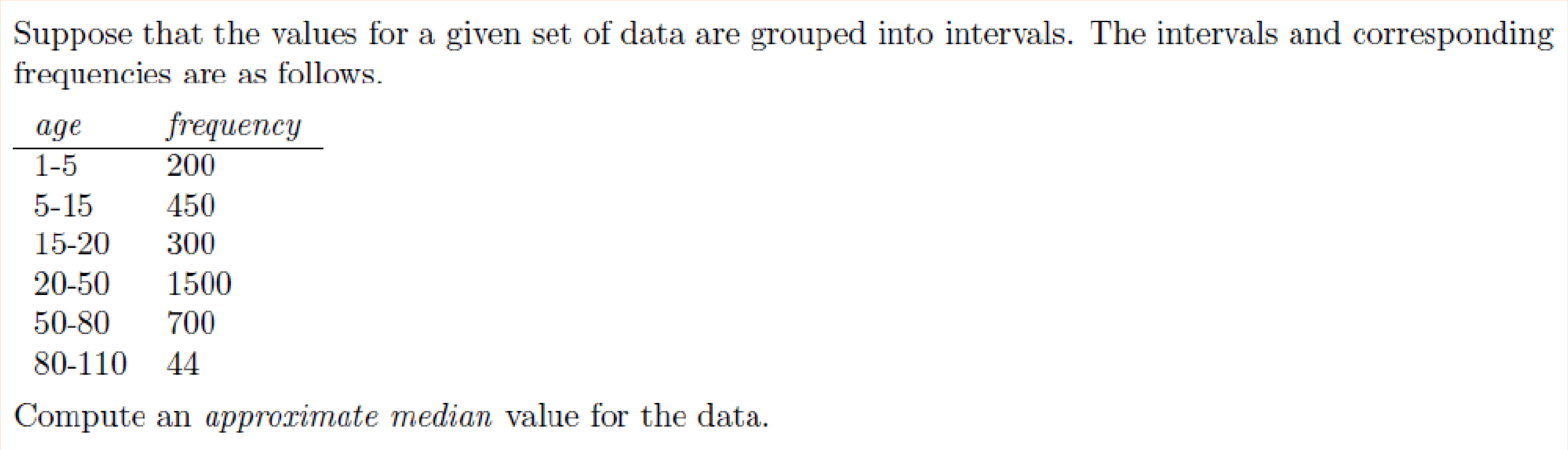
5. The same analyst as discussed in Exercise 4 above finds another database from a different source containing pressure readings. She decides to create a single database containing her own readings and the pressure readings. What is the process of creating such a single database called?



6. It is desired to partition customers into similar groups on the basis of their demographic profile. Which data mining problem is best suited to this task?

7. Suppose in ***Exercise 6***, the merchant already knows for *some* of the customers whether or not they have bought widgets. Which data mining problem would be suited to the task of identifying groups among the remaining customers, who *might* buy widgets in the future?

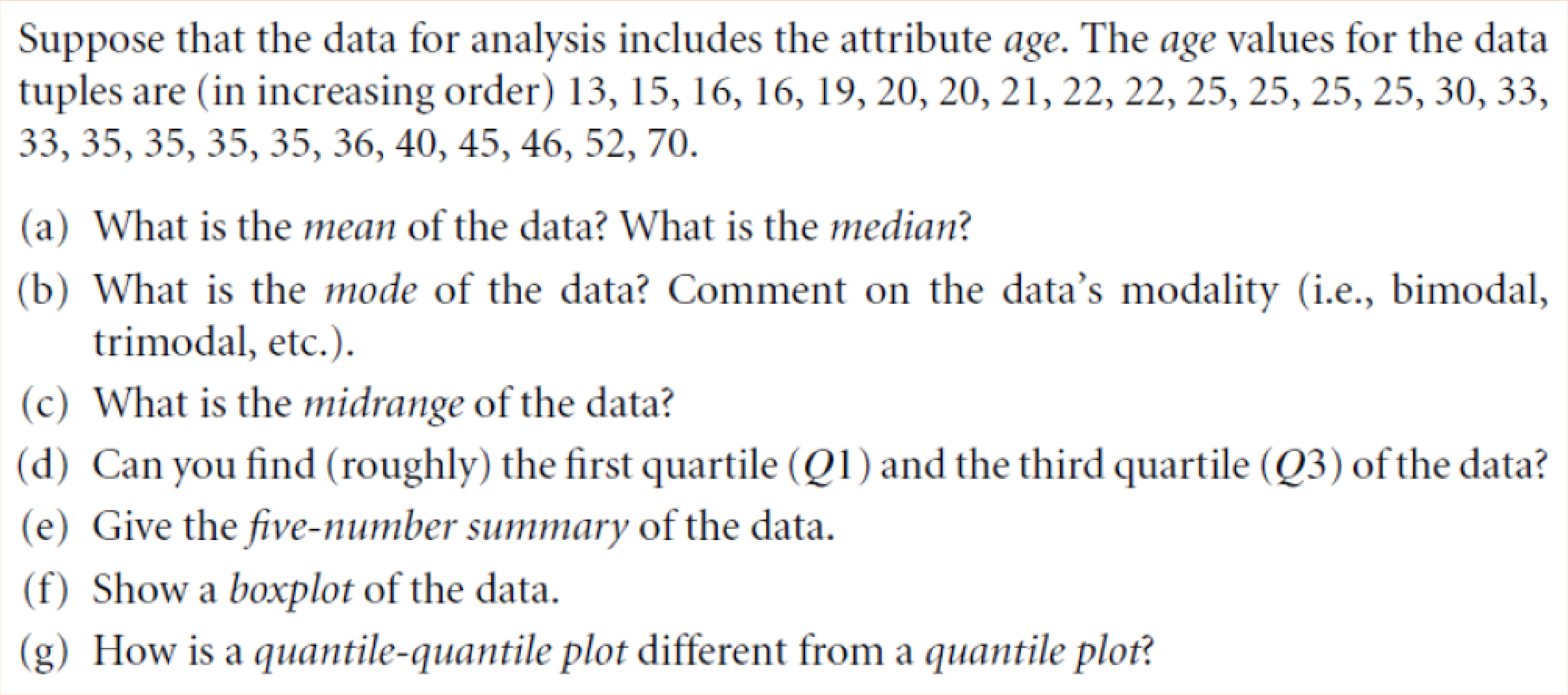
8. Consider the time-series $(-3,-1,\ 1,\ 3,\ 5,\ 7,\ *)$. Here, a missing entry is denoted by $*$. What is the estimated value of the missing entry using linear interpolation on a window of size 3?



Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows.

age	*frequency*
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

Compute an *approximate median* value for the data.



Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the *mean* of the data? What is the *median*?

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

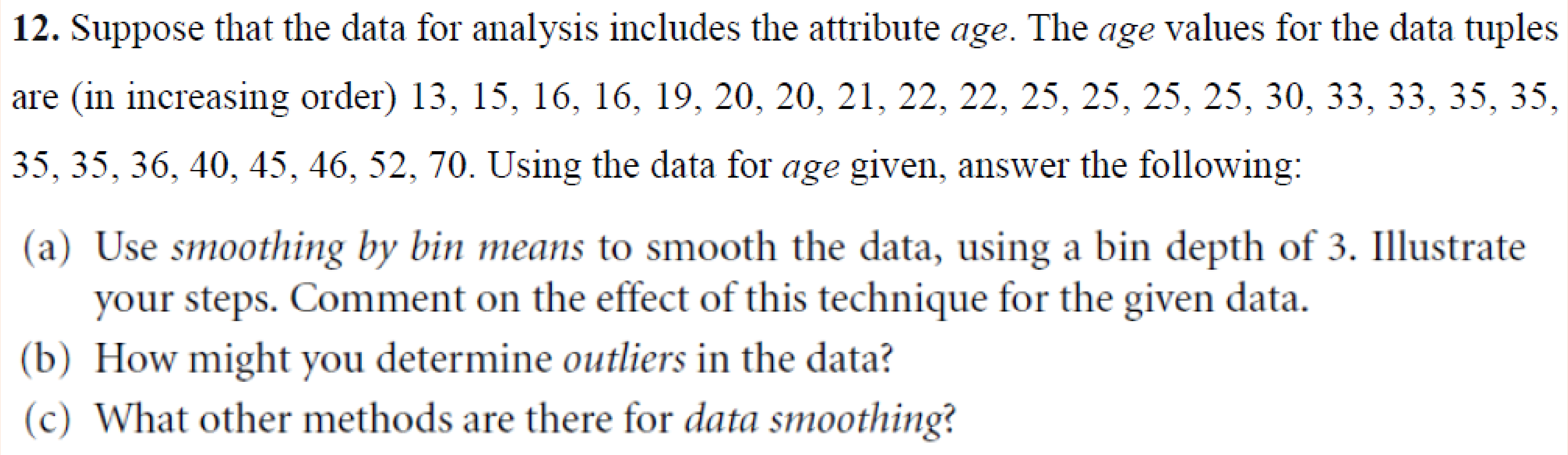
(c) What is the *midrange* of the data?

(d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

(e) Give the *five-number summary* of the data.

(f) Show a *boxplot* of the data.

(g) How is a *quantile-quantile plot* different from a *quantile plot*?

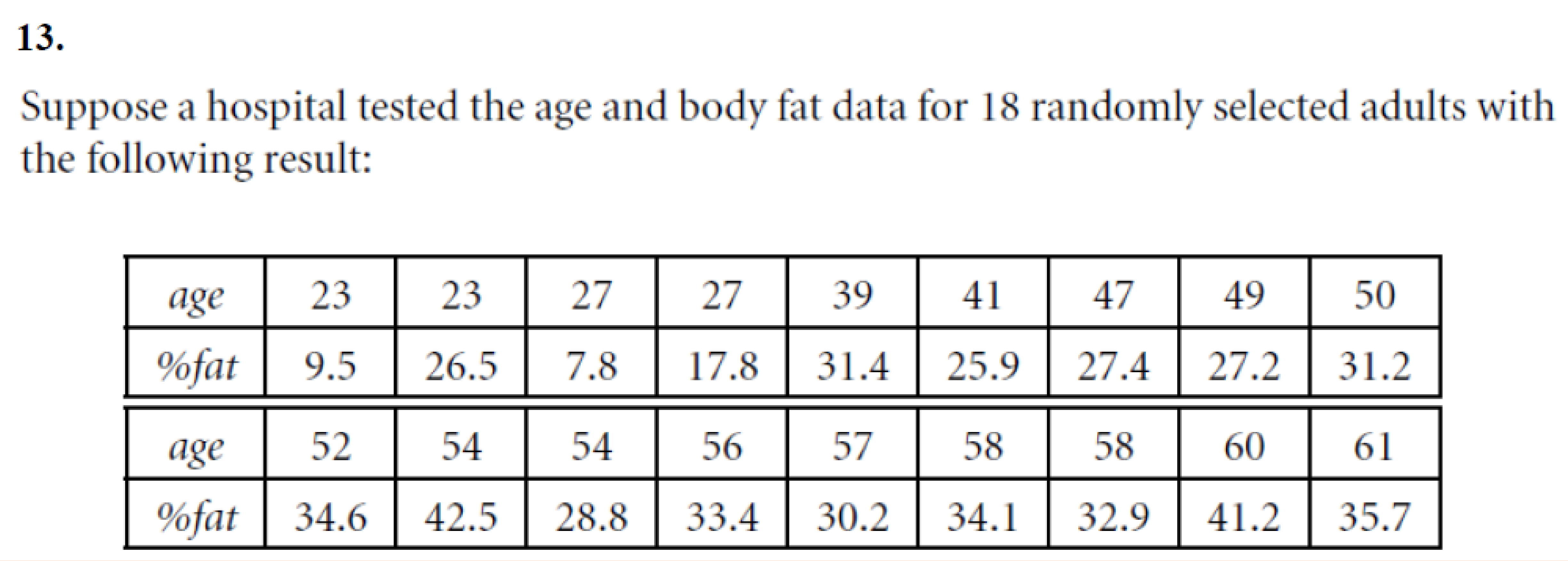


12. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Using the data for *age* given, answer the following:

(a) Use *smoothing by bin means* to smooth the data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

(b) How might you determine *outliers* in the data?

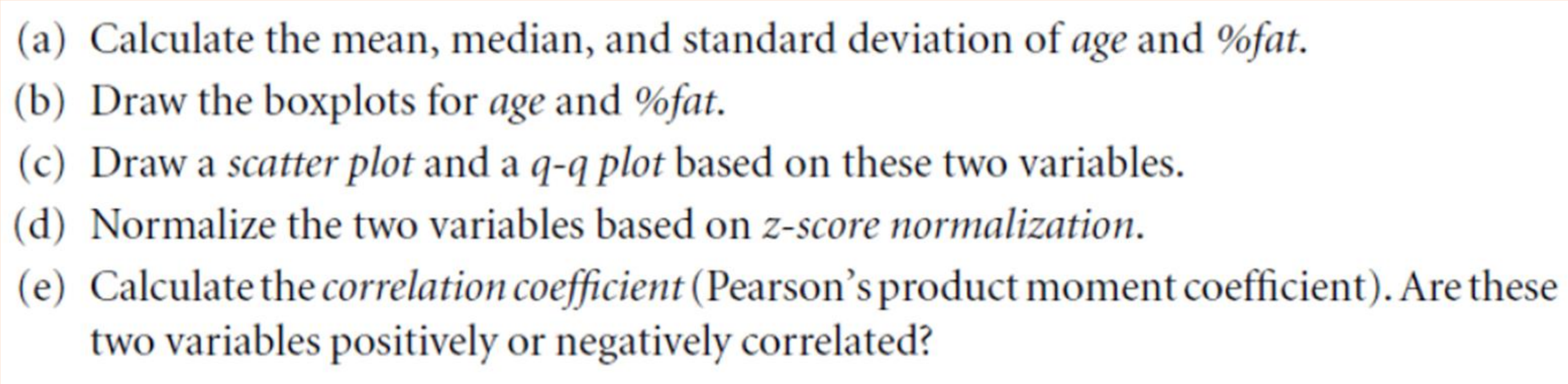
(c) What other methods are there for *data smoothing*?



13.

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result:

age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7



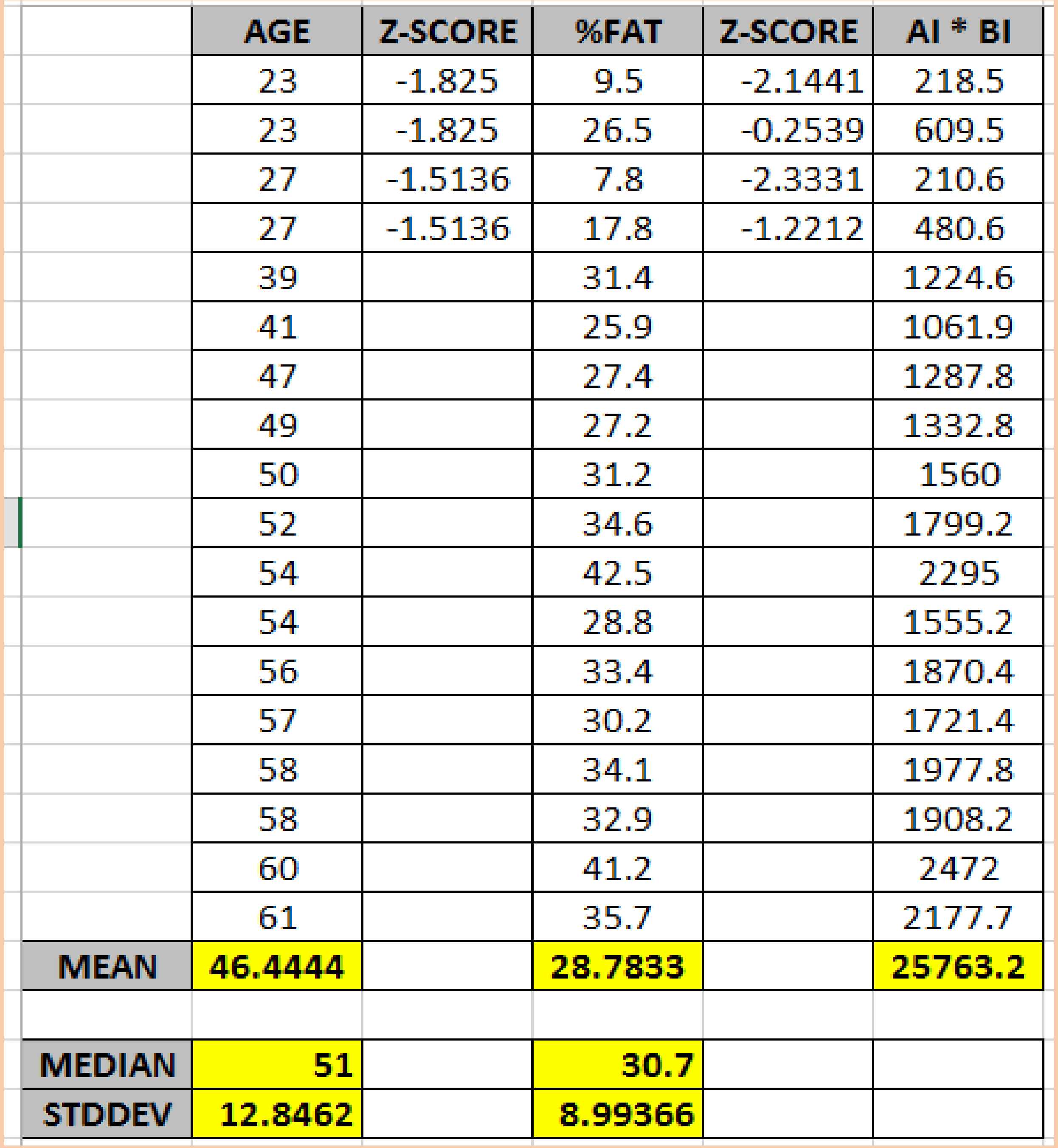
(a) Calculate the mean, median, and standard deviation of *age* and *%fat*.

(b) Draw the boxplots for *age* and *%fat*.

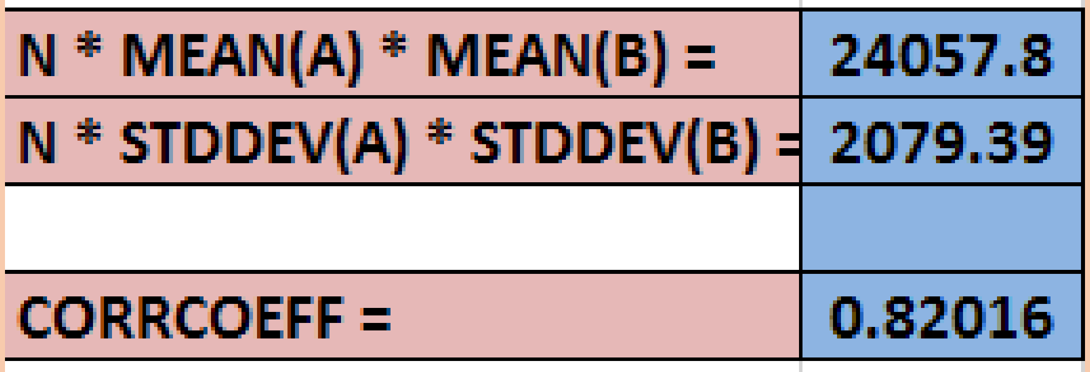
(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

(d) Normalize the two variables based on *z-score normalization*.

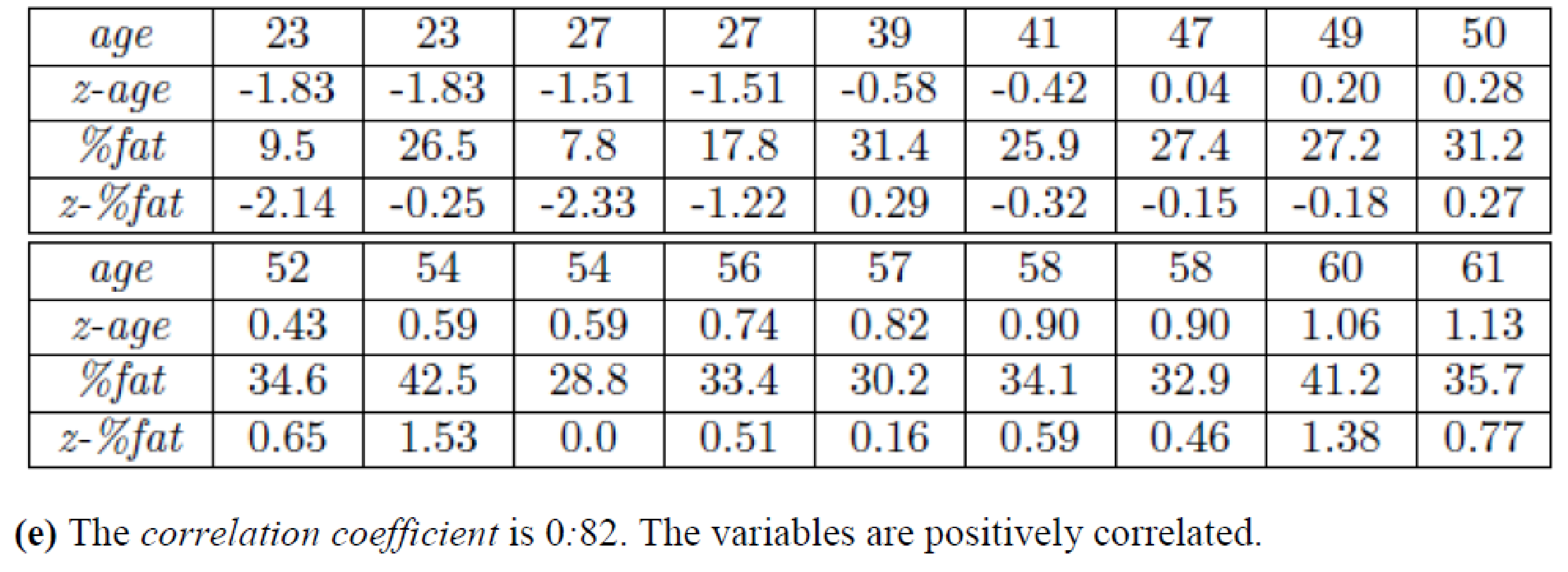
(e) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?



	AGE	Z-SCORE	%FAT	Z-SCORE	AI * BI
	23	-1.825	9.5	-2.1441	218.5
	23	-1.825	26.5	-0.2539	609.5
	27	-1.5136	7.8	-2.3331	210.6
	27	-1.5136	17.8	-1.2212	480.6
	39		31.4		1224.6
	41		25.9		1061.9
	47		27.4		1287.8
	49		27.2		1332.8
	50		31.2		1560
	52		34.6		1799.2
	54		42.5		2295
	54		28.8		1555.2
	56		33.4		1870.4
	57		30.2		1721.4
	58		34.1		1977.8
	58		32.9		1908.2
	60		41.2		2472
	61		35.7		2177.7
MEAN	46.4444		28.7833		25763.2
MEDIAN	51		30.7		
STDDEV	12.8462		8.99366		

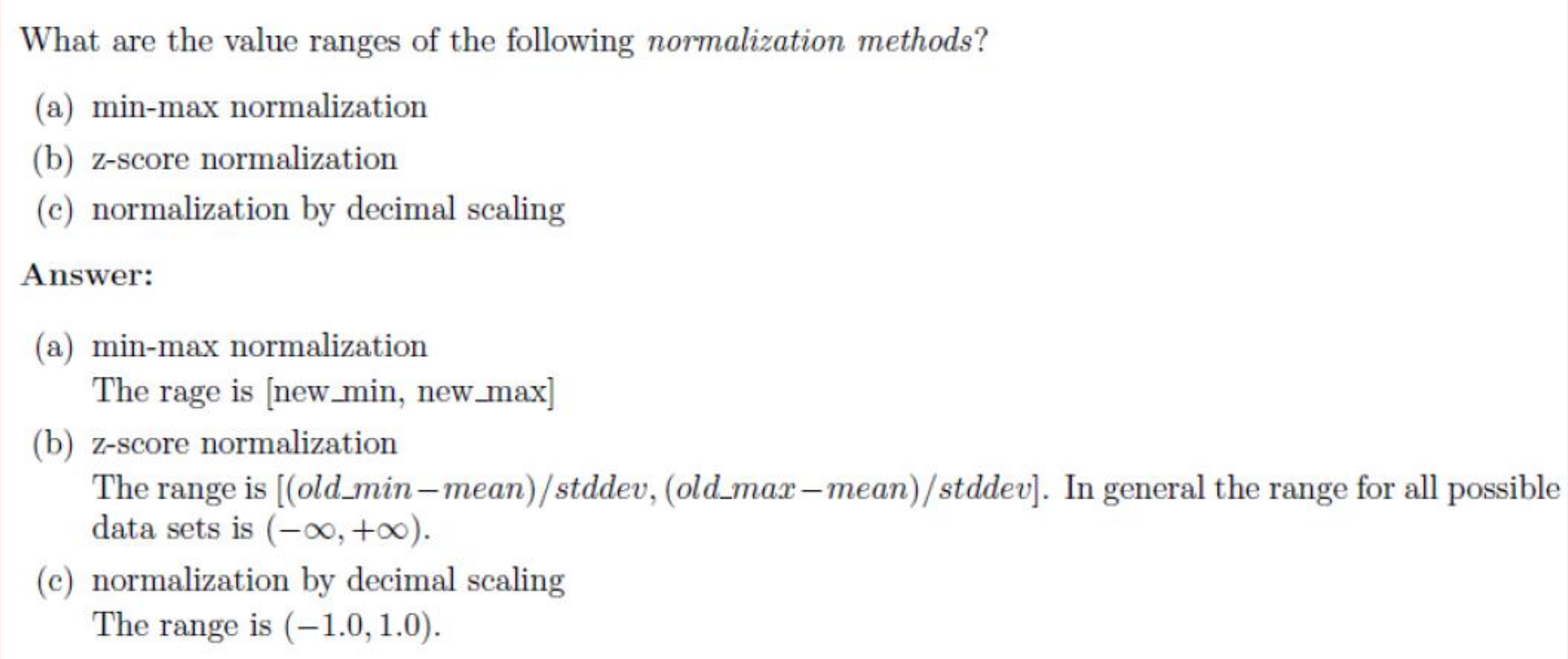


N * MEAN(A) * MEAN(B) =	24057.8
N * STDDEV(A) * STDDEV(B) =	2079.39
CORRCOEFF =	0.82016



age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
z-age	-1.83	-1.83	-1.51	-1.51	-0.58	-0.42	0.04	0.20	0.28	0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
z-%fat	-2.14	-0.25	-2.33	-1.22	0.29	-0.32	-0.15	-0.18	0.27	0.65	1.53	0.0	0.51	0.16	0.59	0.46	1.38	0.77

(e) The *correlation coefficient* is 0:82. The variables are positively correlated.



What are the value ranges of the following *normalization methods*?

(a) min-max normalization

(b) z-score normalization

(c) normalization by decimal scaling

Answer:

(a) min-max normalization

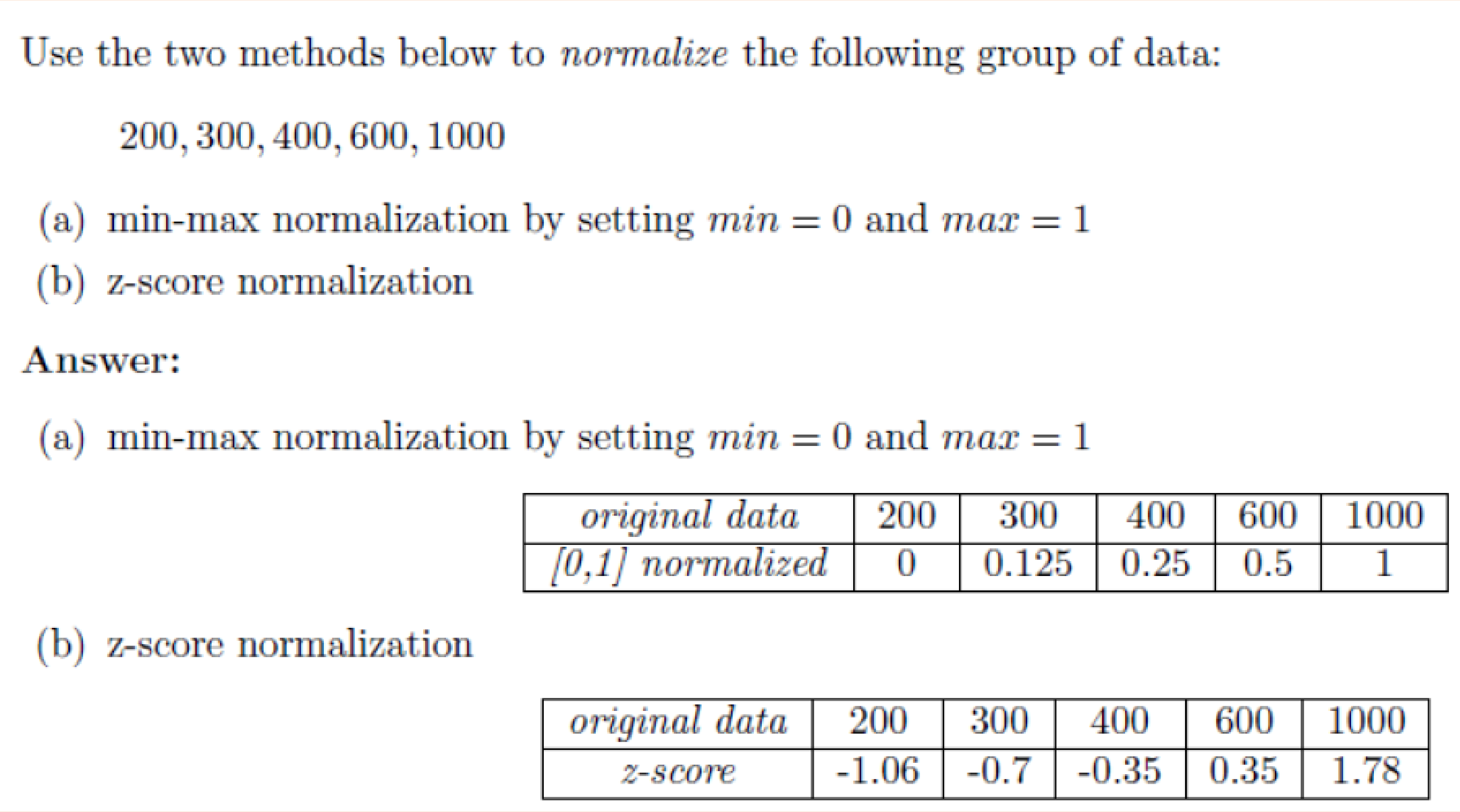
The rage is [new_min, new_max]

(b) z-score normalization

The range is $[(old\_min - mean)/stddev, (old\_max - mean)/stddev]$. In general the range for all possible data sets is $(-\infty, +\infty)$.

(c) normalization by decimal scaling

The range is $(-1.0, 1.0)$.



Use the two methods below to *normalize* the following group of data:

$$200, 300, 400, 600, 1000$$

(a) min-max normalization by setting $min = 0$ and $max = 1$

(b) z-score normalization

Answer:

(a) min-max normalization by setting $min = 0$ and $max = 1$

original data	200	300	400	600	1000
[0,1] normalized	0	0.125	0.25	0.5	1

(b) z-score normalization

original data	200	300	400	600	1000
z-score	-1.06	-0.7	-0.35	0.35	1.78