Chapter 2

2.1 Give three additional commonly used statistical measures that are not already illustrated in this chapter for the characterization of data dispersion. Discuss how they can be computed efficiently in large databases.

Answer:

Data dispersion, also known as variance analysis, is the degree to which numeric data tend to spread and can be characterized by such statistical measures as *mean deviation, measures of skewness*, and the *coefficient of variation*.

The **mean deviation** is defined as the arithmetic mean of the absolute deviations from the means and is calculated as:

$$mean\ deviation = \frac{\sum_{i=1}^{N} |x - \bar{x}|}{N}, \tag{2.1}$$

where $\bar{x}$ is the arithmetic mean of the values and N is the total number of values. This value will be greater for distributions with a larger spread.

A common **measure of skewness** is:

$$\frac{\bar{x} - mode}{s}, \tag{2.2}$$

which indicates how far (in standard deviations, s) the mean ($\bar{x}$) is from the mode and whether it is greater or less than the mode.

The **coefficient of variation** is the standard deviation expressed as a percentage of the arithmetic mean and is calculated as:

$$coefficient\ of\ variation = \frac{s}{\bar{x}} \times 100 \tag{2.3}$$

The variability in groups of observations with widely differing means can be compared using this measure.

2.2 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.

Answer:

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

The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 809/27 = 30$ (Equation 2.1). The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

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

This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35.

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

The midrange (average of the largest and smallest values in the data set) of the data is: $(70+13)/2 = 41.5$

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

The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.

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

The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and for this data is: 13, 20, 25, 35, 70.

(f) Show a *boxplot* of the data. (Omitted here. Please refer to Figure 2.3 of the textbook.)

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

A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.

A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line ($y = x$) can be added to the graph along with points representing where the first, second and third quantiles lie to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

2.3 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
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Compute an *approximate median* value for the data.

Answer:

Using Equation (2.3), we have $L_1 = 20$, $N = 3194$, $(\sum freq)_l = 950$, $freq_{median} = 1500$, $width = 30$, $median = 32.94$ years.

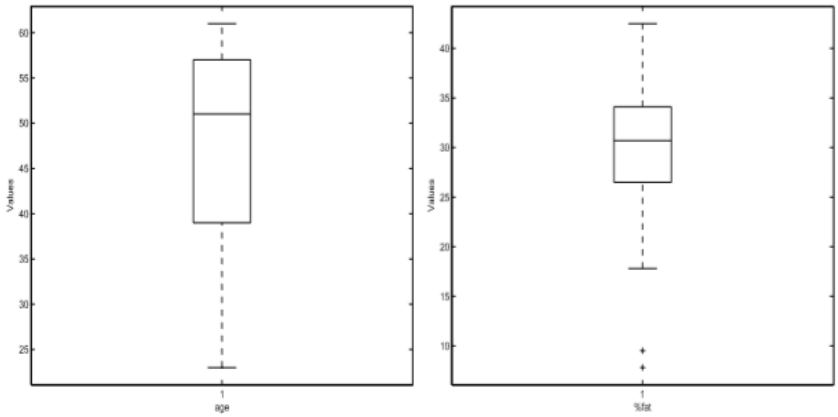
2.4 Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

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

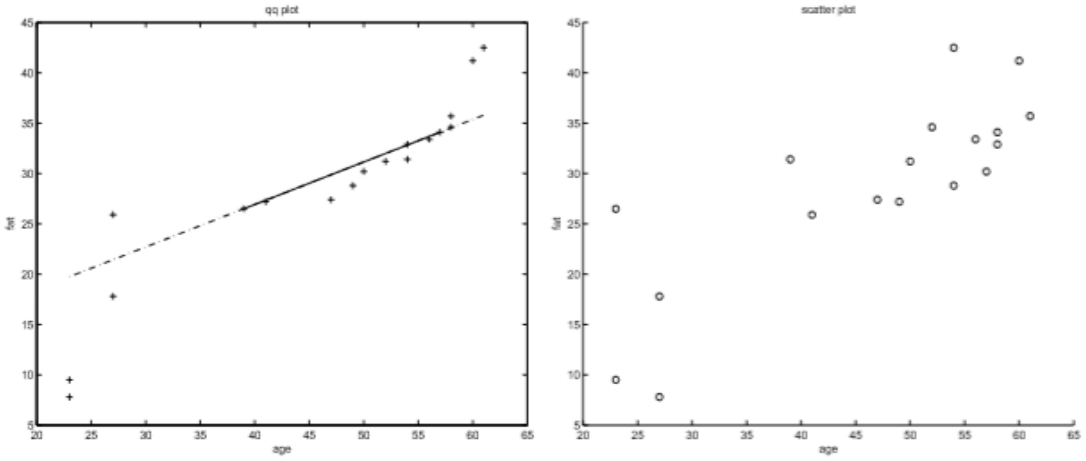
Answer:

(a) Calculate the mean, median and standard deviation of age and %fat. For the variable age the mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable %fat the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.

(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.

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

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

The *correlation coefficient* is 0.82. The variables are positively correlated.

2.5 Briefly outline how to compute the dissimilarity between objects described by the following:

(a) Nominal attributes

Use **Euclidean distance** or **Manhattan distance**. Euclidean distance is defined as

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}.$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{in})$, and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$, are two n-dimensional data objects. The **Manhattan (or city block) distance**, is defined as

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|.$$

(b) Asymmetric binary attributes

If all binary variables have the same weight, we have the contingency Table 7.1.

		object j		
		1	0	sum
object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Table 7.1: A contingency table for binary variables.

In computing the dissimilarity between asymmetric binary variables, the number of negative matches, t, is considered unimportant and thus is ignored in the computation, that is,

$$d(i,j) = \frac{r+s}{q+r+s}. \quad (7.3)$$

2.6 Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the *Euclidean distance* between the two objects.

$$\begin{aligned} d(i,j) &= \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{in}-x_{jn})^2} \\ &= \sqrt{|22-20|^2+|1-0|^2+|42-36|^2+|10-8|^2} = 6.71 \end{aligned}$$

(b) Compute the *Manhattan distance* between the two objects.

$$\begin{aligned} d(i,j) &= |x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{in}-x_{jn}| \\ &= |22-20|+|1-0|+|42-36|+|10-8| = 11 \end{aligned}$$

(c) Compute the *Minkowski distance* between the two objects, using $p = 3$.

$$\begin{aligned} d(i,j) &= (|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\cdots+|x_{in}-x_{jn}|^p)^{1/p} \\ &= (|22-20|^3+|1-0|^3+|42-36|^3+|10-8|^3)^{1/3} = 6.15 \end{aligned}$$

2.7 The median is one of the most important holistic measures in data analysis. Propose several methods for median approximation. Analyze their respective complexity under different parameter settings and decide to what extent the real value can be approximated. Moreover, suggest a heuristic strategy to balance between accuracy and complexity and then apply it to all methods you have given.

Answer:

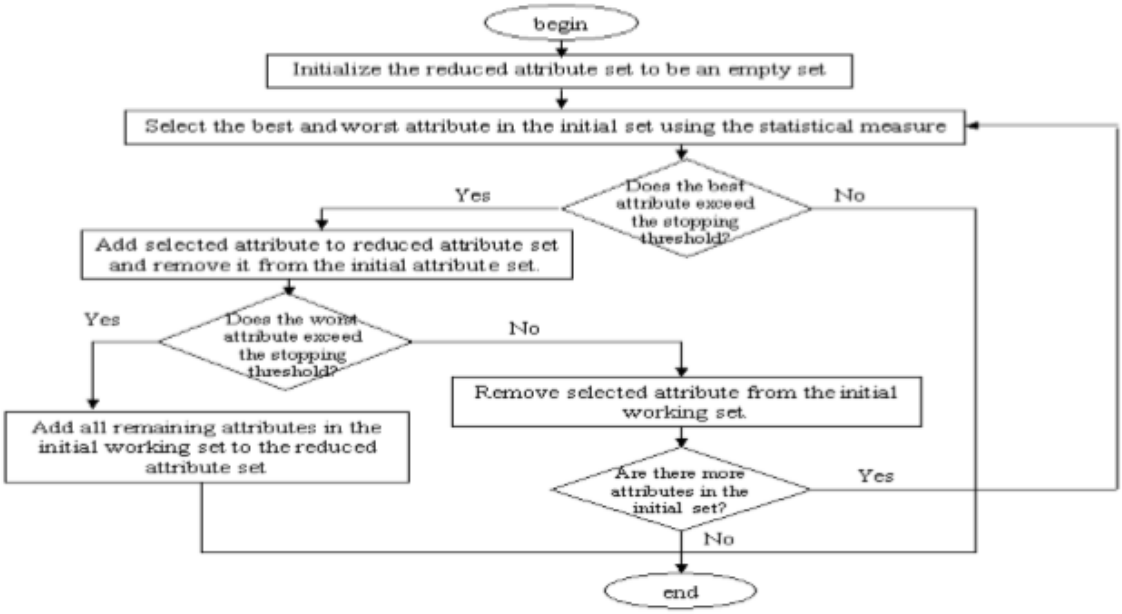
This question can be dealt with either theoretically or empirically, but doing some experiments to get the result is perhaps more interesting.

Given are some data sets sampled from different distributions, e.g., uniform, Gaussian, exponential, and gamma. (The former two distributions are symmetric, whereas the latter two are skewed). For example, if using Equation (2.3) to do approximation as proposed in the chapter, the most straightforward approach is to partition all of the data into k equal-length intervals.

$$median = L_1 + (\frac{N/2 - (\sum freq)_l}{freq_{median}})width, \tag{2.3}$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and $width$ is the width of the median interval.

Obviously, the error incurred will be decreased as k becomes larger; however, the time used in the whole procedure will also increase. The product of error made and time used are good optimality measures. From this point, we can perform many tests for each type of distributions (so that the result won't be dominated



2.8 It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following 2-D data set:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, x = (1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

(b) Normalizethedatasettomakethenormofeachdatapointequalto1.UseEuclidean distance on the transformed data to rank the data points.

(a) Consider the data as two-dimensional data points. Given a new data point, $\boldsymbol{x} = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using (1) Euclidean distance (Equation 7.5), and (2) cosine similarity (Equation 7.16).

The Euclidean distance of two n-dimensional vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, is defined as: $\sqrt[2]{\sum_{i=1}^{n}(x_i - y_i)^2}$. The cosine similarity of $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as: $\frac{\boldsymbol{x}^t \cdot \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||}$, where $\boldsymbol{x}^t$ is a transposition of vector $\boldsymbol{x}$, $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x}$,[1] and $||\boldsymbol{y}||$ is the Euclidean norm of vector $\boldsymbol{y}$. Using these definitions we obtain the distance from each point to the query point.

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$
Euclidean distance	0.14	0.67	0.28	0.22	0.61
Cosine similarity	0.9999	0.9957	0.9999	0.9990	0.9653

Based on the Euclidean distance, the ranked order is $\boldsymbol{x_1}, \boldsymbol{x_4}, \boldsymbol{x_3}, \boldsymbol{x_5}, \boldsymbol{x_2}$. Based on the cosine similarity, the order is $\boldsymbol{x_1}, \boldsymbol{x_3}, \boldsymbol{x_4}, \boldsymbol{x_2}, \boldsymbol{x_5}$.

(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

After normalizing the data we have:

$\boldsymbol{x}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$
0.6585	0.6616	0.7250	0.6644	0.6247	0.8321
0.7526	0.7498	0.6887	0.7474	0.7809	0.5547

The new Euclidean distance is:

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$
Euclidean distance	0.0041	0.0922	0.0078	0.0441	0.2632