Chapter 2

Getting to Know Your Data

2.1 Exercises

1. Give three additional commonly used statistical measures (i.e., not illustrated in this chapter) for the characterization of *data dispersion,* and 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.

   Note that all of the input values used to calculate these three statistical measures are algebraic measures (Chapter 4). Thus, the value for the entire database can be efficiently calculated by partitioning the database, computing the values for each of the separate partitions, and then merging theses values into an algebraic equation that can be used to calculate the value for the entire database.

The measures of dispersion described here were obtained from: Statistical Methods in Research and Production, fourth ed., Edited by Owen L. Davies and Peter L. Goldsmith, Hafner Publishing Company, NY:NY, 1972. ■

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.

   (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*?

   Answer:

   (a) What is the *mean* of the data? What is the *median*?
   The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 809/27 = 30$. 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.
   See Figure 2.1.

   (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, in order 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.

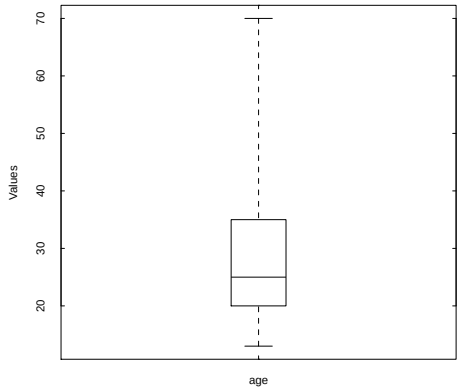

Figure 2.1: A boxplot of the data in Exercise 2.2.

■

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:

$L_1 = 20$, n= 3194, $(\sum_f)_l = 950$, $freq\_median = 1500$, width = 30, median = 30.94 years. ■

4. 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
%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

(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.

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*.
See Figure 2.2.

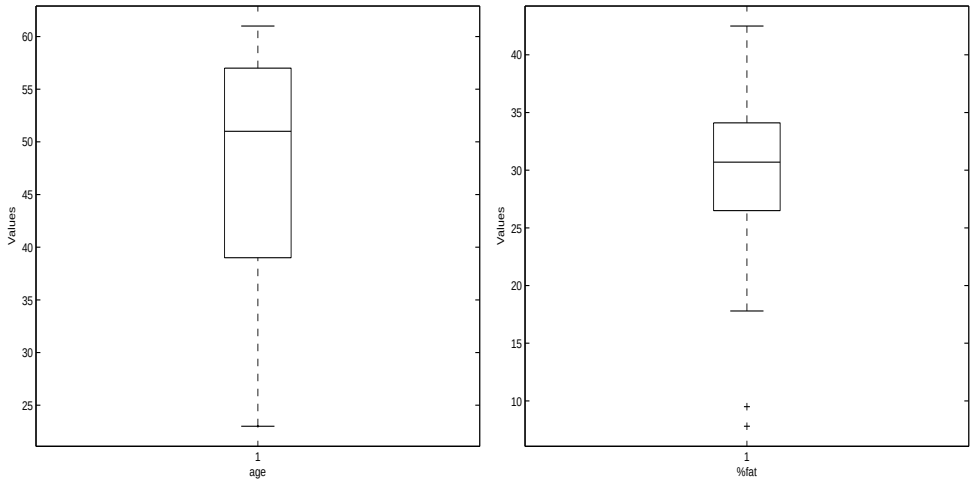

Figure 2.2: A boxplot of the variables *age* and *%fat* in Exercise 2.4.

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.
See Figure 2.3.

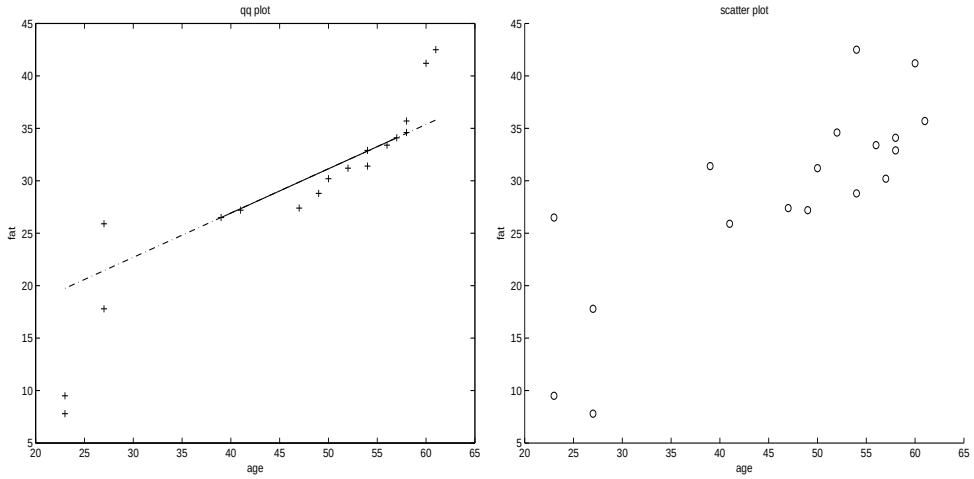

Figure 2.3: A *q-q plot* and a *scatter plot* of the variables *age* and *%fat* in Exercise 2.4.

■

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

   (a) Nominal attributes
   (b) Asymmetric binary attributes
   (c) Numeric attributes
   (d) Term-frequency vectors

Answer:

(a) Nominal attributes

A categorical variable is a generalization of the binary variable in that it can take on more than two states.

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i,j) = \frac{p-m}{p}, \tag{2.4}$$

where m is the number of *matches* (i.e., the number of variables for which i and j are in the same state), and p is the total number of variables.

Alternatively, we can use a large number of binary variables by creating a new binary variable for each of the M nominal states. For an object with a given state value, the binary variable representing that state is set to 1, while the remaining binary variables are set to 0.

(b) Asymmetric binary attributes

If all binary variables have the same weight, we have the contingency Table 2.1.

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

Table 2.1: 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,

$$d(i,j) = \frac{r+s}{q+r+s}. \tag{2.5}$$

(c) Numeric attributes

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

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (x_{in}-x_{jn})^2}. \tag{2.6}$$

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

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{in}-x_{jn}|. \tag{2.7}$$

The **supremum distance** is

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if}-x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if}-x_{jf}|. \tag{2.8}$$

(d) Term-frequency vectors

To measure the distance between complex objects represented by vectors, it is often easier to abandon traditional metric distance computation and introduce a nonmetric similarity function.

For example, the similarity between two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, can be defined as a cosine measure, as follows:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^t \cdot \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||} \tag{2.9}$$

where $\boldsymbol{x^t}$ is a transposition of vector $\boldsymbol{x}$, $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x}$,[1] $||\boldsymbol{y}||$ is the Euclidean norm of vector $\boldsymbol{y}$, and s is essentially the cosine of the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

■

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.

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

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

   (d) Compute the *supremum distance* between the two objects.

   Answer:

   (a) Compute the *Euclidean distance* between the two objects.
   The Euclidean distance is computed using Equation (2.6).
   Therefore, we have $\sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{45} = 6.7082$.

   (b) Compute the *Manhattan distance* between the two objects.
   The Manhattan distance is computed using Equation (2.7). Therefore, we have $|22-20| + |1-0| + |42-36| + |10-8| = 11$.

   (c) Compute the *Minkowski distance* between the two objects, using $h = 3$.
   The Minkowski disance is

   $$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdot + |x_{ip} - x_{jp}|^h} \tag{2.10}$$

   where h is a real number such that $h \geq 1$.
   Therefore, with $h = 3$, we have $\sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3} = \sqrt[3]{233} = 6.1534$.

   (d) Compute the *supremum distance* between the two objects.
   The supremum distance is computed using Equation (2.8). Therefore, we have a supremum distance of 6.

■

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.

   We can give students some data sets sampled from different distributions, e.g., uniform, Gaussian (both two are symmetric) and exponential, gamma (both two are skewed). For example, if we use Equation

[1] The Euclidean normal of vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is defined as $\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$. Conceptually, it is the length of the vector.

(2.11) to do approximation as proposed in the chapter, the most straightforward way is to divide all data into k equal length intervals.

$$median = L_1 + (\frac{N/2 - (\sum freq)_l}{freq_{median}})width, \tag{2.11}$$

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 and larger; however, the time used in the whole procedure will also increase. Let's analyze this kind of relationship more formally. It seems the product of error made and time used is a good optimality measure. From this point, we can do many tests for each type of distributions (so that the result won't be dominated by randomness) and find the k giving the best trade-off. In practice, this parameter value can be chosen to improve system performance.

There are also some other approaches to approximate the median, students can propose them, analyze the best trade-off point, and compare the results among different approaches. A possible way is as following: Hierarchically divide the whole data set into intervals: at first, divide it into k regions, find the region in which the median resides; then secondly, divide this particular region into k sub-regions, find the sub-region in which the median resides; .... We will iteratively doing this, until the width of the sub-region reaches a predefined threshold, and then the median approximation formula as above stated is applied. By doing this, we can confine the median to a smaller area without globally partitioning all data into shorter intervals, which is expensive (the cost is proportional to the number of intervals).

■

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 two-dimensional data set:

	A_1	A_2
$\boldsymbol{x_1}$	1.5	1.7
$\boldsymbol{x_2}$	2	1.9
$\boldsymbol{x_3}$	1.6	1.8
$\boldsymbol{x_4}$	1.2	1.5
$\boldsymbol{x_5}$	1.5	1.0

   (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 Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

   (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.

   Answer:

   (a) Use Equation (2.6) to compute the Euclidean distance, Equation (2.7) to compute the Manhattan distance, Equation (2.8) to compute the supremum distance, and Equation (2.9) to compute the cosine similarity between the input data point and each of the data points in the data set. Doing so yields the following table

	Euclidean dist.	Manhattan dist.	supremum dist.	cosine sim.
$\boldsymbol{x_1}$	0.1414	0.2	0.1	0.99999
$\boldsymbol{x_2}$	0.6708	0.9	0.6	0.99575
$\boldsymbol{x_3}$	0.2828	0.4	0.2	0.99997
$\boldsymbol{x_4}$	0.2236	0.3	0.2	0.99903
$\boldsymbol{x_5}$	0.6083	0.7	0.6	0.96536

These values produce the following rankings of the data points based on similarity:

Euclidean distance: x_1, x_4, x_3, x_5, x_2

Manhattan distance: x_1, x_4, x_3, x_5, x_2

Supremum distance: x_1, x_4, x_3, x_5, x_2

Cosine similarity: x_1, x_3, x_4, x_2, x_5

(b) The normalized query is (0.65850, 0.75258). The normalized data set is given by the following table

	A_1	A_2
$\boldsymbol{x_1}$	0.66162	0.74984
$\boldsymbol{x_2}$	0.72500	0.68875
$\boldsymbol{x_3}$	0.66436	0.74741
$\boldsymbol{x_4}$	0.62470	0.78087
$\boldsymbol{x_5}$	0.83205	0.55470

Recomputing the Euclidean distances as before yields

	Euclidean dist.
$\boldsymbol{x_1}$	0.00415
$\boldsymbol{x_2}$	0.09217
$\boldsymbol{x_3}$	0.00781
$\boldsymbol{x_4}$	0.04409
$\boldsymbol{x_5}$	0.26320

which results in the final ranking of the transformed data points: x_1, x_3, x_4, x_2, x_5

■

2.2 Supplementary Exercises

1. Briefly outline how to compute the dissimilarity between objects described by *ratio-scaled variables*.

   Answer:

   Three methods include:

   - Treat ratio-scaled variables as interval-scaled variables, so that the Minkowski, Manhattan, or Euclidean distance can be used to compute the dissimilarity.
   - Apply a logarithmic transformation to a ratio-scaled variable f having value x_{if} for object i by using the formula $y_{if} = \log(x_{if})$. The y_{if} values can be treated as interval-valued,
   - Treat x_{if} as continuous ordinal data, and treat their ranks as interval-scaled variables.

   ■