Chapter 2

Descriptive Statistics

REVIEW OF KEY CONCEPTS

SECTION 2.1 Measures of Location

2.1.1 Arithmetic Mean

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Consider the data in Table 2.1. They represent serum-cholesterol levels from a group of hospital workers who were regularly eating a standard U.S. diet and who agreed to change their diet to a vegetarian diet for a 6-week period. Cholesterol levels were measured before and after adopting the diet. The mean serum-cholesterol level before adopting the diet is computed as follows:

$$\sum_{i=1}^{24} x_i = 4507\,,\ \bar{x} = \frac{4507}{24} = 187.8 \text{ mg/dL}$$

Advantages

1. It is representative of all the points.
2. If the underlying distribution is Gaussian (bell shaped), then it is the most efficient estimator of the middle of the distribution.
3. Many statistical tests are based on the arithmetic mean.

Disadvantages

1. It is sensitive to outliers, particularly in small samples; e.g., if one of the cholesterol values were 800 rather than 200, then the mean would be increased by 25 mg/dL.
2. It is inappropriate if the underlying distribution is far from being Gaussian; for example, serum triglycerides have a distribution that looks highly skewed (i.e., asymmetric).

Table 2.1 Serum-Cholesterol levels before and after adopting a vegetarian diet (mg/dL)

Subject	Before	After	Before – After
1	195	146	49
2	145	155	–10
3	205	178	27
4	159	146	13
5	244	208	36
6	166	147	19
7	250	202	48
8	236	215	21
9	192	184	8
10	224	208	16
11	238	206	32
12	197	169	28
13	169	182	–13
14	158	127	31
15	151	149	2
16	197	178	19
17	180	161	19
18	222	187	35
19	168	176	–8
20	168	145	23
21	167	154	13
22	161	153	8
23	178	137	41
24	137	125	12
Mean	187.8	168.3	19.5
sd	33.2	26.8	16.8
n	24	24	24

2.1.2 Alternatives to the Arithmetic Mean-Median

One interesting property from the table is that the diet appears to work best in people with high baseline levels versus people with low baseline levels. How can we test if this is true? Divide the group in half, and look at cholesterol change in each half. To do this we must compute the median ≡ 50% point in the distribution. Specifically,

$$\text{Median} = \left(\frac{n+1}{2}\right)\text{th largest point if } n \text{ is odd}$$

$$= \text{average of } \left[\frac{n}{2}\text{th} + \left(\frac{n}{2}+1\right)\text{th}\right] \text{ largest points if } n \text{ is even}$$

For example,

if $n = 7$, then the median = 4th largest point
if $n = 24$, then the median = average of (12th + 13th) largest point

2.1.3 Stem-and-Leaf Plots

How can we easily compute the median? We would have to order the data to obtain the 12th and 13th largest points. An easier way is to compute a *stem-and-leaf plot*. Divide each data value into a leaf (the least-significant digit or digits) and a stem (the most-significant digit or digits) and collect all data points with the same stem on a single row. For example, the number 195 has a stem of 19 and a leaf of 5. A stem and leaf plot of the before measurements is given below.

Cumulative total	Stems	Leaves
24	25	0
23	24	4
22	23	68
20	22	42
	21	
18	20	5
17	19	5277
13	18	0
12	17	8
11	16	698871
5	15	981
2	14	5
1	13	7
	Median=179 mg/dL=$\frac{178+180}{2}$	

We have added a cumulative total column which gives the total number of points with a stem that is $\le$ the stem in that row. It is easy to compute the median from the stem and leaf plot since the median = average of 12th and 13th largest values $= (178+180)/2 = 179$. Note that the leaves within a given row (stem) are not necessarily in order. One use of stem and leaf plots is to provide a visual comparison of the values in 2 different data sets. The stem-and-leaf plots of the change in cholesterol for the subgroups of people below and above the median are given as follows:

≤ 179 mg/dL		≥ 180 mg/dL	
4	1	4	98
3	1	3	625
2	3	2	718
1	3932	1	699
0	28	0	8
–0	8	–0	
–1	03	–1	
–2		–2	
–3		–3	

The change scores in the subgroups look quite different; the subgroup with initial value above the median is showing more change. We will be able to test if the average change score is “significantly different” based on a *t* test (to be covered in Chapter 8 of the text).

2.1.4 Percentiles

We can also use stem-and-leaf plots to obtain percentiles of the distribution.
To compute the p^{th} percentile, if $np/100$ is an integer, then average the

$$\frac{np}{100}\text{th} + \left(\frac{np}{100}+1\right)\text{th largest points}$$

Otherwise, pth percentile $= \{[np/100]+1\}$ the largest point, where $[np/100] =$ largest integer $\le np/100$. For example, to compute the 10th percentile of the baseline cholesterol distribution, also known as the lower decile, we have $n = 24$, $p = 10$, $np/100 = 2.4$, $[np/100] = 2$, lower decile = 3rd largest point = 151 mg/dL. To compute the 90th percentile (or upper decile), $n = 24$, $p = 90$, $np/100 = 21.6$, $[np/100] = 21$. Upper decile = 22nd largest point = 238 mg/dL.

Commonly used percentiles

$10, 20, \ldots, 90\%$ (deciles)	$25, 50, 75\%$ (quartiles)
$20, 40, \ldots, 80\%$ (quintiles)	$33.3, 66.7\%$ (tertiles)

Median

Advantages

1. Always guarantees that 50% of the data values are on either side of the median.
2. Insensitive to outliers (extreme values). If one of the cholesterol values increased from 200 to 800, the median would remain at 179 but the mean would increase from 188 mg/dL to $188+25=213$ mg/dL.

Disadvantages

1. It is not as efficient an estimator of the middle as the mean if the distribution really is Gaussian in that it is mostly sensitive to the middle of the distribution.
2. Most statistical procedures are based on the mean.

We can get an impression of how symmetric a distribution is by looking at the stem and leaf plot. If we look at the stem and leaf plot of the baseline values on the previous page we see that the distribution is only slightly skewed, and the mean may be adequate.

2.1.5 Geometric Mean

One way to get around the disadvantages of the arithmetic mean are to transform the data onto a different scale to make the distribution more symmetric and compute the arithmetic mean on the new scale. The most popular such scale is the ln (natural log or $\log_e$) scale:

$$\ln(x_1), \ldots, \ln(x_n)$$

We can now take an average in the ln scale and denote it by $\overline{\ln x}$:

$$\overline{\ln x} = \frac{\ln(x_1)+\cdots+\ln(x_n)}{n}$$

The problem with this is that the average is in the ln scale rather than the original scale. Thus, we take the antilog of $\overline{\ln x}$ to obtain

$$GM = e^{\overline{\ln x}} = \text{geometric mean}$$

The ERG (electroretinogram) amplitude (μV) is a measure of electrical activity of the retina and is used to monitor retinal function in patients with retinitis pigmentosa, an often-blinding ocular condition. The following data were obtained from 10 patients to monitor the course of the condition over a 1-year period.

	Year 1 ERG amplitude (μV)	Year 2 ERG amplitude (μV)	Absolute change (μV)
1	1.9	1.4	0.5
2	3.9	3.9	0.0
3	64.4	46.2	18.2
4	25.9	19.1	6.8
5	4.0	2.5	1.5
6	0.9	1.6	–0.7
7	2.0	1.8	0.2
8	4.0	3.7	0.3
9	33.8	12.1	21.7
10	6.3	3.5	2.8

The distribution of values at each year is highly skewed, with change scores dominated by people with high year-1 ERG amplitudes. The distribution in the ln scale is much more symmetric. Let's compute the GM for year 1 and year 2.

Year 1

$$\overline{\ln x} = \frac{\ln(1.9) + \cdots + \ln(6.3)}{10} = \frac{0.64 + \cdots + 1.84}{10} = 1.8144$$

$$GM_1 = e^{1.8144} = 6.137\ \mu V$$

Year 2

$$\overline{\ln x} = \frac{\ln(1.4) + \cdots + \ln(3.5)}{10} = \frac{0.34 + \cdots + 1.25}{10} = 1.5508$$

$$GM_2 = e^{1.5508} = 4.715\ \mu V$$

We can quantify the % change by

$$\frac{GM_2}{GM_1} = \frac{4.715}{6.137} = 0.768 \approx 23.2\% \text{ decline } (= 100\% \times (1 - 0.768))$$

Thus, the ERG has declined, on average, by 23.2% over 1 year.

Geometric Mean

Advantages

1. Useful for certain types of skewed distributions.
2. Standard statistical procedures can be used on the log scale.

Disadvantages

1. Not appropriate for symmetric data.
2. More sensitive to outliers than the median but less so than the mean.

SECTION 2.2 Measures of Spread

2.2.1 Range

The range = the interval from the smallest value to the largest value. This gives a quick feeling for the overall spread—but is misleading because it is solely influenced by the most extreme values; e.g., cholesterol data—initial readings; range = (137, 250).

2.2.2 Quasi-Range

A quasi-range is similar to the range but is derived after excluding a specified percentage of the sample at each end; e.g., the interval from the 10th percentile to the 90th percentile. For example, for the cholesterol data

10% point = 3rd largest from bottom = 151 mg/dL
90% point = 3rd largest from top = 238 mg/dL
quasi-range = (151,238)

2.2.3 Standard Deviation, Variance

If the distribution is normal or near normal, then the standard deviation is more frequently used as a measure of spread.

$$s^2 = \text{sample variance} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$$s = \text{sample standard deviation variance} = \sqrt{s^2}$$

Why *s* rather than *s*2?

We want an estimator of spread in the same units as $\bar{x}$; i.e., if units change by a factor of c, and the transformed data is referred to as y, then

$$\bar{y} = c\bar{x} \quad s_y = cs_x \text{ but } s_y^2 = c^2 s_x^2$$

Note that s changes by a factor of c (the same as $\bar{x}$), but s^2 changes by a factor of c^2. Thus, s and $\bar{x}$ can be directly related to each other while s^2 and $\bar{x}$ cannot.

How can we use $\bar{x}$ and s to get an impression of the spread of the distribution? If the distribution is normal, then

$\bar{x} \pm s$ comprises about 2/3 of the distribution
$\bar{x} \pm 2s$ (more precisely, $1.96s$) comprises about 95% of the distribution
$\bar{x} \pm 2.5s$ (more precisely, $2.576s$) comprises about 99% of the distribution

If the distribution is not normal or near normal, then the distribution is not well characterized by $\bar{x}$, s. It is better to use the percentiles in this case (e.g., the median could be used instead of the mean and the quasi-range instead of the standard deviation). For example, for the cholesterol data, the variance and standard deviation of the before measurements are computed as follows:

$$s^2 = \frac{\sum_{i=1}^{24}(x_i - \bar{x})^2}{23} = \frac{25{,}289.96}{23} = 1099.56$$

$$s = 33.2$$

Let's see how normal the distribution looks.

$$\bar{x} \pm 1.96s = 187.8 \pm 1.96(33.2) = (122.8,\ 252.8)$$

includes all points; it should include 95% (or 23 out of 24 points) under a normal distribution.

$$\bar{x} \pm 1s = 187.8 \pm 33.2 = (154.6,\ 221.0)$$

includes $15/24 = 62.5\%$ of points; it should be 2/3 under a normal distribution. The normal distribution appears to provide a reasonable approximation. Note that computer programs such as *Excel* can be used to compute many types of descriptive statistics. See the cd-rom for an example of using *Excel* to easily compute the mean and standard deviation.

2.2.4 Coefficient of Variation (CV)

$$CV = 100\% \times \frac{s}{\bar{x}}$$

The CV is used if the variability is thought to be related to the mean. For the cholesterol data,

$$CV = 100\% \times \frac{33.2}{187.8} = 17.7\%$$

SECTION 2.3 Some Other Means for Describing Data

2.3.1 Frequency Distribution

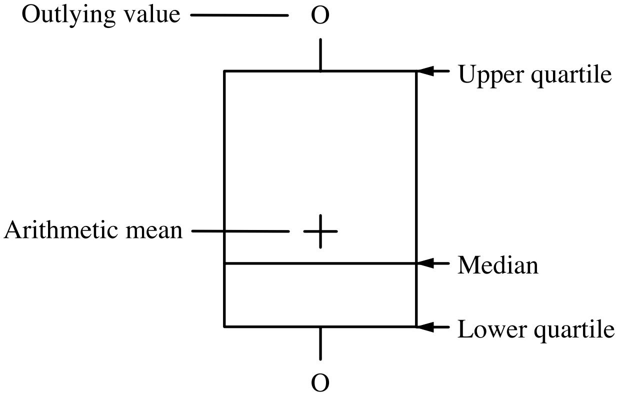
This is a listing of each value and how frequently it occurs (or in addition, the % of scores associated with each value). This can be done either based on the original values, or in grouped form; e.g., if we group the cholesterol change scores by 10-mg increments, then we would have

	Frequency	%
≥ 40.0,	3	13
≥ 30.0, < 40.0	4	17
≥ 20.0, < 30.0	4	17
≥ 10.0, < 20.0	7	29
≥ 0.0, < 10.0	3	13
≥ –10.0, < 0.0	2	8
≥ –20.0, < –10.0	1	4
	24	

This can be done either in numeric or graphic form. If in graphic form, it is often represented as a *bar graph*.

2.3.2 Box Plot

Another graphical technique for displaying data often used in computer packages is provided by a Box plot. The box (rectangle) displays the upper and lower quartiles, the median, arithmetic mean, and outlying values (if any). It is a concise way to look at the symmetry and range of a distribution.



PROBLEMS ....

Suppose the origin for a data set is changed by adding a constant to each observation.

2.1 What is the effect on the median?

2.2 What is the effect on the mode?

2.3 What is the effect on the geometric mean?

2.4 What is the effect on the range?

Renal Disease

For a study of kidney disease, the following measurements were made on a sample of women working in several factories in Switzerland. They represent concentrations of bacteria in a standard-size urine specimen. High concentrations of these bacteria may indicate possible kidney pathology. The data are presented in Table 2.2.

Table 2.2 Concentration of bacteria in the urine in a sample of female factory workers in Switzerland

Concentration	Frequency
10^0	521
10^1	230
10^2	115
10^3	74
10^4	69
10^5	62
10^6	43
10^7	30
10^8	21
10^9	10
10^{10}	2

2.5 Compute the arithmetic mean for this sample.

2.6 Compute the geometric mean for this sample.

2.7 Which do you think is a more appropriate measure of location?

Cardiovascular Disease

The mortality rates from heart disease (per 100,000 population) for each of the 50 states and the District of Columbia in 1973 are given in descending order in Table 2.3 [1].

Consider this data set as a sample of size 51

$$(x_1, x_2, \ldots, x_{51}).$$

If $\sum_{i=1}^{51} x_i = 17{,}409$ and $\sum_{i=1}^{51} (x_i - \bar{x})^2 = 249{,}063.65$ then do the following:

Table 2.3 Mortality rates from heart disease (per 100,000 population) for the 50 states and the District of Columbia in 1973

1	West Virginia	445.4	18	Wisconsin	369.8	35	DC	327.1
2	Pennsylvania	442.7	19	Vermont	369.2	36	South Carolina	322.4
3	Maine	427.3	20	Nebraska	368.9	37	Montana	319.1
4	Missouri	422.9	21	Tennessee	361.4	38	Maryland	315.9
5	Illinois	420.8	22	New Hampshire	358.2	39	Georgia	311.8
6	Florida	417.4	23	Indiana	356.4	40	Virginia	311.2
7	Rhode Island	414.4	24	North Dakota	353.3	41	California	310.6
8	Kentucky	407.6	25	Delaware	351.6	42	Wyoming	306.8
9	New York	406.7	26	Mississippi	351.6	43	Texas	300.6
10	Iowa	396.9	27	Louisiana	349.4	44	Idaho	297.4
11	Arkansas	396.8	28	Connecticut	340.3	45	Colorado	274.6
12	New Jersey	395.2	29	Oregon	338.7	46	Arizona	265.4
13	Massachusetts	394.0	30	Washington	334.2	47	Nevada	236.9
14	Kansas	391.7	31	Minnesota	332.7	48	Utah	214.2
15	Oklahoma	391.0	32	Michigan	330.2	49	New Mexico	194.0
16	Ohio	377.7	33	Alabama	329.1	50	Hawaii	169.0
17	South Dakota	376.2	34	North Carolina	328.4	51	Alaska	83.9

2.8 Compute the arithmetic mean of this sample.

2.9 Compute the median of this sample.

2.10 Compute the standard deviation of this sample.

2.11 The national mortality rate for heart disease in 1973 was 360.8 per 100,000. Why does this figure *not* correspond to your answer for Problem 2.8?

2.12 Does the differential in raw rates between Florida (417.4) and Georgia (311.8) actually imply that the risk of dying from heart disease is greater in Florida than in Georgia? Why or why not?

Nutrition

Table 2.4 shows the distribution of dietary vitamin-A intake as reported by 14 students who filled out a dietary questionnaire in class. The total intake is a combination of intake from individual food items and from vitamin pills. The units are in IU/100 (International Units/100).

Table 2.4 Distribution of dietary vitamin-A intake as reported by 14 students

Student number	Intake (IU/100)	Student number	Intake (IU/100)
1	31.1	8	48.1
2	21.5	9	24.4
3	74.7	10	13.4
4	95.5	11	37.1
5	19.4	12	21.3
6	64.8	13	78.5
7	108.7	14	17.7

2.13 Compute the mean and median from these data.

2.14 Compute the standard deviation and coefficient of variation from these data.

2.15 Suppose the data are expressed in IU rather than IU/100. What are the mean, standard deviation, and coefficient of variation in the new units?

2.16 Construct a stem-and-leaf plot of the data on some convenient scale.

2.17 Do you think the mean or median is a more appropriate measure of location for this data set?

SOLUTIONS

2.1 Each data value is changed from x_i to $x_i + a$, for some constant a. The median also increases by a.

2.2 The mode increases by a.

2.3 The geometric mean is changed by an undetermined amount, because the geometric mean is given by antilog $\left[\sum \ln(x_i + a)/n\right]$ and there is no simple relationship between $\ln(x_i + a)$ and $\ln(x_i)$.

2.4 The range is not changed, since it is the distance between the largest and smallest values, and distances between points will not be changed by shifting the origin.

2.5 The arithmetic mean is given by

$$\frac{10^0(521)+\ldots+10^{10}(2)}{521+\ldots+2}=\frac{3.24\times 10^{10}}{1177}$$
$$=2.757\times 10^7$$

2.6 To compute the geometric mean, we first compute the mean log to the base 10 as follows:

$$\frac{521\log_{10}(10^0)+\ldots+2\log_{10}(10^{10})}{521+\ldots+2}=\frac{521\times 0+\ldots+2\times 10}{1177}$$
$$=\frac{2014}{1177}=1.711$$

The geometric mean is then given by $10^{1.711} = 51.4$.

2.7 The geometric mean is more appropriate because the distribution is in powers of 10 and is very skewed. In the log scale, the distribution becomes less skewed, and the mean provides a more central measure of location. Notice that only 33 of the 1177 data points are greater than the arithmetic mean, while 426 of the 1177 points are greater than the geometric mean.

2.8 We have that $\bar{x} = 17{,}409/51 = 341.4$ per 100,000.

2.9 Since $n = 51$ is odd, the median is given by the $[(51+1)/2]$th or 26th largest value = mortality rate for Mississippi = 351.6 per 100,000.

2.10 We have that

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{50} = \frac{249{,}063.65}{50} = 4{,}981.27$$

Thus, $s = \sqrt{4{,}981.27} = 70.6$ per 100,000

2.11 The national mortality rate is a weighted average of the state-specific mortality rates, where the weights are the number of people in each state. The arithmetic mean in Problem 2.8 is an unweighted average of the state-specific mortality rates that weights the large and small states equally.

2.12 No. The demographic characteristics of the residents of Florida may be very different from those of Georgia, which would account for the difference in the rates. In particular, Florida has a large retiree population, which would lead to higher mortality rates. In order to make an accurate comparison between the states, we would, at a minimum, need to compare disease rates among specific age-sex-race groups in the two states.

2.13

$$\overline{x} = \frac{31.1 + \ldots + 17.7}{14} = \frac{656.2}{14} = 46.9$$

Median = average of the 7th and 8th largest values

$$= \frac{31.1 + 37.1}{2} = 34.1$$

2.14

$$s^2 = \frac{\sum_{i=1}^{14}(x_i - \overline{x})^2}{13}$$

$$= \frac{13{,}158.03}{13} = 1012.16$$

$$s = \sqrt{1012.16} = 31.8$$

$$CV = 100\% \times \frac{s}{\overline{x}}$$

$$= 100\% \times \frac{31.8}{46.9} = 67.9\%$$

2.15 Mean $= 46.9 \times 100 = 4{,}687$ IU,
$s = 31.8 \times 100 = 3{,}181$ IU, $CV = 67.9\%$
(unchanged).

2.16 We will round each number to the nearest integer in constructing the stem-and-leaf plot:

1	938
2	241
3	17
4	8
5	
6	5
7	59
8	
9	6
10	9

2.17 The median is more appropriate, because the distribution appears to be skewed to the right.

REFERENCE

[1] National Center for Health Statistics. (1975, February 10), *Monthly vital statistics report, summary report, final mortality statistics* (1973), **23**(11) (Suppl. 2).