Summarizing Data: Measures of Variation

One aspect of most sets of data is that the values are not all alike; indeed, the extent to which they are unalike, or vary among themselves, is of basic importance in statistics. Consider the following examples:

In a hospital where each patient's pulse rate is taken three times a day, that of patient A is 72, 76, and 74, while that of patient B is 72, 91, and 59. The mean pulse rate of the two patients is the same, 74, but observe the difference in variability. Whereas patient A's pulse rate is stable, that of patient B fluctuates widely.

A supermarket stocks certain 1-pound bags of mixed nuts, which on the average contain 12 almonds per bag. If all the bags contain anywhere from 10 to 14 almonds, the product is consistent and satisfactory, but the situation is quite different if some of the bags have no almonds while others have 20 or more.

Measuring variability is of special importance in statistical inference. Suppose, for instance, that we have a coin that is slightly bent and we wonder whether there is still a fifty-fifty chance for heads. What if we toss the coin 100 times and get 28 heads and 72 tails? Does the shortage of heads — only 28 where we might have expected 50 — imply that the count is not "fair?" To answer such questions we must have some idea about the magnitude of the fluctuations, or variations, that are brought about by chance when coins are tossed 100 times.

We have given these three examples to show the need for measuring the extent to which data are dispersed, or spread out; the corresponding measures that provide this information are called **measures of variation**.

The Range

To introduce a simple way of measuring variability, let us refer to the first of the three examples cited previously, where the pulse rate of patient A varied from 72 to 76 while that of patient B varied from 59 to 91. These extreme (smallest and largest) values are indicative of the variability of the two sets of data, and just about the same information is conveyed if we take the differences between the respective extremes. So, let us make the following definition:

The range of a set of data is the difference between the largest value and the smallest.

For patient A the pulse rates had a range of $76 - 72 = 4$ and for patient B they had a range of $91 - 59 = 32$.

Whereas, the range covers all the values in a sample, a similar measure of variation covers (more or less) the middle 50 percent. It is the **interquartile range**: $Q_3 - Q_1$, where Q_1 and Q_3 may be defined as before. Some statisticians also use the **semi-interquartile range** $\frac{1}{2}(Q_3 - Q_1)$ which is sometimes referred to as the **quartile deviation**.

The Variance and the Standard Deviation

To define the **standard deviation**, by far the most generally useful measure of variation, let us observe that the dispersion of a set of data is small if the values are closely bunched about their mean, and that it is large if the values are scattered widely about their mean. Therefore, it would seem reasonable to measure the variation of a set of data in terms of the amounts by which the values deviate from their mean. If a set of numbers

$$x_1, x_2, x_3, \ldots, \text{ and } x_n$$

constitutes a sample with the mean $\overline{x}$, then the differences

$$x_1 - \overline{x}, x_2 - \overline{x}, x_3 - \overline{x}, \ldots, \text{ and } x_n - \overline{x}$$

are called the **deviation from the mean**, and we might use their average (that is, their mean) as a measure of the variability of the sample. Unfortunately, this will not do. Unless the x's are all equal, some of the deviations from the mean will be positive, some will be negative, the sum of deviations from the mean, $\sum(x - \overline{x})$, and hence also their mean, is always equal to zero.

Since we are really interested in the magnitude of the deviations, and not in whether they are positive or negative, we might simply ignore the signs and define a measure of variation in terms of the absolute values of the deviations from the mean. Indeed, if we add the deviations from the mean as if they were all positive or zero and divide by n, we obtain the statistical measure that is called the **mean deviation**. This measure has intuitive appeal, but because the absolute values if leads to serious theoretical difficulties in problems of inference, and it is rarely used.

An alternative approach is to work with the squares of the deviations from the mean, as this will also eliminate the effect of signs. Squares of real numbers cannot be negative; in fact, squares of the deviations from a mean are all positive unless a value happens to coincide with the mean. Then, if we average the squared deviation from the mean and take the square root of the result (to compensate for the fact that the deviations were squared), we get

$$\boxed{\sqrt{\frac{\sum(x - \overline{x})^2}{n}}}$$

and this is how, traditionally, the standard deviation used to be defined. Expressing literally what we have done here mathematically, it is also called the **root-mean-square deviation**.

Nowadays, it is customary to modify this formula by dividing the sum of the squared deviations from the mean by $n - 1$ instead of n. Following this practice, which will be explained later, let us define the **sample standard deviation**, denoted by s, as

$$\boxed{s = \sqrt{\frac{\sum(x - \overline{x})^2}{n - 1}}}$$

and the **population standard deviation**, denoted by σ, as

$$\boxed{\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}}$$

The **sample variance** is s^2 and the **population variance** is σ^2.

Ordinarily, the purpose of calculating a sample statistics (such as the mean, the standard deviation, or the variance) is to estimate the corresponding population parameter. If we actually took many samples from a population that has the mean μ, calculated the sample means $\overline{x}$, and then averaged all these estimated of μ, we should find that their average is very close to μ. However, if we calculated the variance of each sample by means of the formula $\dfrac{\sum(x - \overline{x})^2}{n}$ and then averaged all these supposed estimates of σ^2, we would probably find that their average is less than σ^2. Theoretically, it can be shown that we can compensate for this by dividing by $n - 1$ instead of n in the formula for s^2.

EXAMPLE: A bacteriologist found 8, 11, 7, 13, 10, 11, 7, and 9 microorganism of a certain kind in eight cultures. Calculate s.

EXAMPLE: A bacteriologist found 8, 11, 7, 13, 10, 11, 7, and 9 microorganism of a certain kind in eight cultures. Calculate s.

Solution: First calculating the mean, we get

$$\overline{x} = \frac{8+11+7+13+10+11+7+9}{8} = 9.5$$

and then the work required to find $\sum(x-\overline{x})^2$ may be arranged as in the following table:

x	$x-\overline{x}$	$(x-\overline{x})^2$
8	-1.5	2.25
11	1.5	2.25
7	-2.5	6.25
13	3.5	12.25
10	0.5	0.25
11	1.5	2.25
7	-2.5	6.25
9	-0.5	0.25
	0.0	32.00

Finally, dividing 32.00 by $8-1=7$ and taking the square root, we get

$$s = \sqrt{\frac{32.00}{7}} \approx 2.14$$

Note in the preceding Table that the total for the middle column is zero; since this must always be the case, it provides a convenient check on the calculations.

It was easy to calculate s in this Example because the data were whole numbers and the mean was exact to one decimal. Otherwise, the calculations required by the formula defining s can be quite tedious, and, unless we can get s directly with a statistical calculator or a computer, it helps to use the formula

$$\boxed{s = \sqrt{\frac{S_{xx}}{n-1}} \quad \text{where} \quad S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

EXAMPLE: Use this computing formula to rework the Example above.

Solution: First we calculate $\sum x$ and $\sum x^2$, getting

$$\sum x = 8+11+7+13+10+11+7+9 = 76$$

and

$$\sum x^2 = 8^2+11^2+7^2+13^2+10^2+11^2+7^2+9^2 = 754$$

then

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 754 - \frac{76^2}{8} = 32$$

so

$$s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{32}{7}} \approx 2.14$$

Applications of the Standard Deviation

For any set of data (population or sample) and any constant k greater than 1, the proportion of the data that must lie within k standard deviations on either side of the mean is at least $1 - \frac{1}{k^2}$.

EXAMPLE: A study of the nutritional value of a certain kind of reduced-fat cheese showed that on the average a one-ounce slice contains 3.50 grams of fat with standard deviation of 0.04 gram of fat.

(a) According to Chebyshev's theorem, at least what percent of the one-ounce slices of this kind of cheese must have a fat content between 3.38 and 3.62 grams of fat?

(b) According to Chebyshev's theorem, between what values must be the fat content of at least 93.75% of the one-ounce slices of this kind of cheese?

Solution:

(a) Since

$$3.62 - 3.50 = 3.50 - 3.38 = 0.12$$

we find that $k(0.04) = 0.12$ and, hence,

$$k = \frac{0.12}{0.04} = 3$$

It follows that at least

$$1 - \frac{1}{3^2} = \frac{8}{9} \approx 88.9\%$$

of the one-ounce slices of the cheese have a fat content between 3.38 and 3.62 grams of fat.

(b) Since $1 - \frac{1}{k^2} = 0.9375$, we find that

$$\frac{1}{k^2} = 1 - 0.9375 = 0.0625 \quad \Longrightarrow \quad k^2 = \frac{1}{0.0625} = 16 \quad \Longrightarrow \quad k = 4$$

It follows that 93.75% of the one-ounce slices of the cheese contain between $3.50 - 4(0.04) = 3.34$ and $3.50 + 4(0.04) = 3.66$ grams of fat.

For distributions having the general shape of the cross section of a bell, we can make the following stronger statements:

About 68% of the values will lie within one standard deviation of the mean, that is between $\overline{x} - s$ and $\overline{x} + s$.
About 95% of the values will lie within two standard deviations of the mean, that is between $\overline{x} - 2s$ and $\overline{x} + 2s$.
About 99.7% of the values will lie within three standard deviations of the mean, that is between $\overline{x} - 3s$ and $\overline{x} + 3s$.

EXAMPLE: Based on 1997 figures, the following are 11.0 "waiting times" (in minutes) between eruptions of the Old Faithful Geyser m Yellowstone National Park:

81	83	94	73	78	94	73	89	112	80
94	89	35	80	74	91	89	83	80	82
91	80	83	91	89	82	118	105	64	56
76	69	78	42	76	82	82	60	73	69
91	83	67	85	60	65	69	85	65	82
53	83	62	107	60	85	69	92	40	71
82	89	76	55	98	74	89	98	69	87
74	98	94	82	82	80	71	73	74	80
60	69	78	74	64	80	83	82	65	67
94	73	33	87	73	85	78	73	74	83
83	51	67	73	87	85	98	91	73	108

(a) Construct a frequency distribution and a histogram.

(b) Find the mean and the standard deviation.

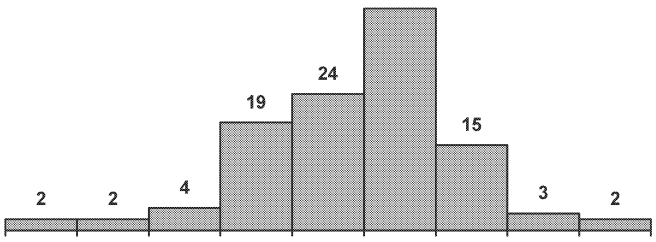
(c) Use these figures to determine from the original data what percentage of the values falls within three standard deviations from the mean.

Solution:

(a) Since the smallest value is 33 and the largest value is 118, we have to cover an interval of 86 values and a convenient choice would be to use the nine classes 30 -39, 40 - 49, 50 - 59, 60 - 69, 70 - 79, 80 - 89, 90 - 99, 100 - 109, and 110-119. These classes will accommodate all of the data, they do not overlap, and they are all of the same size. There are other possibilities (for instance, 25 - 34, 35 - 44, 45 - 54, 55 - 64, 65 - 74, 75 - 84, 85 - 94, 95 - 104, 105 - 114, and 115 - 124), but it should be apparent that our first choice will facilitate the tally.

We now tally the 110 values and get the result shown in the following table:

Waiting between eruption (minutes) Frequency	Tally	
30-39	II	2
40-49	II	2
50-59	IIII	4
60-69	𝍸 𝍸 IIII	19
70-79	𝍸 𝍸 𝍸 𝍸 IIII	24
80-89	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 IIII	39
90-99	𝍸 𝍸	15
100-109	III	3
110-119	II	2
	Total	110



(b) The mean is $\overline{x} = 78.59$ and the standard deviation is $s = 14.35$.

(c) We shall have to determine what percentage of the values falls between

$$78.59 - 3(14.35) = 35.54 \qquad \text{and} \qquad 78.59 + 3(14.35) = 121.64$$

Counting two of the values, 33 and 35, below 35.54 and none above 121.64, we find that 110-2=108 of the values and, hence

$$\frac{108}{110} \cdot 100 = 98.2\%$$

of the original waiting times fall within three standard deviations of the mean. This is fairly close to the expected 99.7%, but then the distribution of the waiting times is not really perfectly bell shaped.

Now suppose that we want to compare numbers belonging to different sets of data. To illustrate, suppose that the final examination in a French course consists of two parts, vocabulary and grammar, and that a certain student scored 66 points in the vocabulary part and 80 points in the grammar part. At first glance it would seem that the student did much better in grammar than in vocabulary, but suppose that all the students in the class averaged 51 points in the vocabulary part with a standard deviation of 12, and 72 points in the grammar part with a standard deviation of 16. Thus, we can argue that the student's score in the vocabulary part is $\frac{66-51}{12} = 1.25$ standard deviations above the average for the class, while her score in the grammar part is only $\frac{80-72}{16} = 0.5$ standard deviation above the average for the class. Whereas the original scores cannot be meaningfully compared, these new scores, expressed in terms of standard deviations, can. Clearly, the given student rates much higher on her command of French vocabulary than on her knowledge of French grammar, compared to the rest of the class.

What we have done here consists of converting the grades into **standard units** or ***z*-scores.** It general, if x is a measurement belonging to a set of data having the mean $\overline{x}$ (or μ) and the standard deviation s (or σ), then its value in standard units, denoted by z, is

$$\boxed{z = \frac{x - \overline{x}}{s} \quad \text{or} \quad z = \frac{x - \mu}{\sigma}}$$

Depending on whether the data constitute a sample or a population. In these units, z tells us how many standard deviations a value lies above or below the mean of the set of data to which it belongs. Standard units will be used frequently in application.

EXAMPLE: Mrs. Clark belongs to an age group for which the mean weight is 112 pounds with a standard deviation of 11 pounds, and Mr. Clark, her husband, belongs to an age group for which the mean weight is 163 pounds with a standard deviation of 18 pounds. If Mrs. Clark weighs 132 pounds and Mr. Clark weighs 193 pounds, which of the two is relatively more overweight compared to his/her age group?

Solution: Mr. Clark's weight is 193 -163 = 30 pounds above average while Mrs. Clark's weight is "only" 132 -112 = 20 pounds above average, yet in standard units we get $\frac{193-163}{18} \approx 1.67$ for Mr. Clark and $\frac{132-112}{11} \approx 1.82$ for Mrs. Clark. Thus, relative to them age groups Mrs. Clark is somewhat more overweight than Mr. Clark.

A serious disadvantage of the standard deviation as a measure of variation is that it depends on the units of measurement. For instance, the weights of certain objects may have a standard deviation of 0.10 ounce, but this really does not tell us whether it reflects a great deal of variation or very little variation. If we are weighing the eggs of quails, a standard deviation of 0.10 ounce would reflect a considerable amount of variation, but this would not be the case if we are weighing, say, 100-pound bags of potatoes. What we need in a situation like this is a **measure of relative variation** such as the **coefficient of variation**, defined by the following formula:

$$\boxed{V = \frac{s}{\overline{x}} \cdot 100 \quad \text{or} \quad V = \frac{\sigma}{\mu} \cdot 100}$$

The coefficient of variation expresses the standard deviation as a percentage of what is being measured, at least on the average.

EXAMPLE: Several measurements of the diameter of a ball bearing made with one micrometer had a mean of 2.49 mm and a standard deviation of 0.012 mm, and several measurements of the unstretched length of a spring made with another micrometer had a mean of 0.75 in. with a standard deviation of 0.002 in. Which of the two micrometers is relatively more precise?

Solution: Calculating the two coefficients of variation, we get

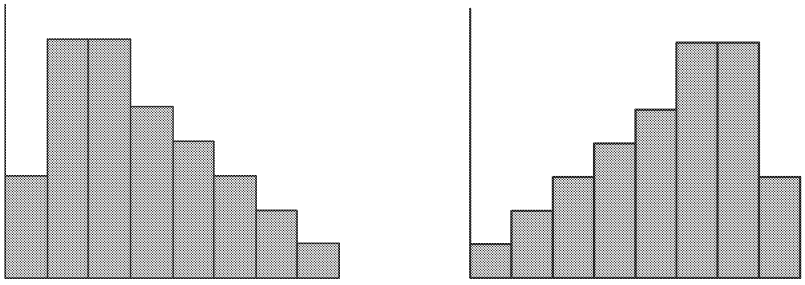
$$\frac{0.012}{2.49} \cdot 100\% \approx 0.48\% \qquad \text{and} \qquad \frac{0.002}{0.75} \cdot 100\% \approx 0.27\%$$

Thus, the measurements of the length of the spring are relatively less variable, which means that the second micrometer is more precise.

Some Further Descriptions

So far we have discussed only statistical descriptions that come under the general heading of measures of location or measures of variation. Actually, there is no limit to the number of ways in which statistical data can be described, and statisticians continually develop new methods of describing characteristics of numerical data that are of interest in particular problems. In this section we shall consider briefly the problem of describing the overall shape of a distribution.

Although frequency distributions can take on almost any shape or form, most of the distributions we meet in practice can be described fairly well by one or another of few standard types. Among these, foremost in importance is the aptly described symmetrical **bell-shaped distribution**. The two distributions shown in the Figures below can, by a stretch of the imagination, be described as bell shaped, but they are not symmetrical. Distributions like these, having a "tail" on one side or the other, are said to be **skewed**; if the tail is on the left we say that they are **negatively skewed** and if the tail is on the right we say that they are **positively skewed**. Distributions of incomes or wages are often positively skewed because of the presence of some relatively high values that are not offset by correspondingly low values.



Positive Skewed Negative Skewed

The concepts of symmetry and skewness apply to any kind of data, not only distributions. Of course, for a large set of data we may just group the data and draw and study a histogram, but if that is not enough, we can use anyone of several statistical **measures of skewness**. A relatively easy one is based on the fact that when there is perfect symmetry, the mean and the median will coincide. When there is positive skewness and some of the high values are not offset by correspondingly low values, the mean will be greater than the median; when there is a negative skewness and some of the low values are not offset by correspondingly high values, the mean will be smaller than the median.

This relationship between the median and the mean can be used to define a relatively simple measure of skewness, called the **Pearsonian coefficient of skewness**. It is given by

$$\boxed{SK = \frac{3(mean - median)}{standard \;\; deviation}}$$

For a perfectly symmetrical distribution, such the mean and the median coincide and $SK = 0$. In general, values of the Pearsonian coefficient of skewness must fall between -3 and 3, and it should be noted that division by the standard deviation makes SK independent of the scale of measurement.

EXAMPLE: Calculate SK for the distribution of the waiting times between eruptions of Old Faithful, using the results of the Example above, where we showed that $\overline{x} = 78.59, \tilde{x} = 80.53$, and $s = 14.35$.

Solution: Substituting these values into the formula for SK, we get

$$SK = \frac{3(78.59 - 80.53)}{14.35} \approx -0.41$$

Which shows that there is a definite, though modest, negative skewness. This is also apparent from the histogram of the distribution.