

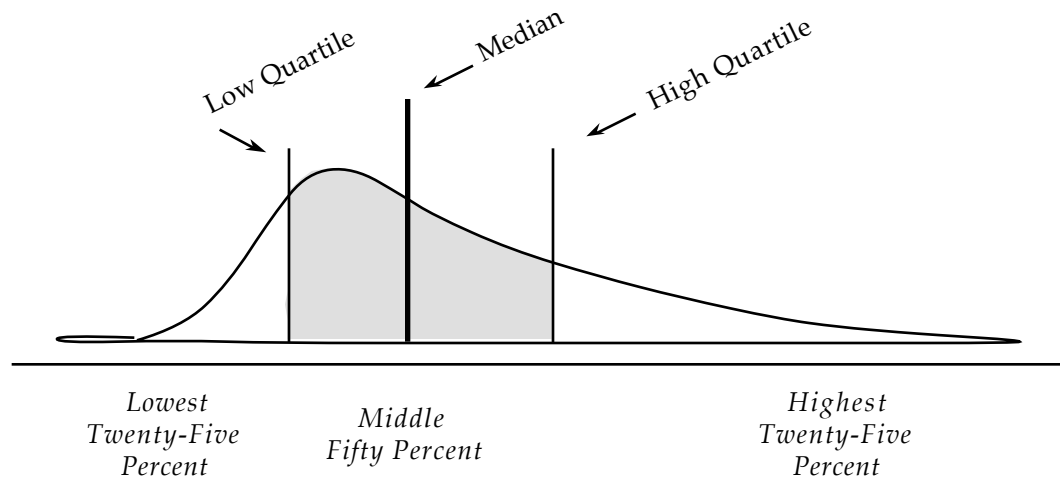
Description: Numbers for the Variation

The Median and The Quartiles

Just as there are several ways of computing the average, there are several ways to compute the variation. But the situation is somewhat simplified because these measures for variation come in pairs. Remember: The median is the center with respect to which variation is minimum in the sense of minimum absolute deviation. So, logically, the median must be paired off with an indicator of absolute deviation. By convention the indicator (or indicators) are the quartiles. (Or, more precisely, by the distances between the medians and the two quartiles.) The two quartiles are average deviations matched to the median: The low quartile is the median of that fifty percent of the data which is below the median. The high quartile is the median of the data that are above the median.) And remember that the mean is the center with respect to which variation is minimum in the sense of least squares. So, logically, if you measure the average as the mean, following the criterion of least squares, then consistency requires that you measure variation by the "variance" which is the mean squared deviation from the mean. (And, in addition, the need to interpret the result in intuition-friendly form will require you to use the standard deviation — which is the square root of the variance -- the square root of the mean squared deviation.) Both these things need to be defined, beginning with the quartiles.

Recall that the median is the middle value. Half of the data are greater than or equal to the median. Half of the data are less than or equal to the median. And now to assess this variation we ask two questions: Among those values that are greater than or equal to the

median, what is the average value? And among those values that are less than or equal to the median, what is the average value? And when we have computed those numbers, then the average of the high values and the average of the low values helps us visualize the spread of the data. So we compute the median of those values that are greater than or equal to the median of all values and call it the high quartile. And we compute the median of those values that are less than or equal to the median of all values and call it the low quartile, using the word "quartile" because these three numbers, the low quartile, the median, and the high quartile divide the data into *four* ranges of values. We use these quartiles to visualize the central "hump" of the data.



Hypothetical Income Distribution Divided Into Quartiles
Showing the Middle Fifty Percent of the Data

The range of values between these two quartiles describes the central range of the data:

The median protein content of breakfast cereals is __ grams of protein, with the typical breakfast cereal providing between __ and ___ grams of protein (specifying the quartiles).

The median personal income of these college graduates is ___ (specifying the median), with typical incomes ranging between ___ and ___ (specifying the quartiles).

That's what we're after, something to express the "middle" of the data, although typically, in print, you will find this information in abbreviated form, simply naming the values: "The median is __, with quartiles at __ and __." That tight little statement, little more than three numbers, presumes that you know what to do with the numbers when you've got them. And what you do with the numbers is to build a mental picture of the center:

The median protein content is 2.3 grams, with quartiles at 2.2 grams and 2.4 grams.

That message gives me a picture of a distribution wrapped tightly around its central value.

The median protein content is 2.3 grams, with quartiles at 1.5 grams and 4 .5 grams.

This message gives me a picture of a distribution that is spread out, and spread more in one direction (toward the high end) than the other.

Customarily we go one step further, adding two more numbers, five in all, to specify the extremes. Thus, completing the description,

The median protein content is 2.3 grams, with typical values lying between the low quartile at 1.5 grams and the high quartile at 4.5 grams. In a few instances the values differ considerably from these typical numbers, ranging as low as 0.8 grams, for rice cereals, and as high as 10.2 grams of protein for Gerbers high protein.

Computing the Quartiles

That's the idea, the rest is detail, important detail, to make sure that we agree on the computation that specifies these quartiles. I will specify a procedure but the important point is the definition: The median divides the data into two sets, high and low. And then the high quartile is the median of the subset of values that are *greater than or equal to* the median. The low quartile is the median of the subset of values that are *less than or equal to* the median.

So to compute these quartiles, we begin as we did with the median, by putting the data in rank order, low to high. Then where "n" is the number of values, the arithmetic is to compute the number $(n+1)/2$. If the result is a whole number, it identifies the location of the median. If the result is a fraction, then it identifies two numbers whose average is the median

n = number of values in the data

$$m = \text{location of median} = (n+1)/2$$

If the result is a whole number then the number of values that are greater than or equal to the median is m . And if the result is a fraction, then the number of values that are greater than or equal to the median is the integer part of m , m . (If m is 10.5, then its integer part is 10, lopping off the fraction.) And thus the location of the quartile is found by computing the number $(m+1)/2$. If the result is a whole number, it identifies the location of the quartile. If the result is a fraction, then it identifies two numbers whose average is the quartile.

$$m = \text{number of values greater than or equal to the median}$$

$$q = \text{location of quartile} = (m+1)/2$$

Exactly the same computation works for the remaining quartile except that you count to q starting at the other end of the distribution. Thus,

$$m = \text{number of values less than or equal to the median}$$

$$q = \text{location of quartile} = (m+1)/2$$

Working it out with eight things: $n = 8$ implies the arithmetic $(n+1)/2 = 4.5$. So, the depth of the median is 4.5 and, using the rank order, the median is the mean of the fourth number and the fifth. The integer part of 4.5 is 4, telling me that the number of values less than or equal to the median is 4.

That gives me $m = 4$. And $m = 4$ implies the arithmetic $(m+1)/2 = 2.5$. So, the depth of the quartile is 2.5 and, using the rank order, the high quartile is the mean of the second and third largest values (in order from large to small) while the low quartile is the mean of the second and third smallest values (in order from small to large).

$$\begin{aligned}n &= 8 \\m &= 4.5 \\m &= 4 \\q &= 2.5\end{aligned}$$

Working it out with nine things: $n = 9$ implies the arithmetic $(n+1)/2 = 5$. So, the depth of the median is 5 and, using the rank order, the median value is the fifth value. The number of values less than or equal to the median is 5.

That gives me $m = 5$. And $m = 5$ implies the arithmetic $(m+1)/2 = 3$. So, the depth of the quartile is 3 and, using the rank order, the high quartile is the third largest values while the low quartile is the third smallest value (in order from small to large).

$$\begin{aligned}n &= 9 \\m &= 5 \\m &= 5 \\q &= 3\end{aligned}$$

Working it out with ten things: $n = 10$ implies the arithmetic $(n+1)/2 = 5.5$. So, the depth of the median is 5.5 and, using the rank order, the median is the mean of the fifth number and the sixth. The integer part of 5.5 is 5, telling me that the number of values less than or equal to the median is 5.

That gives me $m = 5$. And $m = 5$ implies the arithmetic $(m+1)/2 = 3$. So, the depth of the quartile is 3 and, using the rank order, the high quartile is the third largest value (in order from large to small) while the low quartile is the third smallest value (in order from small to large).

n=10
 m = 5.5
 m = 5
 q=3

The Mean and the Standard Deviation

The second way to compute variation is paired with the mean. If you measure the average as the mean, then you measure the variation by computing the standard deviation. The idea for the *standard* deviation begins by defining deviation, any deviation, as the difference between a value found in the data and the mean of all the values found in the data. If I have an income of \$60,000 and the average income is \$50,000, then my deviation is \$10,000.

$$\text{Variance} = \text{Mean Squared Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Deviation} = \text{Observed Value} - \text{Mean Value}$$

Then the basic idea for the *standard* deviation is to compute the mean of the deviations — except that the basic idea doesn't work out. The trouble is that the simple mean of the deviations is a useless number, in fact it is always zero. You can work out this result by simply adding up all the deviations algebraically and dividing by n, computing their mean:

$$\text{Mean Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})$$

Following the algebra in steps: Distributing the summation expands the expression for the average to

$$\text{Mean Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n (x_i) - \frac{1}{n} \sum_{i=1}^n (\bar{x})$$

Evaluating the two expressions on the right, the first is \bar{x} itself, the mean

$$\text{Mean Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \bar{x} - \frac{1}{n} \sum_{i=1}^n (\bar{x})$$

Evaluating the second expression on the right shows that it too is equal to the mean

$$\text{Mean Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \bar{x} - \frac{1}{n} n\bar{x}$$

$$\text{Mean Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \bar{x} - \bar{x}$$

which reduces to zero

$$\text{Mean Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = 0$$

which shows that the average deviation is zero, always zero — so it doesn't tell us anything useful about the data.

The conventional solution is to keep the idea, we are still looking for some sort of average deviation, but modify it by squaring the deviation, computing the mean squared deviation, known as the "variance."

$$\text{Variance} = \text{Mean Squared Deviation} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

That is the basic answer, for a measure of the variation, but we're not quite done with it. There is one more problem: As soon as you try to compute a variance, and then interpret it, you will find that it

measures things in squared units: If the data are measured in grams of protein, then the mean squared deviation will give you a result in *square* grams of protein. That's not usable. I can't say "the variance of the protein content of breakfast cereals is 3 square grams of protein." It makes no sense. So, what we do is take the square root and apply the name "standard deviation" written as s_x . Describing the standard deviation in the jargon of the trade, we use the "root mean squared deviation" as the measure of variation with respect to the mean. You can see each of the terms, the root, the mean, and the square, at work in the formula:

$$\text{Standard Deviation of } X = s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Computing the Quartiles

In practice, the way you compute this thing is in stages. Showing the steps for the data on protein content of wheat cereals, Figure __, the first step is to compute the mean: For these five cereals, the sum is 13.8 grams of protein which, when divided into five equal parts gives the mean of 2.76 grams of protein.

Then computing the deviations from the mean, the first datum, 1.6, deviates from the mean by -1.16, for a squared deviation of 1.35. The sum of these squared deviations is 4.35 grams of protein. The variance (the mean squared deviation) is .87 squared grams of protein. And the standard deviation (the root mean squared deviation) is .93 grams of protein.

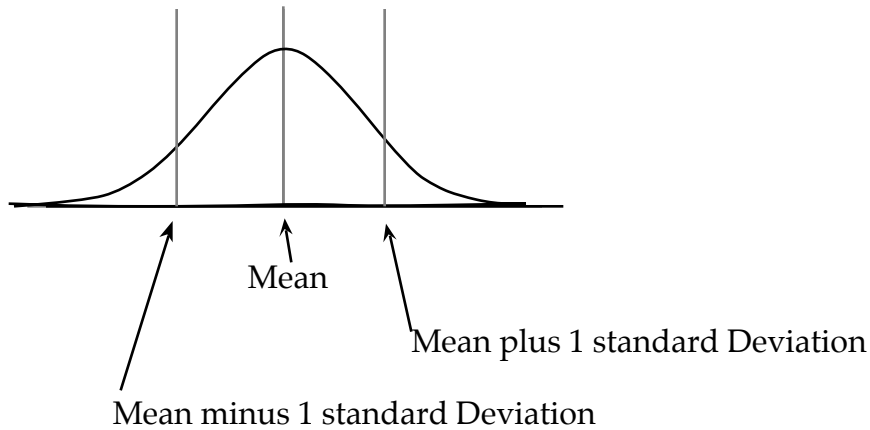
Wheat Cereals	Protein in Grams		Deviations	Squared Deviations
Quaker Puffed Wheat	1.6		-1.16	1.35
Shredded Wheat	2.2		-0.56	0.31
Wheaties	2.8		0.04	0.00
Wheat Chex	2.8		0.04	0.00
Wheat Flakes	4.4		1.64	2.69
<i>Sum</i>	13.8			4.35
<i>Mean/Variance</i>	2.76			0.87
<i>Standard Deviation</i>				0.93

The rendering of this information into English is traditionally a little opaque:

The mean protein content of wheat breakfast cereals is 2.76 grams of protein with a standard deviation of .93 grams.

The mean personal income of these college graduates is ___ specifying the mean, with a standard deviation of ___ (specifying the standard deviation.)

What you are supposed to “see” in that statement is a schematic version of what you actually have. You are supposed to see a distribution of data that is symmetrical and bell shaped: The mean marks the center point. The standard deviation marks off a central region which, schematically, corresponds to the inflection points in the curve of the bell:



Typically, in writing, one standard deviation is used as a yardstick to mark off small variations while two standard deviations are used to mark off large variations: If the difference between the mean incomes of two different populations is less than one standard deviation, that is taken to suggest that the difference between the means is small (which is not to say that it is unimportant). Two standard deviations are used as a yardstick to mark off large variations: If the difference between the mean incomes of two different populations is more than two standard deviations, that is taken to suggest that the difference between the means is large.

But the real cue to writing and using these things is to keep it simple: You are using the mean and the standard deviation to describe a picture of the data. So, provide the picture, your stem and leaf drawing or a histogram, and accompany it with the numbers. Use them all: Use the median, the quartiles, the mean, and the standard deviation. You will learn, with experience, to match the numbers to the picture, matching the numbers to the peculiar things that are likely to show up in real data. But there is no need to speak in code: Speak, and write clearly. Show the picture. Add the numbers.