Chapter 5

Deviating from the Average

In This Chapter

- What variation is all about
- Variance and standard deviation
- Excel worksheet functions that calculate variation
- Workarounds for missing worksheet functions
- Additional worksheet functions for variation

Here are three pieces of wisdom about statisticians:

Piece of Wisdom #1: "A statistician is a person who stands in a bucket of ice water, sticks their head in an oven and says 'on average, I feel fine.'" (K. Dunning)

Piece of Wisdom #2: "A statistician drowned crossing a stream with an average depth of 6 inches." (Anonymous)

Piece of Wisdom #3: "Three statisticians go deer hunting with bows and arrows. They spot a big buck and take aim. One shoots and his arrow flies off ten feet to the left. The second shoots and his arrow goes ten feet to the right. The third statistician jumps up and down yelling, 'We got him! We got him!'" (Bill Butz, quoted by Diana McLellan in *Washingtonian*)

What's the common theme? Calculating the mean is a great way to summarize a group of numbers, but it doesn't supply all the information you typically need. If you just rely on the mean, you might miss something important.

To avoid missing important information, another type of statistic is necessary — a statistic that measures *variation*. It's a kind of average of how much each number in a group differs from the group mean. Several statistics are available for measuring variation. All of them work the same way: The larger the value of the statistic, the more the numbers differ from the mean. The smaller the value, the less they differ.

Measuring Variation

Suppose you measure the heights of a group of children and you find that their heights (in inches) are

48, 48, 48, 48, 48

Then you measure another group and find that their heights are

50, 47, 52, 46, 45

If you calculate the mean of each group, you'll find they're the same — 48 inches. Just looking at the numbers tells you the two groups of heights are different: The heights in the first group are all the same, while the heights in the second vary quite a bit.

Averaging squared deviations: Variance and how to calculate it

One way to show the dissimilarity between the two groups is to examine the deviations in each one. Think of a "deviation" as the difference between a score and the mean of all the scores in a group.

Here's what I'm talking about. Table 5-1 shows the first group of heights and their deviations.

Table 5-1 The First Group of Heights and Their Deviations

Height	*Height-Mean*	*Deviation*
48	48-48	0
48	48-48	0
48	48-48	0
48	48-48	0
48	48-48	0

One way to proceed is to average the deviations. Clearly, the average of the numbers in the Deviation column is zero.

Table 5-2 shows the second group of heights and their deviations.

Table 5-2 The Second Group of Heights and Their Deviations

Height	*Height-Mean*	*Deviation*
50	50-48	2
47	47-48	–1
52	52-48	4
46	46-48	–2
45	45-48	–3

What about the average of the deviations in Table 5-2? That's . . . zero!

Hmmm . . . Now what?

Averaging the deviations doesn't help us see a difference between the two groups, because the average of deviations from the mean in any group of numbers is *always* zero. In fact, veteran statisticians will tell you that's a defining property of the mean.

The joker in the deck here is the negative numbers. How do statisticians deal with them?

The trick is to use something you might recall from algebra: A minus times a minus is a plus. Sound familiar?

So . . . does this mean that you multiply each deviation times itself, and then average the results? Absolutely. Multiplying a deviation times itself is called *squaring a deviation*. The average of the squared deviations is so important that it has a special name: *variance*.

Table 5-3 shows the group of heights from Table 5-2, along with their deviations and squared deviations.

Table 5-3 The Second Group of Heights and Their Squared Deviations

Height	*Height-Mean*	*Deviation*	*Squared Deviation*
50	50-48	2	4
47	47-48	–1	1
52	52-48	4	16
46	46-48	–2	4
45	45-48	–3	9

The variance — the average of the squared deviations for this group — is (4 + 1 + 16 + 4 + 9)/5 = 34/5 = 6.8. This, of course, is very different from the first group, whose variance is zero.

To develop the variance formula for you and show you how it works, I use symbols to show all this. $\bar{X}$ represents the Height heading in the first column of the table and $\bar{X}$ represents the mean. Because a deviation is the result of subtracting the mean from each number,

$$\left(X - \bar{X}\right)$$

represents a deviation. Multiplying a deviation by itself? That's just

$$\left(X - \bar{X}\right)^2$$

To calculate variance you square each deviation, add them up, and find the average of the squared deviations. If *N* represents the amount of squared deviations you have (in our example, five), then the formula for calculating the variance is

$$\frac{\sum\left(X - \bar{X}\right)^2}{N}$$

Σ is the uppercase Greek letter sigma and it stands for the sum of.

What's the symbol for Variance? As I say in Chapter 1, Greek letters represent population parameters and English letters represent statistics. Imagine that our little group of five numbers is an entire population. Does the Greek alphabet have a letter that corresponds to *V* in the same way that μ (the symbol for the population mean) corresponds to M?

As a matter of fact, it doesn't. Instead, we use the *lowercase* sigma! It looks like this: σ. Not only that, but because we're talking about squared quantities, the symbol is σ^2.

So the formula for calculating variance is:

$$\sigma^2 = \frac{\sum\left(X - \bar{X}\right)^2}{N}$$

Variance is large if the numbers in a group vary greatly from their mean. Variance is small if the numbers are very similar to their mean.

The variance you just worked through is appropriate if the group of five measurements is a population. Does this mean that variance for a sample is different? It does, and you'll see why in a minute. First, I turn your attention back to Excel.

VARP and VARPA

Excel's two worksheet functions, VARP and VARPA, calculate the population variance.

Start with VARP. Figure 5-1 shows the Function Arguments dialog box for VARP along with data. Here are the steps to follow:

Figure 5-1: Working with VARP.

1. **Put your data into a worksheet and select a cell to display the result.**

   Figure 5-1 shows that for this example, I've put the numbers 50, 47, 52, 46, 45 into cells B2 through B6 and selected B8 for the result.

2. **From the Statistical Functions menu, select VARP to open the VARP Function Arguments dialog box.**

3. **In the Function Arguments dialog box, enter the appropriate values for the arguments.**

   I entered B2:B6 in the Number1 field. The population variance, 6.8, appears in the Function Arguments dialog box.

4. **Click OK to close the dialog box and put the result in the selected cell.**

Had I defined Score as the name of B2:B6 (see Chapter 2), the formula in the formula bar would be

```
=VARP(Score)
```

When VARP calculates the variance in a range of cells, it only sees numbers. If text or logical values are in some of the cells, VARP ignores them.

VARPA, on the other hand, does not. VARPA takes text and logical values into consideration and includes them in its variance calculation. How? If a cell contains text, VARPA sees that cell as containing a value of zero. If a cell contains the logical value FALSE, that's also zero as far as VARPA is concerned. In VARPA's view of the world, the logical value TRUE is one. Those zeros and ones get added into the mix and affect the mean and the variance.

To see this in action, I keep the numbers in cells B2 through B6 and again select cell B8. I follow the same steps as for VARP, but this time open the VARPA Function Arguments dialog box. In the Value1 field of the VARPA dialog box I type B2:B7 (that's B7, *not* B6) and click OK. Cell B8 shows the same result as before because VARPA evaluates the blank cell B7 as no entry.

Typing TRUE into Cell B7 changes the result in B8 because VARPA evaluates B7 as 1. (See Figure 5-2.)

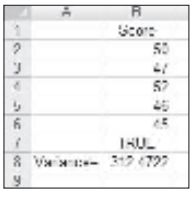

Figure 5-2: VARPA evaluates TRUE as 1.0, changing the variance from the value in Figure 5-1.

Typing FALSE (or any other string of letters except TRUE) into B7 changes the value in B8 once again. This time, VARPA evaluates B7 as zero.

Sample variance

Earlier, I mentioned that you use this formula to calculate population variance:

$$\sigma^2 = \frac{\sum\left(X - \bar{X}\right)^2}{N}$$

I also said that sample variance is a little different. Here's the difference. If your set of numbers is a sample drawn from a large population, you're probably interested in using the variance of the sample to estimate the variance of the population.

The formula you used for the variance doesn't quite work as an estimate of the population variance. Although the sample mean works just fine as an estimate of the population mean, this doesn't hold true with variance, for reasons *way* beyond the scope of this book.

How do you calculate a good estimate of the population variance? It's pretty easy. You just use *N-1* in the denominator rather than *N*. (Again, for reasons way beyond our scope.)

Also, because we're working with a characteristic of a sample (rather than of a population), we use the English equivalent of the Greek letter — *s* rather than σ. This means that the formula for the sample variance is

$$s^2 = \frac{\sum\left(X - \bar{X}\right)^2}{N - 1}$$

The value of s^2, given the squared deviations in our set of five numbers is

(4 + 1 + 16 + 4 + 9)/4 = 34/4 = 8.5

So, if these numbers

50, 47, 52, 46, 45

are an entire population, their variance is 6.4. If they're a sample drawn from a larger population, our best estimate of that population's variance is 8.5.

VAR and VARA

The worksheet functions VAR and VARA calculate the sample variance.

Figure 5-3 shows the Function Arguments dialog box for VAR with 50, 47, 52, 46, 45 entered into cells B2 through B6. Cell B7 is part of the cell range, but I left it empty.

Figure 5-3: Working with VAR.

The relationship between VAR and VARA is the same as the relationship between VARP and VARPA: VAR ignores cells that contain logical values (TRUE and FALSE) and text. VARA includes those cells. Once again, TRUE evaluates to 1.0 and FALSE evaluates to 0. Text in a cell causes VARA to see that cell's value as 0.

This is why I left B7 blank. If you experiment a bit with VARA and logical values or text in B7, you'll see exactly what VARA does.

Back to the Roots: Standard Deviation

After you calculate the variance of a set of numbers, you have a value whose units are different from your original measurements. For example, if your original measurements are in inches, their variance is in square inches. This is because you square the deviations before you average them.

Often, it's more intuitive if you have a variation statistic that's in the same units as the original measurements. It's easy to turn variance into that kind of statistic. All you have to do is take the square root of the variance.

Like the variance, this square root is so important that we give it a special name: *standard deviation*.

Population standard deviation

The standard deviation of a population is the square root of the population variance. The symbol for the population standard deviation is σ (sigma). Its formula is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum\left(X - \bar{X}\right)^2}{N}}$$

For these measurements (in inches)

50, 47, 52, 46, 45

the population variance is 6.8 square inches, and the population standard deviation is 2.61 inches (rounded off).

STDEVP and STDEVPA

The Excel worksheet functions STDEVP and STDEVPA calculate the population standard deviation. After entering your numbers into your worksheet and selecting a cell

1. **Type your data into an array and select a cell for the result.**
2. **In the Statistical Functions menu, select STDEVP to open the STDEVP Function Arguments dialog box.**
3. **In the Function Arguments dialog box, type the appropriate values for the arguments.**

   After you enter the data array, the dialog box shows the value of the population standard deviation for the numbers in the data array. Figure 5-4 shows this.

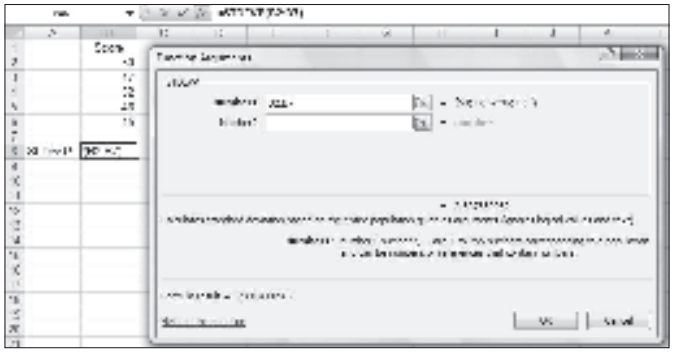

Figure 5-4: The Function Arguments dialog box for STDEVP, along with the data.

4. **Click OK to close the dialog box and put the result into the selected cell.**

Like VARPA, STDEVPA uses any logical values and text values it finds when it calculates the population standard deviation. TRUE evaluates to 1.0 and FALSE evaluates to 0. Text in a cell gives that cell a value of 0.

Sample standard deviation

The standard deviation of a sample — an estimate of the standard deviation of a population — is the square root of the sample variance. Its symbol is *s* and its formula is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum\left(X - \bar{X}\right)^2}{N-1}}$$

For these measurements (in inches)

50, 47, 52, 46, 45

the population variance is 8.4 square inches, and the population standard deviation is 2.92 inches (rounded off).

STDEV and STDEVA

The Excel worksheet functions STDEV and STDEVA calculate the sample standard deviation. To work with STDEV

1. **Type your data into an array and select a cell for the result.**
2. **In the Statistical Functions menu, select STDEV to open the STDEV Function Arguments dialog box.**
3. **In the Function Arguments dialog box, type the appropriate values for the arguments.**

   With the data array entered, the dialog box shows the value of the population standard deviation for the numbers in the data array. Figure 5-5 shows this.
4. **Click OK to close the dialog box and put the result into the selected cell.**

STDEVA uses text and logical values in its calculations. Cells with text have values of 0, and cells whose values are FALSE also evaluate to 0. Cells that evaluate to TRUE have values of 1.0.

Figure 5-5: The Function Arguments dialog box for STDEV.

The missing functions: STDEVIF and STDEVIFS

Here's a rule of thumb: Whenever you present a mean, provide a standard deviation. Use AVERAGE and STDEV in tandem.

Remember that Excel 2007 offers two new functions, AVERAGEIF and AVERAGEIFS, for calculating means conditionally. (See Chapter 4.) Two additional new functions would have been helpful: STDEVIF and STDEVIFS for calculating standard deviations conditionally when you calculate means conditionally.

Excel 2007, however, doesn't provide these functions. Instead, I show you a couple of workarounds that enable you to calculate standard deviations conditionally.

The workarounds filter out data that meet a set of conditions, and then calculate the standard deviation of the filtered data. Figure 5-6 shows what I mean. The data are from the fictional psychology experiment I describe in Chapter 4.

Here, once again, is the description:

A person sits in front of a screen and a color-filled shape appears. The color is either red or green and the shape is either a square or a circle. The combination for each trial is random, and all combinations appear an equal number of times. In the lingo of the field, each appearance of a color-filled shape is called a *trial*. So the worksheet shows the outcomes of 16 trials.

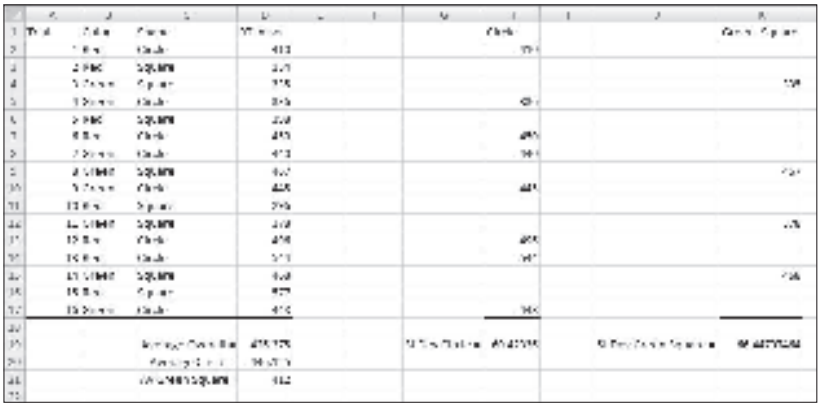

Figure 5-6: Filtering data to calculate standard deviation conditionally.

The person sitting in front of the screen presses a button as soon as he or she sees the shape. Column A presents the trial number. Columns B and C show the color and shape, respectively, presented on that trial. Column D (labeled RT_msec) presents one person's reaction time in milliseconds (thousandths of a second) for each trial. So, for example, row 2 tells you that on the first trial, a red circle appeared and the person responded in 410 msec (milliseconds).

For each column, I defined the name in the top cell of the column to refer to the data in that column. If you don't remember how to do that, reread Chapter 2.

Cell D19 displays the overall average of RT_msec. The formula for that average, of course, is

```
=AVERAGE(RT_msec)
```

Cell D20 shows the average for all trials on which a circle appeared. The formula that calculates that conditional average is

```
=AVERAGEIF(Shape,"Circle",RT_msec)
```

Cell D21 presents the average for trials on which a green square appeared. That formula is

```
=AVERAGEIFS(RT_msec, Color,"Green", Shape,"Square")
```

Columns H and K hold filtered data. Column H shows the data for trials that displayed a circle. Cell H19 presents the standard deviation for those trials and is the equivalent of

```
=STDEVIF(Shape,"Circle",RT_msec)
```

if this function existed.

Column K shows the data for trials that displayed a green square. Cell K19 presents the standard deviation for those trials, and is the equivalent of

```
=STDEVIFS(RT_msec, Color,"Green",Shape,"Square")
```

if *that* function existed.

How did I filter the data? I'll let you in on it in a moment, but first I have to tell you about . . .

A little logic

In order to proceed, you have to know about two of Excel's logic functions: IF and AND. You access them by clicking

Formulas | Logical Functions

and selecting them from the Logical Functions menu.

IF takes three arguments:

- A logical condition to be satisfied
- The action to take if the logical condition is satisfied (that is, if the value of the logical condition is TRUE)
- An optional argument that specifies the action to take if the logical condition is not satisfied (that is, if the value of the logical condition is FALSE)

Figure 5-7 shows the Function Arguments dialog box for IF.

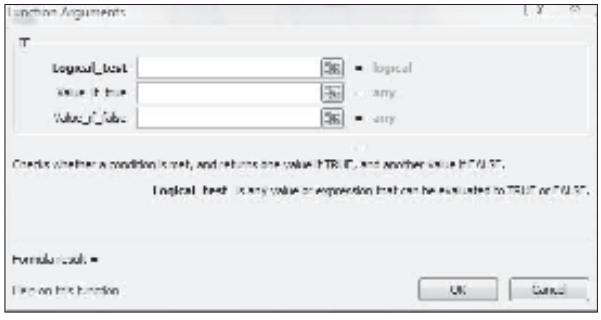

Figure 5-7: The Function Arguments dialog box for IF.

AND can take up to 255 arguments. AND checks to see if all of its arguments meet each specified condition — that is, if each condition is TRUE. If they all do, AND returns the value TRUE. If not, AND returns FALSE.

Figure 5-8 shows the Function Arguments dialog box for AND.

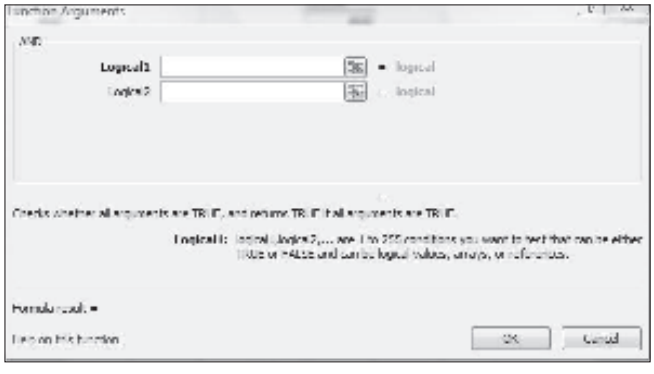

Figure 5-8: The Function Arguments dialog box for AND.

And now, back to the show

In this example, I use IF to set the value of a cell in column H to the corresponding value in column D if the value in the corresponding cell in column C is "Circle". The formula in cell H2 is

```
=IF(C2="Circle",D2," ")
```

If this were a phrase it would be, "If the value in C2 is 'Circle', then set the value of this cell to the value in D2. If not, leave this cell blank." Autofilling the next 15 cells of column H yields the filtered data in column H in Figure 5-6. The standard deviation in cell H19 is the value STDEVIF would have provided.

I could have omitted the third argument (the two double-quotes) without affecting the value of the standard deviation. Without the third argument, Excel fills in FALSE for cells that don't meet the condition instead of leaving them blank.

I use AND along with IF for the cells in column K. Each one holds the value from the corresponding cell in column D if two conditions are true:

- The value in the corresponding cell in column B is "Green"
- The value in the corresponding cell in column C is "Square"

The formula for cell K2 is

```
=IF(AND(B2="Green",C2="Square"),D2," ")
```

If *this* was a phrase it would be, "If the value in B2 is 'Green' and the value in C2 is 'Square', then set the value of this cell to the value in D2. If not, leave this cell blank." Autofilling the next 15 cells in column K results in the filtered data in column K in Figure 5-6. The standard deviation in cell K19 is the value STDEVIFS would have provided.

Related Functions

Before we move on, take a quick look at a couple of other variation-related worksheet functions.

DEVSQ

DEVSQ calculates the sum of the squared deviations from the mean (without dividing by *N* or by *N-1*). For these numbers

50, 47, 52, 46, 45

that's 34, as Figure 5-9 shows.

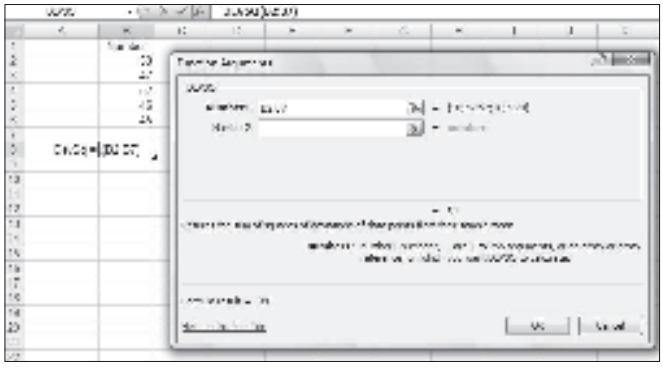

Figure 5-9: The DEVSQ dialog box.

Average deviation

One more Excel function deals with deviations in a way other than squaring them.

The variance and standard deviation deal with negative deviations by squaring all the deviations before averaging them. How about if we just ignore the minus signs? This is called taking the *absolute value* of each deviation. (That's the way mathematicians say "How about if we just ignore the minus signs?").

If we do that for the heights

50, 47, 52, 46, 45

we can put the absolute values of the deviations into a table like Table 5-4.

Table 5-4 A Group of Numbers and Their Absolute Deviations

Height	*Height-Mean*	|*Deviation*|
50	50-48	2
47	47-48	1
52	52-48	4
46	46-48	2
45	45-48	3

In Table 5-4, notice the vertical lines around Deviation in the heading for the third column. Vertical lines around a number symbolize its absolute value. That is, the vertical lines are the mathematical symbol for "How about if we just ignore the minus signs?"

The average of the numbers in the third column is 2.4. This average is called the *average absolute deviation*, and it's a quick and easy way to characterize the spread of measurements around their mean. It's in the same units as the original measurements. So if the heights are in inches, the absolute average deviation is in inches, too.

Like variance and standard deviation, a large average absolute deviation signifies a lot of spread. A small average absolute deviation signifies little spread.

This statistic is less complicated than variance or standard deviation, but is rarely used. Why? For reasons that are (once again) beyond our scope, statisticians can't use it as the foundation for additional statistics you'll meet later. Variance and standard deviation serve that purpose.

AVEDEV

Excel's AVEDEV worksheet function calculates the average absolute deviation of a group of numbers. Figure 5-10 shows the AVEDEV dialog box, which presents the average absolute deviation for the cells in the indicated range.

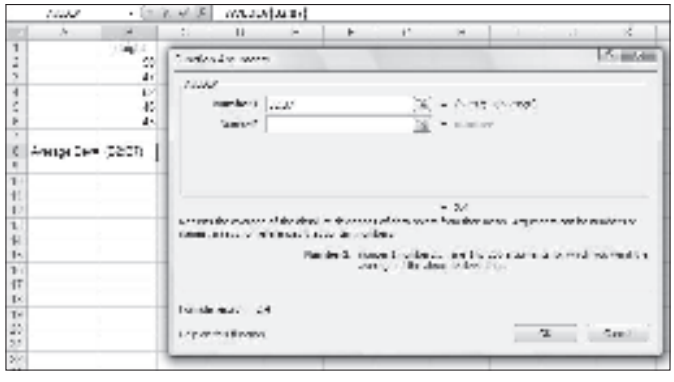

Figure 5-10: The AVEDEV Function Arguments dialog box.