Why divide by (n-1) for sample standard deviation?

Short answer:

Most of the time we do not know μ (the population average) and we estimate it with $\bar{x}$ (the sample average). The formula for s^2 measures the squared deviations from $\bar{x}$ rather than μ. The x_i's tend to be closer to their average $\bar{x}$ rather than μ, so we compensate for this by using the divisor (n-1) rather than n.

Theoretical Answer:

The quantity: $x_i - \bar{x}$ is called the ith deviation from the mean. Note that:

$$\sum_{i-1}^{n}(x_i - \bar{x}) = 0 \quad \textit{(sum of the deviations is zero)}$$

$$\textit{Proof:} \quad \sum_{i-1}^{n}(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

As such, it is customary to refer to s^2 as being based on (n-1) "degrees of freedom. The reason for doing this is that we have already used up one piece of all the information in the dataset in calculating the mean and we should take this into account. This results from the fact that s^2 is based on the n quantities $x_1 - \bar{x}$, $x_2 - \bar{x}$, …, $x_n - \bar{x}$. But the sum of these deviations is zero – thus, specifying the values of any (n-1) of the quantities determines the remaining one. For example, if n = 4 and $x_1 - \bar{x} = 8$, $x_2 - \bar{x} = -6$, and $x_4 - \bar{x} =$ -4, then automatically we know $x_3 - \bar{x} = 2$, so only three of the four values of $x_i - \bar{x}$ are freely determines – hence 3 degrees of freedom.

Empirical Answer:

If we take many samples from a population which has the mean μ, calculate the sample mean $\bar{x}$, and then average all these estimates of μ, we should find that their average is very close to μ. However, if we calculated the variance of each sample by the formula:

$\frac{\sum(x_i - \bar{x})}{n}$ (thus, divide by n), and then average all these supposed estimates of σ^2, we would probably find that their average is less than σ^2. We compensate for this by dividing by (n-1).

See the following web site for an animation:

http://www.uvm.edu/~dhowell/SeeingStatisticsApplets/N-1.html

In the example, consider a population consisting of each of the numbers between 0 and 100. Because we have the whole population, we know that the true mean is $\mu = 50$, and the variance is $\sigma^2 = 850$. The true standard deviation (σ) is thus 29.154. (the website is incorrect in saying that $\sigma^2 = 853$.)