

Variance and Standard Deviation: Where Does the $N - 1$ Term Originate?

Recall that we defined variance as:

$$\text{var}(X) = \mu_2 = \sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 \quad (1)$$

and the *standard deviation* is the square root of this:

$$\text{std}(X) = \sigma_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2} \quad (2)$$

In the variance and standard deviation, we have divided by $N - 1$ rather than N . We do this to obtain *unbiased* estimates of the standard deviation or variance, given that our data set is finite. If N were infinite, subtracting one would make no difference.

Consider the limiting case when N is 1, so that the mean \bar{X} equals our single observation X_1 . We might want to say that our standard deviation was zero, but that would be a little misleading. With only one observation, we can't say anything about the typical deviation from the mean. By using $N - 1$ we will end up computing zero over zero, which gives us an undefined standard deviation, representing exactly what we know.

Mathematically, we can understand where the -1 comes from by working through the derivation carefully. Each of my observations can be represented as $X_i = \hat{X}'_i + \hat{m}$, a deviation plus the observed mean \hat{m} . They can also be written as $X_i = X'_i + m$, a deviation plus the unknown true mean m . Although we don't know what the true mean is, we can represent it as $m = X_i - X'_i = 1/N \sum_{i=1}^N (X_i - X'_i)$. The difference between the true mean and the observed mean can be written $\epsilon = \hat{m} - m = 1/N \sum_{i=1}^N X'_i$. The true variance clearly should be $\mu_2 = 1/N \sum_{i=1}^N X_i'^2$. But since we don't know m , we don't know the true deviations from the mean. For the moment, rather than dividing by N or $N - 1$, we'll assume an unknown normalization term M , in which case our best approximation of the variance is:

$$\begin{aligned} \hat{\mu}_2 &= \frac{1}{M} \sum_{i=1}^N (X_i - \hat{m})^2 = \frac{1}{M} \sum_{i=1}^N (X'_i - \epsilon)^2 \\ &= \frac{1}{M} \sum_{i=1}^N (X_i'^2 - 2X'_i\epsilon + \epsilon^2) \\ &= \frac{1}{M} \sum_{i=1}^N \left(X_i'^2 - 2X'_i \frac{1}{N} \sum_{j=1}^N X'_j + \frac{1}{N^2} \sum_{j=1}^N X'_j \sum_{k=1}^N X'_k \right) \\ &= \frac{1}{M} \sum_{i=1}^N \left(X_i'^2 - \frac{2}{N} X_i'^2 + \frac{1}{N^2} \sum_{j=1}^N X_j'^2 \right) \\ &= \frac{1}{M} \sum_{i=1}^N \left(X_i'^2 - \frac{2}{N} X_i'^2 + \frac{1}{N} X_i'^2 \right) \\ &= \frac{1}{M} \sum_{i=1}^N X_i'^2 \left(1 - \frac{1}{N} \right) \\ &= \frac{1}{M} N \mu_2 \frac{N-1}{N} = \frac{N-1}{M} \mu_2 \end{aligned} \quad (3)$$

This leads to the conclusion that if we want an unbiased estimate of the variance, then $M = N - 1$. Here, we've assumed that X'_i and X'_j are uncorrelated unless $i = j$. Thus

$$\langle X'_i X'_j \rangle = \begin{cases} \mu_2 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \quad (4)$$

Matlab and most statistics packages normalize the standard deviation by $N - 1$ as this discussions suggests. However, if you wanted to have an unbiased estimate of the variance of the standard deviation, you would choose a different normalization.

Probability Density Functions

We have been looking at histograms of data, but histograms have some problems. They aren't universal in character. Figure 1 shows two histograms—one for a small data set, and one for a larger data set. They look similar, but how do we tell if they are really the same? You might be tempted to divide the size of each peak by N , the total number of data points in each plot, in order to make the histograms look the same. That would work for this example, but we would still have problems for cases where we plotted data using different bin widths, as in Figure 2. Clearly we need a generic way to represent data distributions graphically and that is the objective of this discussion.

Probability density functions tells us the probability of observing a value within a specific range. If P is the probability density function (pdf), $P(x)dx$ is the probability of observing a value between x and $x + dx$. This notation can be a little confusing, but it has several important features. The pdf has no dependence on bin width or total sample size. It lets us determine the probability of observing a value in any arbitrary range:

$$\text{Prob}[x_1 < x < x_2] = \int_{x_1}^{x_2} P(x)dx. \quad (5)$$

Accordingly, the probability of observing a value x between $-\infty$ and $+\infty$ is clearly 100% or 1. Thus,

$$\text{Prob}[-\infty < x < \infty] = \int_{-\infty}^{\infty} P(x)dx = 1. \quad (6)$$

The *cumulative distribution function* $C(x)$ is the probability of observing a value less than x . It can be computed by integrating the pdf.

$$C(x) = \int_{-\infty}^x P(x')dx'. \quad (7)$$

$C(x)$ is 0 when x approaches minus infinity, indicating that there's a negligibly small chance of having an infinitely small value of x , and it is 1 when x goes to plus infinity, which says that there is a 100% chance of observing some value. The midpoint, where $C(x) = 0.5$ is the median.

Practical Considerations

Once you've collected data and plotted a histogram, how do you transform this into a pdf. If histogram bin i contains n_i points, then the fraction n_i/N will tell you the probability that an observation appears in

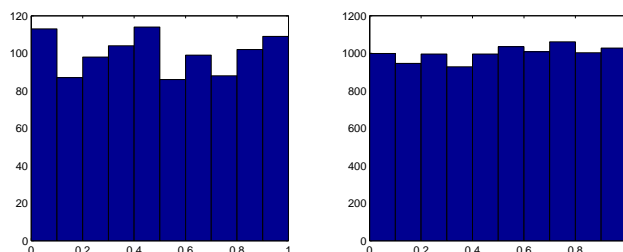


Figure 1: (left) Histogram for uniformly distributed random data, with $N = 1000$. (right) Same thing with $N = 10,000$. Notice the difference in the y-axis scales.

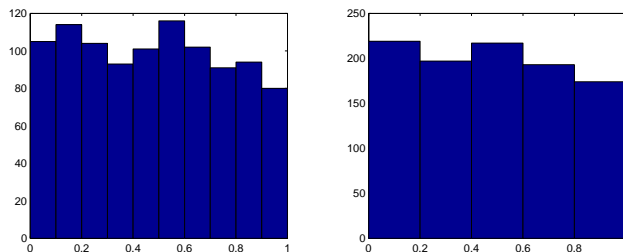


Figure 2: (left) Histogram for uniformly distributed random data, with $N = 1000$, as in left panel of Figure 1 above. (right) Same data plotted using half as many bins. Again notice the difference in the y-axis scales.

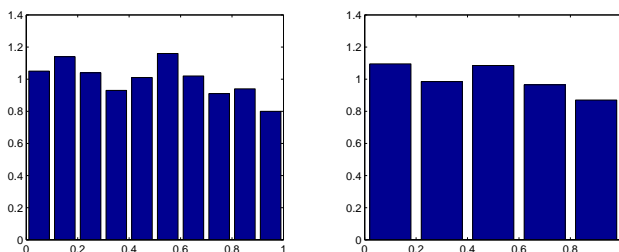


Figure 3: (left) Empirical probability density function for uniformly distributed random data with $N = 1000$ and bin width 0.1. (right) Same thing with bin width 0.2. Note that y-axis scales are the same for both pdfs.

that bin. We also need to divide by the bin width Δx , so that the pdf will integrate to one. Thus the empirical pdf has bins of height $n_i/(N\Delta x)$. Figure 3 shows pdfs derived from the histograms in Figure 2. By eye we can see that these pdfs are very similar. Figure 4 shows sample pdfs for wind data.

Observational data often have *Gaussian* or *normal* distributions—that’s the classic bell-shaped curve that professors sometimes use to fix grades—and most statistical theory assumes that quantities are normally distributed, with

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]. \quad (8)$$

where σ is the standard deviation. The corresponding cumulative distribution function is the error function. These analytic forms are used to derive much of the basic statistical theory that underlies data analysis.

However, other forms of pdfs often appear in observations. Figure 5 shows some common sample pdfs. Velocities, both in the ocean and in the atmosphere, sometimes appear more double exponential than

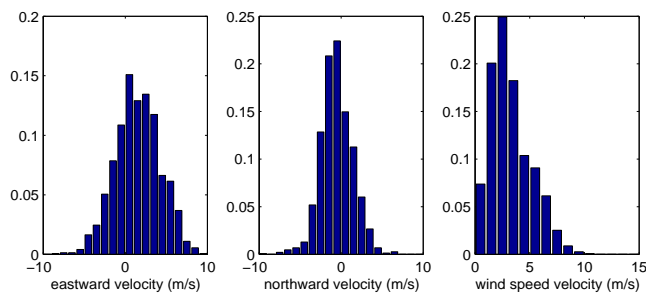


Figure 4: Empirical probability density functions for (left) eastward wind velocity, (center) northward wind velocity, (right) wind speed from the National Centers for Environmental Prediction reanalysis for the year 2000 for a grid point located approximately at San Diego.

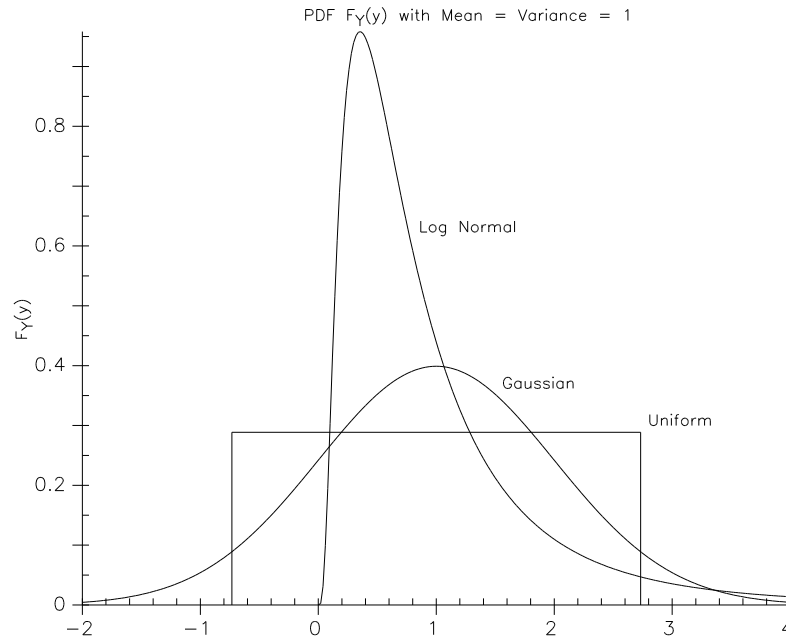


Figure 5: Examples of probability density functions with unit mean and variance.

Gaussian.

$$P(x) = \frac{1}{\sigma\sqrt{2}} \exp\left[-\frac{|x|\sqrt{2}}{\sigma}\right]. \quad (9)$$

Speeds, since they are always positive, roughly follow a Rayleigh distribution (not shown but much like the log normal distribution). Wind directions can be nearly a uniform distribution in cases where the wind is equally likely to blow in any direction between 0 and 2π .