The Central Limit Theorem

Why are data so often assumed to have Gaussian distributions? The *central limit theorem* provides a clear demonstration of why Gaussian distributions are important. Suppose that I have a bunch of sets of variables x_1, x_2, ... x_n, each of which has a uniform distribution (or any other non-Gaussian distribution, $P(x)$), with mean $m = 0$, standard deviation σ, and for convenience a third moment $\mu_3 = 0$. Now I define a variable $y = (x_1 + x_2 + x_3 + ... + x_n)/n$. What is the distribution of y? There are a number of derivations. Here's one.

The mean of y is clearly the mean of each of the x_i's, so $\overline{y} = m$. The variance of y is:

$$\begin{aligned}
\sigma_y^2 &= \int\int\int ... \int \frac{(x_1 + x_2 + x_3 + ... + x_n)^2}{n^2} P(x_1)P(x_2)P(x_3)...P(x_n)\, dx_1\, dx_2\, dx_3 ... dx_n \\
&= \int \frac{nx^2}{n^2} P(x)\, dx \\
&= \frac{1}{n}\int x^2 P(x)\, dx = \frac{\sigma^2}{n} \qquad (1)
\end{aligned}$$

This tells us importantly that the standard deviation of y is $\sigma/\sqrt{n}$. In other words, the standard deviation of the mean is smaller than the standard deviation of each contributing element by a factor $\sqrt{n}$. The third moment of y is

$$\begin{aligned}
\mu_3 &= \int\int\int ... \int \frac{(x_1 + x_2 + x_3 + ... + x_n)^3}{n^3} P(x_1)P(x_2)P(x_3)...P(x_n)\, dx_1\, dx_2\, dx_3 ... dx_n \\
&= \int \frac{nx^3}{n^3} P(x)\, dx \\
&= \frac{1}{n^2}\int x^3 P(x)\, dx = 0. \qquad (2)
\end{aligned}$$

And the fourth moment of y is

$$\begin{aligned}
\mu_4 &= \int\int\int ... \int \frac{(x_1 + x_2 + x_3 + ... + x_n)^4}{n^4} P(x_1)P(x_2)P(x_3)...P(x_n)\, dx_1\, dx_2\, dx_3 ... dx_n \\
&= \int\int \left(\frac{nx_i^4}{n^4} + \frac{n(n-1)6x_i^2 x_j^2}{2n^4} \right) P(x_i)P(x_j)\, dx_i\, dx_j \\
&\approx \frac{3}{n^2}\left[\int x^2 P(x)\, dx\right]^2 = 3\sigma_y^4 \qquad (3)
\end{aligned}$$

Thus in the limit of large n, the kurtosis μ_4/μ_2^2 is 3, as for a Gaussian distribution. We can keep looking at higher order modes, but the fact that the variance and kurtosis are consistent with a Gaussian distribution tells us that y is Gaussian.

Usually we assume that our observed data represent a sum of many random processes. For example, we measure velocities in the ocean, and the velocities are driven by lots of random pushes by winds. This leads to an expectation that geophysical data will have Gaussian distributions. Similarly, we suppose that our errors are the result of many different randomly summed events, so they will also be Gaussian.

Error: Keeping Track of Uncertainty

Data is imperfect. It is corrupted by instrumental errors, measurement noise, and signals that we are not interested in studying (though they may be physically explainable.)

As an example, suppose we measure air temperature from the top of Mount Soledad with the intent of studying long-term climate change in La Jolla. Our estimates might be wrong because the thermometer was

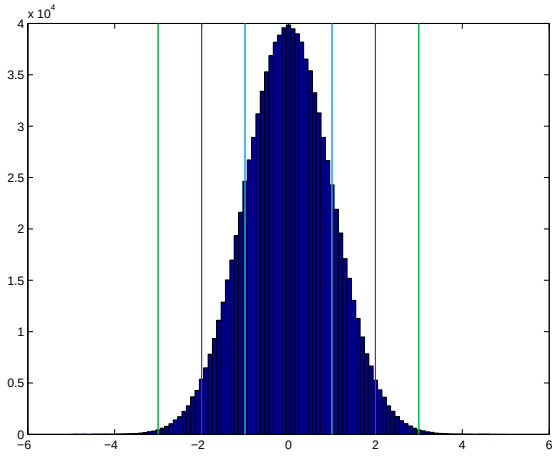

Figure 1: Histogram of random data with Gaussian or normal distribution. Vertical lines indicate 3σ (green), 2σ (red), and 1σ (light blue).

inaccurate—perhaps it is biased, or perhaps it just behaves erratically. They might be wrong because even though the thermometer is essentially accurate, its precision is low, so perhaps the thermometer is reliable to $\pm 0.1^{\circ}$C, but we want to know temperatures to $\pm 0.01^{\circ}$C. And they might be wrong, at least for our purposes, because they measure lots of variability (due to day-night differences, seasons, weather patterns) that has nothing to do with long-term climate change.

How do we keep track of these errors? There are two basic strategies for defining our measurement uncertainty. One is to use an a priori estimate of the measurement error. For example, we read in the manual that came with our thermometer that the instrument is accurate to $\delta T = \pm 0.1^{\circ}$C. Alternatively, we can estimate the uncertainty by looking at the variability in our observations. Suppose I have 100 estimates of temperature. The standard deviation of these estimates tells us how much any one of them might differ from the true value. Recall that the standard deviation is:

$$\mathrm{std}(X) = \sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}. \tag{4}$$

If all my temperature estimates come from the same location at the same instant in time (or at least the same basic conditions), the variance in the data will be determined by the instrumental uncertainty. But if the temperature estimates come from different points in time or space, the spatial and temporal variability of the natural world will control my standard deviation, possibly much more than the instrumental error. If I want to use these measurements to estimate the mean temperature at a fixed location, then the natural variability will look like noise to me, so the standard deviation σ will be a sensible measure of uncertainty.

Usually we assume that measurements have a Gaussian distribution. (This is a reasonable guess, because the central limit theorem dictates that quantities that are determined from multiple independent random processes are likely to have Gaussian distributions, even if each individual process is not Gaussian.) For a Gaussian distribution, 68% of measurements are with $\pm 1\sigma$, 95.4% within $\pm 2\sigma$ and 99.7% within $\pm 3\sigma$, as indicated in Figure 1. Thus typically we use 1σ or 2σ as our error bar.