

FUNDAMENTAL STATISTICS

Let's start with some basic concepts. Measurements have uncertainties. In the laboratory, uncertainties can be caused by instrumental errors or sometimes by variations in laboratory conditions (e.g. an experimental protocol assumes that the laboratory temperature is constant, but in fact the temperature in the lab fluctuates when the air conditioning cycles on and off). In the natural world, instrumental errors are one source of error, but often the biggest sources of errors are related to the fact that we cannot sample every aspect of the natural world that we would like to understand.

Here's some terminology:

random process: A process that does not produce completely predictable results. Examples are a coin toss, or measurements of temperature.

random variable: the measured quantity representing the outcome of a random process. (For a coin toss, this might be X , where $X = 1$ for heads and -1 for tails; for temperature, it would be T , the measured temperature.)

realization: A process that processes one random variable (e.g. one flip of a coin or one measurement of temperature).

ensemble: A collection of realizations (e.g. a whole series of coin tosses; repeated temperature measurements.)

What do we do with our ensemble of measurements? Usually we make measurements for a reason. If I measure the height of everyone in the classroom, perhaps I want to know the typical height of people in the room, or perhaps I want to estimate the typical height of people on the UCSD campus.

We can start by making a histogram of all of our measurements. Figure 1 shows histograms of east-west wind velocity, north-south wind velocity, and wind speed for the year 2000 from the National Centers for Environmental Prediction. A histogram shows us the range of values that we've measured and the frequency with which each occurs, but it's a little inconvenient, so we might want to distill our results into a more compact form.

Characterizing Typical Values

The average or *mean* is probably the most fundamental statistical quantity. We can think of an average in two ways:

- The observed average.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (1)$$

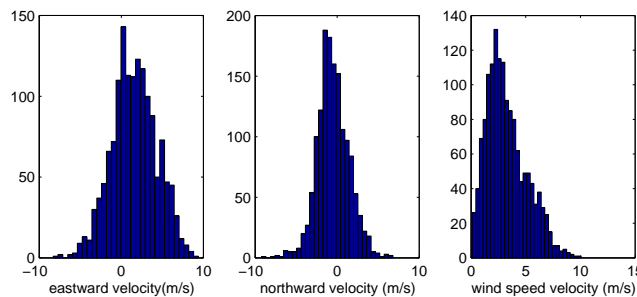


Figure 1: Histograms of (left) eastward wind velocities, (right) northward wind velocities, and (right) wind speed, near San Diego, year 2000.

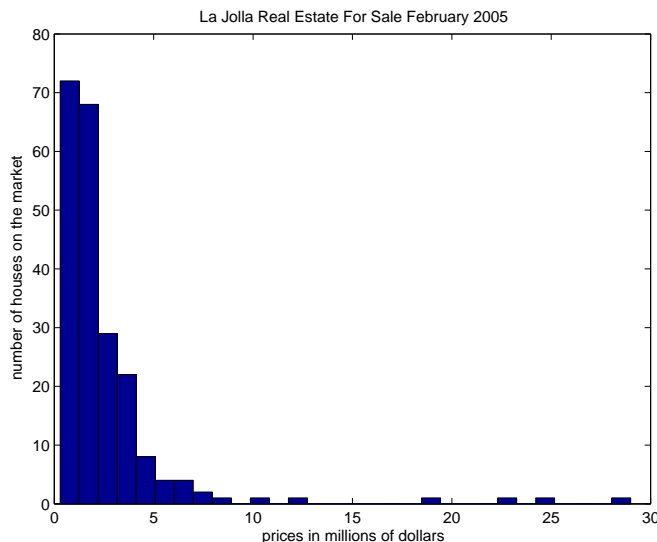


Figure 2: Histogram of listing prices for houses and condominiums on the market in La Jolla, February 2005.

- The true average that we would obtain if we were actually able to measure every possible point.

$$\langle X \rangle = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N X_i \quad (2)$$

This is sometimes called the expected value. You'll see overbars ($\bar{}$) and angular brackets ($\langle \cdot \rangle$) sometimes used interchangeably, but for the moment we'll try to save angular brackets for the expected value and overbars for the observed average.

The average for the wind observations in Figure 1 are 1.58 m s^{-1} , -0.48 m s^{-1} , and 3.26 m s^{-1} for the eastward velocity, northward velocity, and speed, respectively.

Sometimes a mean might not really seem very representative. Economic data can have particularly strange distributions. Figure 2 shows a histogram of housing prices in La Jolla. The data range in value from \$299,900 to \$29,000,000. The mean of these values is about \$2.5 million, but that's clearly skewed by the most expensive properties on the market.

Averages can be deceptive for other reasons. Consider this true factoid: Greater than 95% of Belgians have more than the average number of legs.

Thus sometimes the *median* is a more useful measure. The median is the middle value. To compute it we can sort the data by size, and find the value that falls right in the middle, at the 50th percentile. For the La Jolla housing data, the median is a much more affordable \$1.64 million dollars. And reassuringly, Belgians have a median of two legs each.

The *mode* is another statistical measure. It corresponds to the most probably value—roughly speaking, it's the peak in the histogram. It makes more sense for discrete data (e.g. shoe size, or number of days with rainfall per year). Most environmental data is not discrete, so we don't use modes very often.

Variance and Standard Deviation

How much does a typical measurement differ from the mean? I could try to estimate this by removing the mean from each value before computing the mean:

$$\overline{X - \bar{X}} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X}) = 0 \quad (3)$$

	eastward (m s ⁻¹)	northward (m s ⁻¹)	speed (m s ⁻¹)
mean	2.74	1.93	1.83
$\overline{ X - \bar{X} }_1$	2.21	1.49	1.47
variance	7.51	3.73	3.33
std	2.74	1.93	1.83
median	1.6	-0.6	2.91
skewness	-0.07	0.05	0.73
kurtosis	2.85	3.90	3.06

Table 1: Statistics for wind data shown in Figure 1.

but this clearly is zero, so I haven't achieved much with this calculation.

We might try computing the average of the absolute value of the deviation from the mean:

$$\overline{|X - \bar{X}|}_1 = \frac{1}{N} \sum_{i=1}^N |X_i - \bar{X}| \quad (4)$$

Here, I've used the subscript 1, because this is based on the L^1 -norm, which for a vector is just the sum of all of the absolute values (not divided by N .) This tells us something about typical variations about the mean, but absolute values are hard to work with analytically.

So usually we square the deviation from the mean instead. We'll define the *variance* as:

$$\text{var}(X) = \mu_2 = \sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 \quad (5)$$

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

$$\text{std}(X) = \sigma_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2} \quad (6)$$

These statistics for Figure 1 are summarized in Table 1. In Figure 1, the variances are 7.51 m s⁻¹, 3.73 m s⁻¹, and 3.33 m s⁻¹. Standard deviations are

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. We'll consider this further in lecture 4.

Higher Moments

The variance is the second moment of the data, but there is no reason to stop at the second moment. We can keep going. The third moment is:

$$\mu_3 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^3, \quad (7)$$

the fourth moment is:

$$\mu_4 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^4, \quad (8)$$

and we can keep going indefinitely, although the moments become less useful as we increase. (Some argue that even the third and fourth moments are too noisy to be useful when computed from real data.)

Usually we define these in terms of normalized quantities. Thus, instead of looking at the third moment, we examine *skewness*, which is $\mu_3/\mu_2^{3/2}$. It measures the lopsidedness of the histogram.

Instead of looking at the fourth moment directly, we look at *kurtosis*, which is μ_4/μ_2^2 . This measures the flatness of the histogram, and is often used to interpret the likelihood of having extreme outliers in the tails of the data distribution. But more about this when we talk about probability density functions.