13. Statistics of Measurements

We have finished with time series and we are going to talk in a very general way about some concepts underlying measurements and their statistics. This lecture is based almost entirely on material that you can find in Chapter 1 of Paul Wessel’s notes which you can find on the “class materials” page of the class web site.

Measurements

Types of processes

If we consider a process that we want to measure it might be
(a) Deterministic - it can, at least in principle, be predicted (sometimes after the fact) by physical laws.
(b) Chaotic or Random - it cannot be predicted
(c) Stochastic or Probabilistic - some combination of (a) and (b)

Measurement Errors

Most of the underlying processes that we seek to measure are deterministic, but the measurements have errors for two basic reasons
(i) The measurement apparatus introduces an error.
(ii) The measurements are corrupted by some other processes - “one scientist’s signal is another one’s noise”. As an example, we can predict ocean tides very well but tide gauge which measures water height are also affected by storm surges, variations in atmospheric pressure, and surface waves.

The errors from both (i) and (ii) will be a combination of
(1) systematic - they are offset from the true value in a predictable manner.
(2) stochastic - they are randomly distributed about the true value.

Systematic errors that arise from the measurement apparatus can often be minimized by careful calibration. For example, measurements from mass spectrometers and other geochemical instruments are notoriously prone to systematic errors and so it is important to calibrate the apparatus to known standards. Systematic errors that arise when the measurement process is corrupted by another deterministic process can often be minimized by incorporating corrections for this process into the analysis. For example, we can correct tide gauge data for variations in atmospheric pressure (such corrective measurements are actually built into modern tide gauges).

Stochastic errors cannot be eliminated by calibrations or sophisticated analysis, but they can be minimized if we can understand the statistical properties of errors and make multiple measurements.

It is also important to note that same source of errors can be considered stochastic or systematic depending on the time or space scale. Going back to our tidal example, if we are taking measurements with an old fashioned instrument that does not correct for the atmospheric
pressure. Over a 24-hour period of stable weather, then the atmospheric pressure is primarily a systematic error. If, we look at data over decadal time scales to detect changes in global sea-level or tectonic uplift, then the atmospheric pressure can be treated as a stochastic error that can be minimized (eliminated to all intents and purposes) by averaging.

**Precision and Accuracy**

The presence of both systematic and stochastic errors in measurements is reflected in our terminology. **Precision** refers to the repeatability of measurements - A precise measurement has small stochastic errors but may have large systematic errors. **Accuracy** refers to the how close measurements are to the true value - An accurate measurement has both small systematic and stochastic errors. These two terms are often misused and confused with each other.

**Plotting Data**

The first step after one collects data is to plot it in various ways to understand how it varies with independent variables such as time, space, and any other parameter that is varied during the measurements. This may sound obvious but it is very common mistake to jump into the analysis of data without understanding its characteristics. I can give a few examples from personal experience but I will not write them down.

**Data Statistics**

A (hypothetical) data set comprising all possible observations of a phenomenon is called a population and it may be finite or infinite. A subset of a population is known as a sample.

**Measurements of Central Location**

The arithmetic mean, \( \bar{x} \) (commonly referred to as the ‘average’) of sample of \( n \) observations \( x_i \) is defined

\[
\bar{x} = \frac{1}{n} \sum x_i
\]

(13-1)

\( \bar{x} \) is termed the sample mean to distinguish it for the true mean of the population which is

\[
\mu = \frac{1}{N} \sum x_i
\]

(13-2)

where \( N \) is the number of observations in the population (if it is infinity then the equation would be evaluated in the limit as \( N \) tends to infinity).

The median, \( M \) is the middle value of a sample and for a set of measurements that are in ascending or descending order is simply

\[
M = \text{value of the middle observation}
\]

(13-3)
The mode, the observation that occurs the most frequently. If the data values are not discrete or are discrete with many possible values, we can also define the mode as the mid-value of the range of values that occurs most frequently (i.e., the highest bar on a histogram plot).

**Measurements of spread or variation**

We can define the deviation from the mean of an observation

\[ (13-4) \]

The mean deviation is by definition zero but one useful measure of sample spread is the mean absolute deviation, which is defined

\[ (13-5) \]

Because this quantity is not analytic, it is much more common to consider the mean squared deviation or variance. The population variance is

\[ (13-6) \]

where \( \sigma \) is the standard deviation.

When we are dealing with samples rather than populations the sample standard deviation, \( s \) and the sample variance \( s^2 \) are given by

\[ (13-7) \]

Normally we will not know the population mean, \( \mu \) so will have to use the right hand equality to calculate \( s^2 \). Here, the \( n – 1 \) term in the right hand expression reflects the fact that \( \bar{x} \) is the estimate of the mean determined from the sample and will be biased from the true mean in such a way as to bias the estimate of the sample standard deviation to zero. If we consider each term then \( 1/n \) of \( x \) is just \( x_i \) (see equation (1)) which by definition is zero when it is subtracted from itself so the values of \( x \) will on average be less than \( \bar{x} \). Another way to think about it is to recognize that when \( n = 1 \) but that will not generally be zero.

The sample variance can be calculated without first calculating the sample mean by expanding equation (13-7) and substituting for using equation (13-1)
This expression is useful since it allows one to update the sample variance as additional data is collected without reprocessing all the data.

**Robust Estimation**

In the geosciences, we will commonly deal with data sets that are contaminated by glitches. These are often related to human errors. For example, before the advent of digital data and centralized data processing, arrival times for teleseismic earthquakes were occasionally reported a minute too early or late. Data may also be contaminated if the measurement instrumentation is faulty or is affected by some unwanted process. For example, a towed magnetometer behind a ship may temporarily go haywire as it passes a large metallic buoy.

The first step in robust estimation is to plot the data. If you do not look at the data, you may never know it is contaminated. If you do look at the data, you will often be able to figure out why the data is corrupted and not only remove the bad data but eliminate its source.

Robust statistics provides an objective and potentially automated means to identify and remove contaminated samples. We can introduce the concept of the ‘breakdown point’ which is a measure of the fraction of contaminated data that is required to throw a statistical measure outside reasonable bounds. The breakdown point of the sample mean (equation 13-1) is $1/n$ because one egregious value can affect it dramatically. Conversely, the breakdown point of the median (equation 3) is $\frac{1}{2}$ because it is only dependent on central value and one would need $n/2$ bad measurements to set it to an unreasonable value. The median is thus a much more robust measure of the central location. The breakdown point of the mode will generally be somewhere between that of the mean and the median.

Now clearly the sample standard deviation (equation 13-7) also has a breakdown point of $1/n$. However if we define the absolute deviation from the sample median as

$$ |\Delta x_i| $$

The median absolute deviation is given by

$$ |\Delta x_i| $$

where the values of $|\Delta x_i|$ are in increasing or decreasing order. This quantity again has a breakdown point of $\frac{1}{2}$.

One very simple but effective way of removing bad samples is to normalize all the absolute deviations to the median absolute deviation

$$ |\Delta x_i| $$

Bad samples can then be identified and eliminated by searching for
If we define $x$, this is equivalent to 3 standard deviations for a normal distribution (I suspect you have all come across the term normal distribution but we will describe what this means in detail in lecture 15).

**Standard Deviation of estimates of the mean and standard deviation**

Our estimate of the sample mean $\bar{x}$ (equation 13-1) will typically differ from the population mean $\mu$ (equation 13-2). Similarly the sample standard deviation $s$ (equation 13-6) will differ from the population standard deviation $\sigma$ (equation 13-7). It turns out that if we make multiple estimates of these quantities using samples of $n$ observations, they are both normally distributed with standard deviations given by

$$s_\bar{x} = \frac{\sigma}{\sqrt{n}} \tag{13-12}$$

and

$$s_s^2 = \frac{\sigma^2}{n - 1} \tag{13-13}$$

The quantity $s_\bar{x}$ is termed the standard error. Both these expressions assume the population $N$ is infinite (or very large). Equation (13-12) is important because it tells us that the quality of the estimate of the mean derived from equation (13-1) improves as the square root of the number of observations.

**Moments**

In general the $r$’th moment is defined as

$$m_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^r \tag{13-14}$$

The second moment is the variance and the 3$\text{rd}$ and 4$\text{th}$ moments are termed the skewness and kurtosis. The skewness measures asymmetry of the sample distribution while the kurtosis measures its peakedness (a distribution with a high kurtosis has a sharper peak and fatter tails). In the geosciences, skewness and kurtosis are important in describing the distribution of sediment grain sizes.