

## Mean, Standard Deviation, and Counting Statistics

Whenever you read about an experiment, or perform an experiment of your own, you will find that the data is not perfectly reproducible. There is a little bit of slop that is the result of **systematic** and **random errors**, and some statistical manipulation needs to be done to account for those problems.

When an experiment is performed, a true value is needed. The sample **mean** is the best estimate of the “true” value of a dataset. The **mean** is simply an arithmetic average, or the sum of the observations divided by the number of observations, and is denoted as  $\bar{x}$ .

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N X_i$$

The sample **variance** ( $\sigma^2$ ) and **standard deviation** ( $\sigma$ ) describe the uncertainties associated with experimental attempts to determine “true” values. The variance is the sum of the square of the difference between the observed values and the mean divided by N-1. It is squared so that positive and negative values do not cancel each other out.

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{x})^2$$

The standard deviation is the square-root of the variance.

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{x})^2}$$

The final results can then be reported as:

$$\bar{x} \pm \sigma$$

where the standard deviation is the **absolute error**. The absolute error would have the same units as the mean. In order to compare errors from different sources or different experiments, we must consider the **relative error**, which is simply the standard deviation divided by the mean:

$$\bar{x} \pm (\sigma/\bar{x})$$

The relative error, therefore, is a unitless number.

The Empirical Rule states that if the data can be approximated by a normal curve, then roughly 68% of the observations are within 1 standard deviation of the mean and roughly 95% are within two standard deviations of the mean.

Sometimes you will run across **the standard deviation of the mean** of a number. This number is used when you are combining the data from a number of runs of the same sample.

The **Probability** of an event E, denoted by P(E), is the relative frequency of occurrence of E in a very long series of replications in a chance experiment. After many replications of the experiment, a **probability distribution** can be determined. This distribution shows the probability for any possible value of **x**. It is possible to describe the shape and location of the probability distribution using the **mean** ( $\bar{x}$ ) and the **standard deviation** ( $\sigma$ ). The values reported as  $1\sigma$  or  $2\sigma$  of the mean are this standard deviation.

There are three distributions that are the most important for scientific data analysis: the **Binomial distribution**, the **Poisson distribution**, and the **Gaussian distribution**. Of these, the **Gaussian (or Normal) distribution** is the most popular for the statistical analysis of scientific data because it seems to best describe the distributions of random observations.

The **Binomial distribution** is generally applied to experiments in which the result is one of a small number of possible final states (such as “heads” or “tails” in a coin toss). The **Poisson distribution** is generally appropriate for counting experiments where the data represent the number of items or events per unit interval. Like the Binomial distribution, this is a discrete distribution (defined only at integral values of variable **x**, although  $\bar{x}$  is a positive, real number). This is an appropriate analytical form for counting experiments because it is described in terms of the variable **x** and the parameter  $\bar{x}$ . For this distribution, the standard deviation is equal to the square root of  $\bar{x}$ .

$$\sigma = \sqrt{\bar{x}}$$

This is where the concept of **counting statistics** becomes important. In terms of counting statistics,  $\sigma$  is directly related to the uncertainty of measurement. Even if both systematic and random errors can be greatly reduced, the precision of the measurement may ultimately be determined by counting statistics, i.e. the number of counts.

$$c.s. = \frac{1}{\sqrt{\text{number-of-counts}}}$$

So, if an experiment on a mass spectrometer produces 10,000 counts, it is only possible to get a maximum of 1% precision.

Since mass spectrometer data is frequently represented with  $2\sigma$  error, counting statistics are typically calculated as:

$$c.s. = \frac{2}{\sqrt{\text{number-of-counts}}}$$

Therefore, in order to obtain a maximum precision of 1%, you must manage to count 40,000 ions in the mass spectrometer.