Topic 3.Errors measurement

2. Statistical Representation of Experimental Data

Suppose a large box containing a population of thousands of similar-valued resistors. To get an idea of the value of the resistors, we might measure a certain number of samples from thousands. The result of measured resistor values form a data set, which can be used to imply something about the entire population of the bearings in the box, such as average value and variations.
The probability and statistics allows us to convert (or reduce) raw data (in)to results which provides information for a large class of engineering judgments.

A. Basic Definitions for Data Analysis using Statistics

Basic terms that define the groups that are subject of statistical analysis are as follows:

- **Population:** The entire collection of measurements, not all of which will be analyzed statistically.
- **Sample:** A subset of the population that is analyzed statistically. A sample consists of n measurements.
- **Statistic:** A numerical attribute of the sample (e.g., mean, median, standard deviation).

A population consists of series of measurements (or readings) of some variable x is available. Variable x can be anything that is measurable, such as a length, time, voltage, current, resistance, etc.
Asampling of the variable x under controlled, fixed operating conditions renders a finite number of data points. As a result, a sample of these measurements consists of some portion of the population that is to be analyzed statistically. The measurements are $x_1, x_2, x_3, ..., x_n$, where n is the number of measurements in the sample under consideration.
The following represent some of the statistics that can be calculated:

B.Statistics Definitions Associated with Random Error

Mean: the **sample mean** is simply **the arithmetic average**, as is commonly calculated

$$x_{avg} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where *i* is one of the *n* measurements of the sample. The sample mean, although it is the simplest statistic to calculate, is not always as useful as the sample *median*, which is discussed later since it provides only a limited information about sample.

Deviation: The **deviation** of a measurement is defined as **the difference between a particular measurement and the mean**, i.e., for measurement i,

$$d = x_i - \bar{x}$$

When considering a group or sample of measurements, the deviation of one particular measurement is the same as the precision error or random error of that measurement.
Deviation is not the same as accuracy error. Recall that **accuracy error** (**inaccuracy**) is defined as the difference between a particular measurement and the true value of the quantity being measured:

(accuracy error = $x_i - x_{true}$)

Because of bias (systematic) error, x_{true} is often not even known, and the mean is not equal to x_{true} if there are bias errors.
To get some feel for how much deviation is represented in the sample, we might first think of averaging all the deviations to obtain some kind of mean or average deviation. It turns out that the average of all the deviations is zero! Why? Because by definition, some of the measurements are smaller than the average, and some are larger, and the average deviation turns out to be a meaningless and it is always zero.

Topic 3.Errors measurement

Average absolute deviation: A better measure of deviation is the ***average absolute deviation*** (also called the ***average positive error***), defined as ***the average of the absolute value of each deviation***.

$$\left|\bar{d}\right| = \frac{\sum_{i=1}^{n} |d_i|}{n}$$ where $|d_i|$ is called the **absolute deviation** or the **positive error**.

Sample standard deviation: More accepted measure of how much deviation or scatter occurs in the data is obtained by calculating the ***sample standard deviation***. For *n* measurements,

$$S = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

S is kind of like an average of the deviations, but it is constructed by taking the square root of the average of the *squared* deviations, since d_i can be either positive or negative.

Sample variance: The **sample variance** of the sample is simply **the square of the sample standard deviation:**

$$S^2 = \frac{\sum_{i=1}^{n} d_i^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Relative standard deviation: The relative standard deviation of the sample is simply the sample standard deviation divided by the mean.

$$RSD = \frac{S}{\bar{x}}$$

RSD is nondimensional and it is usually written as a percentage (multiply RSD by 100%); it is then sometimes called %RSD.

Standard error: The standard error is the standard deviation divided by the square root of the number of measurements.

$$\text{standarderror} = \frac{S}{\sqrt{n}}$$

Median: The median of the sample is defined as the value at which half of the measurements are lower and half are higher. A simple way to calculate median is to order all the measurements from lowest to highest. If n is odd, the number in the middle is the median. If n is even, the median is the average of the middle two values.

The median is often a better representative of a typical member of a group. If you take all of the values in a list, and arrange them in increasing order the number at the center will be the median. The median is an actual value belonging to some member of the group depending on the distribution of values, the mean may not be particularly close to the value of any member of the group; and the mean is also subject to skew as few as one value significantly different from the rest of the group can dramatically change the mean. The median gives you a central member of the group without the skew factor introduced by outliers.

Mode: The mode of the sample is the most probable value of the n measurement, the one that occurs most frequently. If none of the measurements are repeated, the mode is undefined.

Mode is not used as often as mean or median because it can be a misleading quantity, especially if the sample size is small and/or the distribution of measurements is not purely random. he mode is the most common member of the group. It doesn’t matter whether it’s the biggest or smallest value in the group whatever value is most common is the mode. The most is the least commonly used of these three average measures, and that’s because it’s generally the least meaningful. But once in a while, it’s useful. If your data is perfectly regular, then the mean, median, and mode will all be the same value. This almost never happens in real life. In general, you’ll find that the median is in between the mode and the mean, closer to the mean.

Example:

Given: Ten length measurements: 12.1, 12.3, 12.2, 12.2, 12.4, 12.3, 12.2, 12.4, 12.2, and 12.5 m.

Topic 3.Errors measurement

To do: Calculate the mean, variance, average absolute deviation, standard deviation, median, and mode.

C. Statistics Definitions Associated with Systematic Error

The following statistics can only be calculated if the true value is known for a sample.

Mean bias error: The ***mean bias error*** of a sample of *n* measurements is defined as,

$$MBE = \frac{\sum_{i=1}^{n} \frac{(x_i - x_{true})}{x_{true}}}{n}$$

MBE is a *nondimensional* measure of overall bias error or systematic error. *MBE* is usually written as a percentage error (multiply *MBE* by 100%).

Another way to calculate MBE is MBE=systematic error / true value.

Root mean square error: *The root mean square error of a sample of n measurements is defined as*

$$MBE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i - x_{true})}{x_{true}}\right)^2}$$

RMSE is a measure of average deviation, somewhat similar to standard deviation, but RMSE is concerned with deviations from the true value whereas S is concerned with deviations from the mean. RMSE is also nondimensional ands usually written as a percentage error (multiply RSME by 100%).