Basic Statistics

Author: John M. Cimbala, Penn State University

Latest revision: 26 August 2011

Introduction

- The purpose of this learning module is to introduce you to some of the fundamental definitions and techniques related to analyzing measurements with ***statistics***.
- In all the definitions and examples discussed here, we consider a collection (sample) of measurements of a steady parameter. E.g., repeated measurements of a temperature, distance, voltage, etc.

Basic Definitions for Data Analysis using Statistics

- First some definitions are necessary:
    - ***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).
- Suppose a *population* – a 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.
- Consider a *sample* of these measurements – 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:
- ***Mean*** – the ***sample mean*** is simply ***the arithmetic average***, as is commonly calculated, i.e., $\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$,

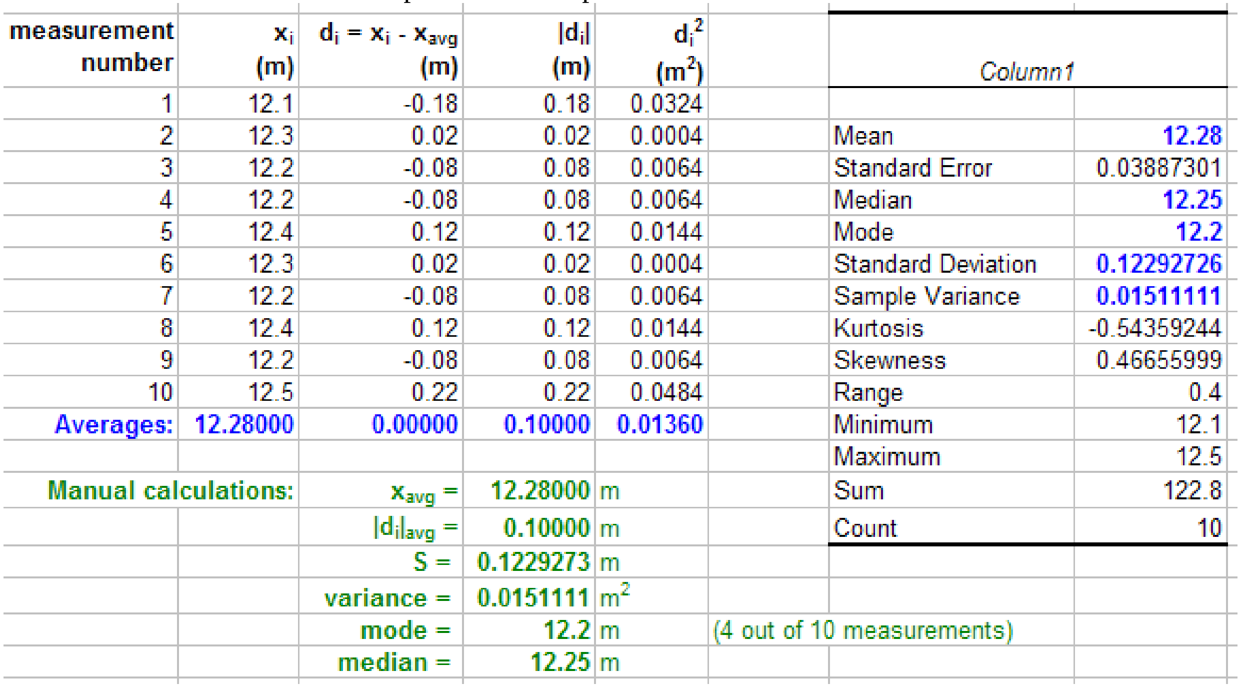
  where i is one of the n measurements of the sample.
    - We sometimes use the notation x_{avg} instead of $\overline{x}$ to indicate the average of all x values in the sample, especially when using Excel since overbars are difficult to add.
    - The sample mean, although it is the simplest statistic to calculate, is not always as useful as the sample *median*, which is discussed later.
- ***Deviation –*** the ***deviation*** of a measurement is defined as ***the difference between a particular measurement and the mean***, i.e., for measurement i, $\boxed{d_i \equiv x_i - \overline{x}}$.
    - 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 = $x_i - x_{\text{true}}$). 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.
- ***Average deviation*** – 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! Try it for any set of numbers, and you will convince yourself that this is true. 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 worthless calculation* – it is always zero.
- ***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***. Mathematically, $\boxed{\overline{|d|} = \frac{1}{n}\sum_{i=1}^{n} |d_i|}$, where $|d_i|$ is called the ***absolute deviation*** or the ***positive error***.
- ***Sample standard deviation*** – an even *better*, and more accepted measure of how much deviation or scatter is in the data is obtained by calculating the ***sample standard deviation***. For n measurements,

$$\boxed{S = \sqrt{\frac{\sum_{i=1}^{n} d_i^{\,2}}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1}}}.$$

    - 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.
    - Notice that the denominator is $n - 1$, not simply n. It turns out that for small sample size (small n), $n - 1$ yields a better estimate of the standard deviation than does n itself. (Details are beyond the scope of this course.) As n gets big, the difference between using n or $n - 1$ in the denominator becomes negligible.

- ***Sample variance*** – the ***sample variance*** of the sample is simply ***the square of the sample standard deviation***, namely, $\boxed{\text{sample variance} = S^2}$.
- ***Relative standard deviation*** – the ***relative standard deviation*** of the sample is simply ***the sample standard deviation divided by the mean***, namely, $\boxed{RSD = \frac{S}{\bar{x}}}$.
    - o *RSD* is *nondimensional*.
    - o *RSD* 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***, namely, $\boxed{\text{standard error} = S/\sqrt{n}}$.
- ***Median*** – the ***median*** of the sample is defined as ***the value at which half of the measurements are lower and half are higher***.
    - o 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.
    - o The median is sometimes more useful than the mean, particularly in cases where one or two values are significantly different than the rest of the values.
- ***Mode*** – the ***mode*** of the sample is ***the most probable value of the n measurement – the one that occurs most frequently***.
    - o 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.
    - o If none of the measurements are repeated, the mode is ***undefined***.

- **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.
  To do: Calculate the mean, variance, average absolute deviation, standard deviation, median, and mode.
  Solution:
    - o We use Excel for convenience. A portion of the spreadsheet is shown below:



measurement number	x_i (m)	$d_i = x_i - x_{avg}$ (m)	$\lvert d_i \rvert$ (m)	d_i^2 (m^2)		*Column1*	
1	12.1	-0.18	0.18	0.0324			
2	12.3	0.02	0.02	0.0004		Mean	12.28
3	12.2	-0.08	0.08	0.0064		Standard Error	0.03887301
4	12.2	-0.08	0.08	0.0064		Median	12.25
5	12.4	0.12	0.12	0.0144		Mode	12.2
6	12.3	0.02	0.02	0.0004		Standard Deviation	0.12292726
7	12.2	-0.08	0.08	0.0064		Sample Variance	0.01511111
8	12.4	0.12	0.12	0.0144		Kurtosis	-0.54359244
9	12.2	-0.08	0.08	0.0064		Skewness	0.46655999
10	12.5	0.22	0.22	0.0484		Range	0.4
Averages:	12.28000	0.00000	0.10000	0.01360		Minimum	12.1
						Maximum	12.5
Manual calculations:		x_{avg} =	12.28000	m		Sum	122.8
		$\lvert d_i \rvert_{avg}$ =	0.10000	m		Count	10
		S =	0.1229273	m			
		variance =	0.0151111	m^2			
		mode =	12.2	m	(4 out of 10 measurements)		
		median =	12.25	m			

    - o In Excel, the simplest way to obtain the statistics is to use the built-in statistical analysis macro:
        - ▪ **Office 2003**: Tools-Data Analysis-Descriptive Statistics-OK.
        - ▪ **Office 2007**: Click the *Data tab* instead of Tools – the rest is the same.
        - ▪ Click in the field called *Input Range*, and highlight all the measurements.
        - ▪ Select Output Range as the *Output Option*, click in the field for *Output Range*, and then click on a cell in some clean (available) portion of the spreadsheet.
        - ▪ Select Summary Statistics, and OK.

- In the Excel file that goes along with this learning module, we also calculate the statistics “by hand” (manually) using the definitions provided above as a check. To calculate the median, we first sort the data. The middle two values are 12.2 and 12.3, so the median is 12.25, the average of these two.

Comments:

- If “Data Analysis” is not available on your computer [you should have to do this *only the first time*]:
  - **Office 2003**: Tools-Add-Ins…-Analysis ToolPak-OK.
  - **Office 2007**: Click the *Office button*-Excel Options instead of Tools – the rest is the same.
- The mean, mode, and median are not identical. In a system in which all the deviations are purely random, and the number of samples is very *large*, the *mean, mode, and median become nearly identical*.
- The average absolute deviation is not the same as the standard deviation. As mentioned above, standard deviation is a more useful measure of the “average” amount of deviation in the sample.

Statistics Definitions Associated with Systematic Error

- ***Mean bias error*** – the ***mean bias error*** of a sample of *n* measurements is defined as $\boxed{MBE = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - x_{\text{true}}}{x_{\text{true}}}}$.

  Some comments about mean bias error:
  - *MBE* is a measure of overall bias error or systematic error.
  - *MBE* is *nondimensional*.
  - *MBE* cannot be calculated unless the true value is known.
  - *MBE* is usually written as a percentage error (multiply *MBE* by 100%).
  - Another way to calculate MBE is $\boxed{MBE = \text{systematic error / true value}}$, as was mentioned in a previous learning module.
- ***Root mean square error*** – the ***root mean square error*** of a sample of *n* measurements is defined as $\boxed{RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - x_{\text{true}}}{x_{\text{true}}}\right)^2}}$.

  Some comments about root mean square error:
  - *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 *nondimensional*.
  - *RMSE* cannot be calculated unless the true value is known.
  - *RMSE* is usually written as a percentage error (multiply *RSME* by 100%).