Errors and Calibration

Author: John M. Cimbala, Penn State University
Latest revision: 27 August 2009

Random vs. Systematic Errors

- There are two general categories of error: *systematic* (or *bias*) *errors* and *random* (or *precision*) *errors*.
- ***Systematic errors*** (also called ***bias errors***) are ***consistent, repeatable errors***. For example, suppose the first two millimeters of a ruler are broken off, and the user is not aware of it. Everything he or she measures will be too short by two millimeters – a systematic error.
- ***Random errors*** (also called ***precision errors***) are caused by ***a lack of repeatability in the output of the measuring system***. The most common sign of random errors is ***scatter*** in the measured data. For example, background electrical noise often results in small random errors in the measured output.

Systematic (or Bias) Errors

- **Systematic errors are consistent, repeatable errors** in a set of measurements
- The ***systematic error*** is defined as the ***average of measured values minus the true value***.
- Systematic errors arise for many reasons. Here are just a few:
  - ***calibration errors*** – perhaps due to nonlinearity or errors in the calibration method.
  - ***loading*** or ***intrusion errors*** – the sensor may actually *change* the very thing it is trying to measure.
  - ***spatial errors*** – these arise when a quantity varies in space, but a measurement is taken only at one location (e.g. temperature in a room - usually the top of a room is warmer than the bottom).
  - ***human errors*** – these can arise if a person consistently reads a scale on the low side, for example.
  - ***defective equipment errors*** – these arise if the instrument consistently reads too high or too low due to some internal problem or damage (such as our defective ruler example above).
- A *nondimensional* form of bias error is the ***mean bias error***, defined as ***MBE = systematic error / true value***.

Random (or Precision) Errors

- **Random errors are unrepeatable, inconsistent errors** in the measurements, resulting in *scatter* in the data.
- The ***random error*** of one data point is defined as the ***reading minus the average of readings***.
- **<u>Example</u>**:

  Given: Five temperature readings are taken of some snow: -1.30, -1.50, -1.20, -1.60, and -1.50^{o}C.

  To do: Find the maximum magnitude of random error in ^{o}C.

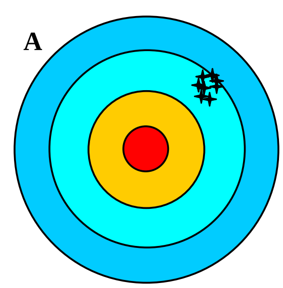
  Solution: The mean (average) of the five temperature readings is -1.42^{o}C. The largest deviation from this is the reading of -1.20^{o}C. The random (or precision) error for this data point is defined as the reading minus the average of readings, or -1.20 - (-1.42) = 0.22^{o}C. Thus, the maximum absolute value of random error is **0.22^{o}C**. You can verify that the magnitude of the random error for any of the other data points is less than this.

  Comment: Five readings are not sufficient to measure a good average reading, especially in this example since the individual readings vary so widely.

Accuracy

- Accuracy is the closeness of agreement between a measured value and the true value.
- ***Accuracy error*** is formally defined as the ***measured value minus the true value***.
- The accuracy error of a reading (which may also be called ***inaccuracy*** or ***uncertainty***) represents a combination of bias and precision errors.
- The ***overall accuracy error*** (or the ***overall inaccuracy***) of a set of readings is defined as the ***average of all readings minus the true value***. Thus, *overall accuracy error is identical to systematic or bias error*.

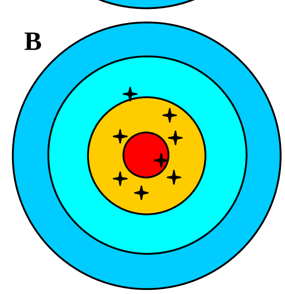
A



Precision

- Precision characterizes the *random error* of the instrument's output.
- ***Precision error*** (of one reading) is defined as the ***reading minus the average of readings***. Thus, *precision error is identical to random error*.
- The drawings to the right show the difference between precision and accuracy. Two people, A and B, shoot guns at targets. Both people shoot eight times. Each plus sign marks the spot where a bullet hits the target. It is proper to say:
  - A is more precise than B, but not as accurate.
  - On the other hand, B is more accurate than A, but not as precise.
- Instrument precision is often associated with instrument resolution, but these are *not* the same thing. An instrument can have great resolution, but poor precision.

B

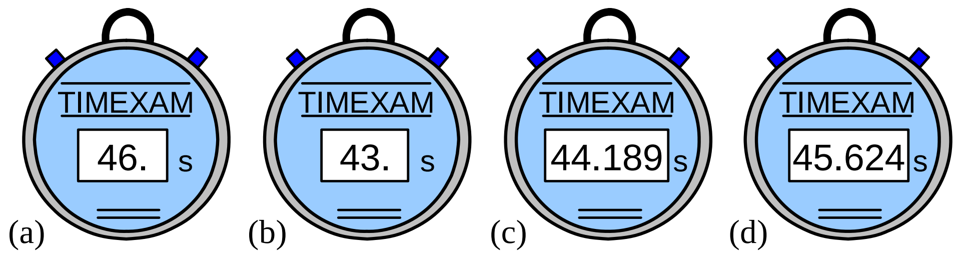


Resolution

- Resolution characterizes the ability of instrument's output or display to show changes in the quantity.
- ***Resolution*** is formally defined as ***the smallest change or increment in the measured quantity that the instrument can detect.*** For digital instruments, resolution is usually associated with the number of digits displayed on the output. E.g., a voltmeter with 5 digits has better resolution than one with 4 digits.
- Be careful – an instrument can be very precise, but not very accurate, and vice-versa. Furthermore, a high-resolution instrument may be neither precise nor accurate!

- **Example**:
  Given: Consider the same five temperature readings as in the previous example, i.e., -1.30, -1.50, -1.20, -1.60, and -1.50°C. Also suppose that the true temperature of the snow was -1.45°C.
  To do: Calculate the accuracy error of the third data point in °C. What is the overall accuracy error?
  Solution: The accuracy error (that is, the *inaccuracy*) of this data point is defined as the reading minus the true value, or -1.20 - (-1.45) = **0.25°C**. The overall accuracy error is the same as the systematic error or bias error, which is the average reading minus the true value, i.e., -1.42 - (-1.45) = **0.03°C**.

- **Example**: Four stopwatches are used to measure a time span. The exact, true, or actual time span is **45.623451 … s**. What can you say about the accuracy, precision, and resolution of each stopwatch?



  - Stopwatch (a) has poor resolution (only two digits and Δt = 1 s), but does the best job possible within its limits – it is accurate since its reading matches the true reading rounded off to two digits.
  - Stopwatch (b) is not very accurate since it displays an inaccurate reading compared to the true value, *and* it has poor resolution (only two digits of display and Δt = 1 s).
  - Stopwatch (c) has excellent resolution with 5 digits and Δt = 0.001 s, but the reading is very inaccurate.
  - Stopwatch (d) is both accurate (reading close to true value) *and* has excellent resolution.
  - *We cannot conclude anything about precision*, because we have only one reading. However, bear in mind that many people equate precision with resolution; but technically they are *not* the same thing.
  - If we repeat the measurement many times, we suspect that stopwatch (c) or (d) would be the most precise as well – they have the most *potential* to be precise – but we cannot tell from the given information.

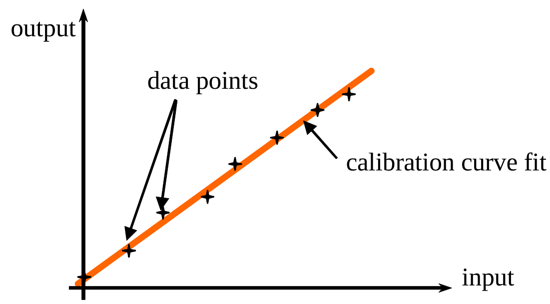
Other Errors

- There are many other errors, which all have technical names, as defined here:
  - ***zero error***: The instrument does not read zero when the input is zero. Zero error is a type of bias error that offsets all measurements taken by the instrument, but can usually be corrected by some kind of zero offset adjustment. ***Zero balance*** is a term used by manufacturers to indicate the maximum expected zero error of their instrument.
  - ***linearity error***: The output deviates from the calibrated linear relationship between the input and the output (see further discussion in the next section on calibration). Linearity error is a type of bias error, but unlike zero error, the degree of error varies with the magnitude of the reading.
  - ***sensitivity error***: The slope of the output vs. input curve is not calibrated exactly in the first place. Since this affects all readings by the instrument, this is a type of systematic or bias error.
  - ***resolution error***: The output precision is limited to discrete steps (e.g., if one reads to the nearest millimeter on a ruler, the resolution error is around +/- 1 mm). Resolution error is a type of random or precision error.
  - ***hysteresis error***: The output is different, depending on whether the input is increasing or decreasing at the time of measurement. [This is a separate error from instrument repeatability error.] For example, a motor-driven traverse may fall short of its reading due to friction, and the effect would be of opposite sign when the traverse arrives at the same point from the opposite direction; thus, hysteresis error is a systematic error, not a random error.

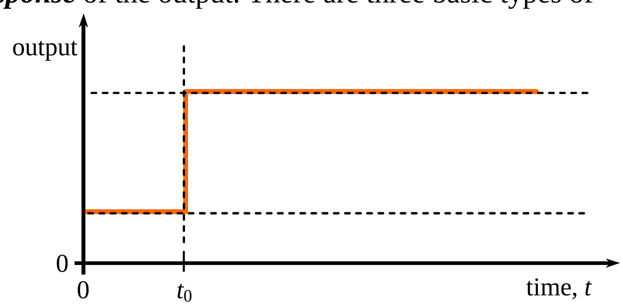
- o ***instrument repeatability error***: The instrument gives a different output, when the input returns to the same value, and the procedure to get to that value is the same. The reasons for the differences are usually random, so instrument repeatability error is a type of random error.
- o ***drift error***: The output changes (drifts) from its correct value, even though the input remains constant. Drift error can often be seen in the zero reading, which may fluctuate randomly due to electrical noise and other random causes, or it can drift higher or lower (***zero drift***) due to nonrandom causes, such as a slow increase in air temperature in the room. Thus, drift error can be either random or systematic.

Calibration

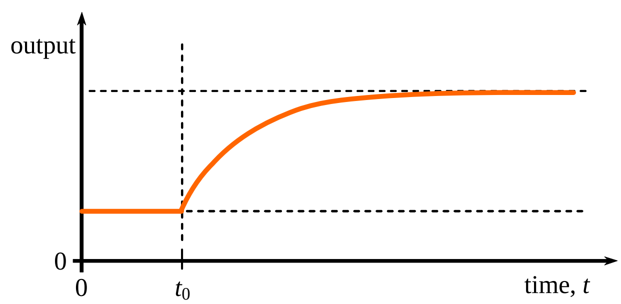
- There are two types of calibrations: static calibration and dynamic calibration.
- ***Static calibration*** is performed when ***time is not relevant in the measurement.***



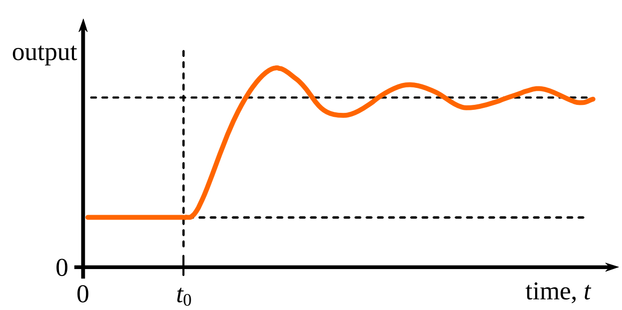
  - o Normally, some output (a voltage, current, etc.) is plotted as it varies with some known reference input, as sketched.
  - o Here, several data points are taken at known input values, and a calibration curve fit is drawn through the points. The curve fit can be any function (linear, parabolic, exponential, etc.), but most instruments are designed to have a *linear* behavior. In the example here, a straight line has been fitted through the data.
- ***Dynamic calibration*** is required when ***time is relevant to the measurement.***
  - o Normally, the time response of the system or instrument can be found by suddenly increasing or decreasing the input, and then recording the ***time response*** of the output. There are three basic types of time response:



  - o ***Zero order*** (ideal) – In this case, the output increases instantaneously with the input. Here, the input is suddenly increased at time t_0.
  - o No real system has a perfect, ideal response like this, but this is the goal towards which designers of instruments shoot.



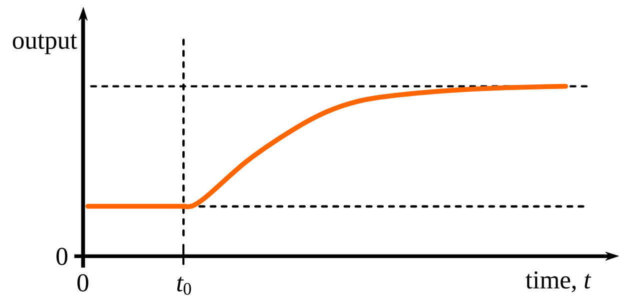
  - o ***First order* –** For the same instantaneous increase of the input at time t_0, a first-order system rises smoothly to its final value, as sketched, with no overshoot.
  - o As can be seen, there is a time lag in the instrument – some time elapses before the output shows its final steady-state value. A good example is sticking a thermometer into a pot of boiling water. It takes some time for the thermometer to read the correct temperature of the water, and the thermometer behaves as a first-order measurement system.
  - o ***Second order* –** For the same instantaneous increase of the input at time t_0, a second-order system also rises to its final value after some time lag, but there may or may not be overshoot, depending on how much ***damping*** is available in the measurement system. There are generally three possibilities:



    - ▪ ***Underdamped* –** With *too little* damping, the signal rapidly overshoots, and then oscillates many times before settling to the final reading.

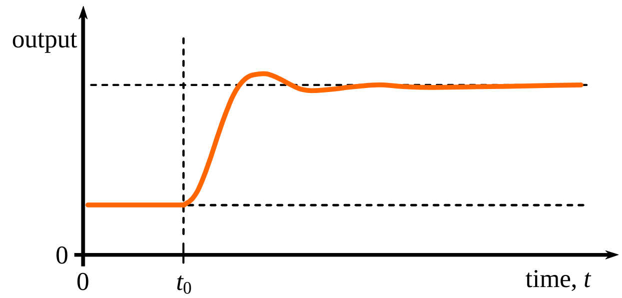
      A good example is seen when a person jumps onto a common bathroom scale. The scale behaves as a second-order measurement system, oscillating several times before settling down.

- ***Overdamped*** **–** With *too much* damping, the signal does not overshoot at all, but in general takes much more time to settle to the final reading.

  From these sketches one is able to distinguish between a first-order system and a second-order system with overdamping. The main difference is that at time t_0, the slope of the second-order curve is zero, while that of the first-order curve suddenly changes from zero to some finite value. The result is that the second-order overdamped system curve is *smoother* than the first-order curve around time t_0. [This can be seen in the sketches.]

- ***Optimally damped*** **–** With *just the right amount* of damping, the signal overshoots a little (to about 5% overshoot), and then quickly settles to the final reading with very small amplitude oscillations.

  The optimally damped system settles down the quickest of the three. For best time response, many measurement systems can be adjusted so as to behave with optimal damping. We discuss dynamic systems such as these in more detail later in the course.