COMP 333 — Week 4 Numerical Data

Numerical Data

For data analysis in science and engineering the use of numerical measurements is central.

Even in other domains, there is widespread use of numbers,
counts of items or occurrences, and
measurements of length, weight, volume, frequency, etc.

Engineers and scientists use scientific notation for numbers.
but all use of numbers separate the "number" from the "unit of measurement".

We will look at some aspects of numerical data that are important:

- accuracy
- precision
- significant digits
- normalization

Note that when we get to Machine Learning, the terms *accuracy* and *precision*

will have a different meaning that the meaning here for numerical measurements.

Measurements

Measurements form a major part of the data in data analysis.

Each measurement comes from an instrument.

The instrument may be

- the human eye, with the brain estimating size or distance or colour
- a rule marked each cm and each quarter of a cm
- a micrometer capable of measuring to mm precision
- the CCD of the Hubble telescope recording light intensity

The proper use of the numbers in data analysis

is important to making valid conclusions from the analysis.

Accuracy and Precision

Accuracy describes the difference between the measurement and the part's actual value.

For example, a weight measurement of 3.2 kg for a given part of actual weight 10 kg,

then your measurement is not accurate.

Precision describes the variation you see when you measure the same part repeatedly with the same device

Using the example above, if you weigh a given part five times, and get 3.2 kg each time,

then your measurement is very precise.

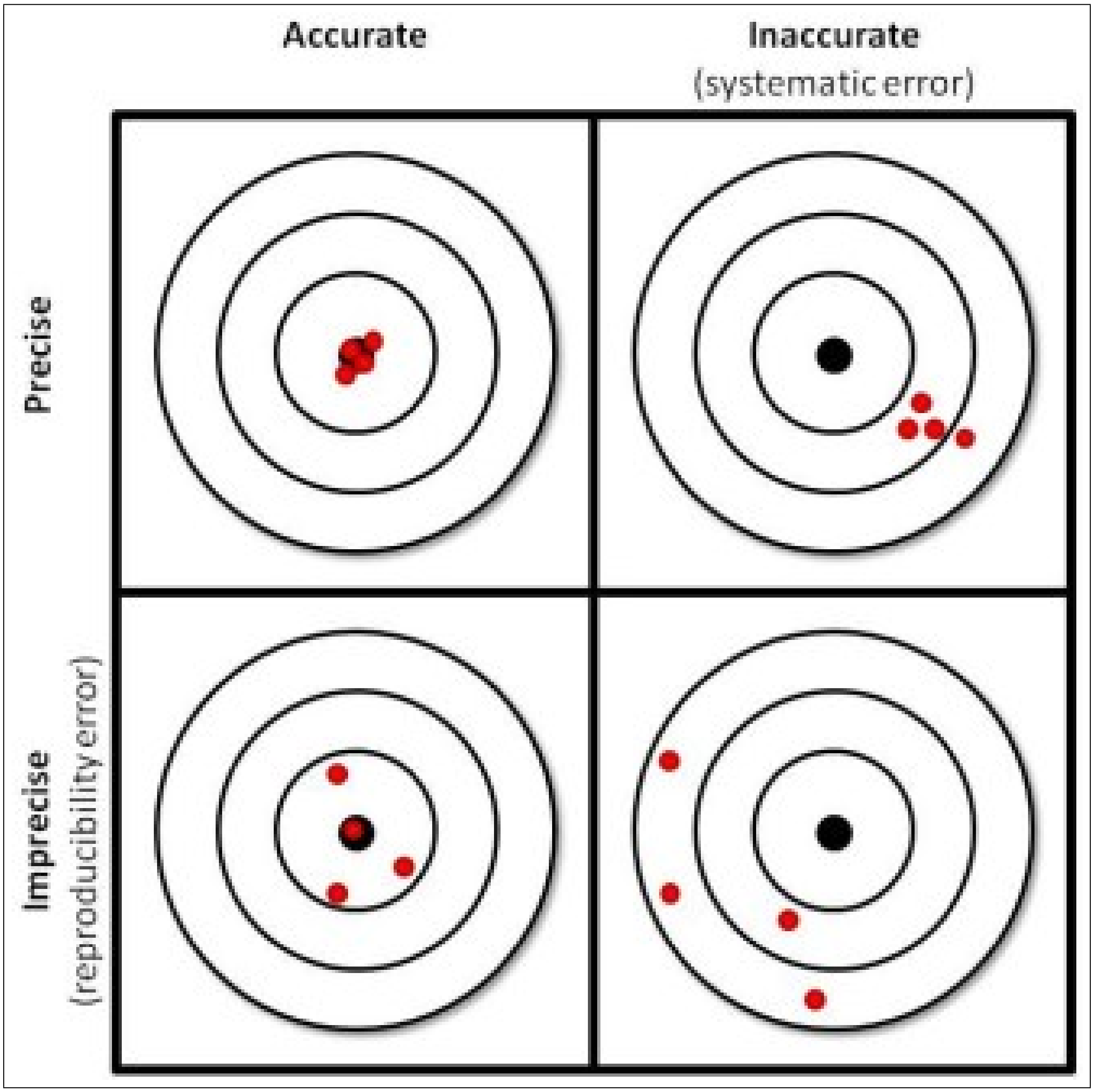
Precision is independent of accuracy.

You can be very precise but inaccurate, as described above.

You can also be accurate but imprecise.

This is best illustrated by a picture,

where the centre of the target is the actual value:



http://www.geographer-miller.com/accuracy-vs-precision/

Significant Digits

This material comes from `https://www.physics.uoguelph.ca/tutorials/sig_fig/SIG_dig.htm`

A common problem with numerical data is that computers
show more digits in a number than are meaningful,
especially in decimal component.

A number should only use those digits which are supported by the system of measurement,
that is, the instrument used for the measurement
and the resolution of that instrument to distinguish between close measurements.

The digits used should be *significant*
that is, meaningful,
and if you show a digit then you are claiming
that the digits is significant.

Some examples:

0.046 has two significant digits
4009 kg has four significant digits
7.90 has three significant digits
8200 has 2, 3, or 4 significant digits (***unclear***)
8.200 $\times 10^3$ has four significant digits
8.20 $\times 10^3$ has three significant digits
8.2 $\times 10^3$ has two significant digits

There are three aspects of your computation where you need to consider the significant digits:

- Need to know the significant digits for input data
- Need to keep track of significant digits during arithmetic
- Be careful formatting numbers in the output

Decimal Points Consider the above example 8200 that we said was unclear.

The use of the decimal point is important for indicating whether or not zeros
are significant, or not.

8200. means that zeros are significant, so 4 significant digits
8200 means that zeros are not significant, so 2 significant digits

Calculating Number of Significant Digits Basically,

the number of significant digits in the result of a calculation
is never more than
the smallest number of significant digits amongst the inputs

See `https://www.saddleback.edu/faculty/jzoval/worksheets_tutorials/ch1worksheets/sig_figs_in_calc_rules_7_1_09.pdf`

Normalization

Normalization is about bringing data into a common form
so that values can be compared.

It requires data to be in the same units
for example, grams, kilograms, or pounds

You cannot be directly compared: 5.3 kg is not the same as 5.3 g
even if 5.3 equals 5.3

In data wrangling, normalization is an important step.
Not only for numerical data,
but also for dates, time, currency, and names.

In data wrangling, the term *unification* is often used.
For example, date unification to unify the formats
05-11-2020, 11-05-2020, 2020-05-11 or 11-May-2020.

Unification of strings is often an important step
that simplifies comparison of strings:

- a string *" the Happiest day of My Life "*
- to all lower case
- and without leading or trailing blanks
- and only one blank between words
- *"the happiest day of my life"*

Normal Forms

A *normal form ...*

is a unique representation for an entity.

A normal form allows a simple test for equality:

A equals B if and only if normal form A is (literally) identical to normal form of B.

Normalization is the process that creates a normal form.