COMP 499 Introduction to Data Analytics

Lecture 2 — Numbers and Data

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Overview of Lecture

1. Measurement Scales
2. Normalization
3. Accuracy & Precision
4. Significant Digits
5. Data Formats
6. Data Schemas
7. Metadata
8. Self-Descriptive Data

Data Scales

Categorical

Nominal

Values have *names* as in enum or scalar type
equality testing allowed
mode is measure of central tendency

Ordinal

Ranked values, such as *good, better, best*
equality and comparison allowed
median is measure of central tendency
mean and deviation do not make sense

Continuous

Interval

Difference between values can be determined, eg integers
has no absolute zero
equality, comparison, $+$, $-$ allowed
mean is measure of central tendency; deviation makes sense

Ratio

Value is a ratio of continuous values, eg real number
has absolute zero
also $\times$, $/$ allowed
geometric mean is measure of central tendency

Data Scales

See video from UoVirginia: `https://www.youtube.com/watch?v=zHcQPKP6NpM`

Robust Statistics

median and Inter-Quartile Range (IQR) are robust to outliers

Outliers — John Tukey's Definition

Outlier is more than 1.5 times IQR from Q1 or Q3
Extreme value is more than 3.0 times IQR from Q1 or Q3

Plots — Categorical Data

Bar chart shows frequency, so shows modes (one or more)

Plots — Continuous Data

Histogram shows frequency, so shows modes (one or more)
Box plot shows median, Q1, Q3 box and whiskers to min and max
if outliers then shows fences at Q1-1.5IQR and Q3+1.5IQR
Both show central tendency, variability, and skewness; not modes

Contingency Tables and Scatter Plots

Normalization

A normal form ...

is a unique representation for an entity

Examples

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"

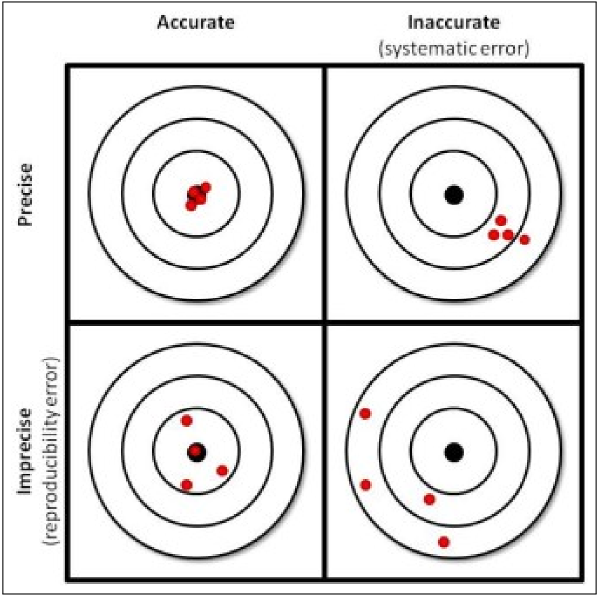
Normalization creates a normal form

allows simple test for equality

More Examples

Names
Dates
Currency
Metric vs Imperial measurements

Accuracy and Precision



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

Significant Digits

Problem

Showing more digits in a number than are meaningful

Especially in decimal component

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**)

8.200×10^3 has four significant digits

8.20×10^3 has three significant digits

8.2×10^3 has two significant digits

Problem

Need to know significant digits for input data

Need to keep track of sig. digits in arithmetic

Be careful formatting output

Reference

`https://www.physics.uoguelph.ca/tutorials/sig_fig/SIG_dig.htm`

Significant Digits

Decimal Point Convention

8200. means that zero's are significant, so 4 significant digits
8200 means that zero's are not significant, so 2 significant digits

Calculating Number of Significant Digits

Basically, never more than 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`

Data Formats

comma-separated values (csv)

Tab-separated values (tsv)

Attribute-Relation File Format (ARFF)

XML

RDF

Binary files (BLOBs)

HDF5 (Hierarchical Data Format version 5)

Data Formats — ARFF — Weka

ARFF files

ASCII files: Header followed by Data

Header

- the name of the relation,
- a list of the attributes (columns in data),
- their types

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Data Formats — ARFF

Data looks like

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Data Schemas

Tidy Data Schema in R

Tabular format with properties

1. Each variable is saved in its own column
2. Each observation is saved in its own row
3. Each type of observation is stored in its own (single) table

See video

`https://www.youtube.com/watch?v=1ELALQlO-yM&list=PL9HYL-VRXOoQOWAFoKHFQAsWAI3ImbNPk&index=2`

Metadata

Metadata

is data that provides information about other data

For example

Means of creation of the data
Purpose of the data
Time and date of creation
Creator or author of the data
Location on a computer network where the data was created
Standards used
File size
Data quality
Source of the data
Process used to create the data

Provenance of Data

is the origin and/or history of an object (that is, data, in our case).

Self-Descriptive Data

You can make sense of the file as a stand-alone.

therefore human-readable

ARFF
XML
HDF