COMP 333 — Week 3 Types of Data

Types of Data

It is important to understand the type of data that has been collected for each variable.

This affects the descriptive statistics that you can use for the data

when you want to present central tendency or variation.

And it affects which plots make sense for the data!

It also affects how you can compare one variable to another.

Categorical vs Continuous Data

The top-level distinction to be made is between *categorical* and *continuous* data.

Categorical variable *"can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property."*

Examples are race, sex, age group, and educational level.

Continuous variable is a numerical variable representing discrete numbers such as counts, or measurements on an infinite scale.

Examples are age, temperature, salary, and tip size.

Stevens Types of Data

In 1946 psychologist Stanley Smith Stevens developed the best-known classification of measurement with four scales of measurement:

- nominal
- ordinal
- interval
- ratio

See the wikipedia article: `https://en.wikipedia.org/wiki/Level_of_measurement`

As examples of these measurement scales

- nominal values represent discrete units.
  Values have *names* as in enum or scalar type
  Examples: hair colour, gender, race, religion
- ordinal values represent discrete units with a natural rank-order
  Values are ranked values, such as *good, better, best*
  Examples: grade letter, Likert scale, race finish position
- interval values represent ordered units with intermediate values,
  and the distance between units is the same
  Values allow the difference between values can be determined, eg integers
  but have no absolute zero
  Examples: celsius, fahrenheit, normalized scores
- ratio values are interval values that have an absolute zero.
  Value is a ratio of continuous values, eg real number
  Examples: Kelvin, weight

Relevant Arithmetical Operations

You need to be aware of how you can manipulate values of the four scales of measurement because you are going to be computing with them!

The wikipedia article discusses this.

To summarize:

- nominal: equality testing allowed
- ordinal: equality and comparison allowed
- interval: equality, comparison, $+$, $-$ allowed
- ratio: also $\times$, / allowed

Relevant Statistics

Prof Meyer discussed the applicability of measures of central tendency and variability.

To summarize:

- nominal: mode is measure of central tendency
- ordinal: median is measure of central tendency
  Note that mean and standard deviation do not make sense
- interval: mean is measure of central tendency; standard deviation makes sense
- ratio: geometric mean is measure of central tendency

Overview

From the wikipedia article we can summarize as a table.

Note that the complete list includes the values of previous levels.

This is inverted for the "Measure property".

Incremental progress	Measure property	Mathematical operators	Advanced operations	Central tendency
Nominal	Classification, membership	=, ≠	Grouping	Mode
Ordinal	Comparison, level	>, <	Sorting	Median
Interval	Difference, affinity	+, −	Yardstick	Mean, Deviation
Ratio	Magnitude, amount	×, /	Ratio	Geometric mean, Coefficient of variation

Relevant Plots

Prof Meyer discussed the applicability of different plots.

Plots — Categorical Data Bar charts are applicable.

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

Plots — Continuous Data Histograms, boxplots, and violin plots are applicable

Histogram shows frequency, so shows modes (one or more)

Box plot shows median, Q1, Q3 box and whiskers to min and max
and if outliers then shows fences at Q1-1.5IQR and Q3+1.5IQR

The violin plot combines boxplot and the frequency distribution like a histogram.

Both show central tendency, variability, and skewness.

Histogram and violin plot show modes.

Boxplot does not show modes.

Likert Scale

The Likert scale is the common five-value scale used as answers to questions on surveys.
See the examples in the figure below:

The Likert scale is an *ordinal* scale.
– The values are discrete units.
– The units have a natural rank-order.
– The units have no intermediate values,
– and are not the same distance apart,
so the Likert scale is **not** an interval scale.

So the computation of mean and standard deviation do not make sense for a Likert scale.

Note that **encoding** a Likert scale into five integers
such as (0, 1, 2, 3, 4) or (-2, -1, 0, 1, 2)
does **not** change the fact that it is an ordinal scale!

Beware because many people see the integers and automatically start calculating mean and standard deviation.
They are wrong to do so.