# Lecture 1 (CR.1)

Two Types of Statistics:

| Descriptive | Inferential |
|---|---|
| mean | To infer something |
| median | about a (population) |
| mode | from a / single sample. |
| range | |
| histogram | |
| scatterplot | |

A set (finite or infinite)
of objects to which we have
incomplete access.

population

Sample
(a subset to which
we have complete access)
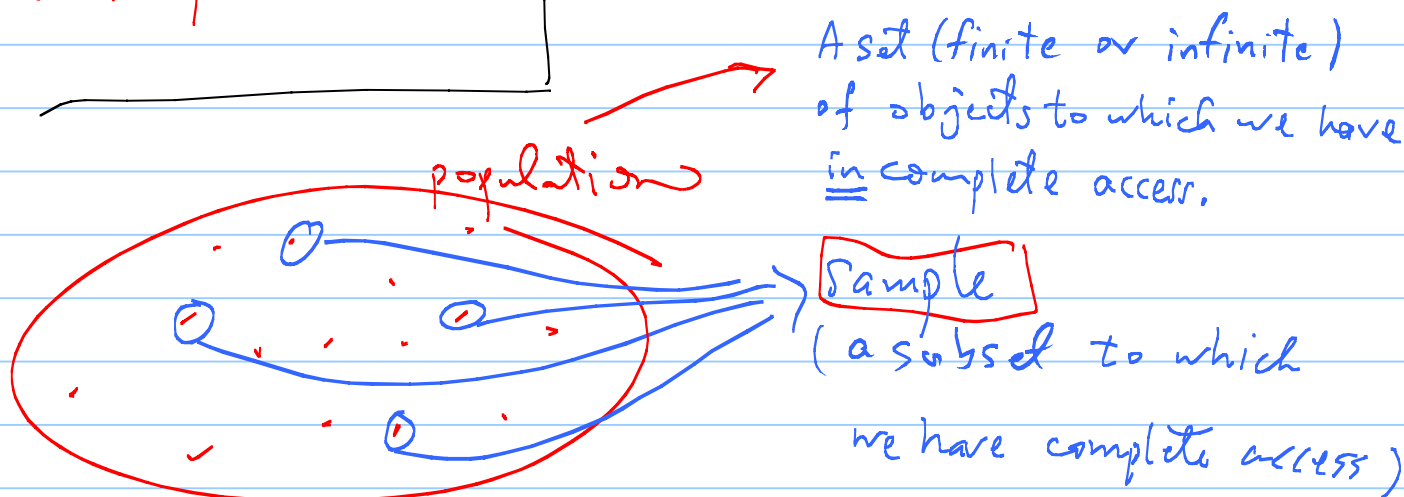
E.g. From a sample, we can compute sample mean,
sample range, sample --- .
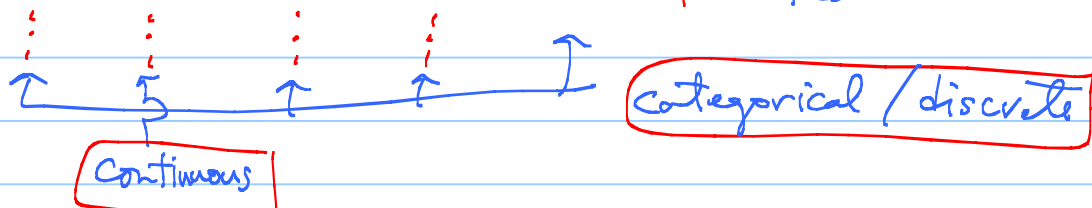
What do These say about The population mean?

# Data:

Imagine "data" as columns of things:

| Case | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|------|-------|-------|-------|-------|-------|
| 1 | 150 | 3.1415 | A | 0 | Apple |
| 2 | 153 | 2.7968 | C | 0 | Orange |
| 3 | 151 | --- | B | 1 | Banana |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |

Continuous

Categorical / discrete

There are other "finer" types of variables, too: Ordinal/Nominal/--- But we don't deal with them in this course.

Even though multiplying $x_2$ by 10,000 converts the numbers to integers (i.e. something that looks discrete), the decimal point suggests cont.

Categ./discrete variables can be <u>qualitative</u> (non-numeric), e.g. $x_3, x_5$ or <u>quantitative</u> (numerical), e.g. $x_1, x_4$.

The distinction is important from a practical point of view, because the techniques for handling them are different.

The diff. between cont. and discrete/categ is easy when the latter is qualitative, but it's very fuzzy when it's quantitative.

To further illustrate the fuzziness, consider a variable that represents time; e.g. time it takes to execute a code. On the one hand, we should treat it as a continuous variable, because we know that it is! On the other hand, if you have 100 cases in your data, and they are only 175 sec, 180 sec, and 182 sec, then you have to treat that variable as a single categorical variable with 3 levels (or categories).

$\boxed{\text{hw\_lect 1}}$

Construct a data set with the following specifications.
Any source is allowed: web, books, papers,
your own work, etc. Specifications:

1) number of cases: 30 or more

2) 2 categorical / discrete (qualitative or quantitative) vars.
   with no more than 6 levels for each variable.
   See part b) for a requirement on the hist.

3) 2 continuous / quantitative variables.

4) The 4 variables must relate to a common
   problem; not 4 unrelated variables.

a) print the data in the following format, and turn it in:

   Variable 1      Variable 2      Variable 3      Var. 4

30 or   ⎧   :              ′              ′              ′
more  ⎨   :              :              ′              ′
cases  ⎩   :              :              |              |
          ′              /              |              |

Keep a copy of the data set because you will
need it for other hw problems.

b) plot histograms for each of the 4 variables. (By hand)
   For the continuous vars. pick an appropriate # of bins.
   For the discrete vars. it is important for the hist
   to have at least 2 bars with more than 1 count.

For part b) You may want to wait until after the lect on histograms.