

MATH 203: PRINCIPLES OF STATISTICS 1

FALL 2024

DATA ANALYSIS ELEMENTS & TERMINOLOGY

We have identified seven elements of a statistical analysis:

1. **Objective:**

- what is our scientific objective in collecting and analyzing data ?
- what conclusions do we aim to draw ?

2. **Experimental units:**

- what are the 'objects' or 'things' on which we will be making measurements ?
- we identify the experimental units **before** we actually take any measurements;
- here 'make a measurement' could mean make a real physical measurement using a measurement device, but could also mean 'take a count' or 'record a value' by observation.

3. **Population:**

- the entire collection of experimental units;
- could be a finite collection, or could be large enough to be regarded as infinite.

4. **Characteristic(s) of interest:**

- what are we going to measure on each of the experimental units ?
- could be a single measurement, or be multiple measurements of different features.

5. **Sample:**

- the collection of **actual** experimental units on which we **do** make measurements;
- 'the sample' is also used to refer to the recorded measurements themselves;
- we are only able to take a **finite** sample;
- we will usually need to make some assumptions about **how** the sample is collected.

6. **Statistical inferences:**

- the conclusions we draw about the **population** from the **sample**;
- the term also captures the way we **formulate** the scientific question in mathematical terms (for example, the features of the population we are trying to learn about, the specific questions we are attempting to address).

7. **Measures of reliability:**

- given that we only have access to a **finite** sample of measurements, we also need to report the reliability of our statistical inferences for learning about the population;
- we want to capture the idea that inferences from a **large** sample should be more 'reliable' than those from a **small** sample.

Example	Element	
Bananas	1. Objective	To discover the average weight (in grammes) of bananas sold in Montreal supermarkets.
	2. Experimental unit	An individual banana.
	3. Population	The collection of all bananas that could be sold over a long time period in Montreal supermarkets; the size of this collection is large enough to be considered infinite.
	4. Characteristic	The weight of a banana.
	5. Sample	A collection of bananas bought (and their measured weights) in Montreal supermarkets in a given week. We assume that all the bananas bought are comparable, and representative of the population; there are no other sources of systematic variability (by region, importer etc.).
	6. Statistical inferences	Conclusions about the average weight: eg does the average weight exceed 100g ?
	7. Measure of reliability	Report a plausible range of variation of banana weights.
	Notes	In a more advanced analysis, we might be interested in comparing geographical regions, or comparing different time periods, etc; this can be thought of as an attempt to compare populations.
Gas prices	1. Objective	To discover the average price of a litre of gas on a given day at gas stations in an urban region.
	2. Experimental unit	A gas station in the region.
	3. Population	The collection of all gas stations in the region (assumed large enough to be considered infinite).
	4. Characteristic	The recorded gas price.
	5. Sample	A collection of observed gas prices. We might assume that the price is constant across a day, and the price at one gas station does not influence the price at another.
	6. Statistical inferences	Conclusions about the average price: eg is the average price equal to \$1.65 ?
	7. Measure of reliability	Report a plausible range of variation of per-litre gas prices.
	Notes	Our simplifying statistical assumptions may be too strong: gas prices undergo variation within a day, and also it seems that gas prices at nearby stations do influence one another.

Example	Element	
CGPA	1. Objective	To study the CGPA of students who enrol in a given MATH class.
	2. Experimental unit	An individual student in the class.
	3. Population	The finite collection of all 300 enrolled students.
	4. Characteristic	The CGPA at enrolment.
	5. Sample	A subset of the classlist selected by randomly choosing 50 ID numbers, and their CGPA at enrolment.
	6. Statistical inferences	Conclusions about the average CGPA: eg does the average CGPA exceed 3.0 ?
	7. Measures of reliability	Report a range of variation of CGPAs represented in the population.
	Notes	In this finite population setting, it is possible to compute the average CGPA precisely by recording the CGPA of all students in the population. However, if we only take a sample of a size (50) which is smaller than the population size (300), we still have incomplete information, and our goal is to learn about the population characteristics from the sample.
Foot traffic	1. Objective	To assess whether the number of people passing through the Roddick Gates between 8am and 9am is greater on a Tuesday than a Monday.
	2. Experimental unit	A single one hour period between 8am and 9am on a Monday or a Tuesday.
	3. Population	The collection of all one hour periods (8am–9am) on Monday and Tuesday over a large number of weeks.
	4. Characteristics	The count of the number of people passing through the Roddick Gates within each individual 8am–9am slot, and the day of the week (Mon or Tue).
	5. Sample	Observations of the characteristics made over a specific 15 week period; we assume that all Mondays and Tuesdays in this period are representative of the population.
	6. Statistical inferences	Conclusions about the average counts on the two days eg is the average count lower on a Monday than on a Tuesday ?
	7. Measures of reliability	Report a plausible range for the difference in average counts.
	Notes	The statistical assumptions, and data collection, inform about the sub-population corresponding to the specific time of year when the data were collected; it is plausible that Mondays and Tuesdays in September are not reflective of Mondays and Tuesdays in June.

Example	Element	
Drug efficacy	1. Objective	To assess whether a new drug reduces pain more than a generic drug.
	2. Experimental unit	An individual patient who is taking, or who has been assigned to take, one of the two drugs.
	3. Population	All possible patients who might take the drugs.
	4. Characteristics	The difference in reported pain score (measured on a standard scale) before and after taking standard doses of the two drugs, and the drug that was taken.
	5. Sample	32 individuals who have taken the new drug, and 27 who took the generic, and their reported differences in pain scores.
	6. Statistical inferences	Conclusions about the relative efficacy of two drugs: eg is pain relief provided by the new drug greater than that provided by the generic ?
	7. Measures of reliability	Report a range of variation for the differences in pain relief score provided by the two drugs.
	Notes	We need to consider carefully the manner in which the drugs were allocated to individuals; in an experimental study, the drugs are allocated randomly by the researcher to a group of comparable patients, irrespective of the patient's characteristics; in an observational study, the patients choose the drug themselves, perhaps in a way that is determined by their own preferences, characteristics etc.
Ice cream	1. Objective	To determine the most popular flavour(s) of ice cream.
	2. Experimental unit	A person who is willing to express their ice cream preference.
	3. Population	All people willing to express an ice cream preference (presumed an infinite population).
	4. Characteristic	The reported ice cream preference (either from a fixed list of flavour options, or a list that includes 'Other' or 'None of the above', or an open choice of flavours).
	5. Sample	The reported flavour preferences of 2517 people filling out an online survey.
	6. Statistical inferences	Report a ranked list of ice cream flavours.
	7. Measures of reliability	Report whether there are any flavours that are clearly preferred, or by how much one specific flavour is preferred to another.
	Notes	The sample might be considered as only representative of people who will fill in an online survey, and therefore may not be representative of the general population.