Mathematical Statistics

MAS 713

Chapter 1.1

This subchapter

1.1 Introduction

- 1.1.1 What is statistics ?
- 1.1.2 The statistical process
- 1.1.3 Population and samples
- 1.1.4 Random sampling

What is statistics ?

What is statistics ?

In order to learn about something, you must first collect observations, referred to as **data**

Definition

Statistics is the science of the

(i) collection,

(ii) processing,

(iii) analysis,

(iv) and interpretation of data.

What is statistics ?

In short, statistics is the art of learning from data

It allows to gain new insights into the behaviour of many phenomena

$\rightsquigarrow$ statistical concepts and methods are not only useful but indeed often indispensable in understanding the world around us

Further, it allows to turn observational evidence into information for decision making

What is statistics ?

Includes diversified tasks like for example

- calculating the average length of the downtimes of a computer
- checking whether the level of lead in the water supply is within safety standard
- collecting and presenting data on the number of persons attending seminars on solar energy

What is statistics ?

Statistics is mainly concerned with numbers and figures, and is therefore a branch of mathematics.
However, what makes statistics different is that it considers the presence of randomness, uncertainty and variation, which are everywhere in real life.

What is statistics ?

- If each computer had exactly the same length of downtime,
- If the level of lead was exactly identical everywhere and everytime in the water supply,
- If each seminar attracted the same number of people,

And in addition,

- if those values were known with absolute accuracy,

$\Longrightarrow$ then a single observation would reveal all desired information

$\Longrightarrow$ we would not need statistics.

What is statistics ?

Statistics

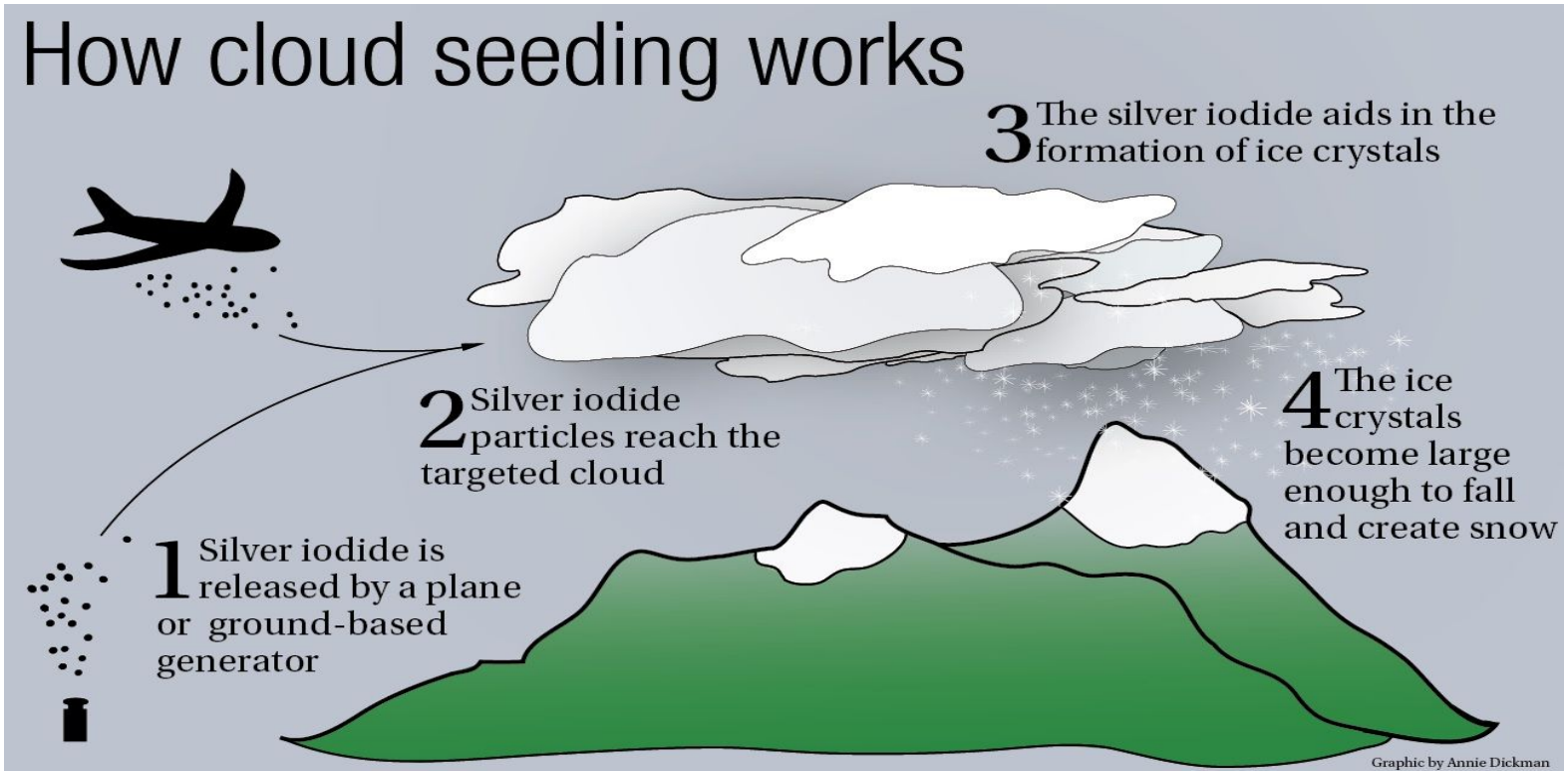
Statistics allows to describe, understand and control the variability insofar as possible and to **take this uncertainty into account** when making judgements and decisions

Example 1:
Does cloud seeding work ?

Example 1 : does cloud seeding work ?

Cloud seeding is the attempt to change the amount of precipitation that falls from clouds, by dispersing substances into the air that serve as cloud condensation

The usual intent is to increase precipitation (rain or snow)



Example 1 : does cloud seeding work ?

1. A natural question may be

   "does cloud seeding using a given substance (say, silver iodide) really work ?"

   $\rightsquigarrow$ research question

How can we answer this question ?

2. First, we should observe the amount of precipitation that falls from seeded clouds, as well as from unseeded clouds

   $\rightsquigarrow$ experiment, collection of data

Example 1 : does cloud seeding work ?

We run the following experiment :
we observe 52 clouds, 26 of which were chosen at random and seeded with silver iodide

Example 1 : does cloud seeding work ?

The following rainfall (in acre-feet) are recorded :

Unseeded Clouds
1202.6, 830.1, 372.4, 345.5, 321.2, 244.3, 163.0, 147.8, 95.0, 87.0, 81.2, 68.5, 47.3, 41.1, 36.6, 29.0, 28.6, 26.3, 26.1, 24.4, 21.7, 17.3, 11.5, 4.9, 4.9, 1.0.

Seeded Clouds
2745.6, 1697.8, 1656.0, 978.0, 703.4, 489.1, 430.0, 334.1, 302.8, 274.7, 274.7, 255.0, 242.5, 200.7, 198.6, 129.6, 119.0, 118.3, 115.3, 92.4, 40.6, 32.7, 31.4, 17.5, 7.7, 4.1.

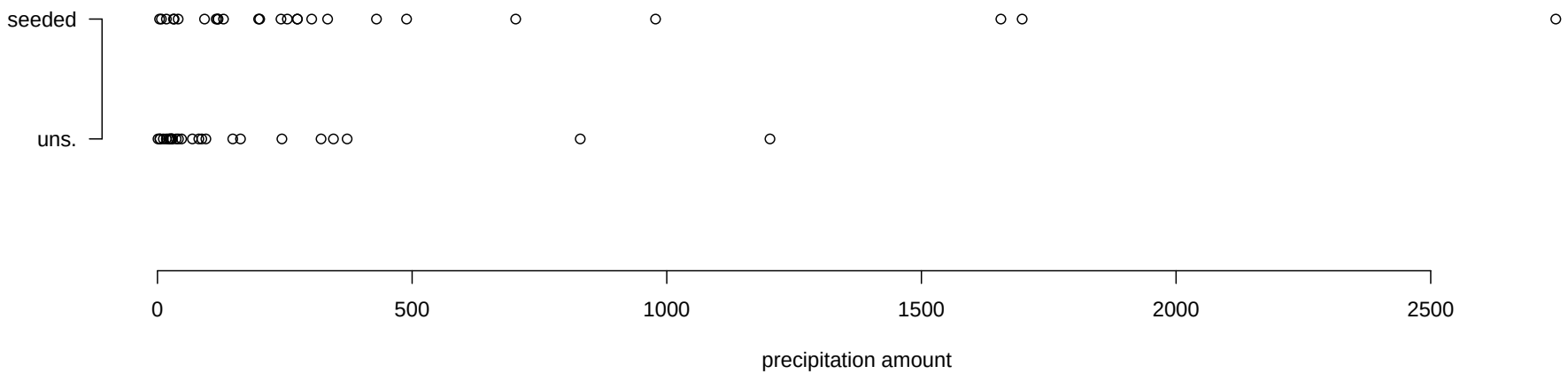
$\rightsquigarrow$ those figures are our data

Example 1 : does cloud seeding work ?

However, these long series of numbers do not really speak for themselves

We should present the data so that they are readily comprehensible

This includes graphical representations



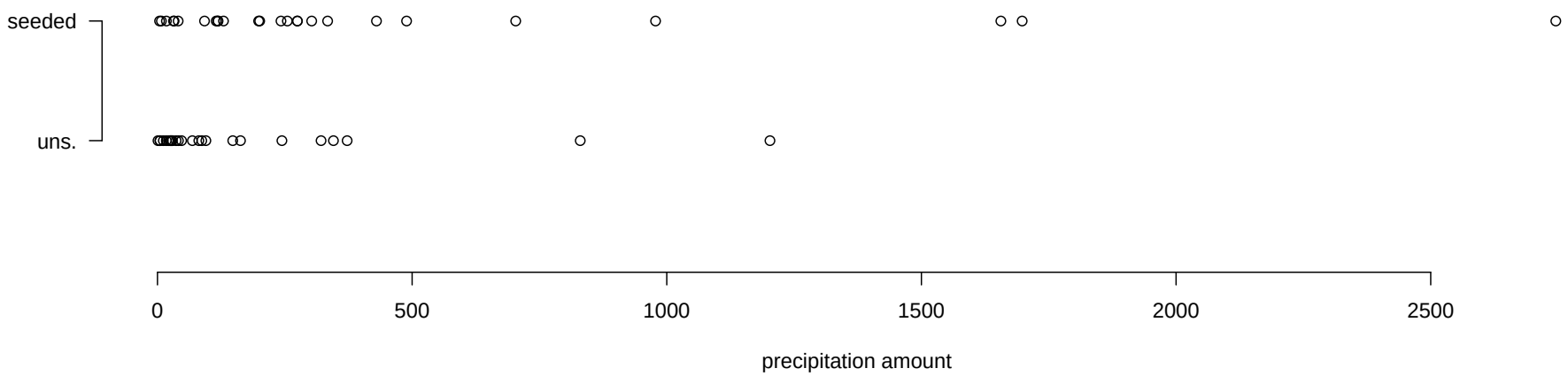
as well as numerical summary measures

average prec. seeded $= 441.98$ average prec. unseeded $= 164.58$

This part of statistics, concerned with the description and summarisation of data, is called **descriptive statistics**

Example 1 : does cloud seeding work ?

At first sight, seeded clouds seem to give more precipitation than unseeded clouds



average prec. seeded $= 441.98$ average prec. unseeded $= 164.58$

Careful! - We must take into account the possibility of chance :

- we only observed 52 clouds. If we had observed 52 (or more) other clouds, we would have observed different rainfall amounts
- due to chance only, the 26 clouds that we seeded might be the clouds that would have given more rainfall anyway
- due to chance only, the 26 clouds that we did not seed might be the clouds that would have given less rainfall anyway

$\rightsquigarrow$ can we really conclude that the observed higher amount of rainfall for seeded clouds is due to seeding? Or is it possible that the seeding was not responsible for that but rather that the higher rainfall amount was just a chance occurrence ?

$\rightsquigarrow$ can we really generalise what we are seeing on a particular data set beyond that data set ? How risky is it ?

We should analyse and interpret the data bearing in mind that the observed features may be consequences of chance only

Example 1 : does cloud seeding work ?

This part of statistics is called **inferential statistics** or **statistical inference**

One must decide how far to go in generalising from an observed data set, whether such generalisations are reasonable or justifiable, whether it might be reasonable to collect more data, etc.

Some of the most important problems in inference concern the appraisal of the risks and consequences of making wrong decisions

$\rightsquigarrow$ risks are often appraised by calculating **probabilities** of some events occurring

Example 1 : does cloud seeding work ?

Probability and statistics are closely related and each depends on the other in a number of different ways, so that they are traditionally studied together

$\rightsquigarrow$ we will have an insight into **probability theory** in the next chapter.

Example 1 : does cloud seeding work ?

5. Finally, we should draw conclusions from our investigations, that is, we should answer the question

   "does cloud seeding using silver iodide really work ?"

The statistical process

The statistical process

The **typical statistical procedure** :

1) set clearly defined goals for the investigation;
   formulate the research question
2) decide what data is required/appropriate and how to collect them;
   collect the data
3) display, describe and summarise the data in an efficient way;
   check for any unusual data features
4) choose and apply appropriate statistical methods to extract intelligent information from the data
5) interpret the information, draw conclusions and communicate the results to others

Fact

Every step in this process requires understanding statistical principles and concepts as well as knowledge and skills in statistical methods

The statistical process

Nowadays, the ideas of statistics are everywhere :

- descriptive statistics are featured in every newspaper and magazine
- statistical inference has become indispensable to public health and medical research, to marketing and quality control, to education, to accounting, to economics, to meteorological forecasting, to polling and surveys, to sports, to insurance, to gambling and obviously to engineering

$\rightsquigarrow$ **to all research that makes any claim to being scientific**

Statistics has indeed become ingrained in our intellectual heritage

Example 2:
Hair colour and pain tolerance

Example 2 : Hair colour and pain tolerance

An experiment conducted at the University of Melbourne suggests that there may be a difference in pain threshold for blonds and brunettes. So,

1) The research question is :

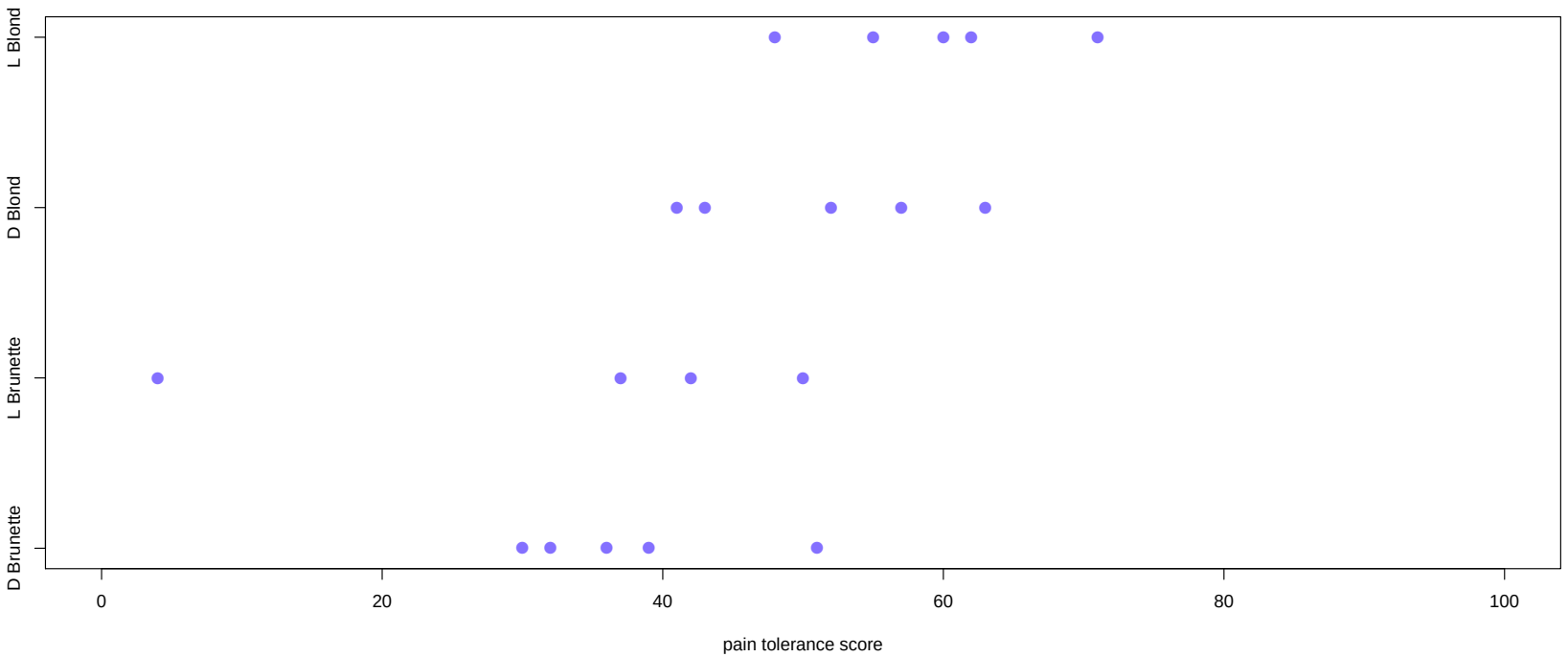
"is pain threshold related to hair colour?"

Example 2 : Hair colour and pain tolerance

2) A group of 19 subjects was divided into light blond, dark blond, light brunette and dark brunette groups and a pain threshold score was measured for each subject

- A higher score means a higher pain threshold

3) Below are the data (already visualized)



Example 2 : Hair colour and pain tolerance

4) Pain threshold seems to increase with lighter hair colour, but is this effect real or just due to chance ? (the number of observations is quite small : we have 4 or 5 observations per hair colour)

$\rightsquigarrow$ we have to apply some inferential method to come to a conclusion :

5) either

it is clear from the data that pain tolerance is related to hair colour

or

the data do not allow to conclude that pain tolerance is related to hair colour

depending on what you have observed

Remark : in the latter case we won't say "pain tolerance is not related to hair colour"; it might still be the case but with such a small number of observations (that is, with such a small amount of information in hand) we are not sure and it would be too risky to affirm it is

Populations and samples

Population

Most of the time, we are interested in obtaining information about a total collection of elements, which will refer to as the **population**

The elements are often called **individuals** (or units)

Population

Given the research question, we have observed some characteristic for each individual. This characteristic, which could be quantitative or qualitative, is called a **variable** (it varies from one individual to another)

Example 1

In Example 1 (clouds seeding), the population consists of all the clouds of the sky. An individual is a cloud and the variable of interest is the rainfall amount fallen from the cloud

Example 2

In Example 2 (pain tolerance), the population consists of all blonds and brunettes of the world. An individual is one of those girls and the variable of interest is the pain threshold score

Sample

It is easily understood from the previous two examples that it is often physically impossible or infeasible from a practical standpoint to obtain data on the whole population

Think also of very expensive, or very time-consuming, or destructive experiments

$\rightsquigarrow$ in most situations, we can only observe a subset of the population, that is, we must work with only partial information

The subset of the population which is effectively observed is called the **sample**

The **data** are the measurements that are actually collected over the sample in the course of the investigation

Note : sometimes, we may use “population” to designate the set of all potential measurements and “sample” to designate the subset of measurements actually observed (i.e., the data)

Sample

In Example 1, the sample consists of the 52 clouds whose rainfall amounts have been recorded

(We can also say that we have two samples : 26 seeded clouds and 26 unseeded clouds)

In Example 2, the sample consists of the 19 girls whose pain thereshold scores have been measured

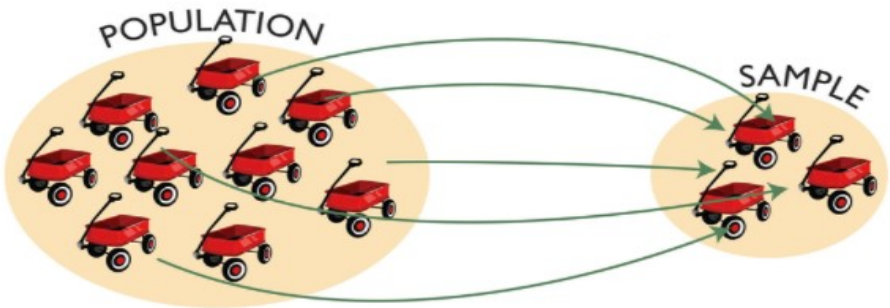
Sample

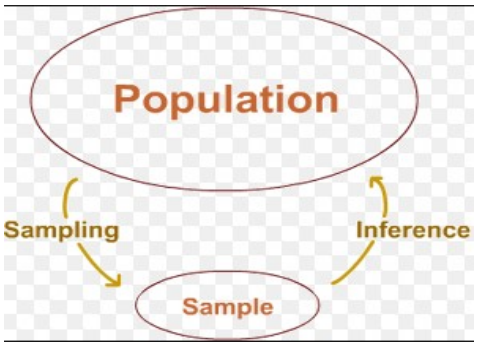
Definition

Population is the entire set of individuals or objects having some common characteristics selected for a research study.

Definition

Sample is a sub-set of the population.





Random sampling

Sampling

The process of selecting the sample is called the sampling

If the sample is to be informative about the total population, it must be representative of that population.

Sampling

Suppose you are interested in the average height of the NTU students.

Would it make sense to select the sample as the NTU basketball team?

Suppose you are interested in the average age of the NTU students, would it make sense to select a sample made up of postgraduate students only?

It is quite clear that the quality of the data is paramount in an efficient statistical study :

Your results are only as good as your data !

Also rendered by the acronym GIGO : Garbage In, Garbage Out

$\rightsquigarrow$ the sampling must be carefully done, impartially and objectively

Random sampling

In practice, the only sampling scheme that guarantees the sample to be representative of the population is the **random sampling**

$\rightsquigarrow$ the individuals of the sample are selected in a totally random fashion, without any other prior consideration

Indeed, any specific nonrandom rule for selecting a sample often results in one which is inherently biased toward some values as opposed to others

$\rightsquigarrow$ we need not attempt to deliberately choose the sample according to some criterion

Instead, we should just leave it up to “chance” to obtain a sample which correctly covers the underlying population

The importance of random sampling

Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population, taking the randomness into account (using probabilities)

$\rightsquigarrow$ not possible if the sample is not random !

Information drawn in nonrandom sample cannot, as a rule, be generalised to larger populations

The importance of random sampling

Fact

The statistical procedures presented in this course may not be valid when applied to nonrandom samples

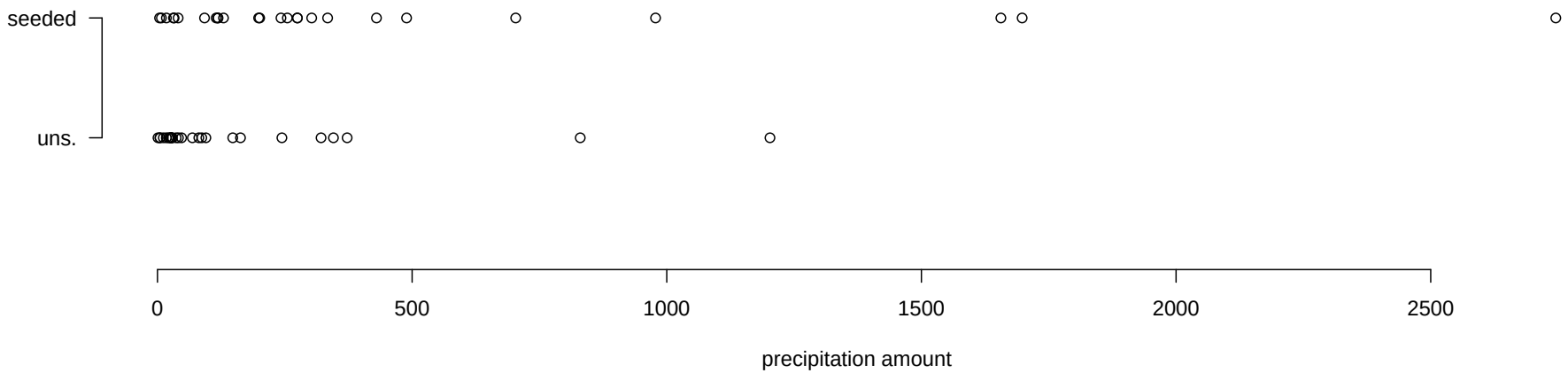
$\rightsquigarrow$ never unquestioningly accept samples without knowing how the data have been generated

The importance of random sampling

Let's revisit Example 1 : does cloud seeding work ?

What if a company wants to promote their product for cloud seeding.

They might use an *Post hoc alteration*, or *Cherry picking* to show that their product increase precipitation



The importance of random sampling

Some examples:

A 1996 study on the effects of nicotine on cognitive performance revealed that findings of nicotine or smoking improving performance were more likely to be published by scientists who acknowledged tobacco industry support.

source: Christina Turner; George J Spilich (1997). "Research into smoking or nicotine and human cognitive performance: does the source of funding make a difference?".

The importance of random sampling

Some examples:

A 2006 review of experimental studies examining the health effects of cell phone use found that studies funded exclusively by industry were least likely to report a statistically significant result.

source: Anke Huss; Matthias Egger; Kerstin Hug; Karin Huwiler-Muntener; Martin Roosli (2006-09-15). "Source of Funding and Results of Studies of Health Effects of Mobile Phone Use: Systematic Review of Experimental Studies"

The importance of random sampling

Some examples:

Two opposing commercial sponsors can be at odds with the published findings of research they sponsor.

A 2008 Duke University study on rats, funded by the Sugar Association, found adverse effects of consuming the artificial sweetener Splenda. The manufacturer, Johnson & Johnson subsidiary McNeil Nutritionals LLC, responded by sponsoring its own team of experts to refute the study.

source: Stephen Daniells (2009-09-25). "Splenda study: Industry and academia respond.

The importance of random sampling

Some examples:

A 2012 analysis of outcomes of studies pertaining to drugs and medical devices revealed that manufacturing company sponsorship "leads to more favorable results and conclusions than sponsorship by other sources."

source: Lundh, A; Sismondo, S; Lexchin, J; Busuioc, OA; Bero, L (Dec 12, 2012). "Industry sponsorship and research outcome.". The Cochrane database of systematic reviews 12.

Objectives

Now you should be able to :

- identify the role that statistics can play in problem-solving process
- discuss how variability affects the data collected and used for making decisions
- discuss how probability theory is used in science
- discuss the importance of random sampling