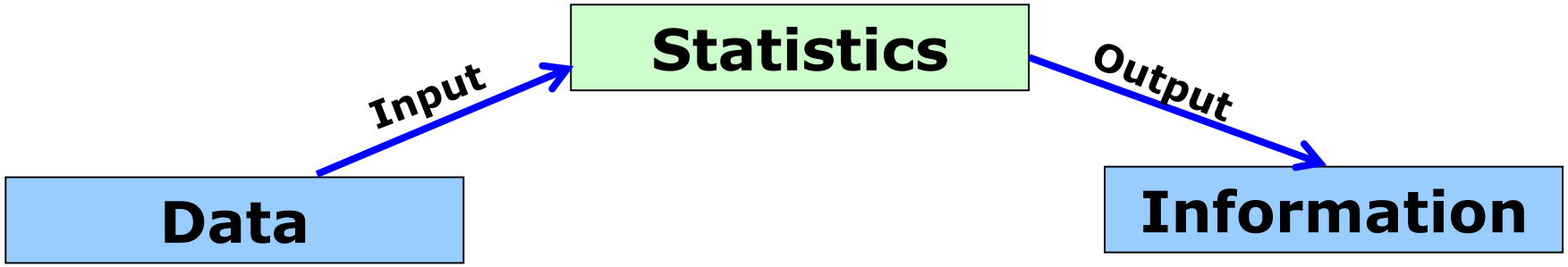
Data Analytics

ECON 1008, Semester 1, 2019

Giulio Zanella
University of Adelaide
School of Economics

What is statistics?

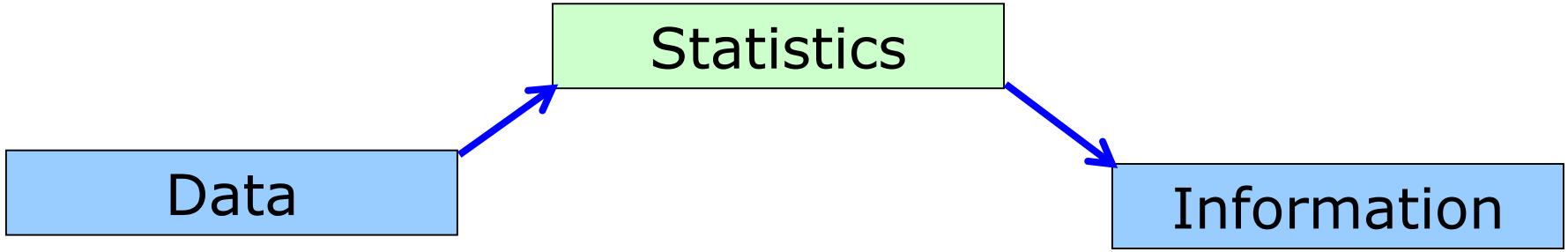
'Statistics is a way to get information from data to make informed decisions.'



Data: Mostly numerical facts collected from direct observations, measurements or surveys.

Information: Knowledge communicated concerning some particular fact, which can be used for decision making.

Example 1: Chickens and hunger



Quantity of chickens eaten in Australia every year and total population

Summary information derived about the statistics

500 million chickens

25 million people

Each of us eats 500/25 =

20 chickens every year

Example 1: Chickens and hunger

'Typical intake'

Mean (average intake)

Mean = 20

Is this enough information?

Vegans eat zero!

Hungry Jack eats one per week (52)

Measures of **Dispersion** (inequality) must be considered, e.g. large difference between those eating 0 or 40.

Example 2: Stats anxiety

A student enrolled in ECON 1008 is attending his second lecture. The student is somewhat apprehensive because he believes the myth that the course is difficult. To alleviate his anxiety, the student asks the lecturer about last year's mid-semester exam marks. The lecturer happily obliges and provides a **random sample of 76 observations** from the final marks.

What information can the student obtain from the list?

Example 2: Stats anxiety…

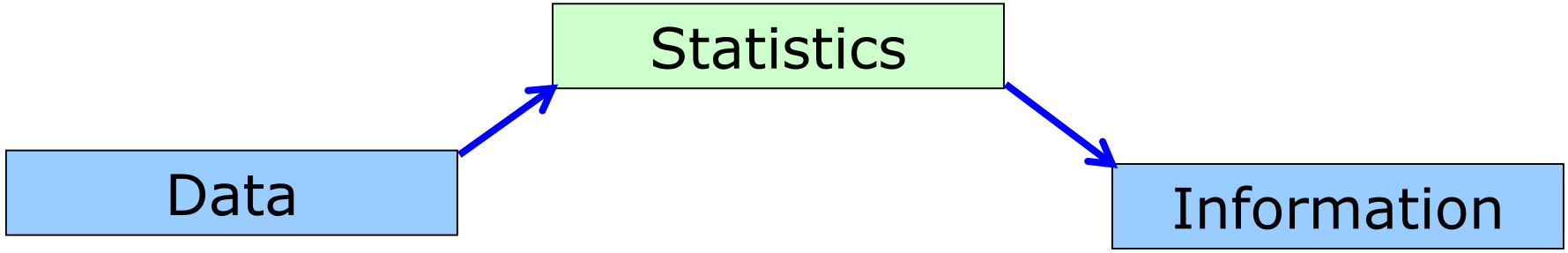
List of data provided by the lecturer to the student.

40	32	44	46	32	36	32	36	26	36	36	42	36	28	22
38	38	36	34	38	40	34	40	38	36	30	26	36	34	42
38	36	42	32	40	42	40	28	38	36	28	34	30	42	30
32	28	38	38	42	44	40	28	44	42	38	28	44	32	18
40	40	30	30	32	26	40	26	34	38	16	40	40	32	34
34														

These data are not very useful per se, even if I tell you that there were 50 points in total.
(and it's only 76 observations, imaging having 76K or 76M)

We need some data analysis.
Getting the data is only the first step.

Example 2: Stats anxiety…



List of last year's statistics marks

40
38
38
38
32
40
:

Summary information derived about the statistics class.

e.g., Class average, proportion of class receiving F's, most frequent mark, Highest and lowest marks, distribution, etc.

Example 2: Stats anxiety…

'Typical mark'

- Mean (average mark)
- Median (mark such that 50% above and 50% below)
- Mode (most frequent mark)

- Mean = 35.1 (70.2% or total points)
- Median = 36 (72% of total points)
- Mode = 40 (72% of total points)

Is this enough information?

Example 2: Stats anxiety…

Are most of the marks clustered around the mean or are they more spread out?

Range = Maximum – minimum = 46 – 16 = 30

Standard deviation = 6.1

(average distance from the average mark 35.1, more later)

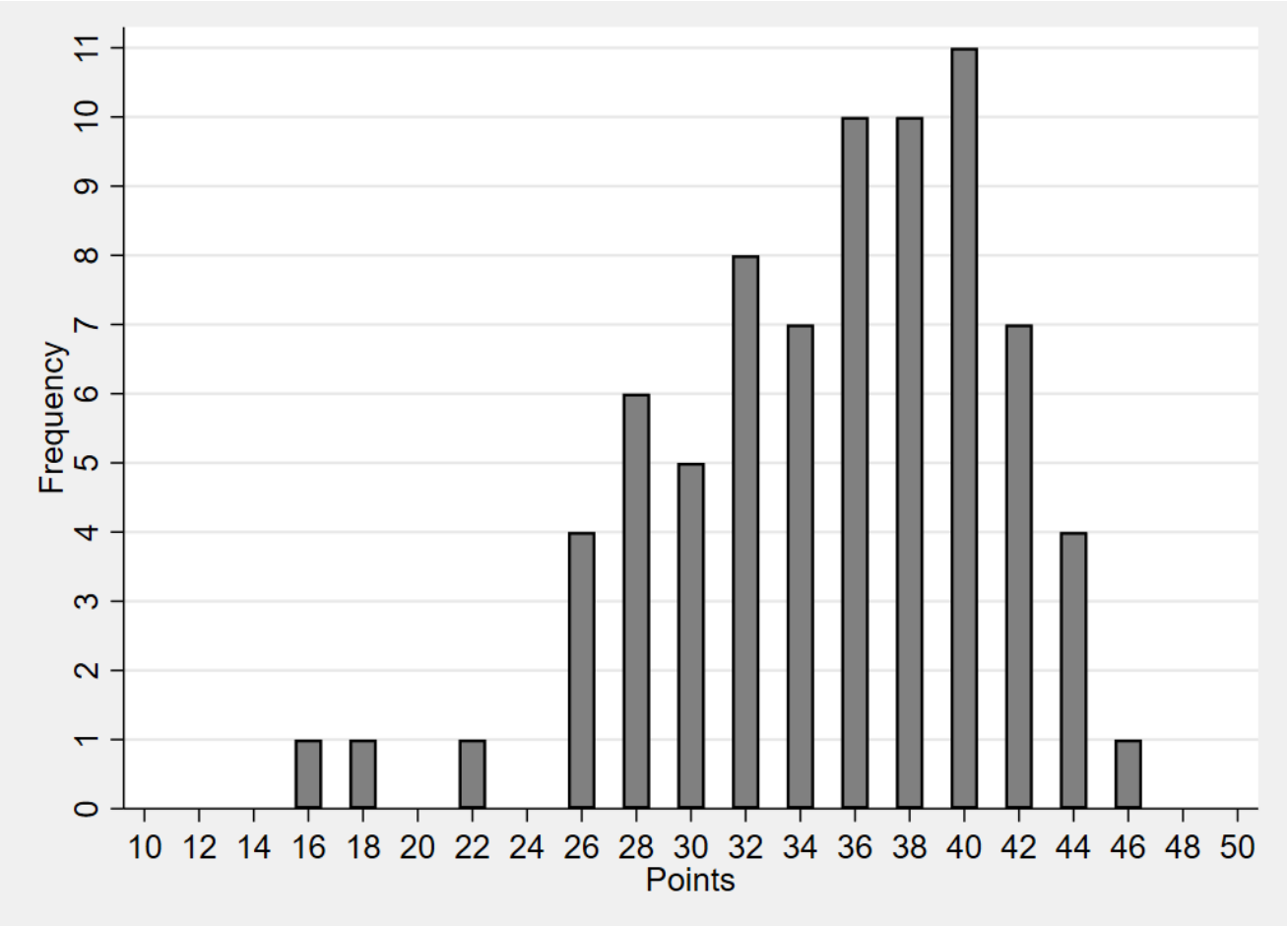
Example 2: Stats anxiety…

Are there many marks below 25 (50% of total points)
or above 40 (80% of total points)?

What proportion are “HD, D, P and F”-like grades (distribution of grades)

A graphical technique – **histogram** – can provide us with a lot of useful information

Example 2: Stats anxiety…



What information do we get and how can the instructor use it to make better decisions?

Example 2: Stats anxiety…

- A few students scored less than 25, so the instructor and the tutors must help them (they are here to help!)
- No student scored more than 46, so a couple of questions were too hard
- Most students are concentrated in the middle of the histogram so the exam was overall “fair”: not too easy, not too hard, apart from those couple questions.

Two major branches of Statistics

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics

Descriptive statistics deals with methods of organising, summarising, and presenting data in a convenient and informative way (what we did in the 2 examples)

One form of descriptive statistics uses graphical techniques, which allow statistics practitioners to present data in ways that make it easy for the reader to extract useful information.

Chapters 3 and 4 introduce several graphical methods.

Descriptive Statistics

Another form of descriptive statistics uses numerical measures to summarise data.

The mean and median are popular numerical measures to describe the location of the data.

The range, variance and standard deviation measure the variability of the data

Chapter 5 introduces several numerical statistical measures that describe different features of the data.

Inferential Statistics

Descriptive statistics describe the data set that is being analysed, but does not provide many tools for us to **draw any conclusions or make any inferences about the data**. Hence we need another branch of statistics: ***inferential statistics***.

Inferential statistics is also a set of methods, but it is used to draw conclusions or inferences about characteristics of ***populations*** based on sample statistics calculated from ***sample data***.

Chapters 9 and over introduce several techniques in inferential statistics.

1.1 Key statistical concepts

Population

A ***population*** is the group of **all** items (data) of interest.

Population is frequently very large.

E.g. 1. All students enrolled in ECON 1008

2. All current 2 million or so members of an automobile club (Example 1.4 in the textbook).
3. All oysters available in South Australian oyster farms.
4. All people living in Australia

Key statistical concepts

Sample

A ***sample*** is a set of items (data) drawn from the population of interest.

Sample could potentially be very large, but much less than the population.

E.g. 1. A sample of 76 students

2. A sample of 500 members of the automobile club selected.
3. A sample of 1000 oysters from Port Lincoln.
4. A sample of 100,000 Australian residents

Key statistical concepts

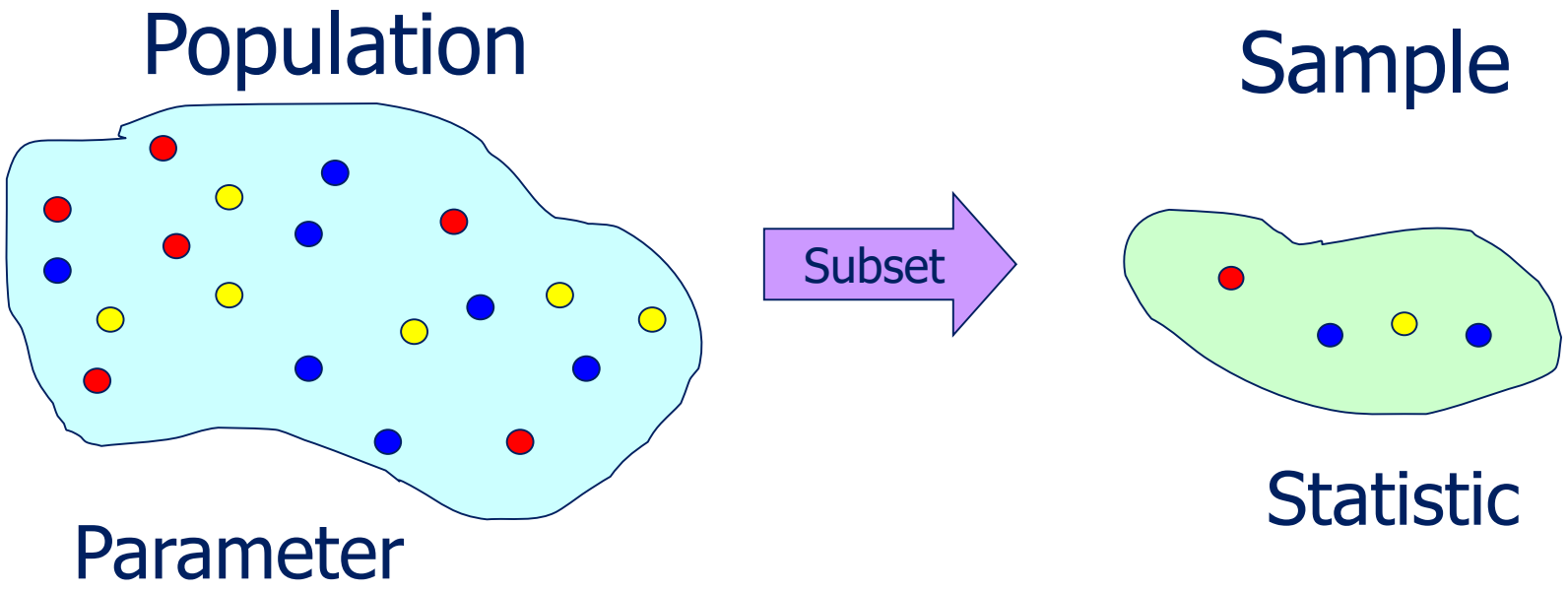
Parameter

A descriptive measure of a ***population***.

Statistic

A descriptive measure of a ***sample***.

Key statistical concepts

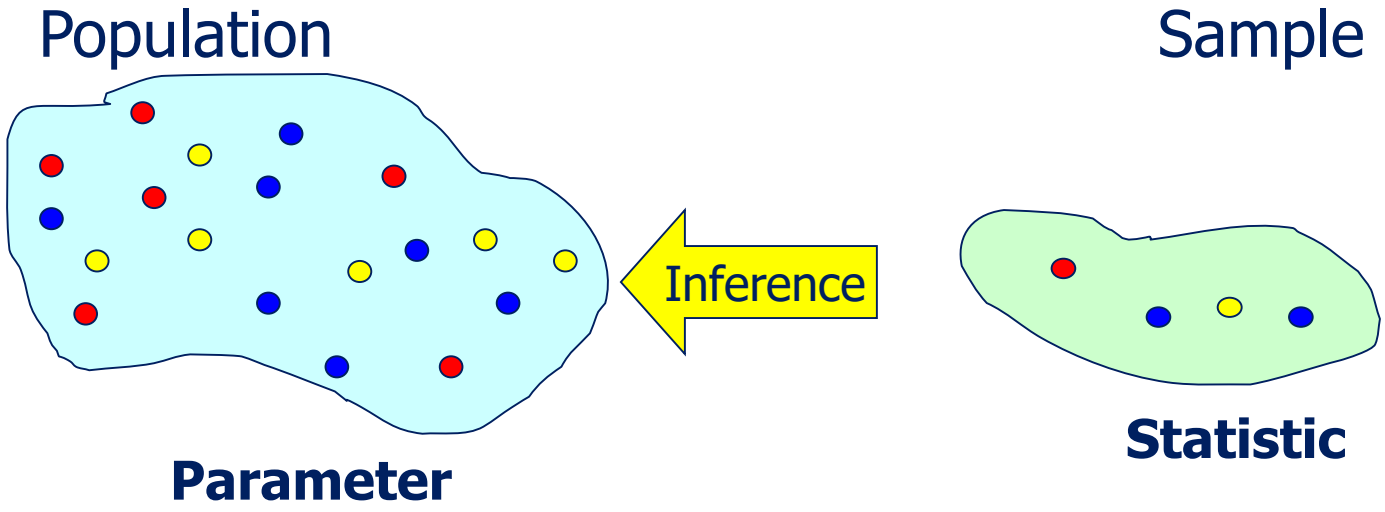


A descriptive measure of a population is called a ***parameter*** (e.g. Population mean)

A descriptive measure of a sample is called a ***statistic*** (e.g. Sample mean)

Statistical inference

Statistical inference is the *process* of making an estimate, prediction, or decision about a population based on a sample.



What can we ***infer*** about a population's parameter based on a sample's statistic?

Statistical inference

We use **sample statistics** to make inferences about **population parameters**.

Therefore, we can produce an estimate, prediction, or decision about a **population** based on **sample** data.

Thus, we can apply what we know about a sample to the larger population from which it was drawn!

Statistical inference

Rationale:

- Large populations make investigating each member impractical and expensive.
- Easier and cheaper to take a sample and make estimates about the population from the sample.

However:

- Such conclusions and estimates are not always going to be correct.
- For this reason, we build into the statistical inference 'measures of reliability', namely **confidence level** and **significance level.**

Example: Exit polls

When an election for political office takes place, the television networks cancel regular programming and instead provide election coverage.

The television networks often compete on the evening of an election day to be the first to correctly identify the winner of the election.

One commonly used technique is through **exit polls**, wherein a random sample of voters who exit the polling booths is asked for whom they voted.

Example: Exit polls…

Suppose that in the Brisbane electorate, 500 voters from various booths were asked to whom they voted.

From the data, the sample proportion of voters supporting the candidates is computed.

A statistical technique is applied to determine whether there is enough evidence to infer that the Labor party candidate will garner enough votes to win.

Example 3: Exit polls…

Suppose that the results were coded on a two-party preferred basis as 1 = Liberal/National candidate and 2 = Labor candidate.

The network analysts would like to know whether they can conclude that the Labor party candidate will win.

Voter	Response
1	1
2	2
3	2
4	1
5	2
. .	
495	2
496	1
497	1
498	1
499	2
500	1

Example 3: Exit polls…

This example describes a very common application of statistical inference.

The population the television networks wanted to make inferences about is the approximately 87 000 who voted in the electorate of Brisbane.

The sample consisted of the 500 people **randomly selected** by the polling company who voted for either of the two main candidates.

Example 3: Exit polls…

The characteristic of the population that we would like to know is the proportion of the total electorate that voted for Labor after preferences (on a two-party preferred basis).

Specifically, we would like to know whether more than 50% of the electorate voted for Labor (after preferences) in the electorate of Brisbane.

Example 3: Exit polls…

Because we will not ask every one of the 87 000 actual voters for whom they voted, we cannot predict the outcome with 100% certainty.

A sample that is only a small fraction of the size of the population can lead to correct inferences only a certain percentage of the time.

You will find that statistics practitioners can control that percentage and usually set it between 90% and 99%.