S T A T I S T I C S

1) Data Collection

1.1- Populations and samples

A population is the whole set of items that are of interest.
A census observes or measures every member of a population
A sample is a selection of observations taken from a subset of the population which is used to find out information about the population as a whole.

Normally, larger the sample → more accurate / IF the population is very varied, you need a larger sample than IF the population were uniform.

	Advantages	Disadvantages
-Census	accurate result	time consuming; expensive; cannot be used when the testing process destroys the item; hard to process large quantity of data
-Sample	less time consuming and expensive; fewer people have to respond; less data to process	data not as accurate; sample not representative (sub-groups)

Individual units of a population are known as sampling units.

Often sampling units of a population are individually named or numbered to form a list called a sampling frame.

1.2) Sampling

A simple random sample of size n is one where every sample of size n has an equal chance of being selected. → lottery / computer random number generator

- Ad → free of bias; easy & cheap to implement for small populations and small samples; each unit has a known and equal chance of selection
- Dis → Not suitable IF population size or the sample size is large; sampling frame needed

Large data set (LDS)

- Some readings are <u>not available</u> (n/a) → You might not be able to generate the data size you want.
- You can improve data set study by → take random sample of days from a month (or whole year) for different locations of the UK.
- The large data set only contains weather information from <u>May - October</u>, so using the data to make conclusions about the whole year round weather patterns may not be entirely reliable.
- The LDS has weather information on <u>5 UK locations and 3 overseas locations.</u>

· <u>UK</u> →		· <u>Overseas</u> →	
	Camborne	Beijing	(China)
	Heathrow	Jacksonville	(Florida)
	Hurn	Perth	(Australia)
	Leeming		
	Leuchars		

(Perth only one in southern hemisphere)

- <u>Cloud cover</u> is a discrete variable. It is measured in <u>oktas.</u>
- <u>Relative humidities above 95%</u> are associated with mist and fog.
- A <u>trace of rainfall</u> indicates a recorded variable of rainfall less than 0.05mm.
- 1 knot → 1.15 mph (1 m → 1.6 km)

mean ↗ 0.025

The Large Data Set contains 11 variables:

+ - Daily maximum temperature. / In °C to 1.d.p
+ - Daily Total Rainfall. / In mm (<0.05mm → tr)
- Daily maximum sunshine. / In hours to 1.d.p
- Daily maximum relative humidity. / In %. (>95% → mist / fog)
+ - Daily mean windspeed / In knots (1 knot, 1.15 mph)
- Daily maximum gust (ráfaga / racha de viento) / avg. over 24h.
- Daily mean wind direction / In ° relative to north
- Daily maximum gust direction / In ° relative to north
- Cloud cover / In oktas (measured in 1/8 → 0 clear sky / 8 complete overcast)
- Visibility / In dm, measured horizontally.
+ - Pressure / In hPa (1 millibar = 1 hPa = 0.001 bar)

↙ only these available for abroad.

In Systematic sampling, the required elements are chosen at regular intervals from an ordered list

- Ad → simple and quick to use; suitable for large samples and large populations.
- Dis → sampling frame needed; it can introduce bias if the sampling frame is not random

In Stratified sampling, the population is divided into mutually exclusive strata (males and females for example) and a random sample is taken from each.

- Ad → sample accurately reflects the population structure; guarantees proportional representation of groups within a population.
- Dis → population must be clearly classified into distinct strata; selection within each stratum suffers from the same disadvantages as simple random sampling.

The number sampled in a stratum: $\frac{\text{number in stratum}}{\text{number in population}}$ x overall sample size.

1.3) Non-random sampling

In quota sampling, an interviewer or researcher selects a sample that reflects the characteristics of the whole population

- Ad → allows a small sample to still be representative of the population; no sampling frame required; quick, easy and inexpensive
- Dis → can introduce bias; population must be divided into groups which can be costly or inaccurate; increasing scope of study increases number of groups, which adds time and expense.

Opportunity sampling consists of taking the sample from people who are available at the time the study is carried out and who fit the criteria you are looking for. (convenience sampling)

- Ad → easy to carry out; inexpensive.
- Dis → unlikely to provide a representative sample; highly dependent on individual researcher