Chapter 3: Collecting Data

A **population** is a collection of objects, items, humans/animals ("units") about which information is sought.

A **sample** is a part of the population that is observed.

A **parameter** is a numerical characteristic of a *population*, e.g. percent unemployment in the US.

A **statistic** is a numerical function of the *sampled data*, used to estimate an unknown parameter, e.g., percent unemployment in a sample.

Don't get confused between the population and the sample! There is separate notation for the population and the sample – make sure you know what you can calculate and what you can't calculate! You can't calculate anything from the population if you only have a sample.

A **sampling frame** is a list of all units in a finite population. Often, we do not have a sampling frame and we need to choose which units to sample.

A **representative sample** does not differ in systematic and important ways from the population.

Convenience sampling involves using a sample that is easily available.
Judgment sampling involves a trained sample collector (who may use **quota sampling**). Bias is possible with these sampling methods. To avoid bias, sample randomly from the population.

A **simple random sample** (SRS) of size n from a population of size N is drawn without replacement so that each possible sample of size n has the same chance of being chosen.

An SRS is not always practical to obtain, for instance for a highly diverse population, or a really large population.

Stratified random sampling involves dividing the population into homogeneous subpopulations and drawing and SRS from each one. This is useful when you want to do statistics on subpopulations as well as on the whole population.
e.g. customers stratified by race (some races are rarer than others)

Multistage cluster sampling – "Tree structured" sampling, units are different at each stage. Useful for sampling large populations.

e.g. 1) Draw SRS of states
2) Draw SRS of counties from the states
3) Draw of people from each county

1-in-k systematic sampling consists of selecting every kth unit. Useful for sampling items coming off assembly lines.