Statistical thinking is a way of understanding a complex world through using simple terms for essential structure, acknowledging and assessing a degree of uncertainty.
It includes using data to challenge intuition, which may be wrong

Statistics can do three main things

- Describe
  - The reduction to simple relative risk summaries (such as in that for food) describes a complex dataset
- Decide
  - Did the results occur by chance?
- Predict
  - We generalise from the data to a new situation
  - To do this well we need a large and representative sample

Fundamental concepts of statistics:

- Learning from data
  - We start with a set of hypotheses to do so
- Aggregation
  - Reducing the data to summary statistics
- Uncertainty
  - The aggregate findings don't apply in all cases
- Sampling
  - We need a representative sample to summarise an entire population
  - The way a sample is obtained is critical
- Causality
  - It is not enough to just detect patterns in the data as these may be due to some other underlying factor or an error in the sample
  - We need to control and manipulate the specific factor of interest to understand causal relationships

We can generate hypothetical independent samples that allows us to have a sampling distribution to find the population mean and standard deviation.
If we independently repeat the same sampling experiment and use the same sample size and population a large number of times and produce a large number of sample mean estimates we get the sampling distribution of the mean.

We can also use central limit theorem which shows that the sampling distribution becomes normal as the sample size becomes large. This occurs even if the data from the original population is not normal distributed. We can also scale this to get a standard normal distribution.

SEM=$\frac{\sigma}{\sqrt{n}}$
We can use the sample standard deviation for this if the population one is unknown. The larger the standard deviation, the larger the standard error

We can also use the central limit theorem to calculate confidence intervals. If the population standard deviation is unknown we use the student's t distribution and if it is we use a normal critical value. These are equal if n is large. These are symmetric about the sample mean. $\left(\bar{X}_n + t_{\frac{\alpha}{2}, n-1}\frac{s_n}{\sqrt{n}}, \bar{X}_n + t_{1-\frac{\alpha}{2}, n-1}\frac{s_n}{\sqrt{n}}\right)$

Confidence intervals are interpreted as the interval will capture the true population mean (1-α)*100% of the time.

We can also do bootstrap confidence intervals in which we resample hypothetical data sets from the original sample with replacement, calculate the sample average for each and get a hypothetical sample as an approximate sampling distribution and use the quantiles of this to build a confidence interval.
This doesn't rely on the central limit theorem or 'asymptotic' results and relates to the sampling distribution of $\bar{X}_n$ for a sample size n.
To bootstrap the sample mean we take a sample of observations and take R bootstrap samples and for each draw n values with replacement, and calculate $\bar{x}_n^r$ with the same sample size.

Independent samples occurs when there is no connection between individual respondents in each group.

We might want to characterise the difference between two sub populations. We want to ask if the group means are equal. To do this we do a confidence interval. $\left(\bar{X}_1 - \overline{X_2} + t_{\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{X}_1 - \overline{X_2} + t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$