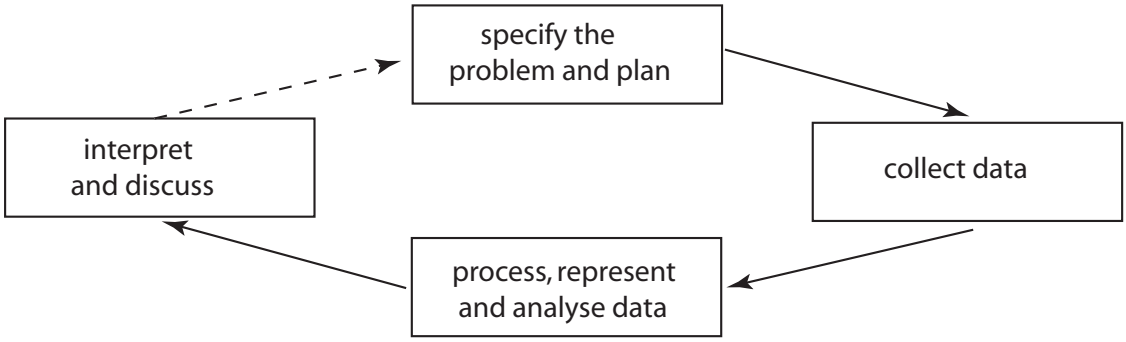
# The statistical problem solving cycle

Data are numbers in context and the goal of statistics is to get information from those data, usually through *problem solving*. A procedure or paradigm for statistical problem solving and scientific enquiry is illustrated in the diagram. The dotted line means that, following discussion, the problem may need to be re-formulated and at least one more iteration completed.




## Descriptive statistics

Given a sample of $n$ observations, $x_1, x_2, \ldots, x_n$, we define the **sample mean** to be

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

and the *corrected* sum of squares by

$$S_{xx} = \sum (x_i - \bar{x})^2 \equiv \sum x_i^2 - n\bar{x}^2 \equiv \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$\dfrac{S_{xx}}{n}$ is sometimes called the *mean squared deviation*.

An **unbiased estimator** of the population variance, $\sigma^2$, is $s^2 = \dfrac{S_{xx}}{(n-1)}$ . The **sample standard deviation** is $s$. In calculating $s^2$, the divisor $(n-1)$ is called the **degrees of freedom (df)**. Note that $s$ is also sometimes written $\hat{\sigma}$ .

If the sample data are ordered from smallest to largest then the:

minimum (Min) is the smallest value;
lower quartile (LQ) is the $\frac{1}{4}(n+1)$-th value;
median (Med) is the middle [or the $\frac{1}{2}(n+1)$ -th] value;
upper quartile (UQ) is the $\frac{3}{4}(n+1)$-th value;
maximum (Max) is the largest value.

These five values constitute a **five-number summary** of the data. They can be represented diagrammatically by a *box-and-whisker plot*, commonly called a *boxplot*.