Estimating the population mean from a sample

In this unit we will introduce continuous data, such as the weight of a squirrel (in grams). We'll draw random samples from a population and calculate the sample mean for each sample.

When we looked at the ages of the people in the room, we were able to get data for the whole population; we didn't need to sample. We can't catch all the squirrels in the forest and weigh them, so we want to make inferences about the population from the sample. For that we use mean and standard deviation (SD) rather than median and interquartile range(IQR).

Drawing the sample

The plastic chips in the bag represent weights of a simulated population of squirrels.

Shake the bag well and take 10 chips at random. Open a new spreadsheet and enter the 10 values in one column of the spreadsheet. We'll denote these values by "*x*". Return the chips to the bag.

Notice that the weight of the randomly-drawn squirrel is different each time: *x* is a **random variable**. We deal with the squirrel weights by describing the distribution of the weights, usually in terms of the mean and spread (variance or standard deviation).

Estimating the mean from the data

In the spreadsheet, use the AVERAGE function to calculate the arithmetic mean. The **sample mean** is referred to as $\bar{x}$, "x bar".

Compare the values of $\bar{x}$ with other members of the group (we'll put them all up on the whiteboard).

Notice that $\bar{x}$ is different for each randomly-drawn sample: $\bar{x}$ is also a random variable with its own distribution.

We don't usually know the true mean of the population we took samples from, but because this is a simulated population, we know the true mean, $\mu = 1000.01$.

- Did anyone get a value of $\bar{x}$ equal to μ?
- How many sample means were higher? How many were lower?
- How low was the lowest? How high was the highest?

$\bar{x}$ is rarely equal to μ because of **sampling errors**. These errors are not mistakes, they are simply due to the sampling process and the variability in the population sampled.

However, the values of $\bar{x}$ cluster around μ, with the number bigger than μ roughly equal to the number smaller than μ. In fact, if we had a huge number of values of $\bar{x}$, the mean of the $\bar{x}$'s would equal μ. We say that the **expected value** of $\bar{x}$ is μ:

$$E(\bar{x}) = \mu$$

The best estimate of μ (for which we use the symbol $\hat{\mu}$, "mu hat") based on our sample is $\bar{x}$:

$$\hat{\mu} = \bar{x}$$

In the next unit you will fit a Bayesian model to your data and get a posterior distribution for the mean and the SD of the population in the forest.