

MATH 203: PRINCIPLES OF STATISTICS 1

FALL 2024

POPULATIONS, EXPERIMENTAL UNITS, SAMPLES

1. Experimental units:

- the ‘objects’ or ‘things’ on which we will be making measurements ?
- we identify the experimental units **before** we actually take any measurements;

2. Population:

- the entire collection of experimental units;
- could be a finite collection, or could be large enough to be regarded as infinite.

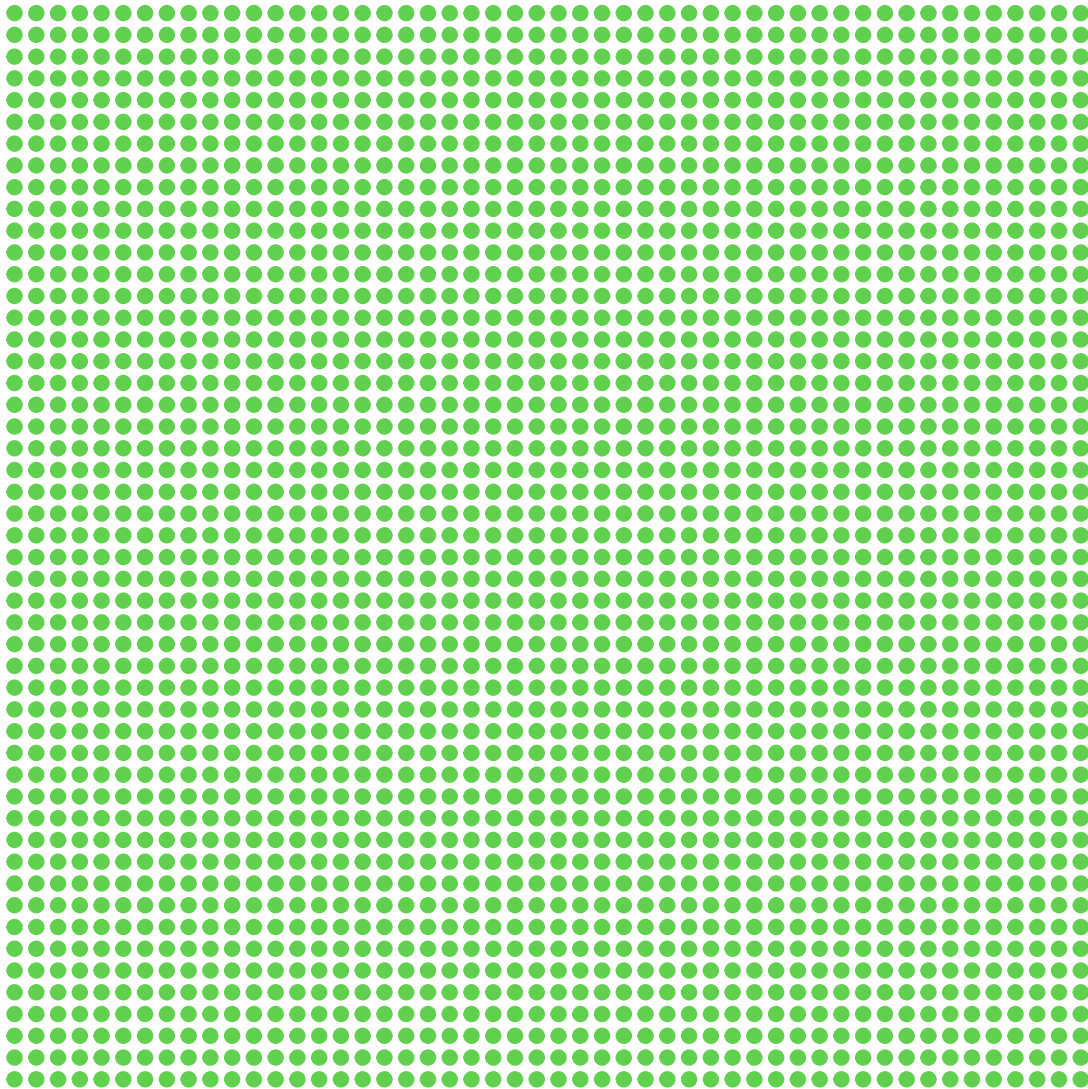
3. Sample:

- the collection of **actual** experimental units on which we **do** make measurements;
- ‘the sample’ is also used to refer to the recorded measurements themselves;
- we are only able to take a **finite** sample;
- we will usually need to make some assumptions about **how** the sample is collected.

In the following plots we have a reasonably large population of size $N = 2500$ where the experimental units are represented using a green circle.

Note: here the population is large but finite, but sometimes we can conceptualize the population as infinite (if we can imagine collecting data indefinitely on different experimental units), or to be so large we can treat it as infinite.

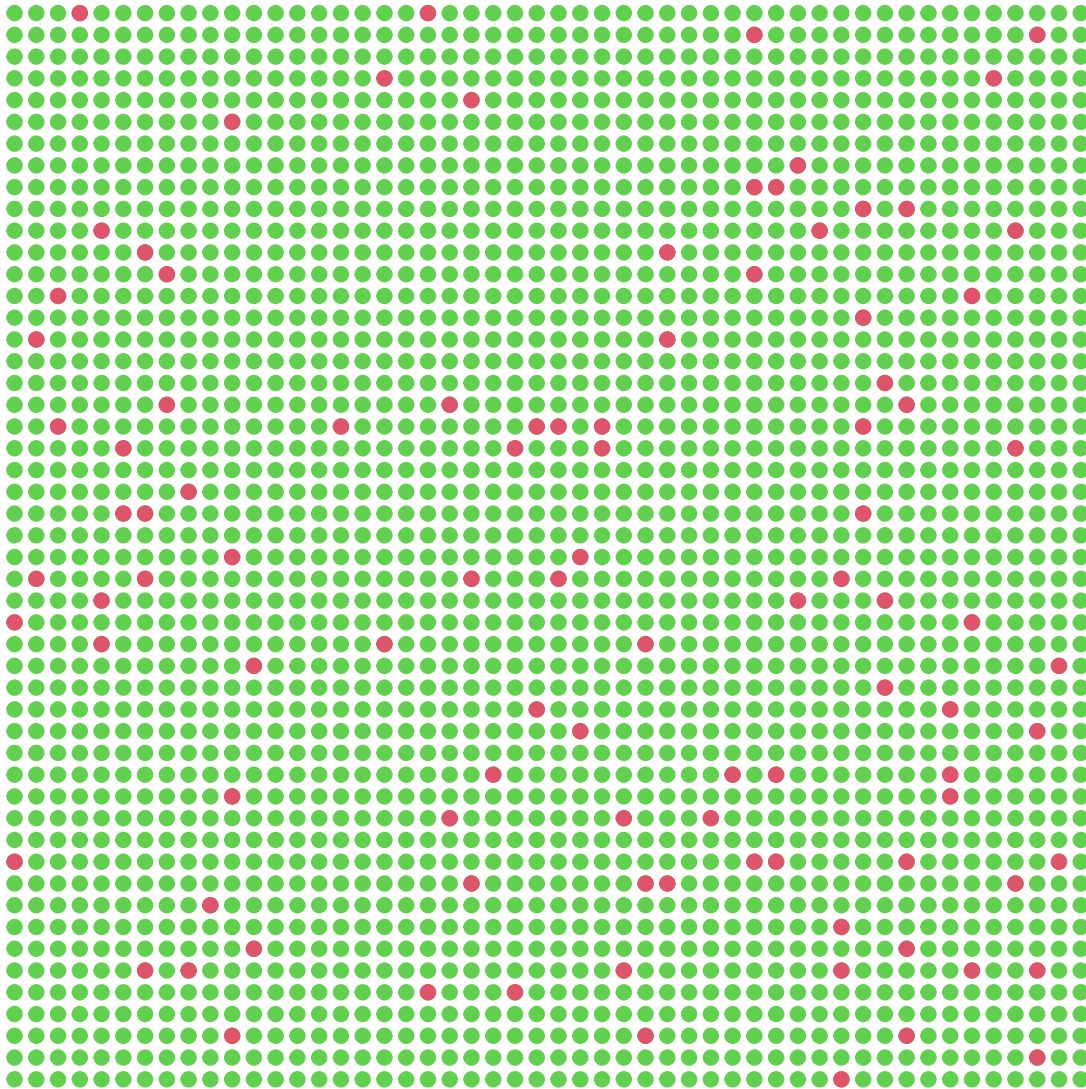
The population



The population is the entire collection of experimental units on which we could take measurements to study the variable of interest.

The population comprises experimental units that (with respect to the variable of interest) are believed to be comparable. Here the population is large, but finite.

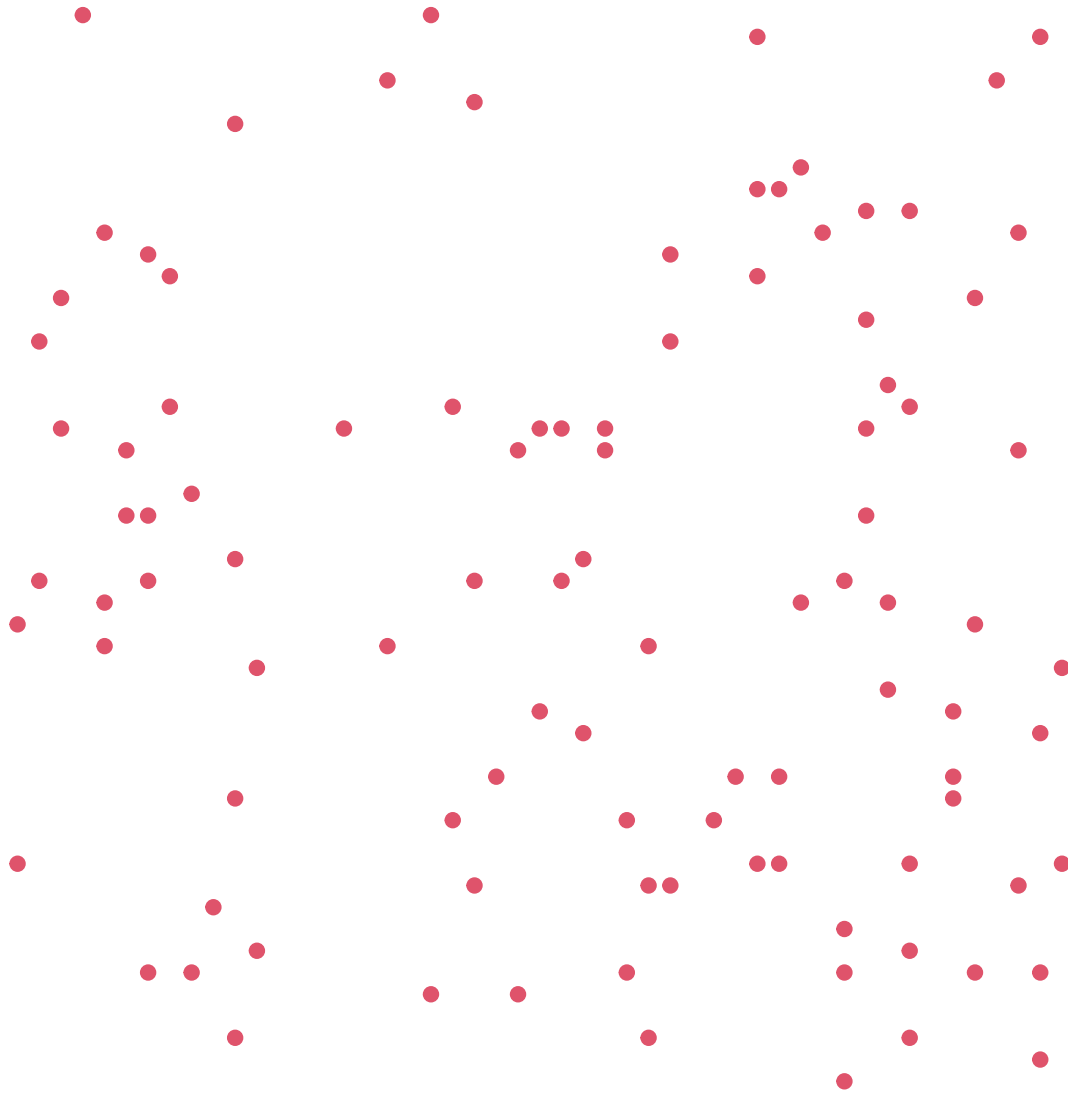
The sample



In the sample of size $n = 100$, we select a collection of the experimental units and make the measurements.

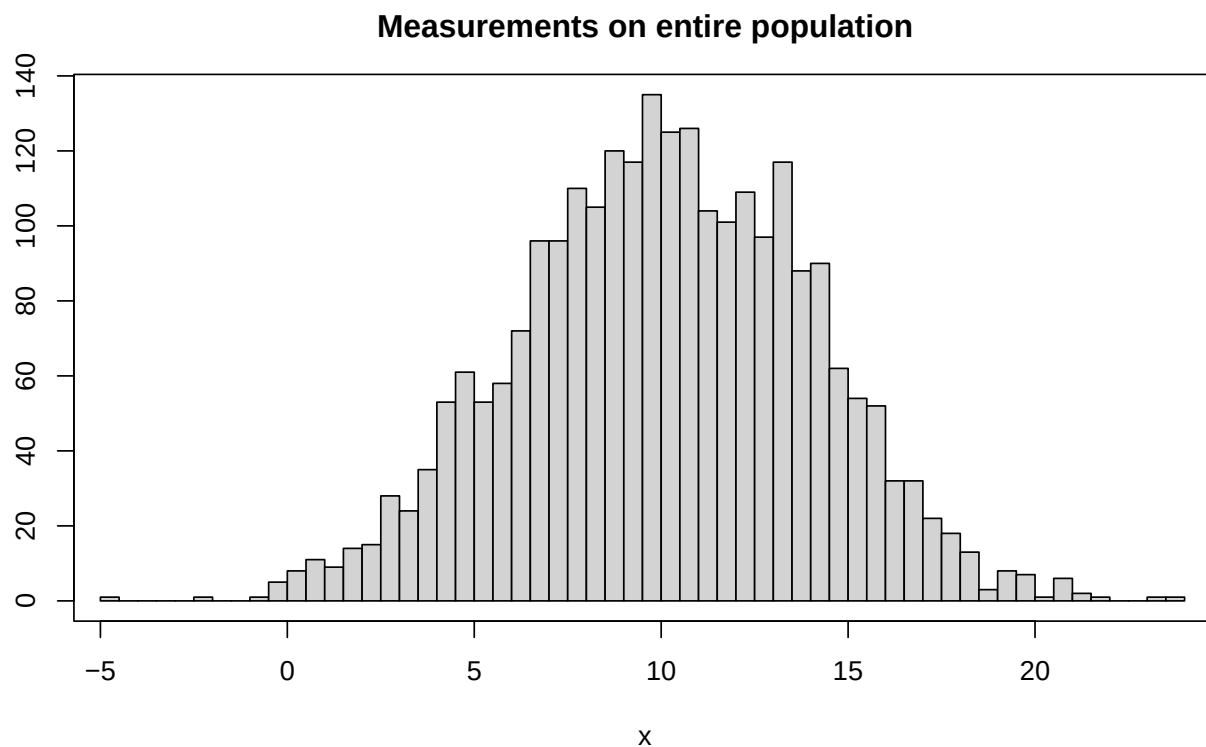
Here, the sample comprises measurements on typical experimental units; the sampled units are representative of units in the population.

The sample

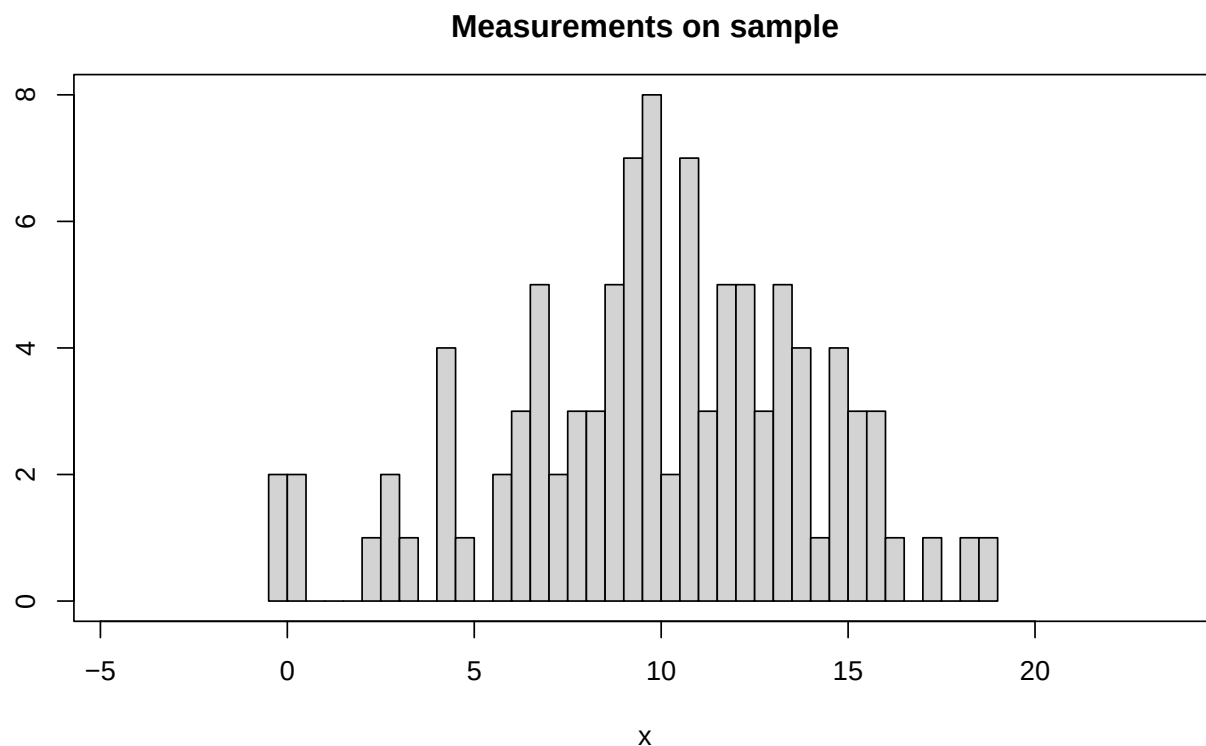


Sometimes, we start by collecting the sample, and considering what population it is representative of.

If we were able to make measurements on the **entire** population, we would know with certainty all of the characteristics of the population.

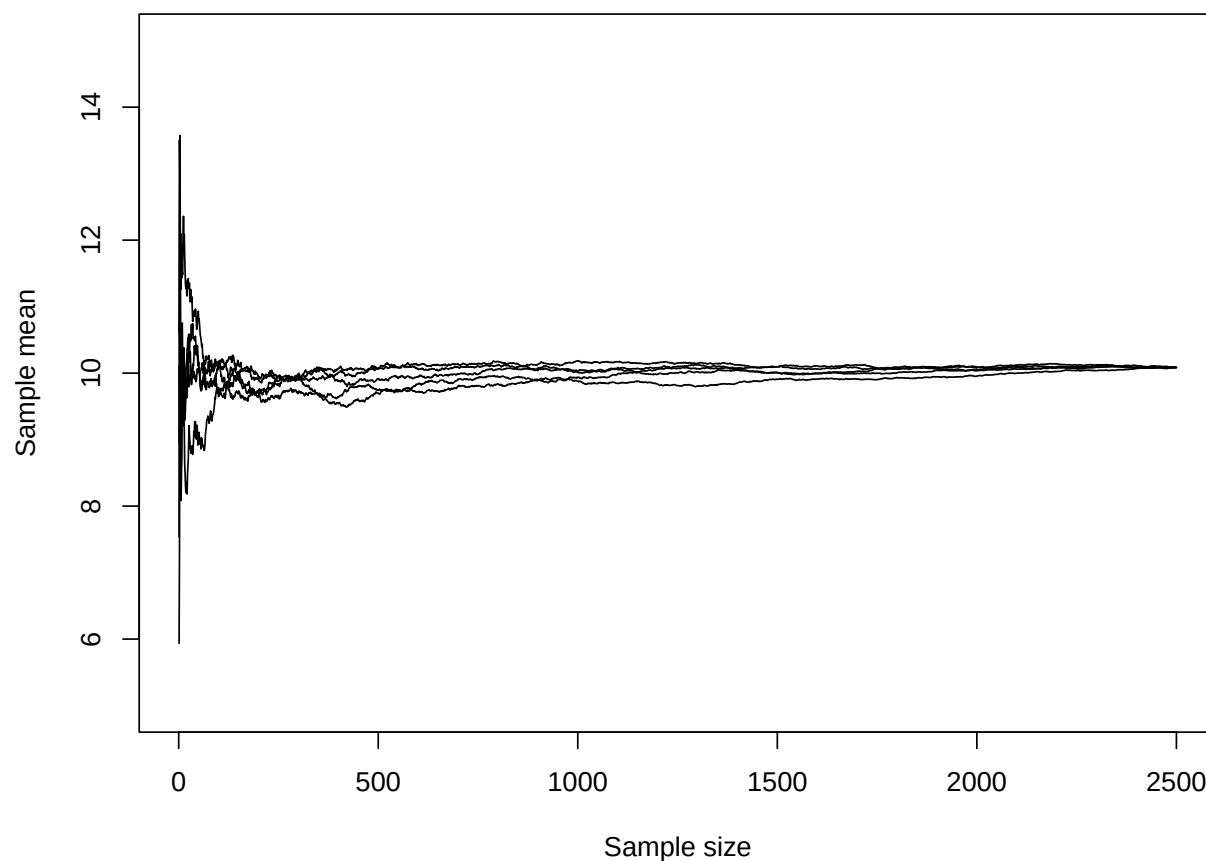


We might wish to report the population mean: in this case it is $\mu = 10.085$.



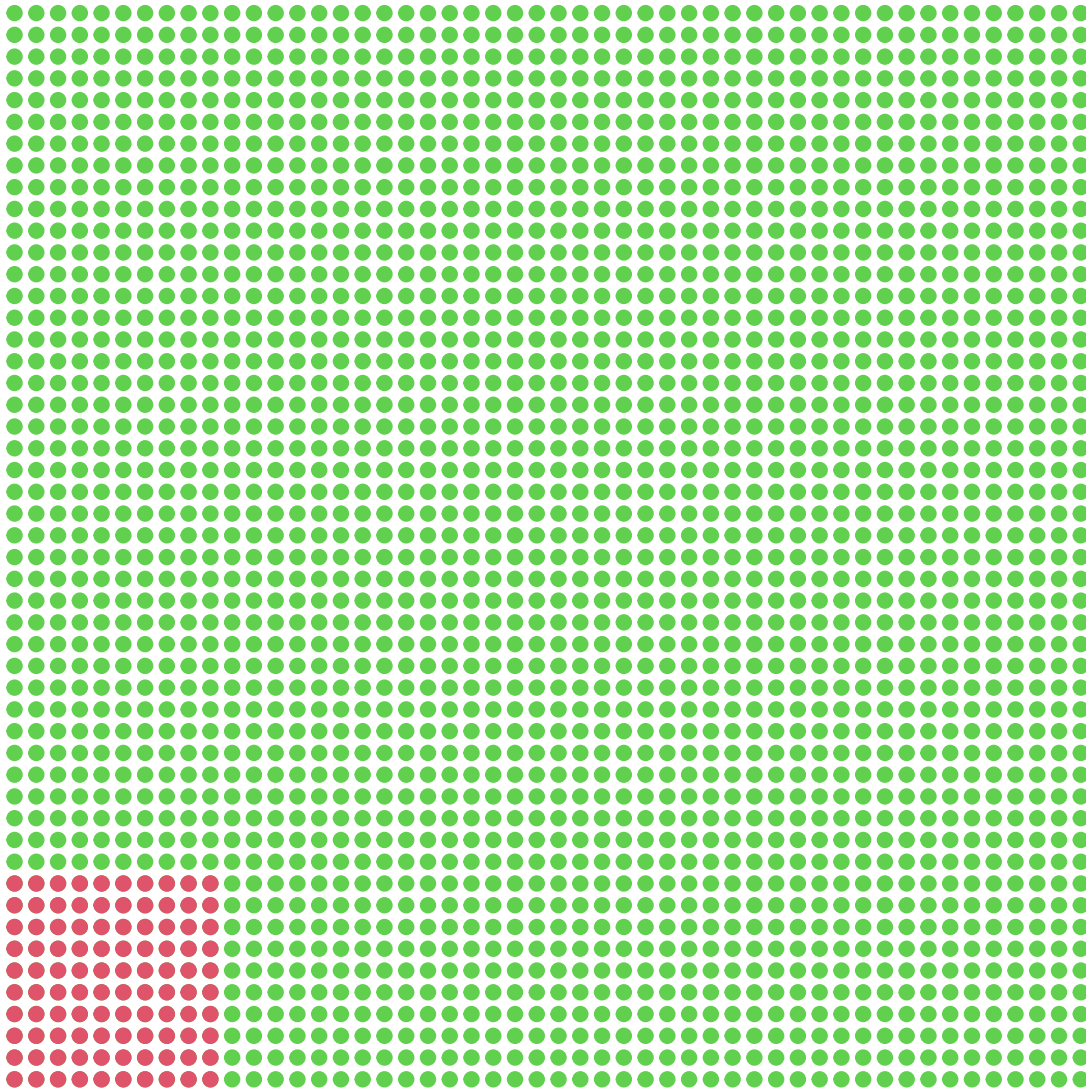
In the sample, we only see a subset of the measured values. The sample mean is $\bar{x} = 9.923$

If we could ultimately collect measurements from **all** experimental units, then we see that the sample mean will 'converge' to the population mean; in the next plot, the running sample means from five studies that eventually collect data from all experimental units are plotted.



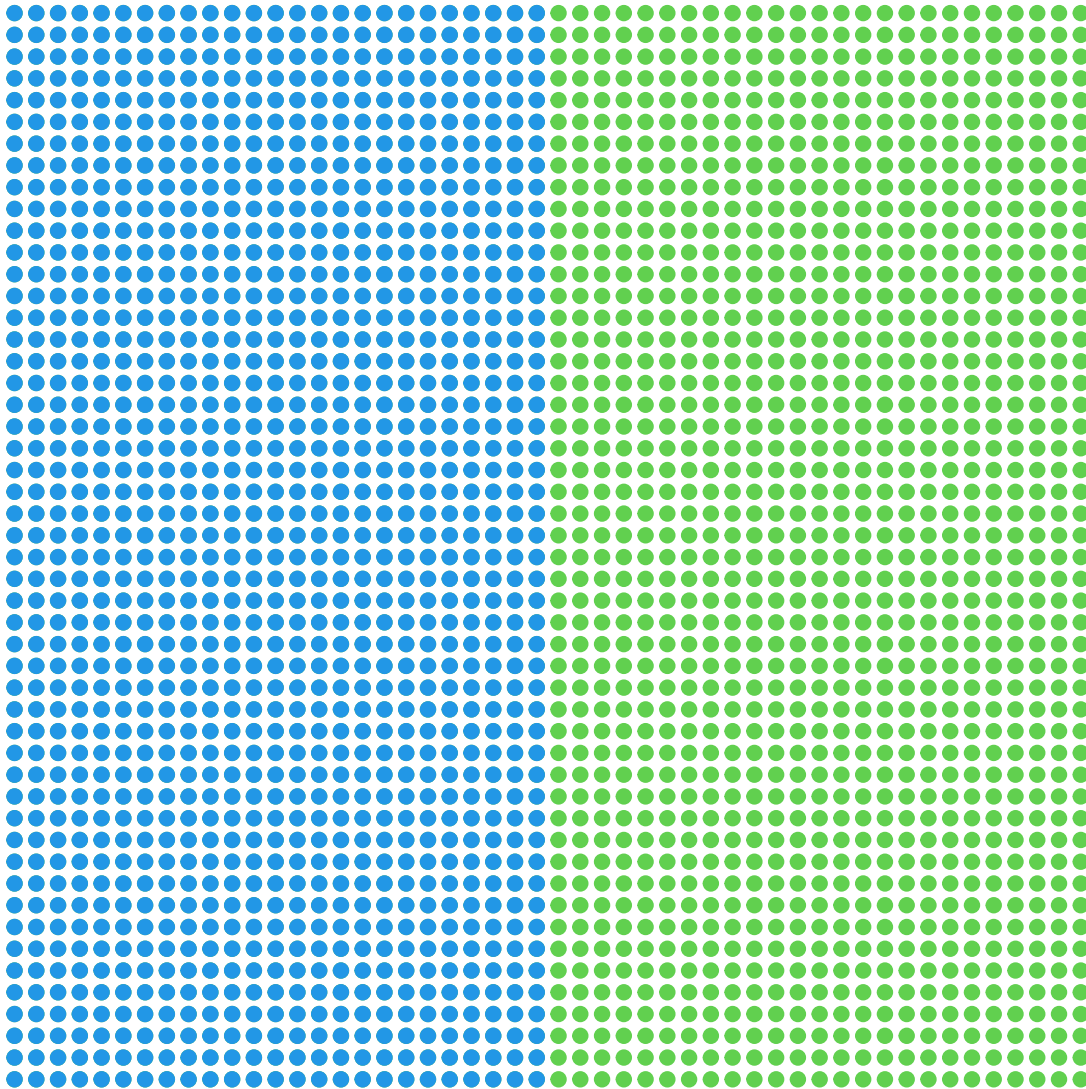
Ultimately, we see that all five studies yield $\bar{x} = \mu$, when the sample size n reaches the population size N .

Another sample



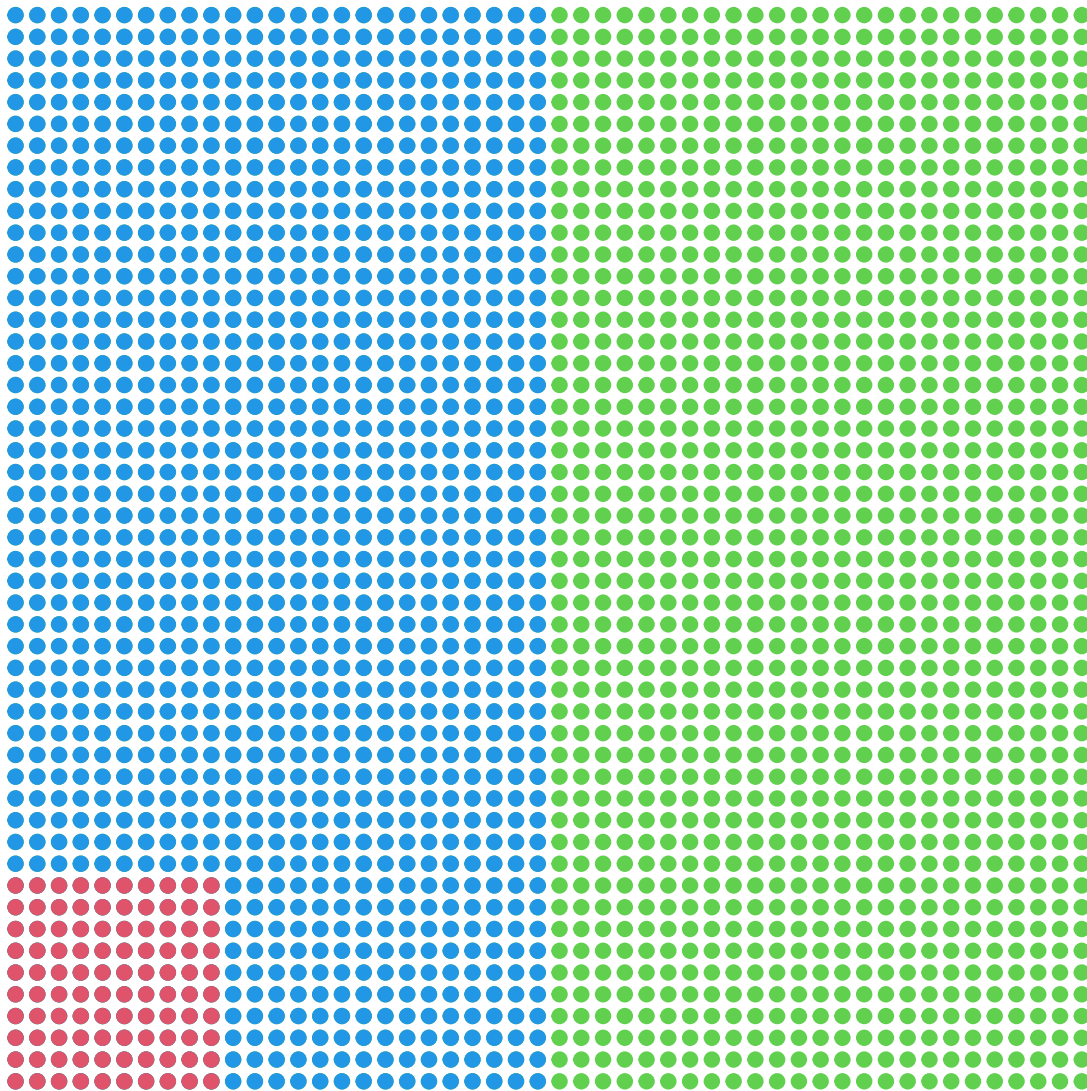
In this case, the sample is still representative of the population. Here, the sample mean is $\bar{x} = 9.560$.

Two populations



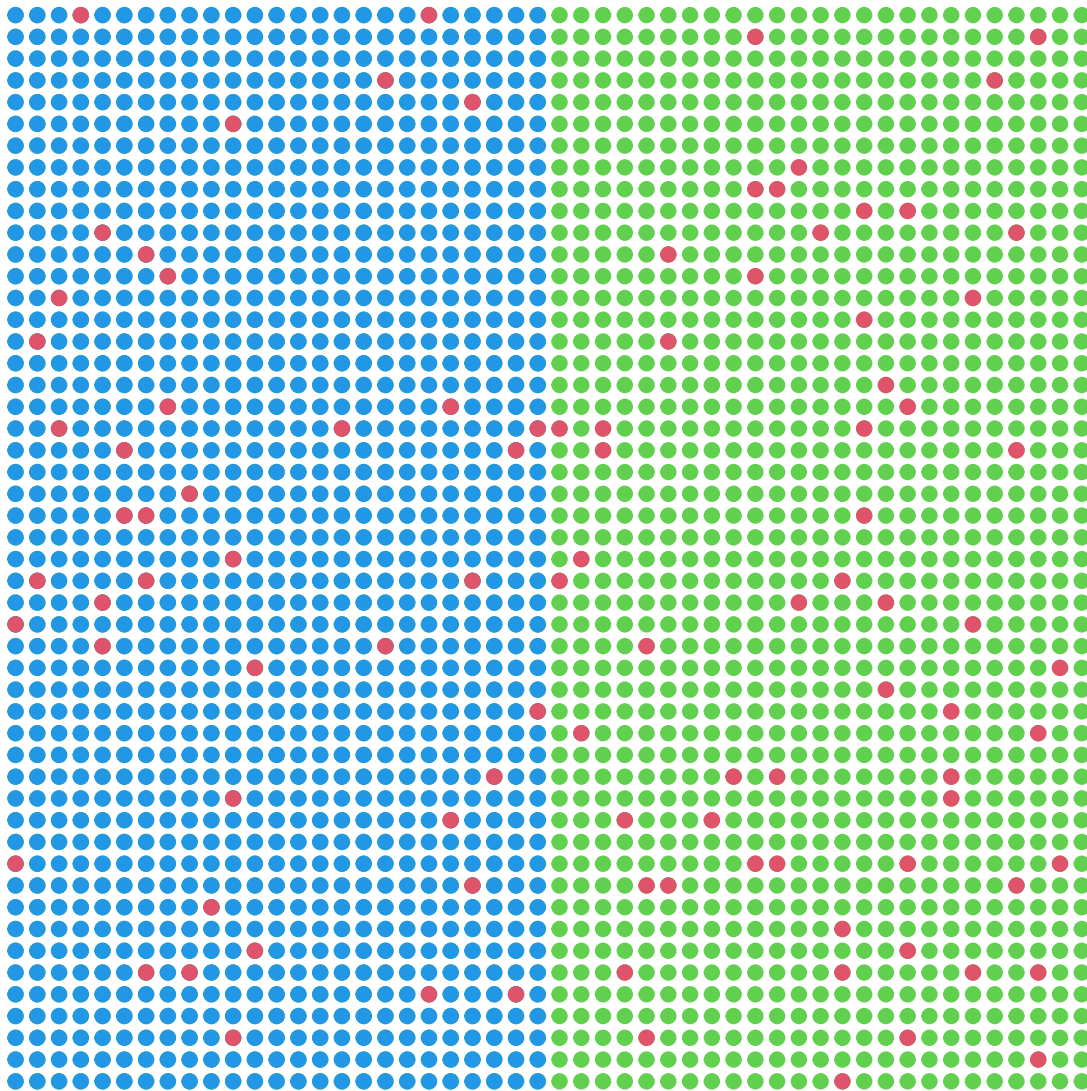
Here there are two large populations that have different characteristics. Suppose the population means are 14.953 (for the blue circles) and 10.218 (for the green circles).

Sample from one of the populations



Here the sample is representative of the blue population: the sample mean is $\bar{x} = 14.560$.

Sample from the combined populations



Here the sample is not representative of either population: the sample mean is $\bar{x} = 12.023$