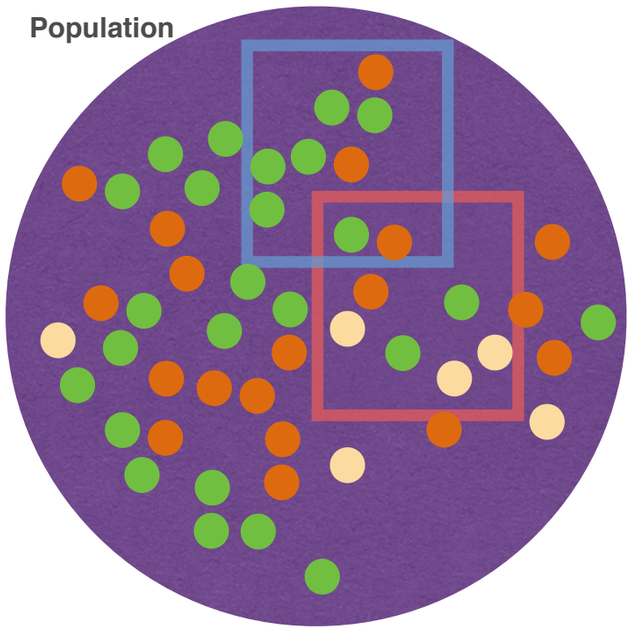
Inferential Statistics

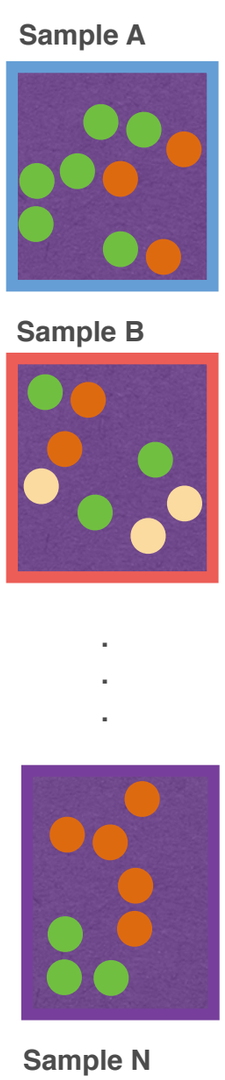
Inference

Inferential statistics allows us to draw conclusions from a **sample** and generalise them to a **population**.



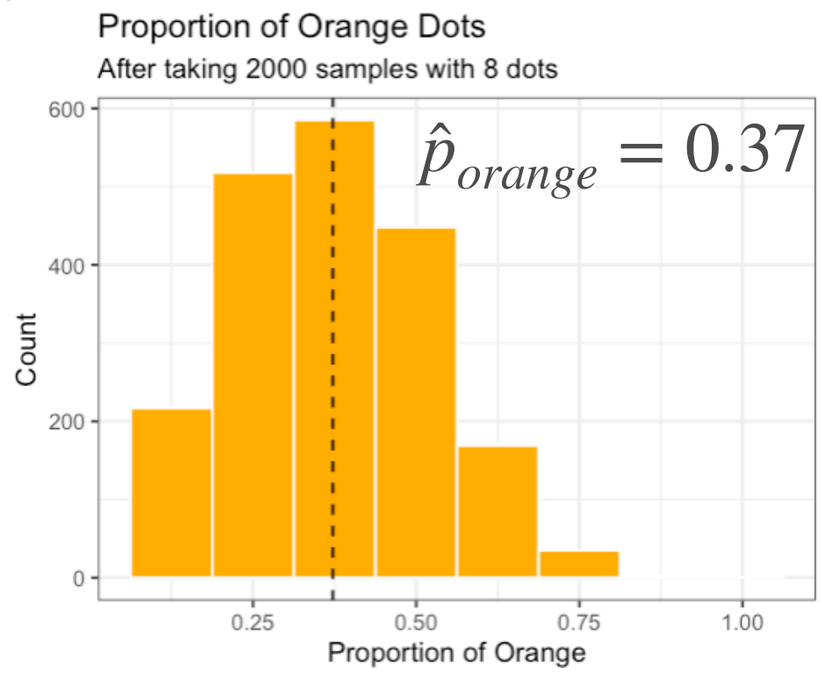
Colour	N dots	Prop
Green	25	0.51
Orange	16	0.37
Pale Yellow	6	0.12

$$\pi_{orange} = 0.37$$



replicate	colour	n	prop
1	Green	5	0.625
1	Orange	3	0.375
1	Pale yellow	0	0
⋮	⋮	⋮	⋮
2000	Green	3	0.375
2000	Orange	5	0.625
2000	Pale Yellow	0	0

Sampling Distribution - Distribution of Sample Statistic (Proportion)



Population	Sample
All the objects, elements, or entities conceivably of interest for an investigative study. We will rarely have access to the data for an entire population.	A part or subset of the population. We typically draw samples to calculate a sample statistic/point estimate and infer a population parameter.
Parameters: properties of the population such as	**Statistics/Point Estimates**: properties of a sample, such as the mean or standard deviation
μ **Mean** σ **Standard Deviation** π **Proportion**	$\bar{x}$ **Mean** S **Standard Deviation** $\hat{p}$ **Proportion**

Central Limit Theorem

The **central limit theorem** is a mathematical theorem describing

the tendency for the **distribution of sample statistics** calculated from multiple sample distributions to resemble the **normal distribution** provided the sample size is large enough ($n >= 30$).

This is true even if the underlying population isn't normally distributed! And forms the foundation for a lot of inferential statistics.

Inferential Statistics

Sampling Distribution

The distribution of a calculated sample statistic after taking multiple samples from a population.

The narrower a sampling distribution: the more certainty in our sample statistic being close to the true/population value.

Standard Error

The **standard deviation** of the **sampling distribution** and so is a measure of how **spread** the different calculated sample statistics are. For our sampling distribution of 2000 samples, our standard error was 0.151.

A small standard error indicates that the **sample** is representative of the **population**.

Standard Error can be reduced by increasing the sample size (n).

Confidence Intervals

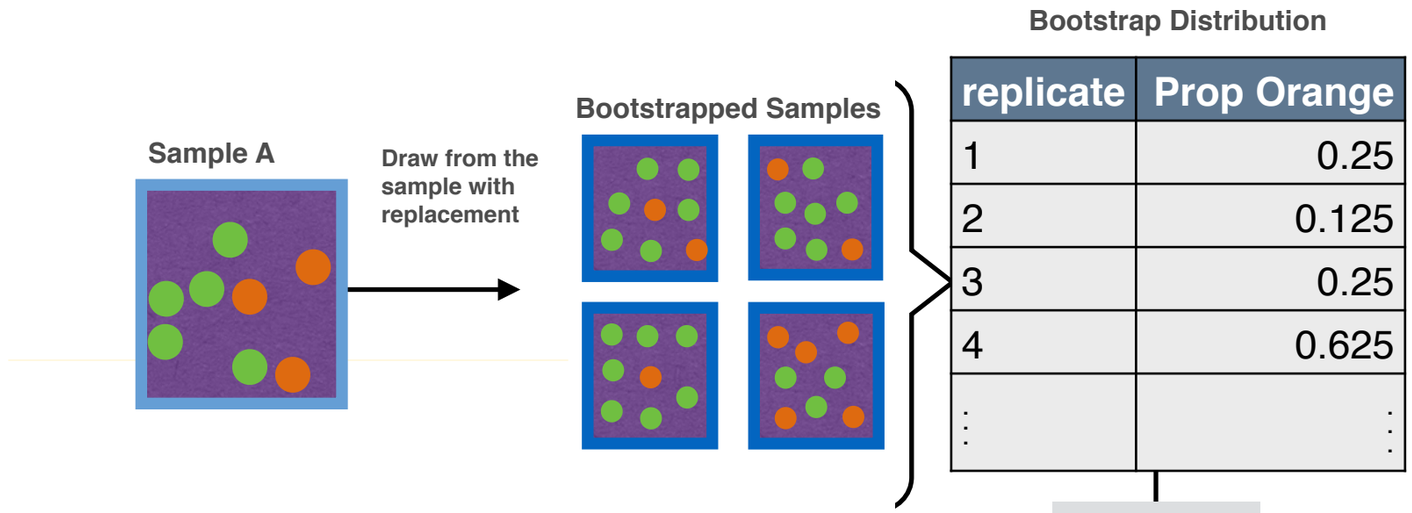
A range of values likely to contain the **population parameter**.

Like the standard error, the width of CIs can be reduced by increasing the sample size (n).

Bootstrapping/Bootstrap Resampling

A process where we simulate taking multiple samples from our population and replicate this many times. We do this by repeatedly drawing from our initial sample with replacement multiple times.

By doing this we can generate an approximation of the sampling distribution generated from taking multiple samples.



Interpretation

From bootstrap resampling our Sample A ($n = 8$): we estimate the proportion of orange dots to be 0.382.

Using this method with this sample, we state with 95% confidence that the value for the population parameter lies between 0.125 and 0.7.

Larger Sample Size:

From bootstrap resampling a larger sample ($n = 24$): we estimate the proportion of orange dots to be 0.333.

Using this method with this sample, we state with 95% confidence that the value for the population parameter lies between 0.165 and 0.55.