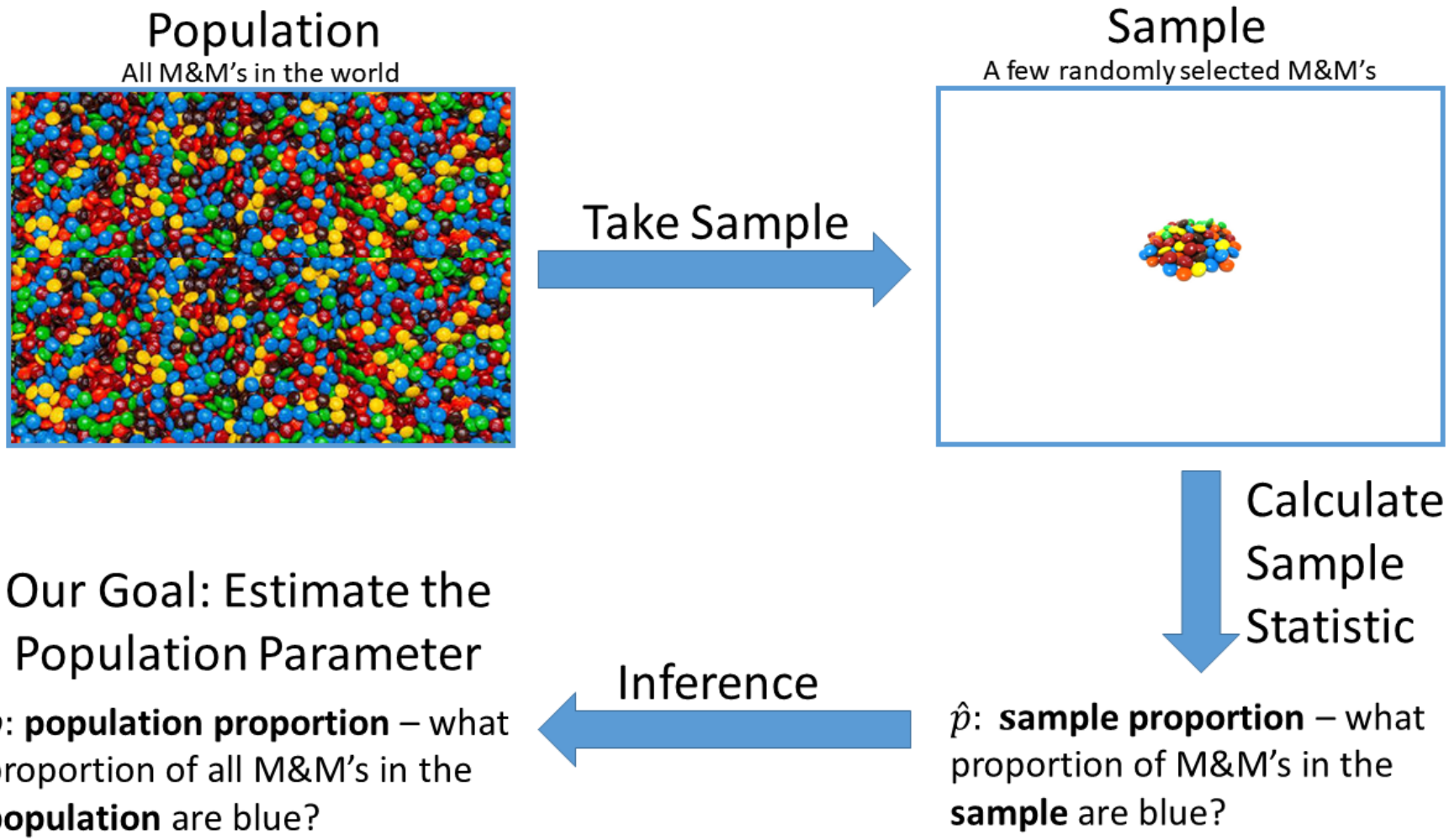
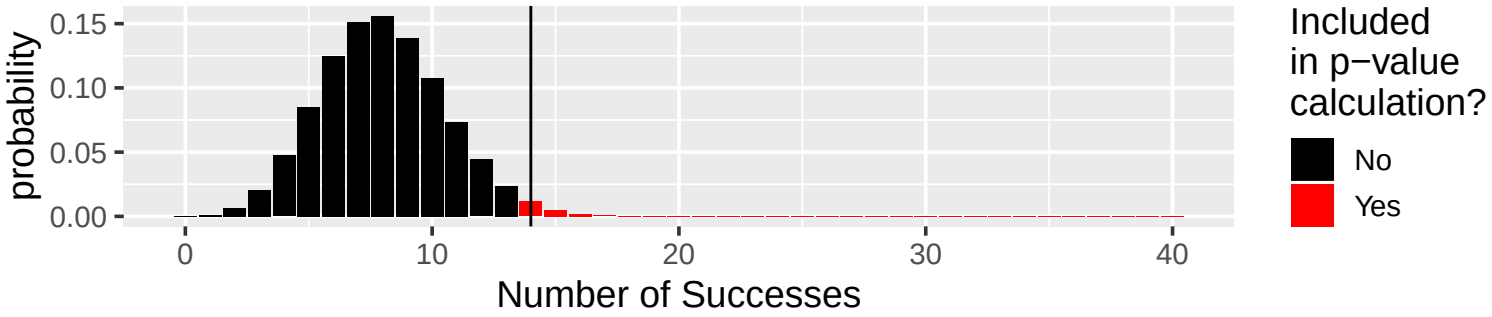
Summary of Recent Stuff



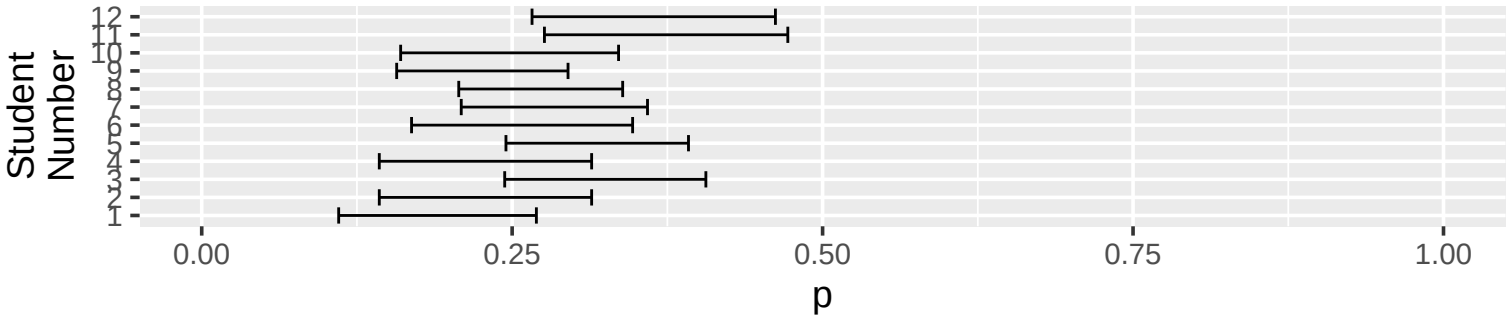
- Population Parameter (p): What proportion of the population are in a certain category of a categorical variable?
- Sample Statistic (X): How many observational units in our sample are in that category?

Summary of Inference for Population Proportion

- Sampling Distribution:
  - Distribution of values of sample statistic, across all possible samples of size n.
  - $X \sim \text{Binomial}(n, p)$
- Hypothesis tests:
  - Calculations based on what the sampling distribution would be if the null hypothesis was true.
  - Example: H_0: p = 0.2 vs. H_A: p > 0.2. Suppose $n = 40$ and $x = 14$.
  - p-value: Probability of getting a test statistic at least as extreme as what we observed in our sample, if H_0 is true. A small p-value is evidence against H_0.
  - If H_0 is true, then $X \sim \text{Binomial}(40, 0.2)$



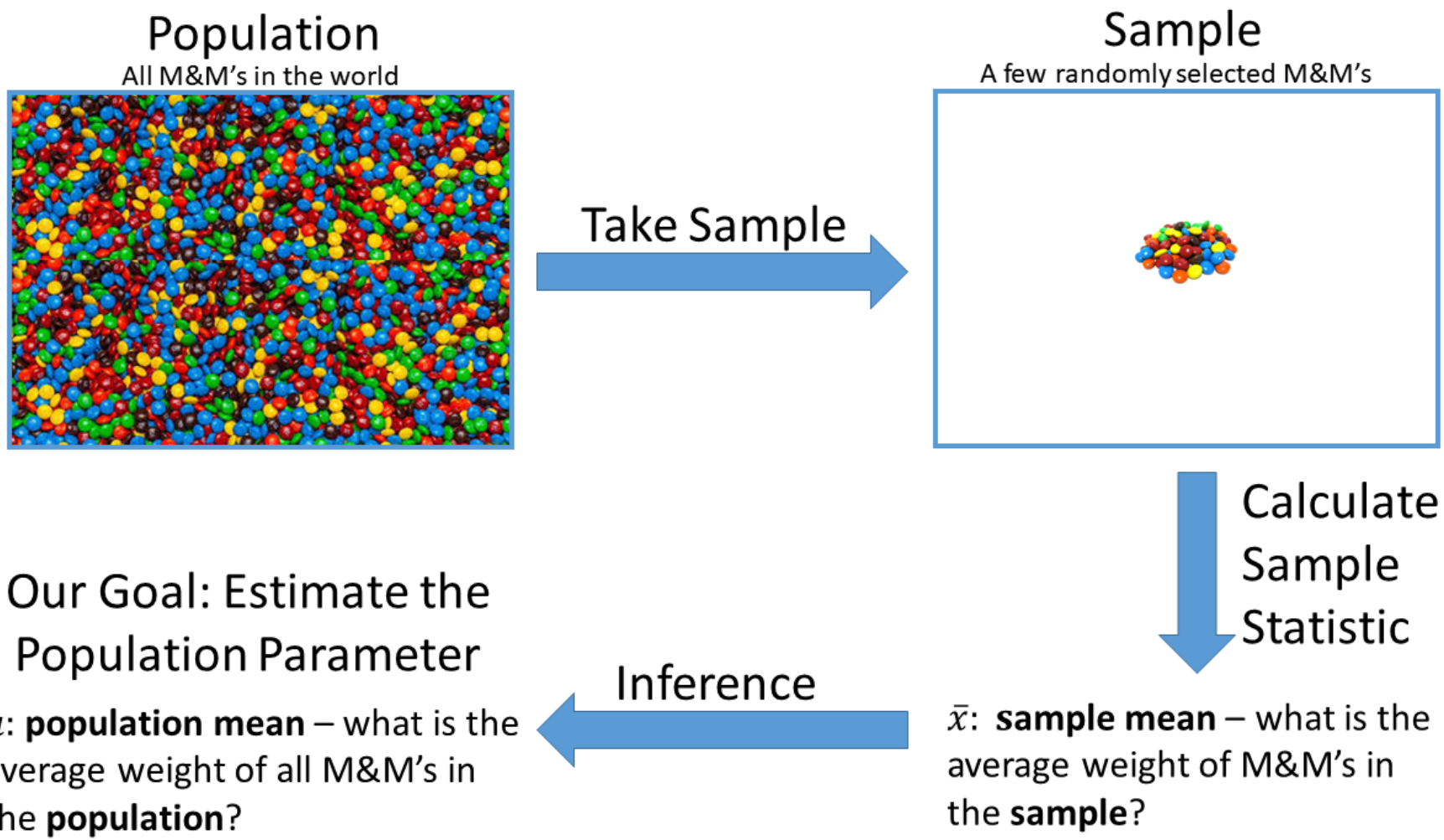
- Confidence Intervals:
  - Also based on the Binomial sampling distribution, but we did not explore the details.
  - If we take a lot of samples and calculate a different confidence interval based on each sample, about 95% of those confidence intervals will contain the true value of the population proportion.



The Course So Far: Describe Observations in a Sample, Inference for a Proportion

Variable Type(s)	Plot	Numeric Summary	Population Parameter	Sample Statistic	Sampling Distribution
1 Categorical	Bar	(Marginal) distribution	p (proportion in a category)	X (count in a category)	$X \sim \text{Binomial}(n, p)$
2 Categorical	Bar	Joint Distribution, Conditional Distribution			
1 Quantitative	Histogram or Density	mean, median, quantiles, standard deviation, variance, IQR			
1 Categorical, 1 Quantitative	Density or Box	summary statistics of the quantitative variable for each level of the categorical variable			
2 Quantitative	Scatter	correlation			

Next Up: Inference for the Mean of a Quantitative Variable



- Population Parameter (μ): What is the average value of a quantitative variable, across the whole population?
- Sample Statistic ($\bar{x}$ or $\hat{\mu}$): What is the average value of a quantitative variable, across the sample?

First Step: What is the sampling distribution of the sample mean?