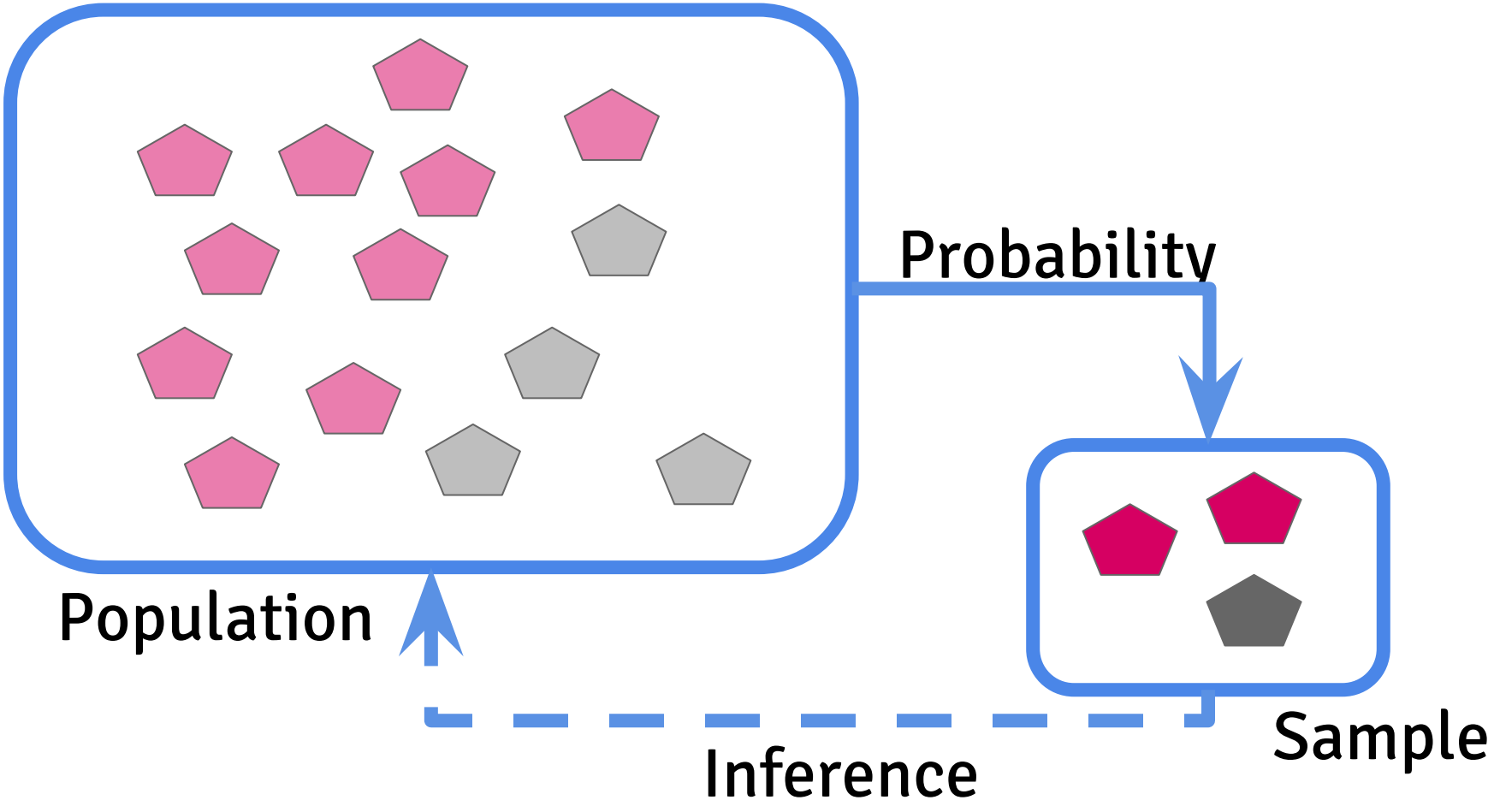
Representing data

Jeff Leek

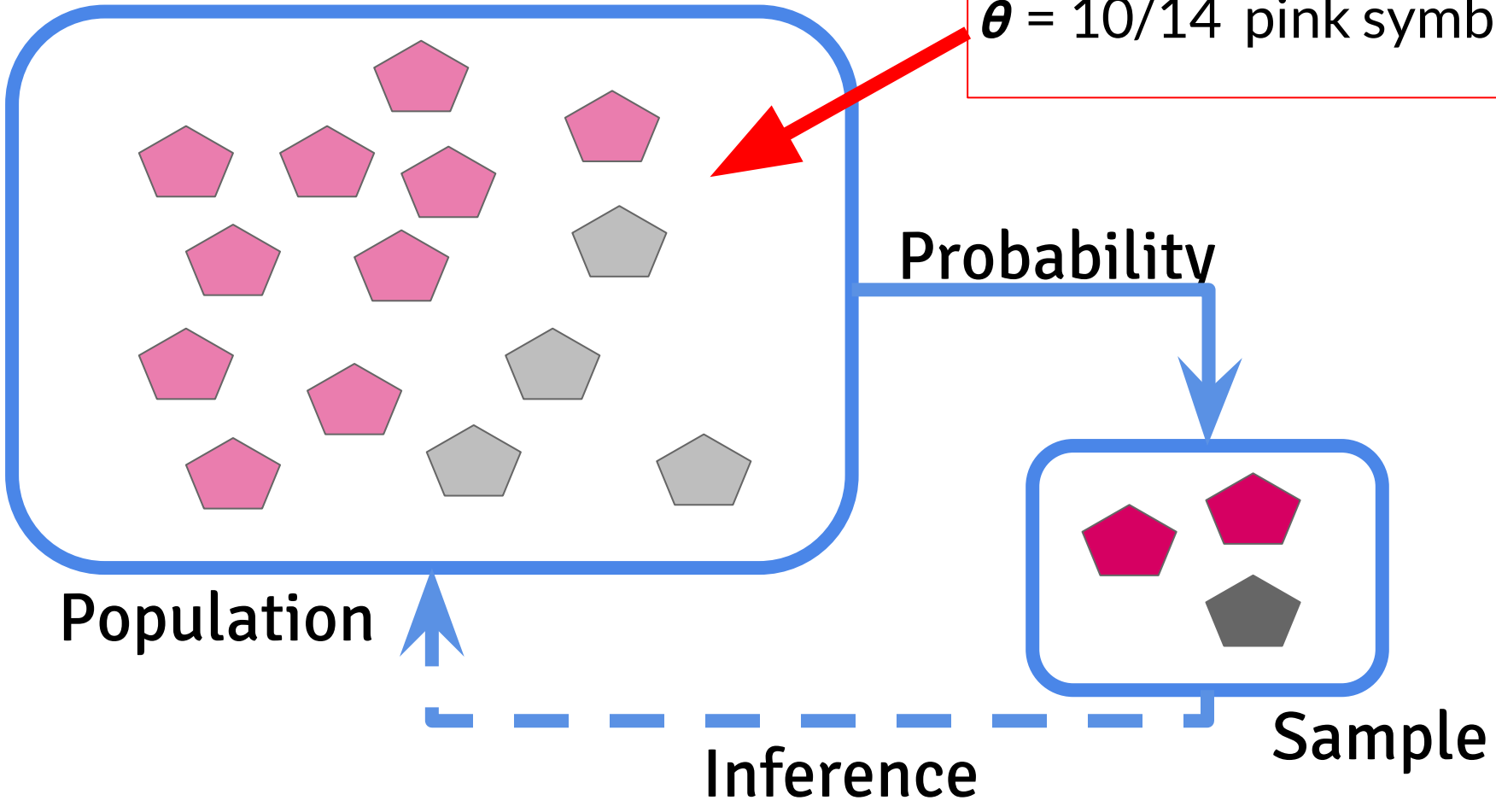
@jtleek

www.jtleek.com

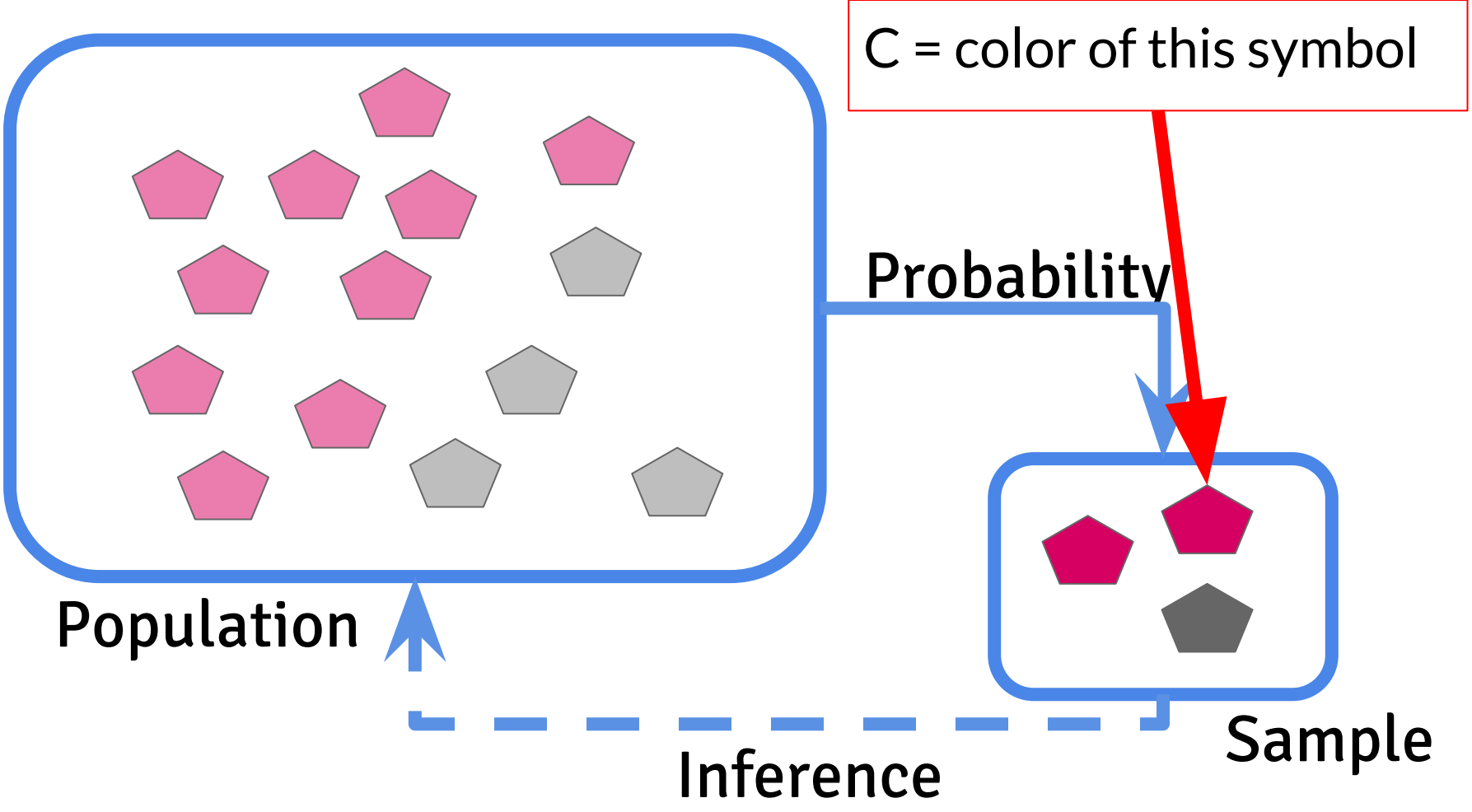
Central dogma of statistics


Probability
Population
Sample
Inference

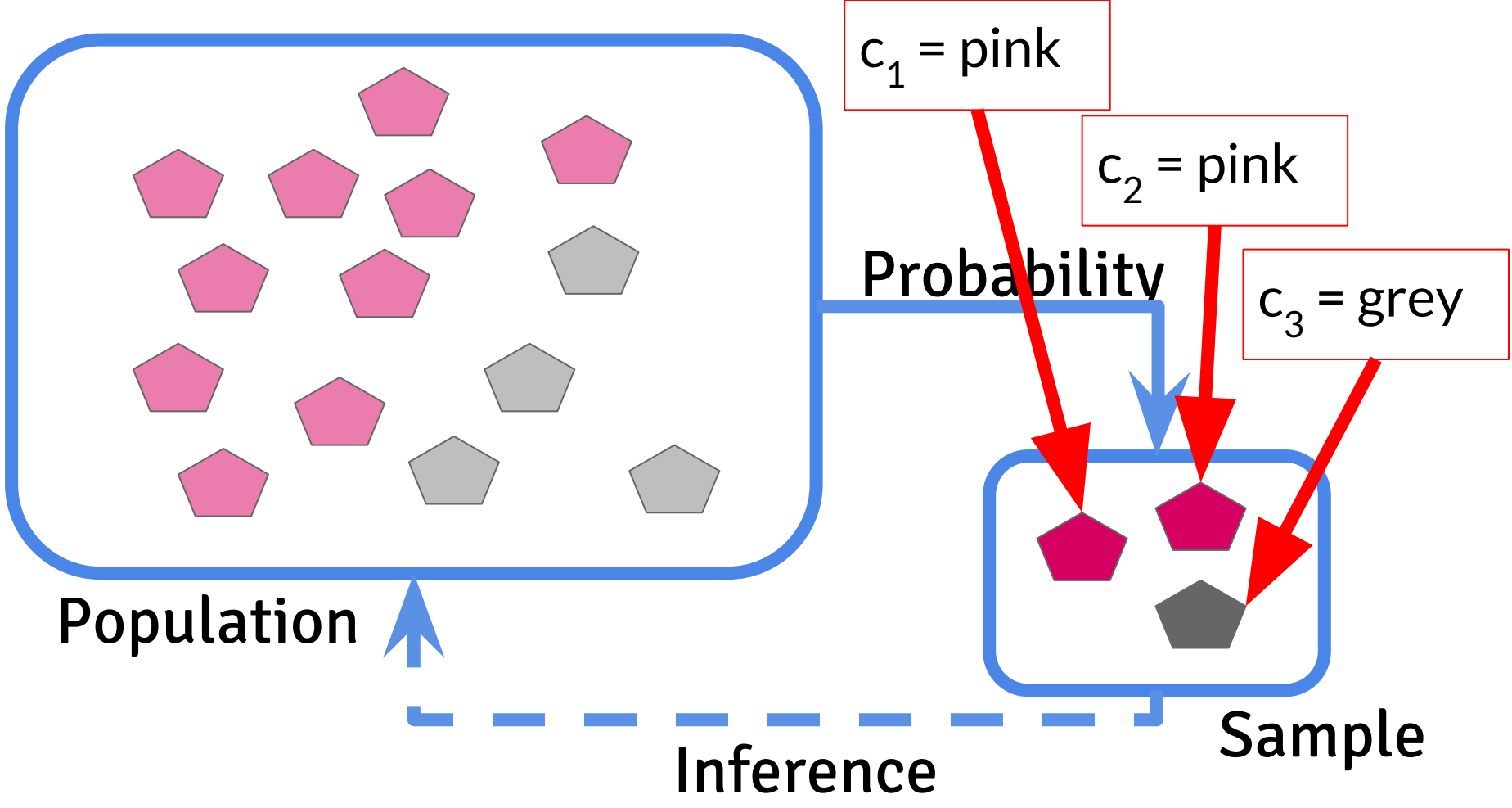
Parameters are characteristics of the population


θ = 10/14 pink symbols
Probability
Population
Sample
Inference

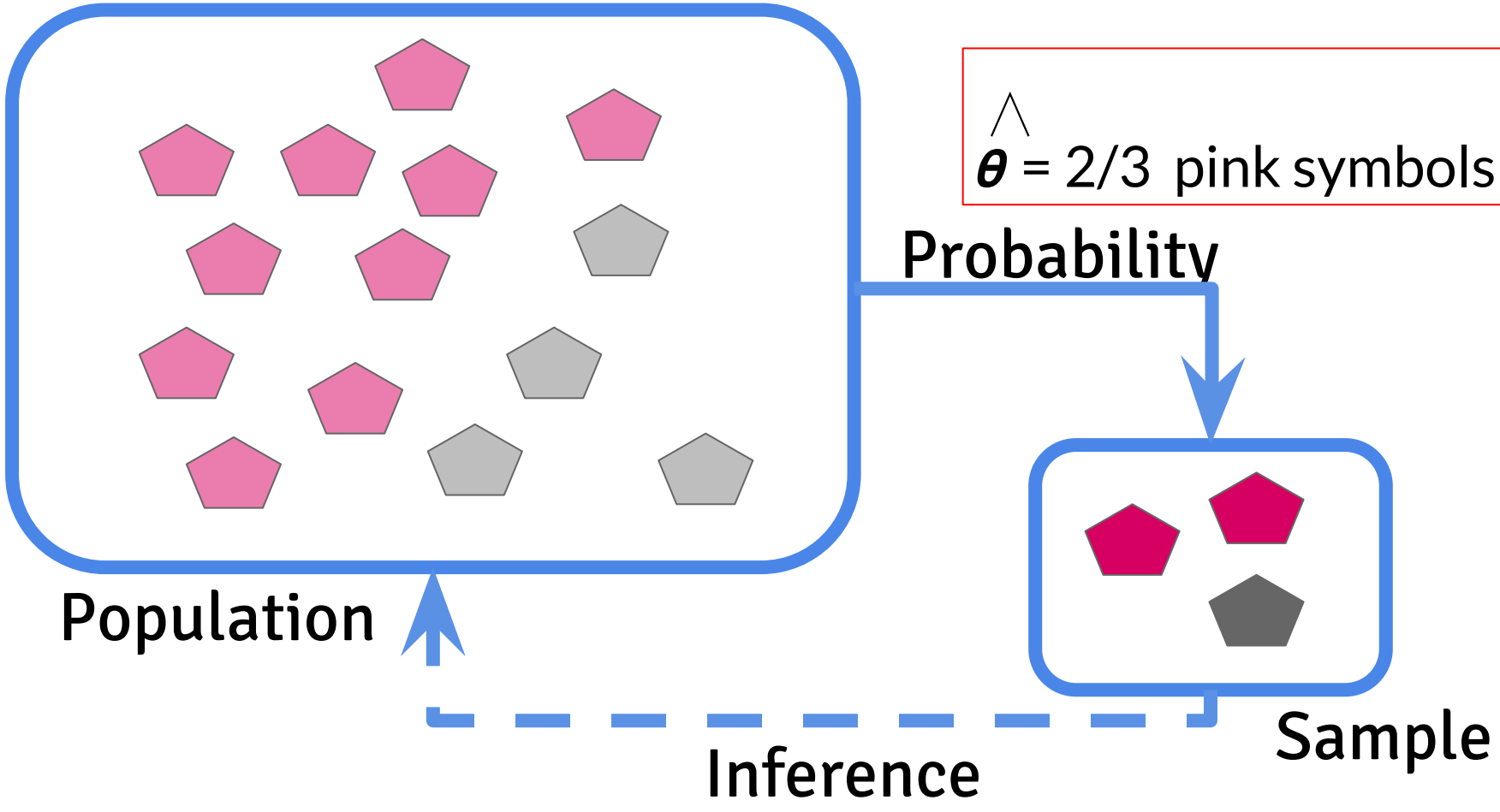
Data points are values we measure


C = color of this symbol
Probability
Population
Sample
Inference

Multiple values have subscripts


c_1 = pink
c_2 = pink
c_3 = grey
Probability
Population
Inference
Sample

We estimate population parameters with the data


$\hat{\theta}$ = 2/3 pink symbols
Probability
Population
Sample
Inference

Summary of notation conventions

- Data points are represented by letters
  - H for height, W for weight, C for count etc.
- Subscripts are used for different data points
  - C_1, C_2, C_3 are the counts for people 1,2,3
- Sometimes we write X for all values
  - X_1, X_2, X_3 are the counts for people 1,2,3
- We may need another subscript
  - X_{11} for the count for gene 1 on person 1

- Parameters are Greek letters
  - θ is average height in population
- Hats are used for estimates
  - $\hat{\theta}$ is our estimate of average height in population
- Y is usually outcome, X is usually covariate