ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 10: Sampling

William Marsh

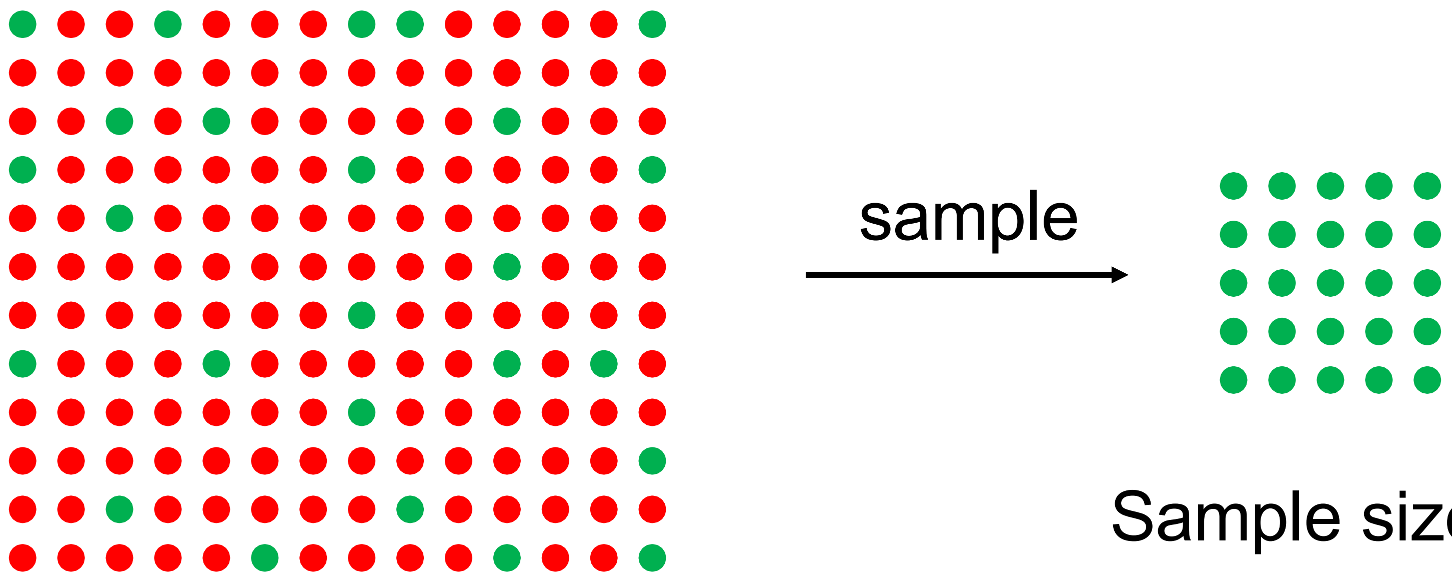
Outline

- Aim: Understand the difference between a population and a sample and potential problems of sampling

- Populations, samples and uncertainty
- Statistical inference: estimate & uncertainty
  - Estimation: Maximum likelihood example
  - Sample distribution simulation
- Sample mean and variance
- Central limits theorem

What is a Sample?

Population and Sample



Population size N

Sample size n

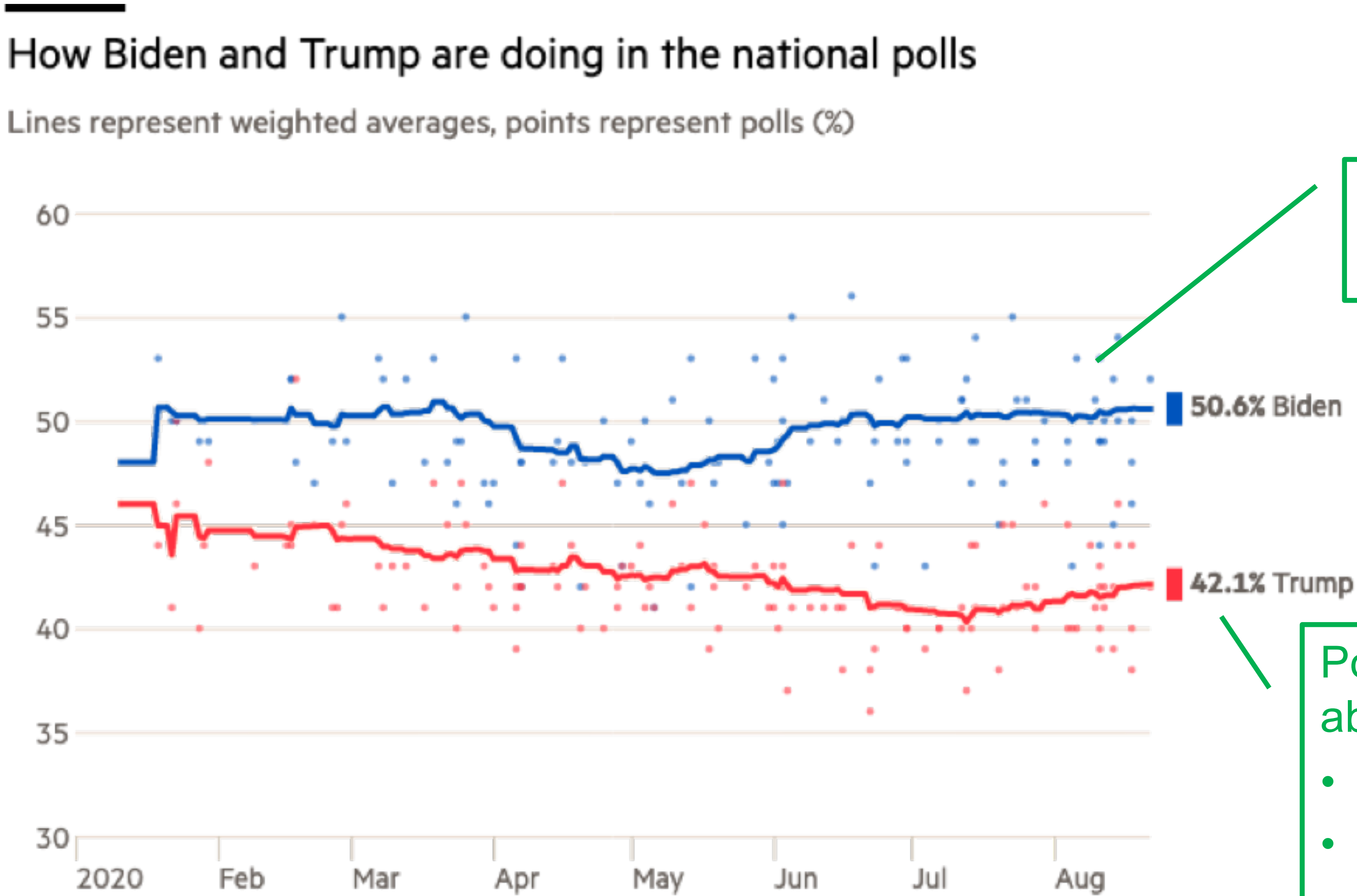
- Sample from a population
- Measure the sample (e.g. political preference)
- Statistical inference about population

Types of Sampling

- Random
  - Select people randomly
- Stratified
  - Strata: people with common characteristics (e.g. age range)
  - Sampling from each strata

- Bias
  - Bias is a systematic (cf. random) error
  - Sample bias: sample does not represent population

Chance and Bias



Chance: poll results vary

Pollsters worry about bias:

- Demographics
- Location
- Truth

Source (26/8) https://ig.ft.com/us-election-2020/

Household Survey of Covid-19 (England)

- 6 week period

(15 July to 25 August)		
	Number testing positive	**Sample size**
Participants	71	73,176
Tests	79	151,440
Households	68	36,348

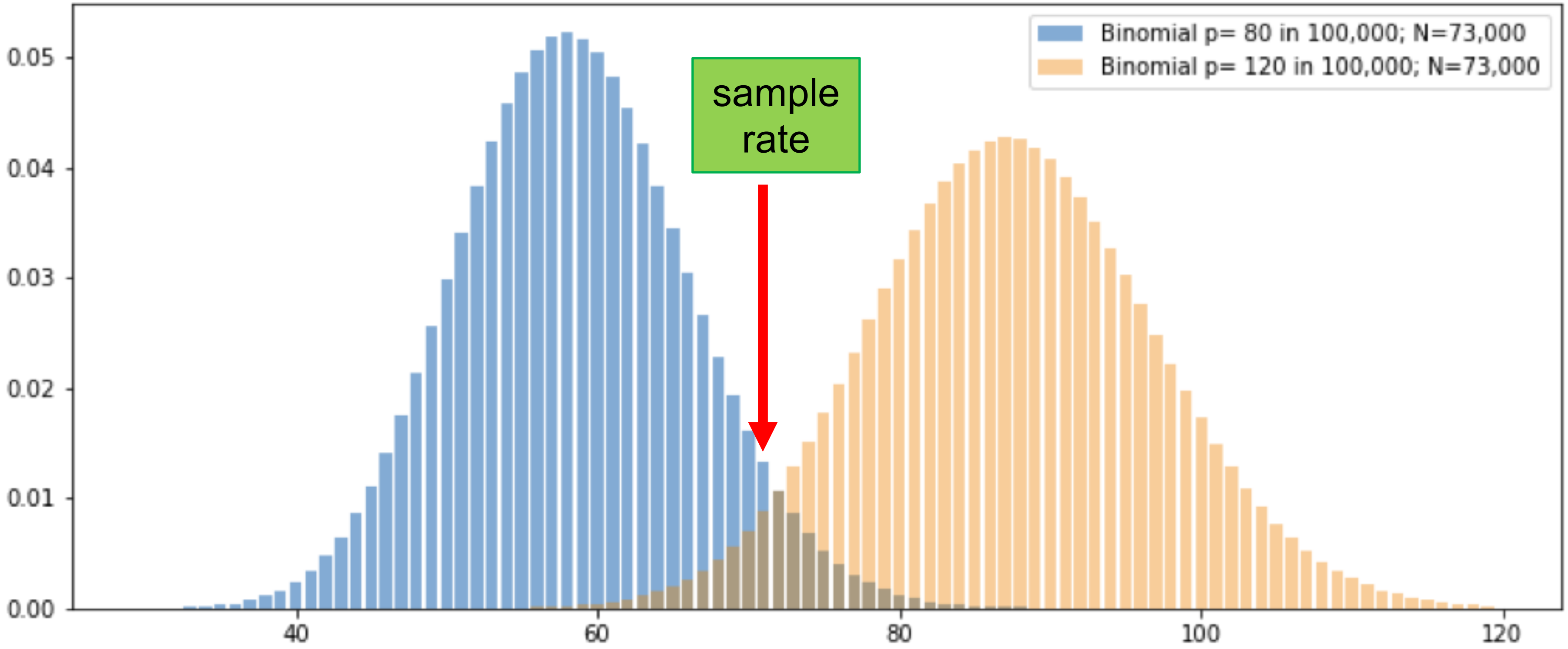
1. *Infections in private households; excludes hospitals, care homes etc*
2. *A small proportion of samples are excluded from this analysis due to missing age, sex or region data*
3. *This table is based on nose and throat swabs taken.*

Chance Problem

- Covid-19: 71 out of 73,176
  - Sample rate is 97 in 100,000)
- What is the impact of chance? Could the population rate be e.g.
  - lower: e.g. 80 in 100,000 or
  - higher: e.g. 120 in 100,000?

Chance Problem – II

- What is the impact of chance? Could the population rate be e.g.
  - lower: e.g. 80 in 100,000 or
  - higher: e.g. 120 in 100,000?



Statistical Inference: Two Problems

- Estimate a parameter from a sample
  - The mean and variance
  - A rate (probability in a binomial)
  - A regression coefficient

- Say how certain we can be that the estimate is near the true value (in the population)

Every lecture will have a ‘learning reflection’ slide

Going Deeper with Python Libraries

How to get a deeper understanding of Python libraries?

Deeper Understanding of Pandas

The Challenge

- Complex libraries
- Starting from examples
  - **What does it do?**
  - **Can I change it?**
  - **How do I do ... ?**
- Top-down learning

Techniques

- Documentation
  - *Best for Pandas*
- Looking at types
  - *What is the type of a column?*
- Error messages
- Evaluating sub-expressions

Estimates: Problem 1

Possible Approaches (Include)

- Unbiased estimate
  - Equally likely to be too low and too high
- Maximum likelihood
  - Makes probability of data highest

- *Others*

Maximum Likelihood Example

- Bernoulli trial
  - Data: n successes in N flips
  - Parameter: success probability p
- Estimate we have used is $\hat{p}$ = n / N

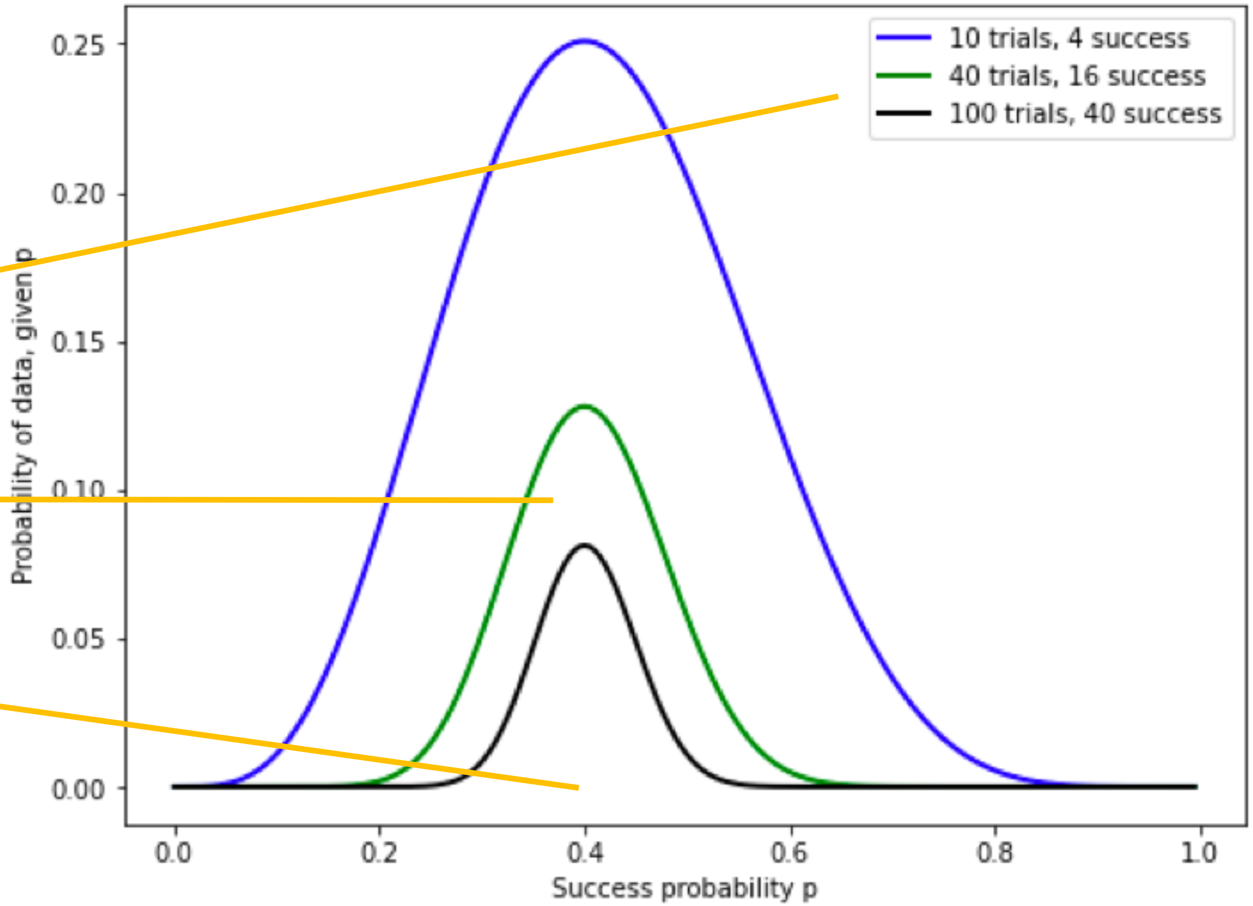
- Likelihood(p) = Pr(Data | p)

$$\mathcal{L}(p) = [\text{Constant}]\, p^n (1-p)^{(N-n)}$$

Note: the probability of any data low

Graph of $\mathcal{L}(p)$ (Bernoulli Trial)

- Can be shown that $\hat{p}$ = n / N is the estimate that maximises the probability of the data
  - Called an MLE
  - Also unbiased



All curves have same $\hat{p}$ (i.e. n/N)

More trials: less spread (smaller probabilities)

Maximum likelihood for n / N

Sampling Distribution

Tackling problem 2:

Using simulation

Simulation Concept

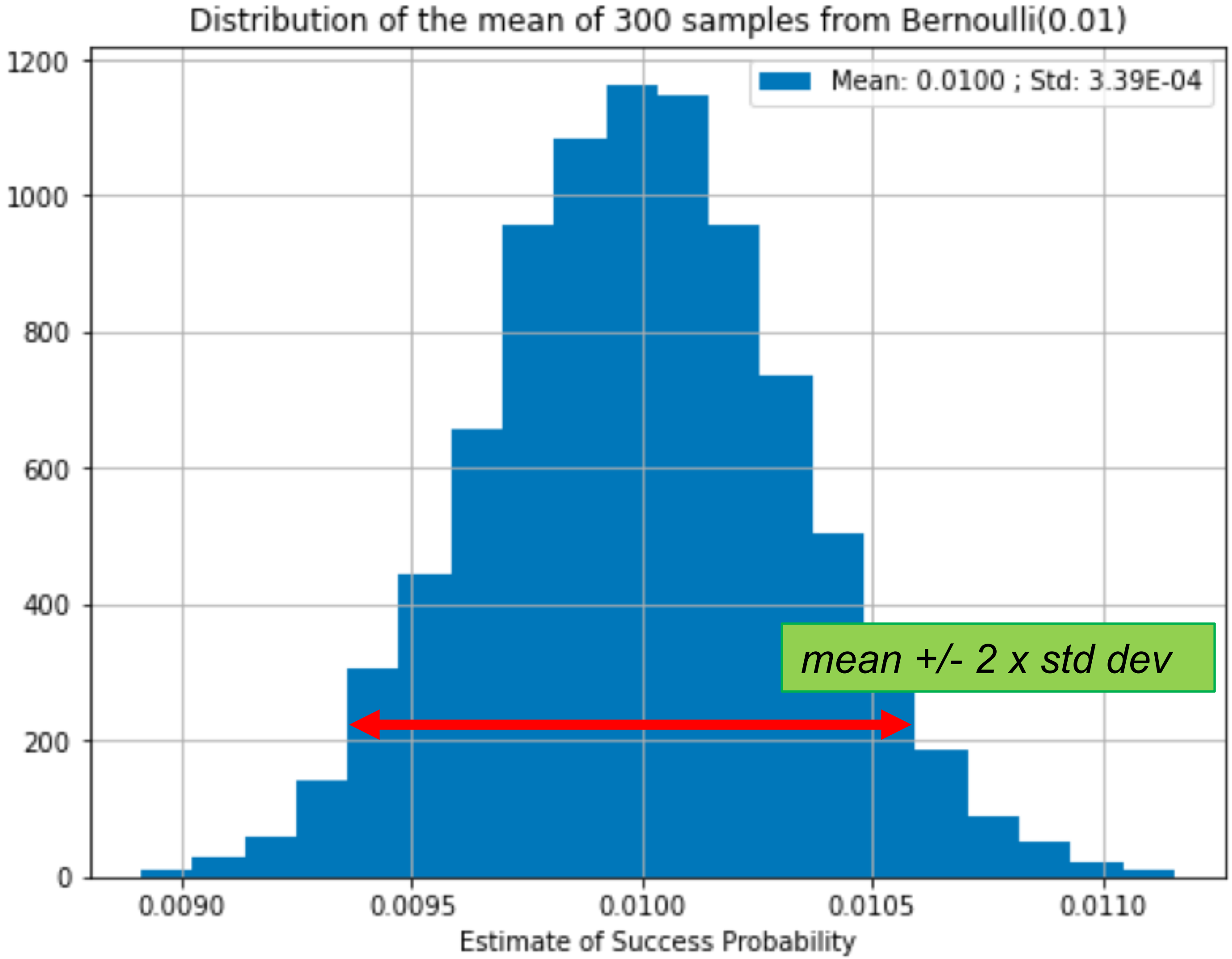
- Bernoulli trial – find ‘p’ by sampling

- Set up:
  - Assume population probability p is 1%
  - Repeatedly** take a sample N=300
    - success $\sim$ Binomial(N, p)
    - Estimate $\hat{p}$ (sample rate) by success / N
  - Look at the distribution of the sample rate

  - *****repeated 5,000 times***

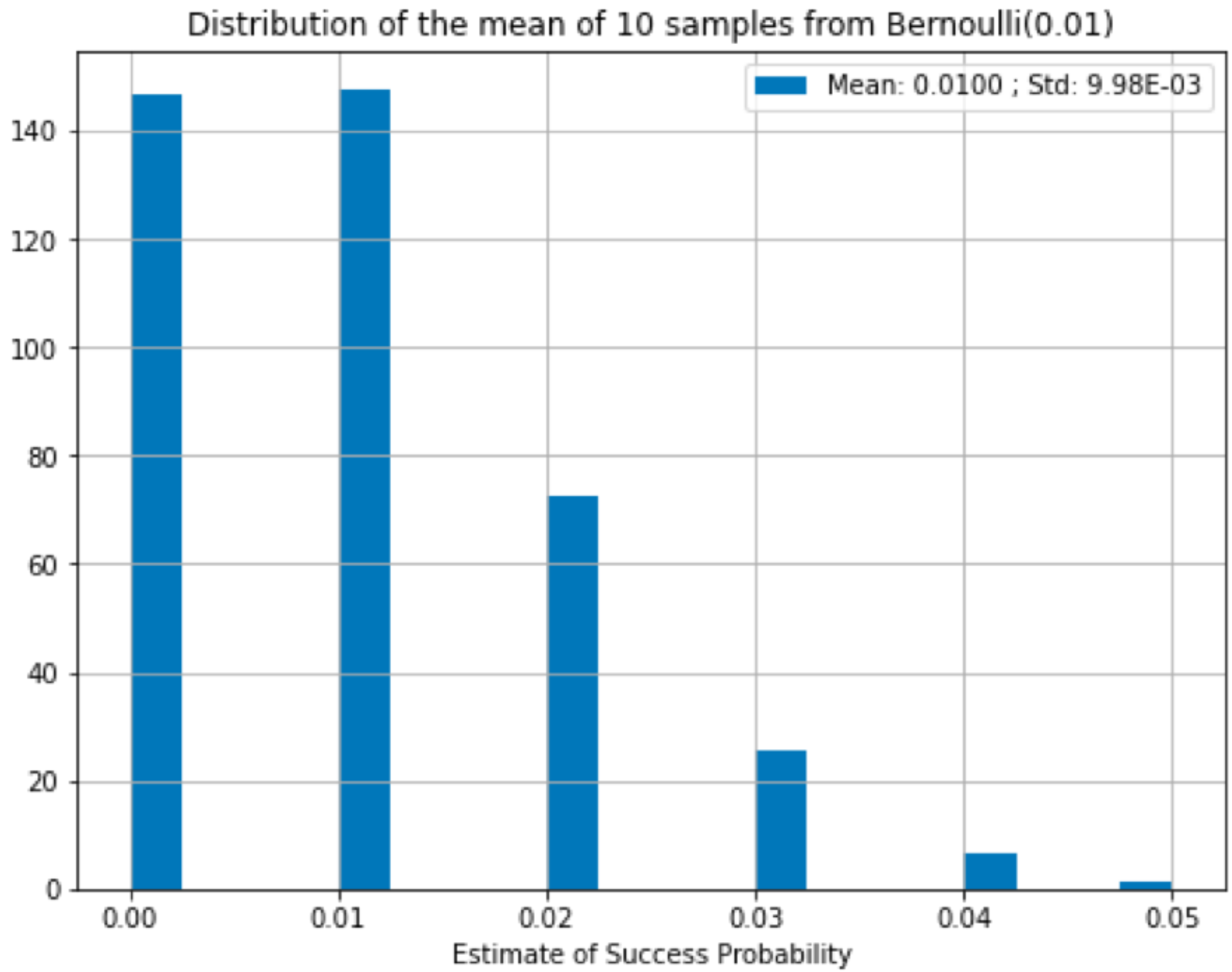
Distribution of $\hat{p}$

- Means estimate is 1%
- Range is 0.93% to 1.07%



Smaller Sample?

- Range of estimate increases



<u>Simulation</u> versus <u>Reality</u>

- Population rate known
- Repeated sampling
- Simulate sample rates

- Sample rate known
- Sample once
- Infer populate rate

We do not get exact results from a sample. We need a way to estimate the uncertainty in the sample results

Different Approaches

- Sampling distribution and confidence intervals
- Computational approaches (bootstrap)
- Bayesian inference
  - calculate Pr(parameter | data)

Quiz

Sample Mean and Variance

Sample and Population Statistics I

- Mean
  - Population mean is unknown: μ
  - Sample mean is calculated: $\bar{x}$
- $\bar{x}$ is an estimate of μ
  - Unbiased: population mean just as likely to be > or < the sample mean

Sample and Population Statistics II

- Standard deviation
  - Population standard deviation is unknown: $\boldsymbol{\sigma}$
  - Sample standard deviation is calculated: s
  - s estimates $\boldsymbol{\sigma}$
- Sample deviation calculated with (n-1):

$$s^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

Sample mean

Degrees of freedom: n-1 not n

If we have samples 1 to (n-1) and the sample mean then we know that last sample value

Central Limit Theorem

Central Limit Theorem

- Sample a population with mean μ and standard deviation σ
  - If the sample is sufficiently large then
  - ... the distribution of the sample means will be approximately normally distributed.

- Holds even if the source population is skewed
  - ... provided the sample size is sufficiently large

- *We can use the normal distribution to quantify uncertainty when inferring a population mean from the sample mean*

How Many Sample?

- In general, standard deviation (breadth) of the distribution of the sample mean is:

$$\frac{\sigma}{\sqrt{n}}$$

Population standard deviation:

- The larger it is, the **wider** the distribution of the sample mean

The number of samples:

- The larger it is, the **smaller** the distribution of the sample mean

Summary

- Usually measure only a <u>sample</u> of a population
  - Random sampling avoid possible ‘sample bias’

- Estimate properties of population (a ‘statistic’ e.g. a mean) from the sample

- Uncertainty about true value of population statistic
  - Reduces as the sample size increases
  - Proportional to population standard deviation

$$\boxed{\frac{\sigma}{\sqrt{n}}}$$

- We have illustrated the problem: need a practical solution for a single sample