UG2 Semester 1 survey

Fill in by 17 January 2025
to enter a draw to win
one of two £25 vouchers

https://forms.office.com/e/Rd25UeYKec

Inf2 - Foundations of Data Science: Randomness, sampling and simulation - Sampling, statistics, simulations

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

So far...

1. Intro to inferential stats
   - Estimation
   - Hypothesis testing
   - Comparing two samples (A/B testing)

2. Two examples of inference on coins
   - Estimate the average year of a coin
     - we have an estimate, but we don't know how precise it is
   - Test the hypothesis that the coins are unbiased
     - we think the coins are unbiased, but we can't prove it

Today

- Big idea: method to determine if the coin is biased: Statistical simulation of what we expect to happen if the coin isn't biased

- Steps:
  1. sampling, both random and non-random
  2. definition of a "statistic"
  3. statistical simulation

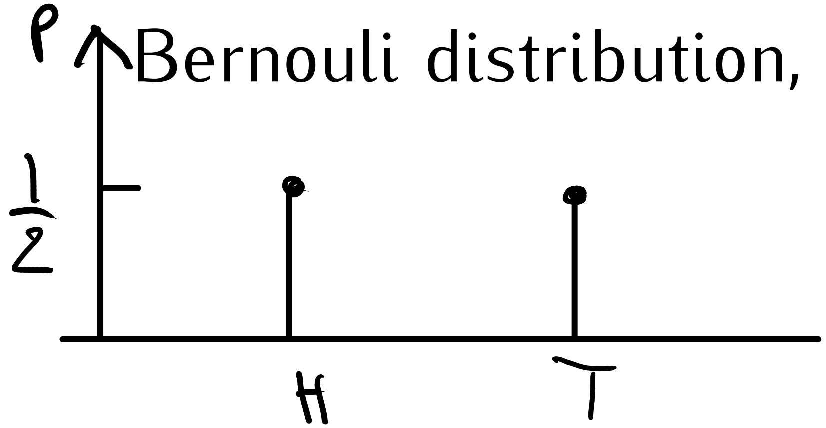
- Then get intuition about what happens as sample size changes
  1. distribution of statistics from small samples
  2. distribution of statistics from large samples

Statistical simulation overview

Reality

Model of unbiased coin

Bernouli distribution, p=1/2



Experiment

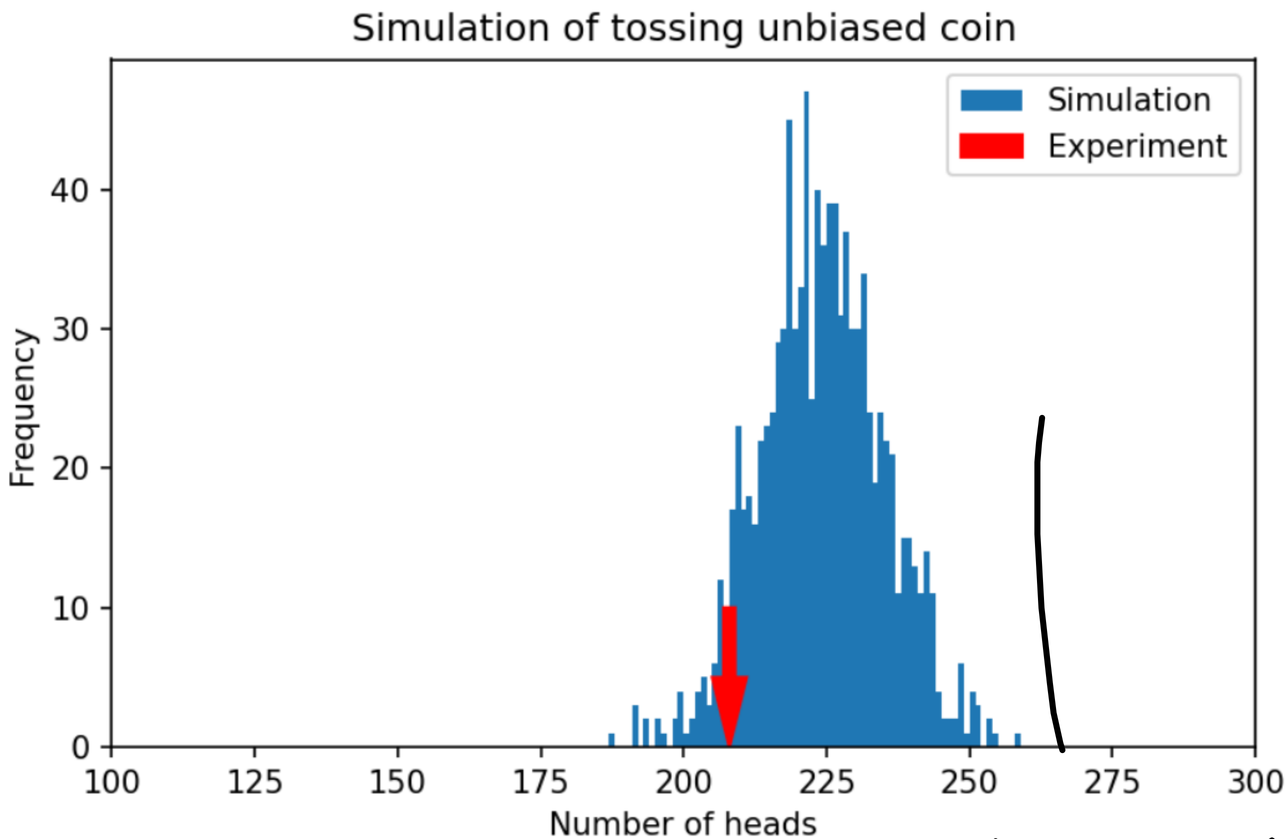
448 tosses, of which
208 Heads and 240 Tails

Computational simulation

448 samples, of which
220 Heads and 228 Tails
231 217

Statistical simulation overview

1000 repetitions later... consistent with experiment?



Data consistent with hypothesis of unbiased coin.

Definition of a random sample (Strictly, an "independent and identically distributed" (iid) random sample)

In a random sample of size n from either

- a probability distribution
- or a finite population of N items

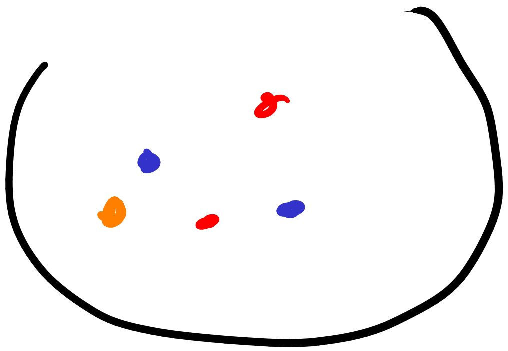
the random variables $X_1, \ldots, X_n$

comprising the sample are all

1. independent and
2. have the same probability distribution

Sampling from a finite population of discrete items without replacement

Discrete items

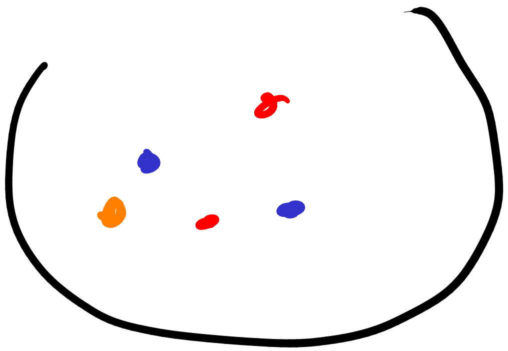


N = 5

n	Samples	Counts R	B	Y
3		1	1	1
3		0	2	1
5		2	2	1
5				

Sampling from a finite population of discrete items with replacement

Discrete items



N = 5

		Counts		
n	Samples	R	B	Y
3		0	1	2
3				
5				
5				

Questions

1. Is sampling with replacement an iid random sample?

2. Is sampling without replacement an iid random sample?

If $\frac{n}{N} < 0.05$, sampling without replacement is approximately a random sample

Why are random samples good?

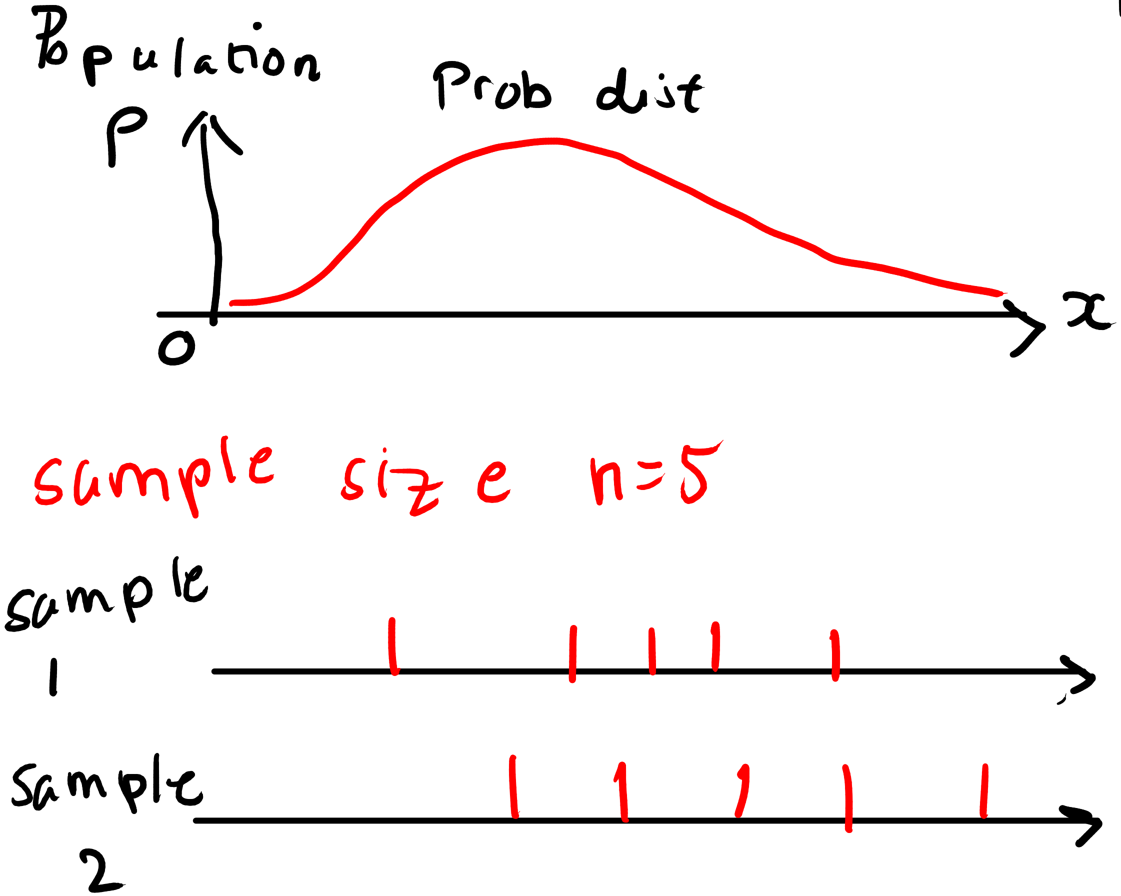
Consider non-random samples

Day	$
Mon	100
Tue	120
Wed	130
Thu	140
Fri	150
Sat	130
Sun	120
Mon	100
⋮	⋮

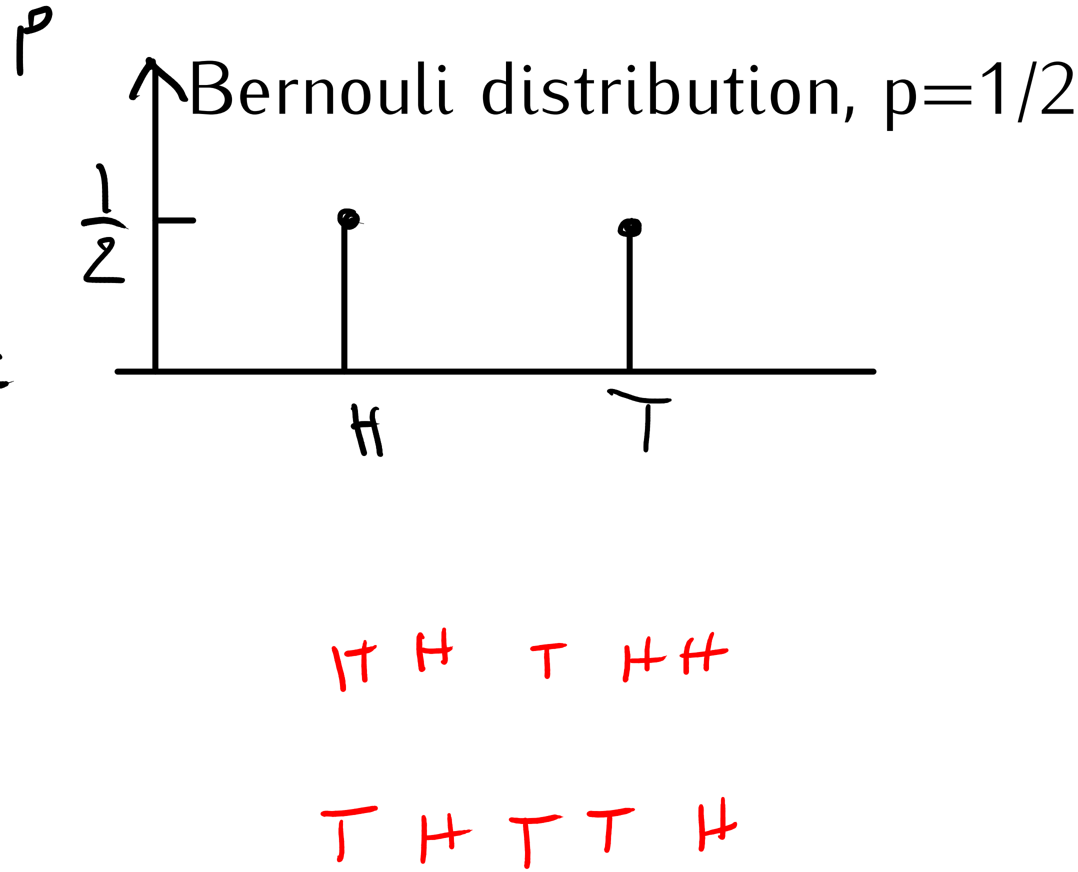
Non-random samples can be biased

Sampling from a probability distribution

Continuous



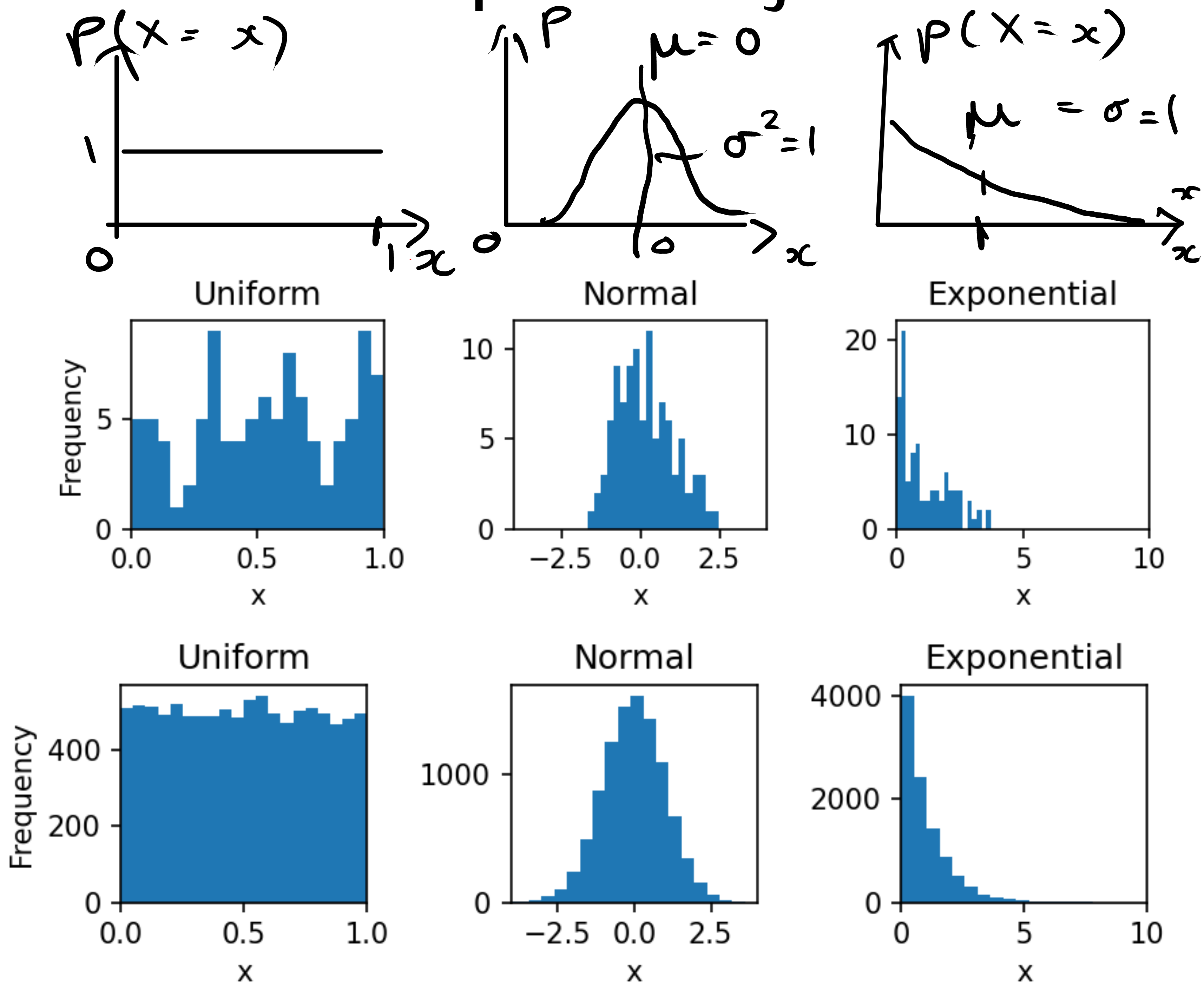
Discrete



Sampling from continuous probability distributions

100 samples

10000 samples



RNGs

Definition of a statistic

A statistic is any quantity whose value can be calculated from sample data

Example: Number of heads from sequence of coin tosses

Treat statistics from simulations as random variables and denote with upper case: H

Denote observed sample statistic with lower case: h

Recipe for a statistical simulation

A. Decide on

- Statistic of interest H num. heads
- Population distribution or set of items   Bernouli $p = 0.5$
- Sample size $n = 448$
- Number of repetitions $k = 1000$

B. Simulation procedure

1. For i in $1, \ldots, k$
   a. Sample n items from the population distribution or set
   b. Compute and store statistic of interest

2. Generate histogram of the k stored sample statistics

Statistical simulation applied to Swain versus Alabama

8 out of 100 people selected for a jury panel were black
26% of population of Alabama were black
How do we simulate unbiased jury selection?

Statistic: T_0 # black people on panel of n=100 members

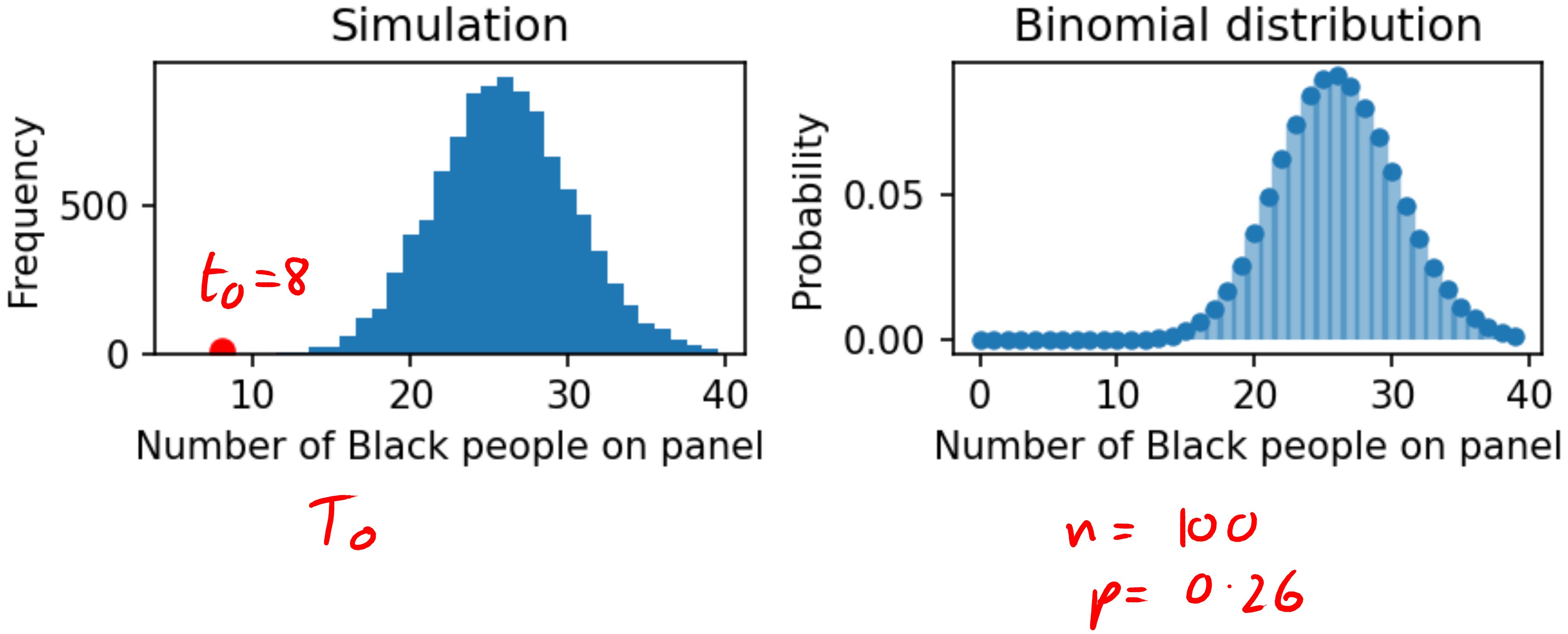
Pouplation: Bernouli dist with sample space {Black, White} p(Black) = 0.26

BBBB
WWWWWWWW
WWWW

Sample size: n = 100

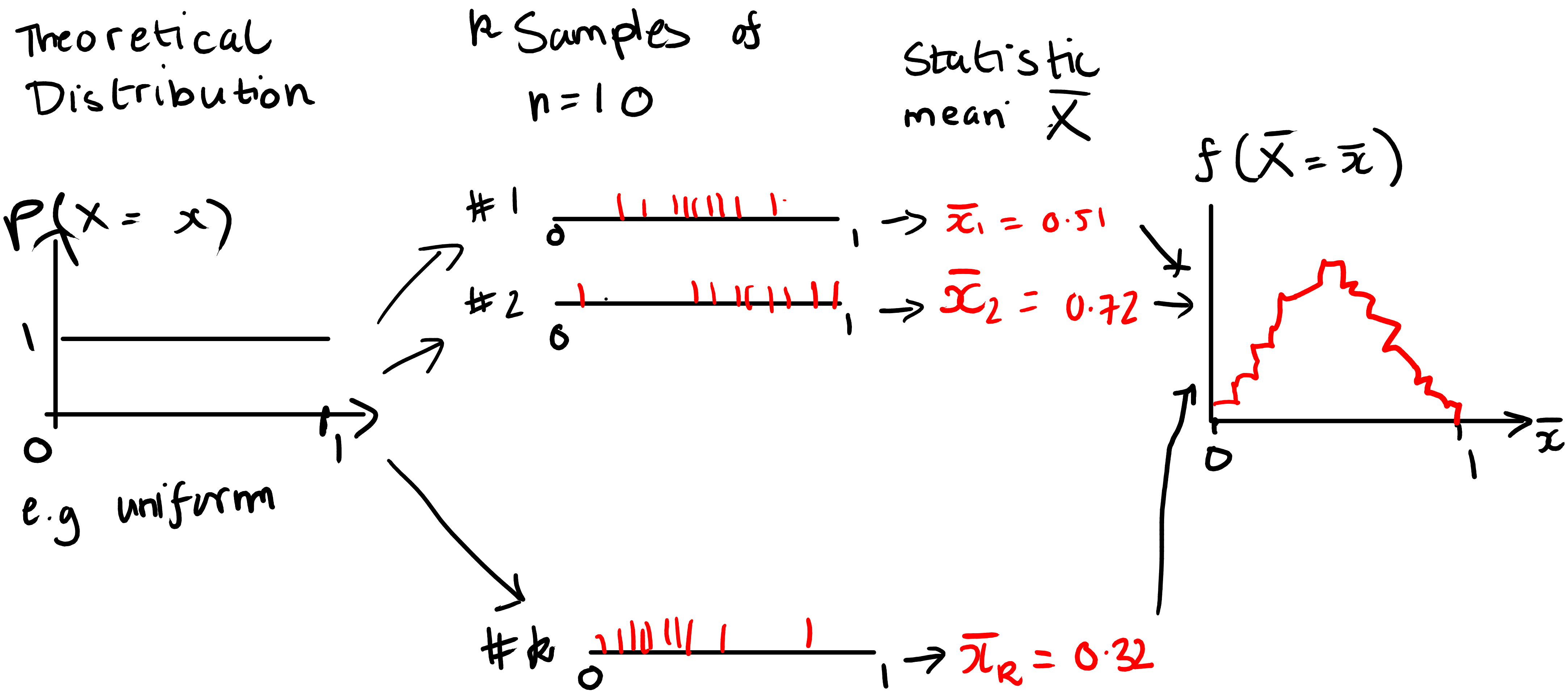
Num. repetitions: k = 10,000

Swain versus Alabama simulation results

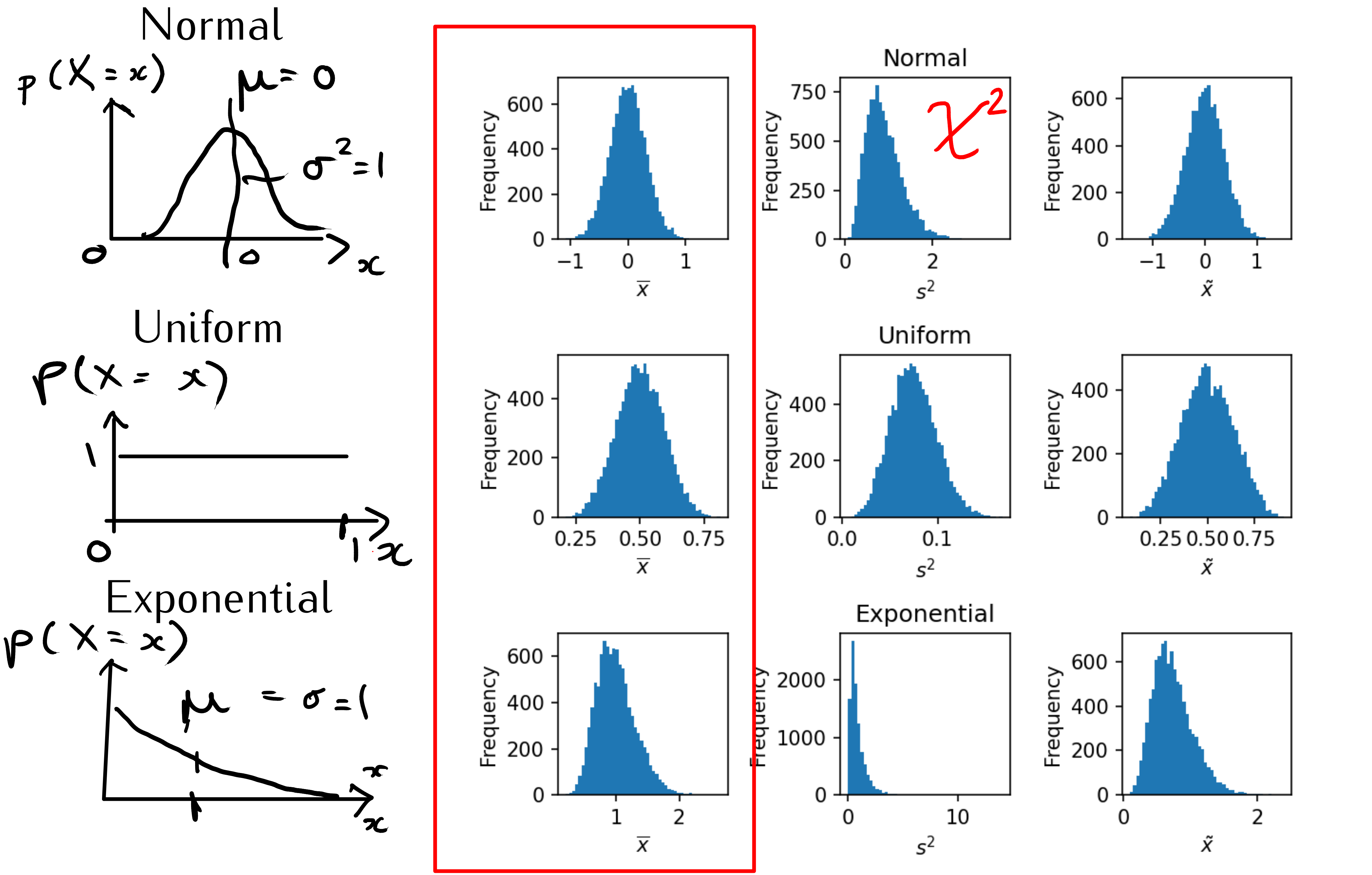


Inf2 - Foundations of Data Science: Randomness, sampling and simulation - Distributions of sample statistics from small samples

Example: Sampling statistics from continous distributions

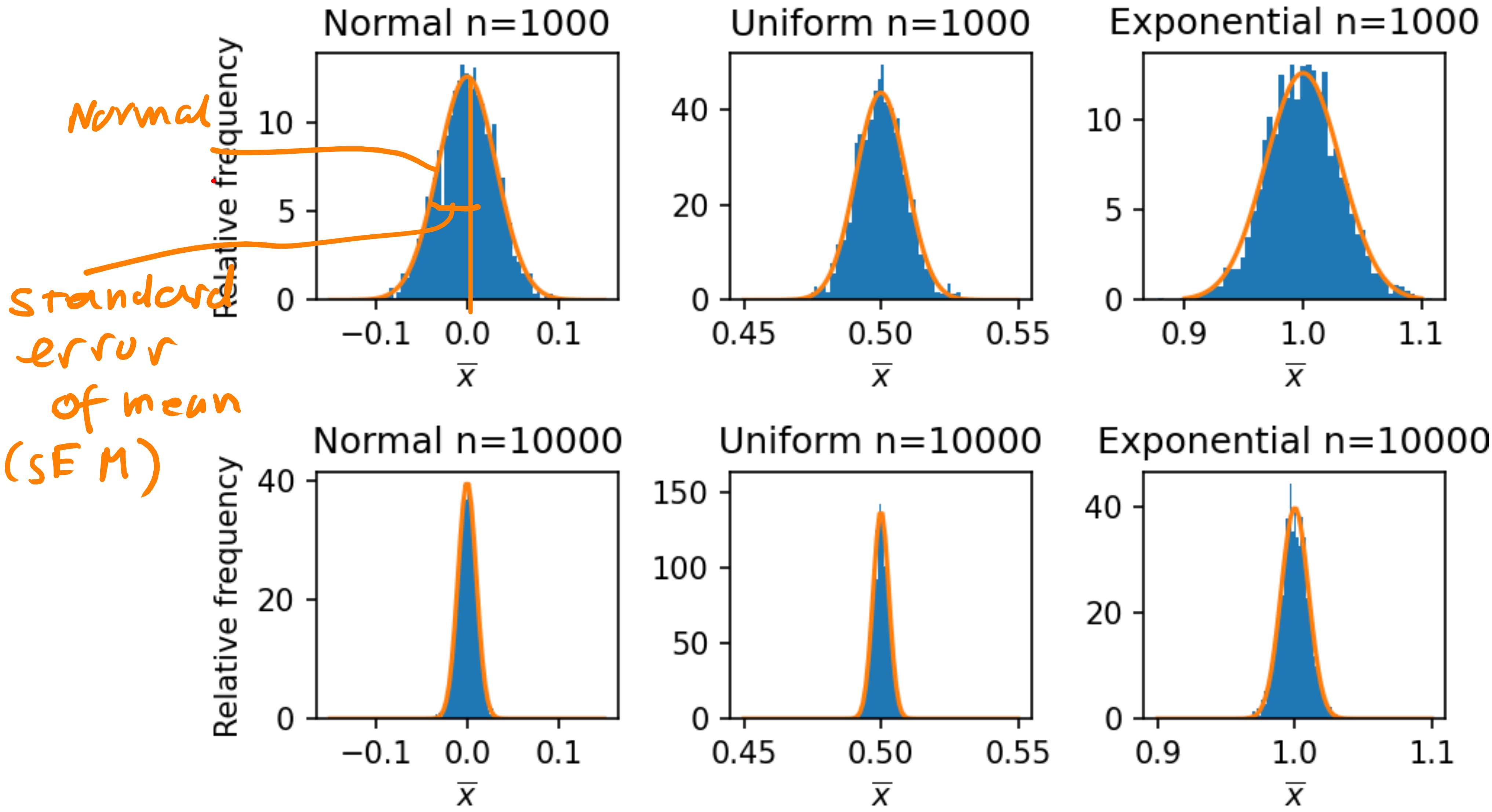


Distribution Mean Variance Median



Inf2 - Foundations of Data Science: Randomness, sampling and simulation - Distributions of sample statistics from large samples

Distribution of sample mean from large samples



Central Limit Theorem

Distribution of the mean (or the sum) of a random sample drawn from any distribution will converge on a normal distribution

If the population distribution mean is μ and variance is σ^2 and sample size is n then:

Expected value of sample mean is the same as the mean of the population distribution

$$\mu_{\bar{x}} = E[\bar{X}] = \mu$$

Expected variance of the mean

$$\sigma^2_{\bar{x}} = E[(\bar{X} - E[\bar{X}])^2] = \frac{\sigma^2}{n}$$

Standard error of the mean (SEM) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Law of large numbers

In the limit of infinite sample size n , the expected value of the sample mean $\bar{x}$ tends to the population mean μ and the expected value of the sample variance σ^2 tends to 0.

Not the same as the "law of averages" AKA "the gambler's fallacy".

Summary

- Statistical simulations
  - Sampling
  - Statistics

- Distributions of common statistics for small sample sizes

- Sampling distribution of the mean is normal for large samples from any distribution (Central Limit Theorem)

UG2 Semester 1 survey

Fill in by 17 January 2025
to enter a draw to win
one of two £25 vouchers

https://forms.office.com/e/Rd25UeYKec