

UG2 Semester 1 survey

Fill in by Friday 16 January 2026
to enter a draw to win
one of two £25 vouchers .



<https://edin.ac/47Z4zAU>



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
 - The Central Limit Theorem & Law of Large Numbers

Statistical simulation overview

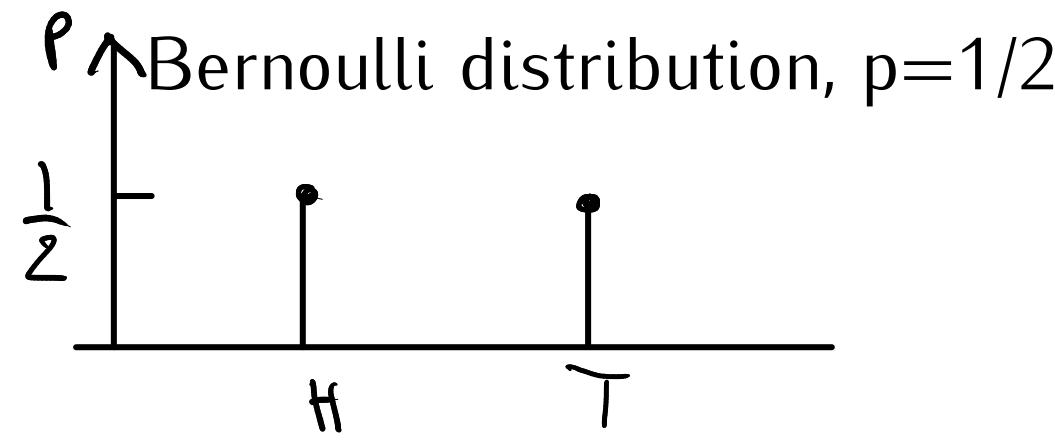
Reality



Experiment

67 students made
536 tosses, of which
264 Heads and 272 Tails

Model of unbiased coin



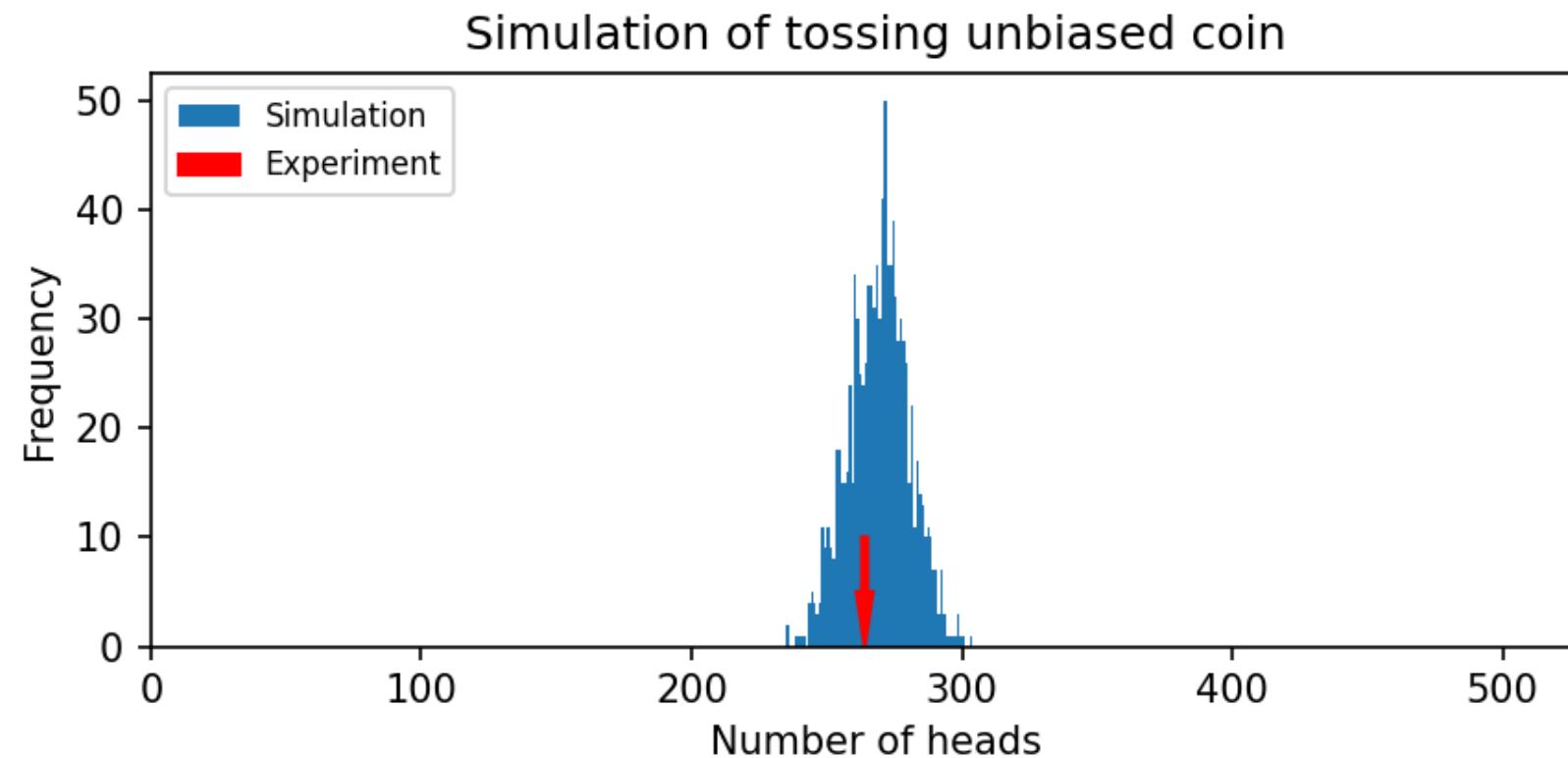
Computational simulation

536 samples, of which
Heads and Tails

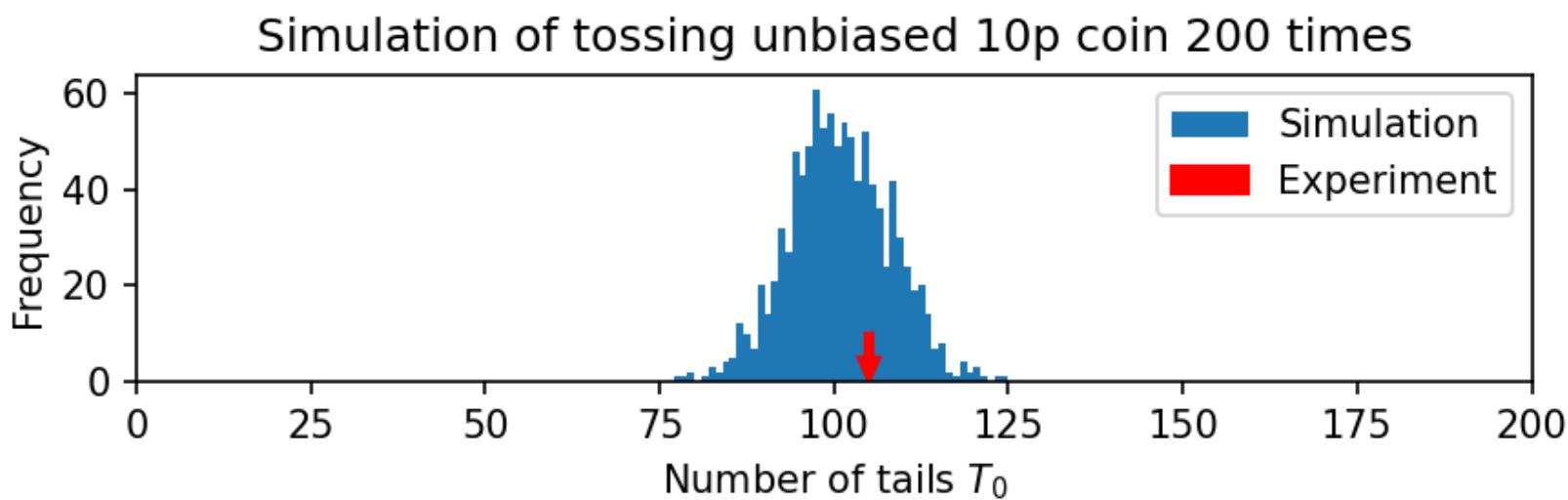
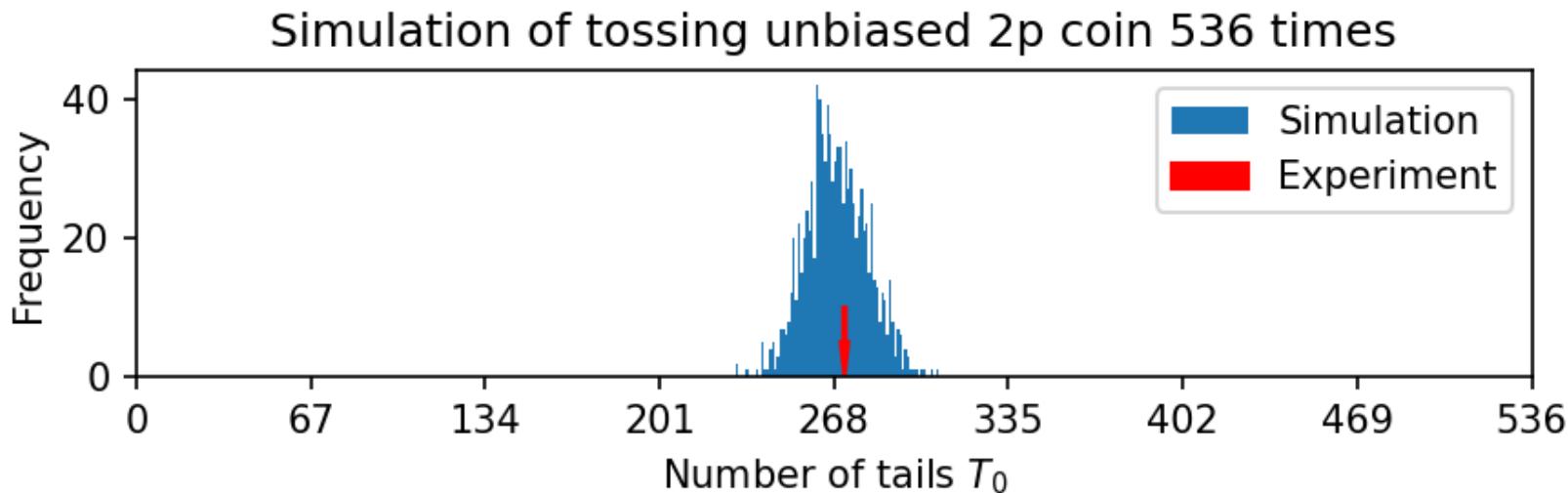


Statistical simulation results

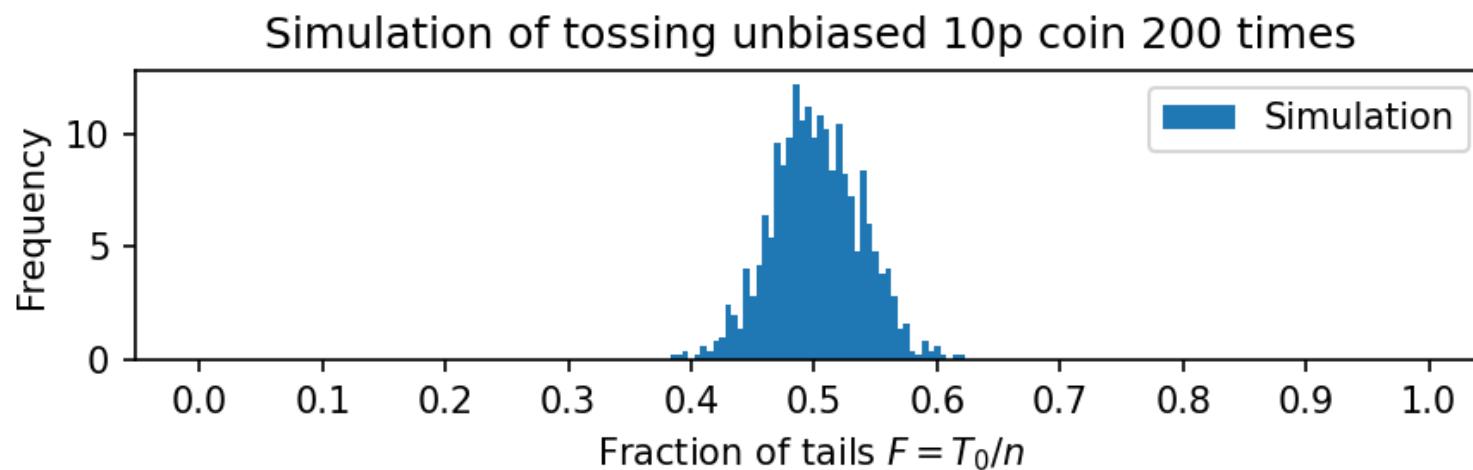
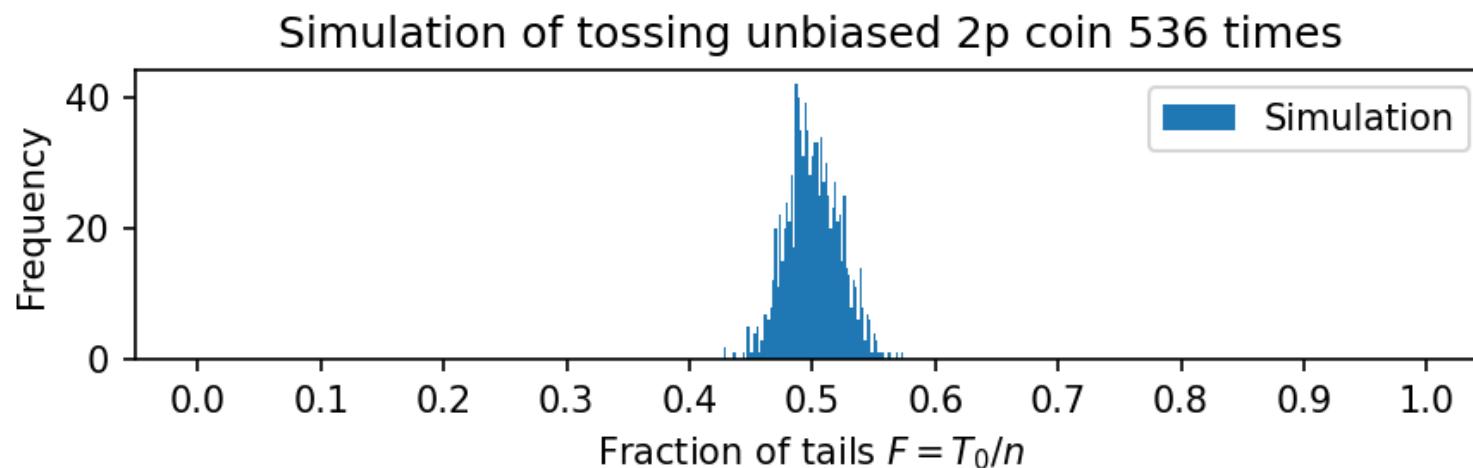
1000 repetitions later... consistent with experiment?



Statistical simulation results - for the 10p coins too



Plot fractions of tosses that are tails to help comparison

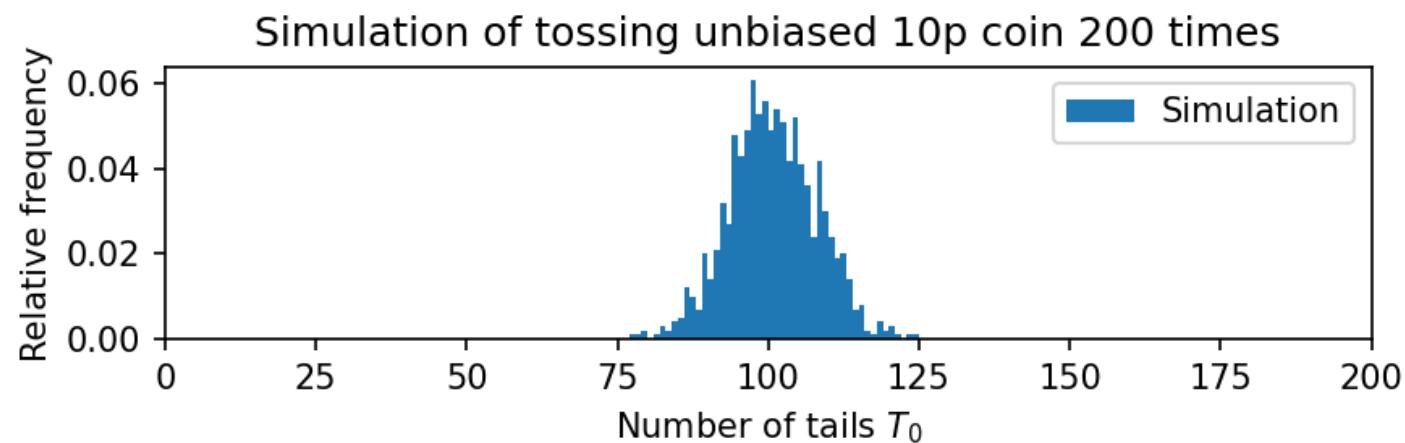
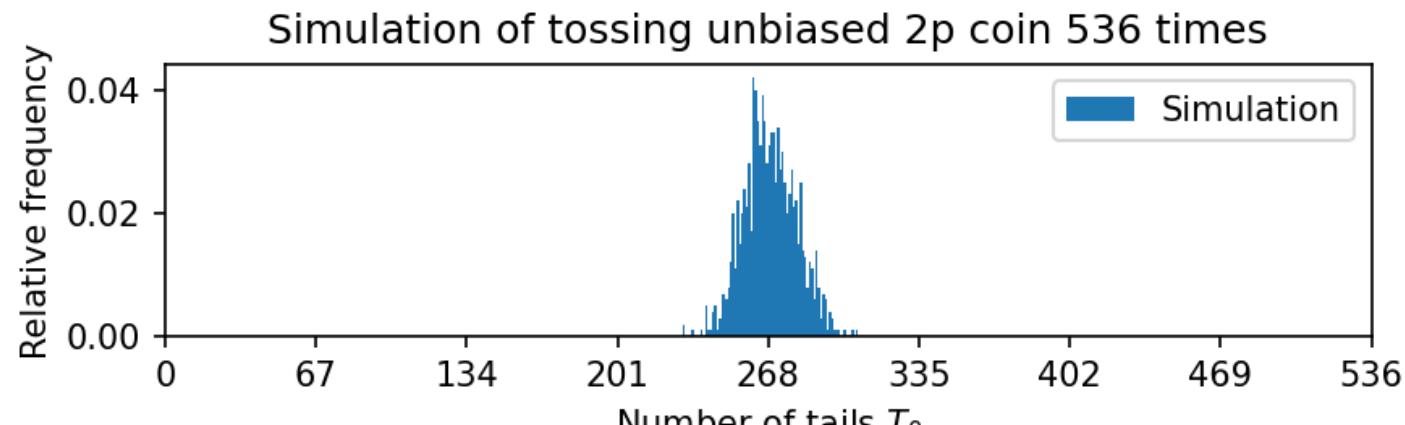


Are the sample means of the fraction of the number of tails in each simulation similar or different?

Are the sample variances of the fraction of the number of tails in each simulation similar or different?

Note on relative frequency

Relative frequency of event A = $\frac{\text{number of outcomes in which A occurs}}{\text{total of number of outcomes in experiment}}$



e.g. A = "275 Tails"

$$\Rightarrow f_A = \frac{n_A}{n} = \frac{25}{1000} = 0.025$$

Also called empirical probability

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



THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS
OF
DATA
SCIENCE

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

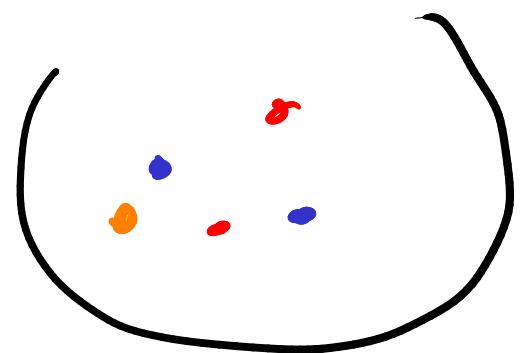
the random variables

comprising the sample are all

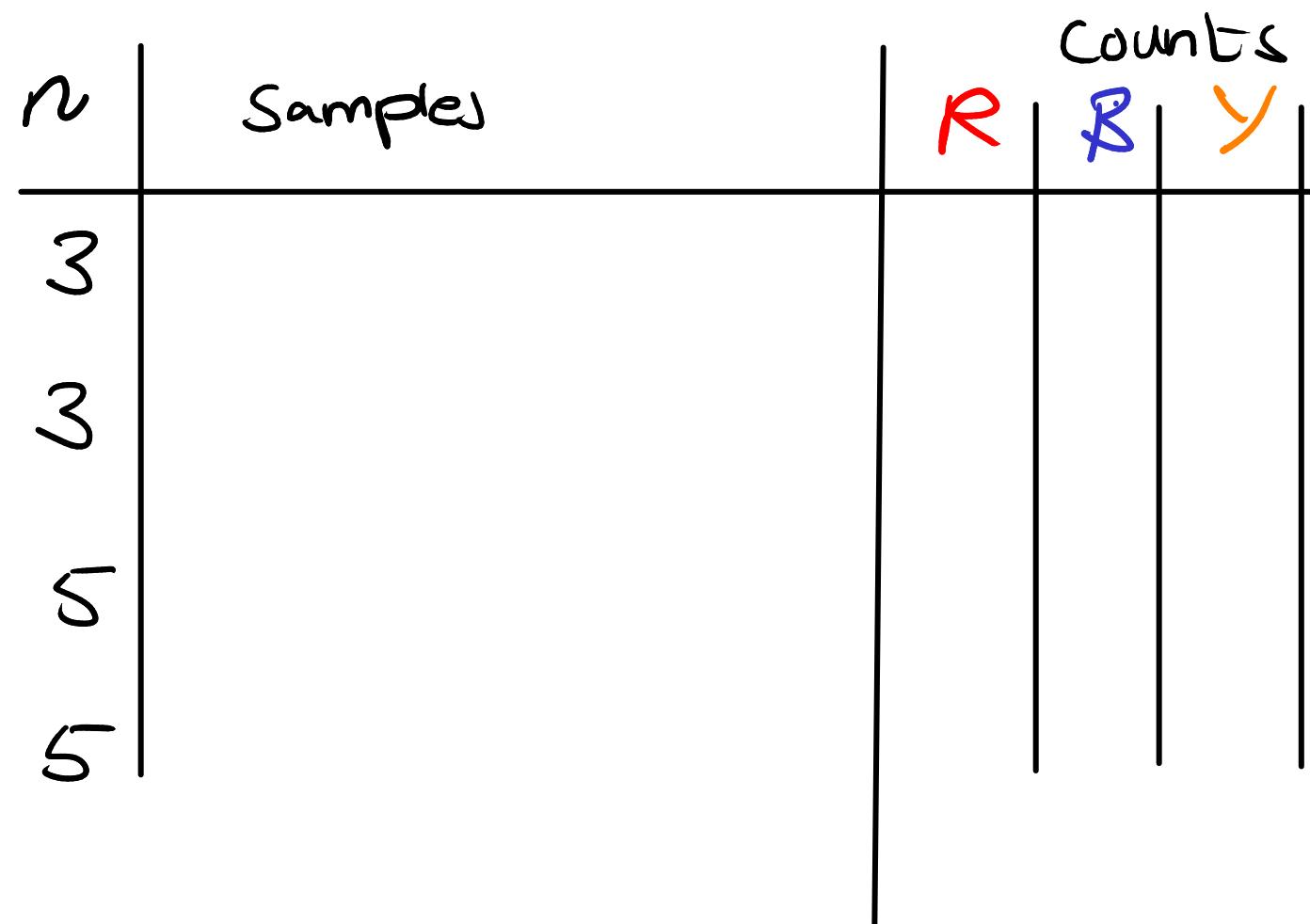
1. independent and
2. have the same probability distribution

Sampling from a finite population of discrete items without replacement

Discrete items n

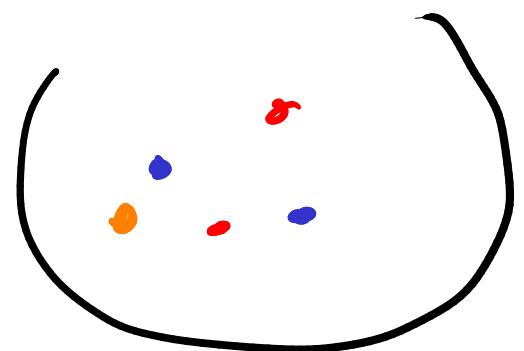


$$N = 5$$

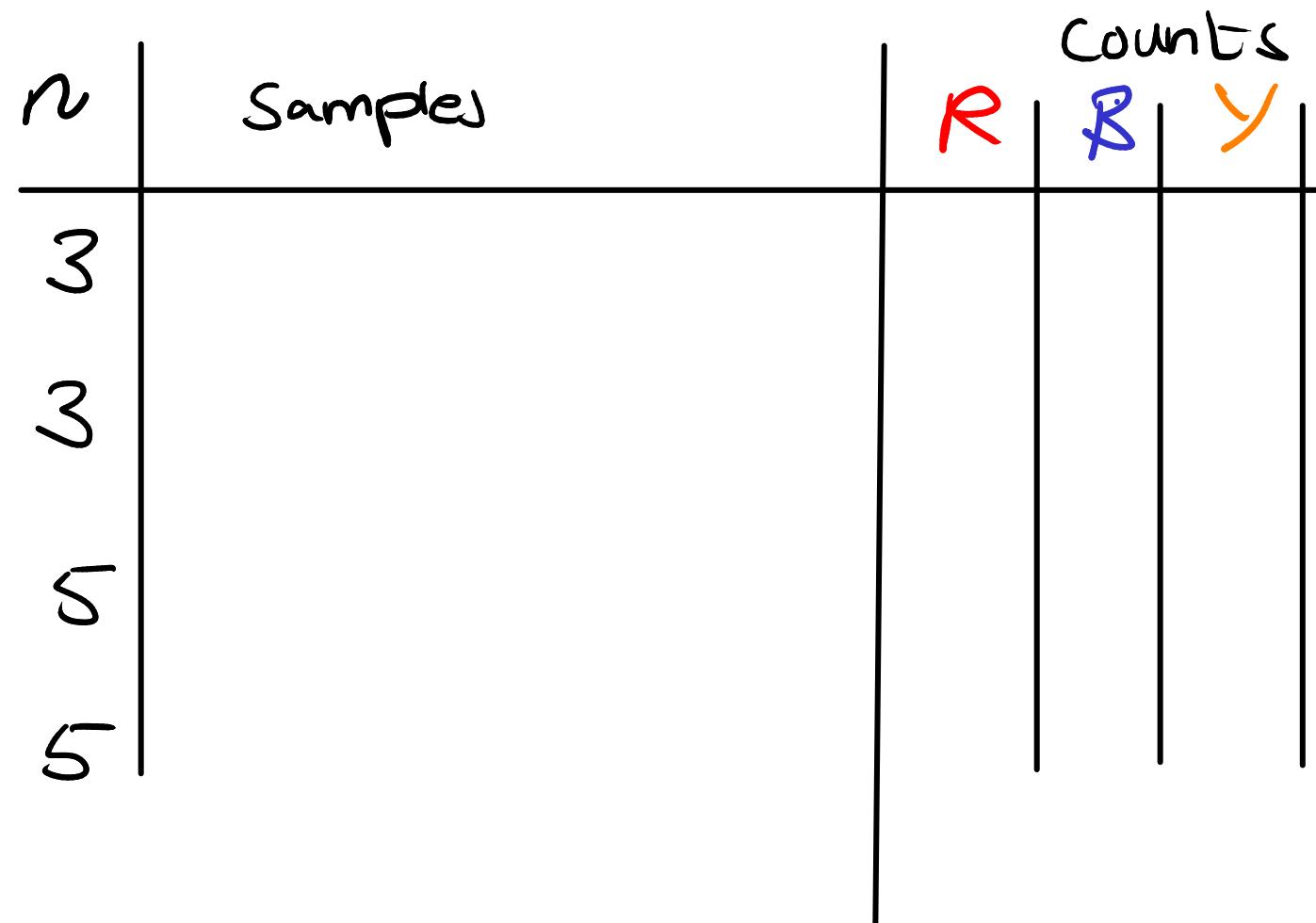


Sampling from a finite population of discrete items with replacement

Discrete items



$$N = 5$$



Practical sampling

For each set of S2 comprehension questions:

I am going to enter every consenting student who has done the questions by 12 noon of the day of the next lecture into a prize draw, and announce the winner in that day's lecture (unless they elect to remain anonymous).

Your probability of winning will be in proportion to the number of points you get on your last attempt of the questions (you can take the questions as many times as you want).

Prizes to be determined! (Though they won't be big.)

Questions

1. Is sampling with replacement an iid random sample?
2. Is sampling without replacement an iid random sample?

Why are random samples good?

Consider non-random samples

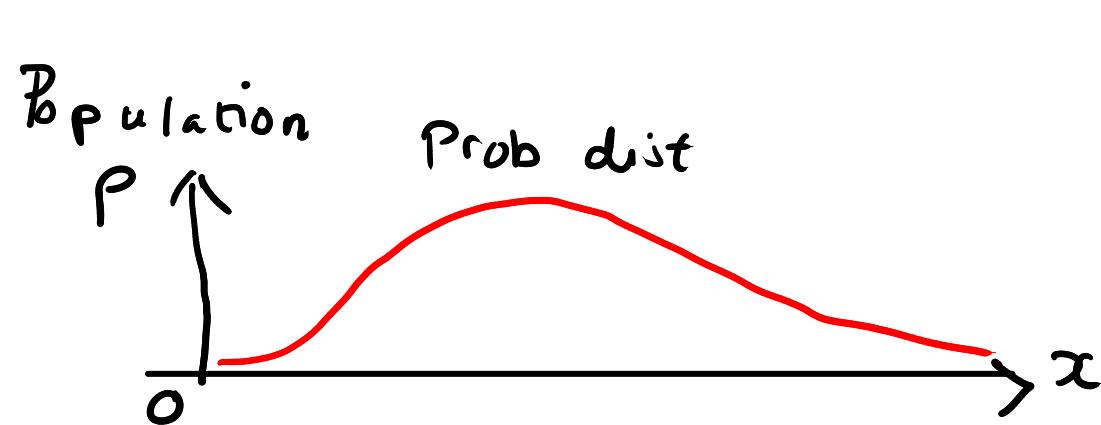
Day	\$
Mon	100
Tue	120
Wed	130
Thu	140
Fri	150
Sat	130
Sun	120
Mon	100
!	!



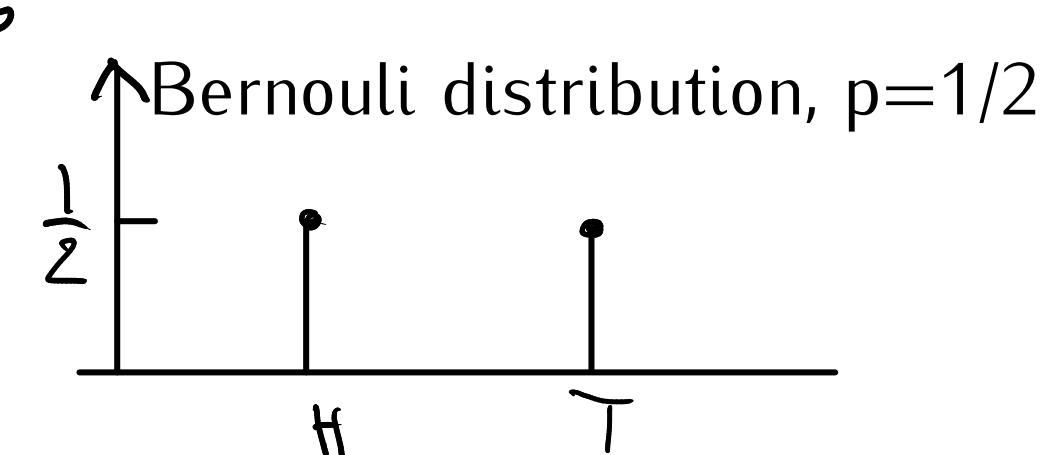
Mstyslav Chernov, Wikimedia Commons, CC BY SA 3.0

Sampling from a probability distribution

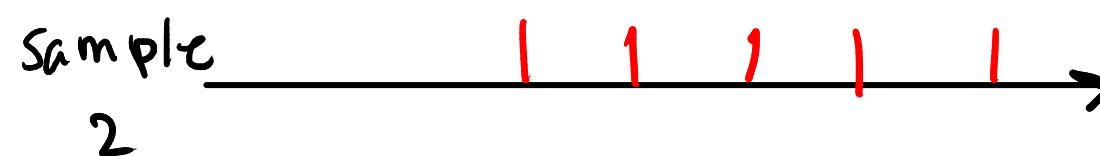
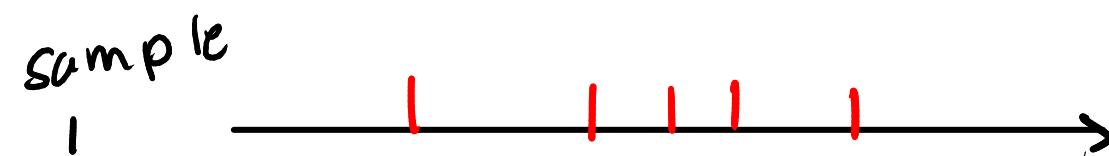
Continuous



Discrete



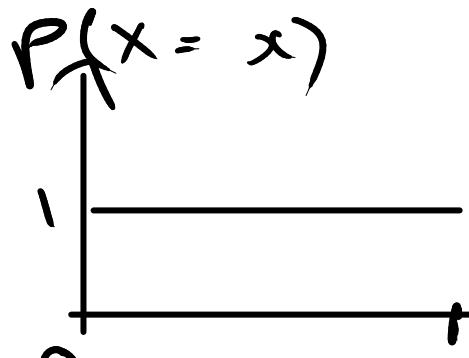
sample size $n=5$



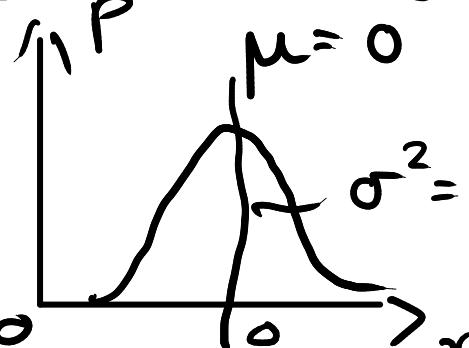
HTH THT

THTTT H

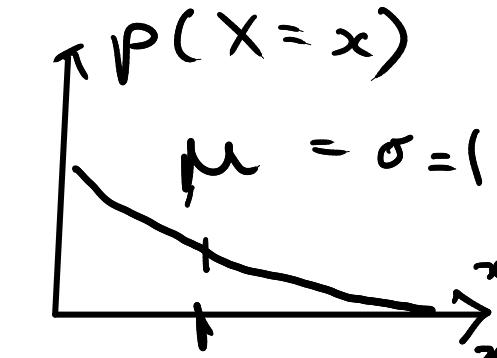
Sampling from continuous probability distributions



Uniform

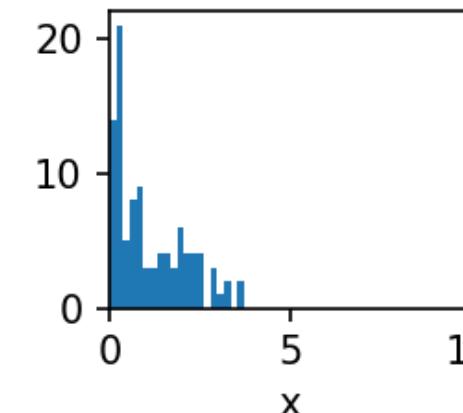
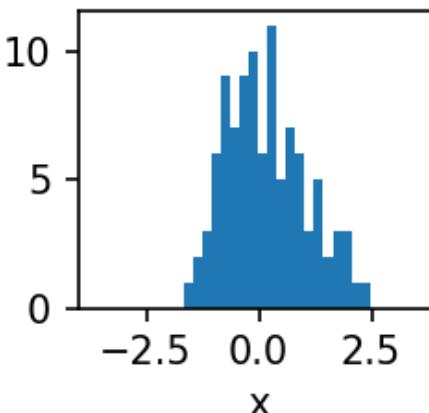
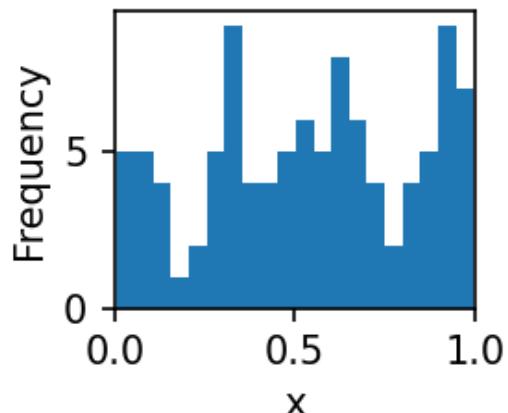


Normal

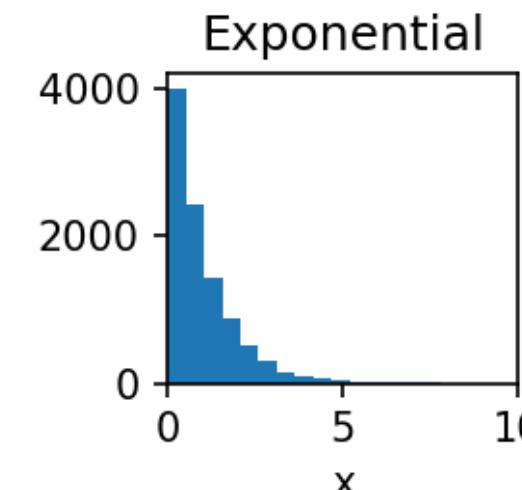
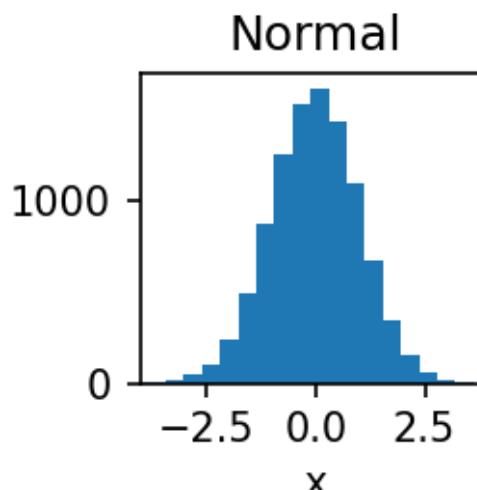
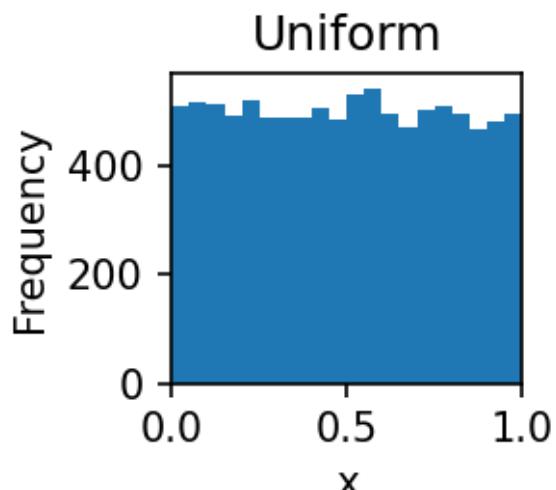


Exponential

100 samples



10000 samples



RNGs

Inf2 - Foundations of Data Science: Randomness, sampling and simulation - Statistics and statistical simulation



THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS
OF
DATA
SCIENCE

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:

Denote observed sample statistic with lower case:

Recipe for a statistical simulation

A. Decide on

- Statistic of interest
- Population distribution or set of items
- Sample size
- Number of repetitions

B. Simulation procedure

1. For i in

- a. Sample i items from the population distribution or set
- b. Compute and store statistic of interest

2. Generate histogram of the i 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?

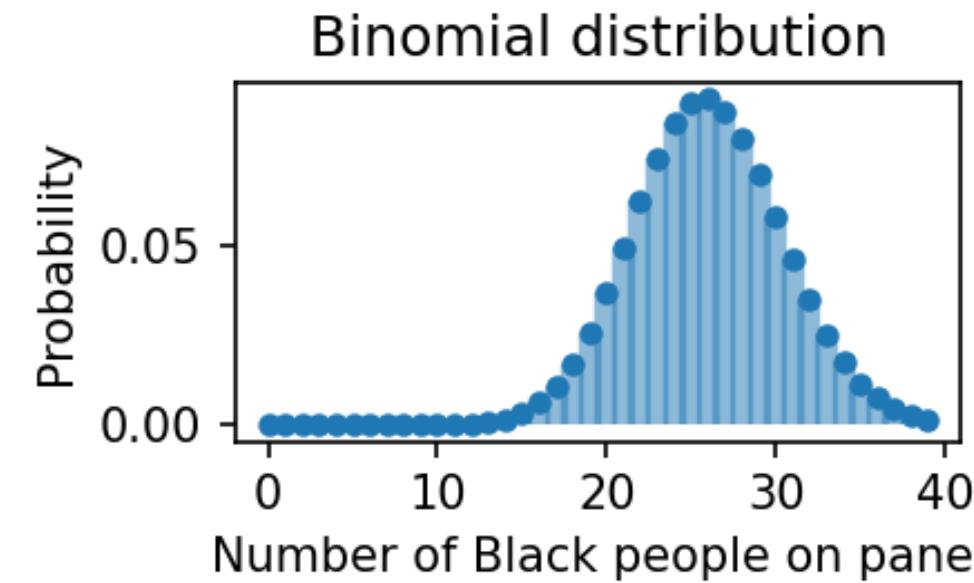
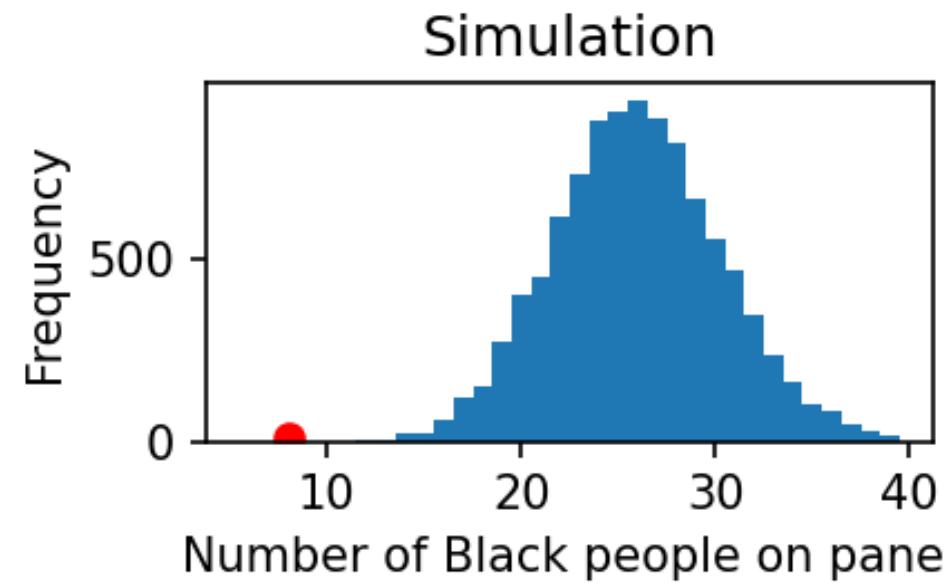
Statistic:

Population:

Sample size:

Num. repetitions:

Swain versus Alabama simulation results



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

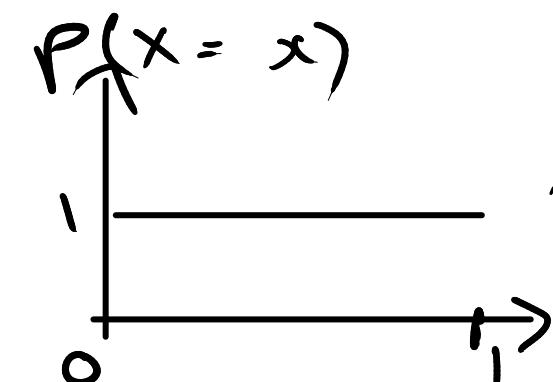


THE UNIVERSITY *of* EDINBURGH
informatics

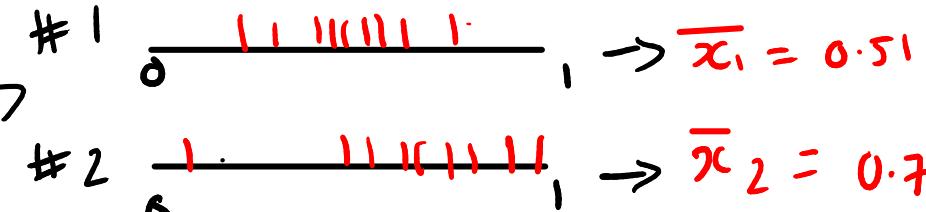
FOUNDATIONS
OF
DATA
SCIENCE

Example: Sampling statistics from continuous distributions

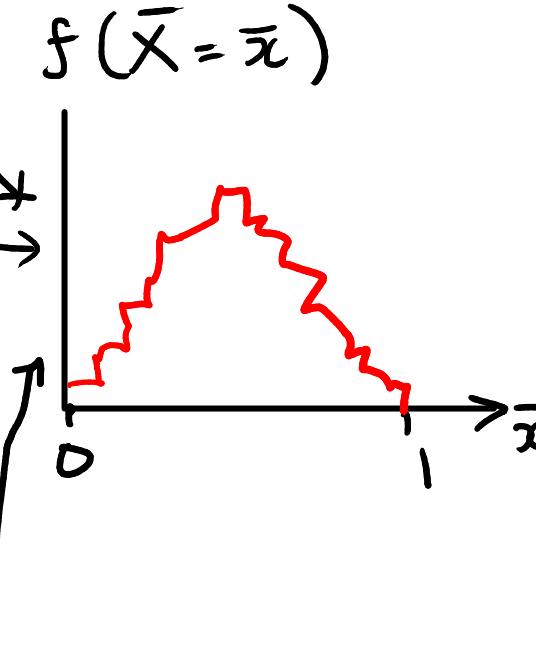
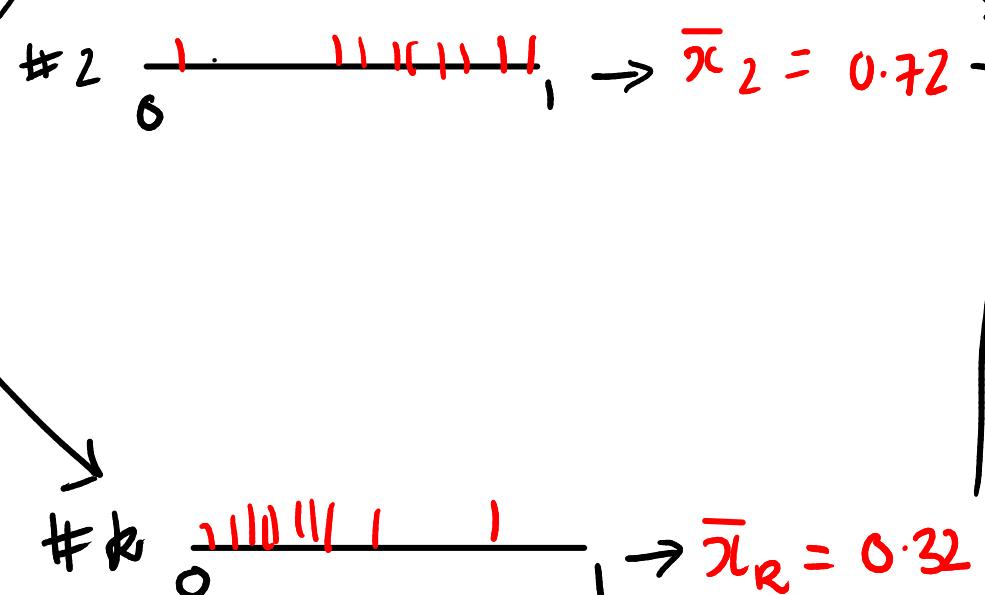
Theoretical
Distribution



k Samples of
 $n=10$

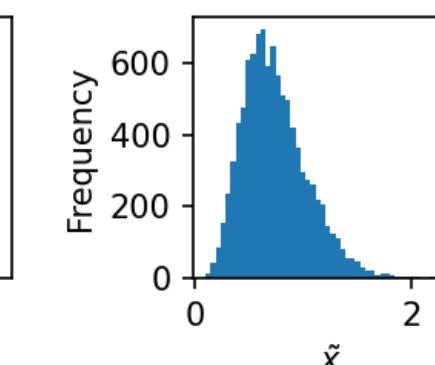
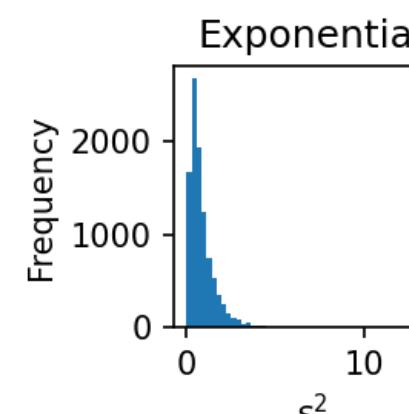
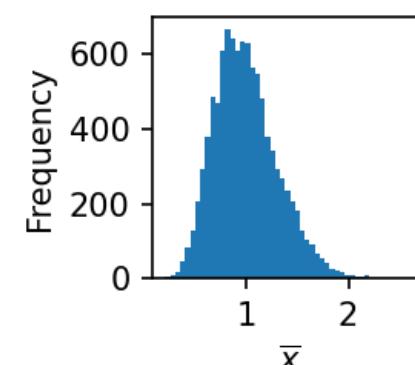
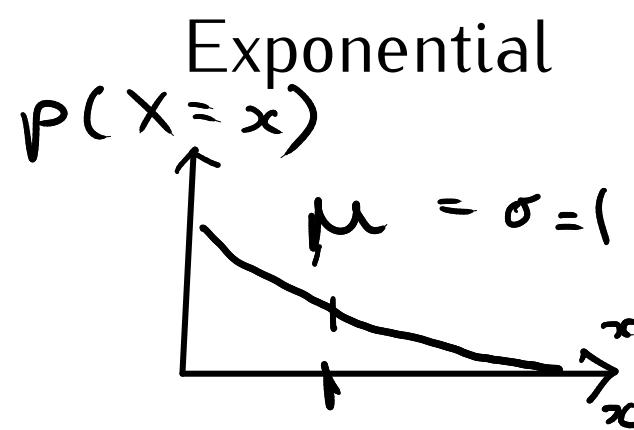
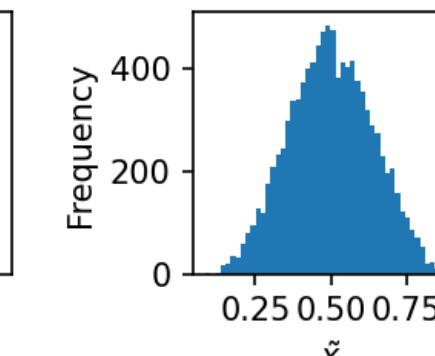
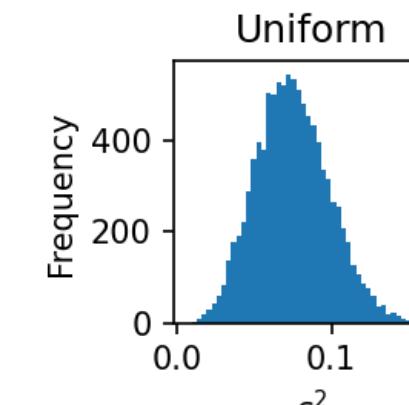
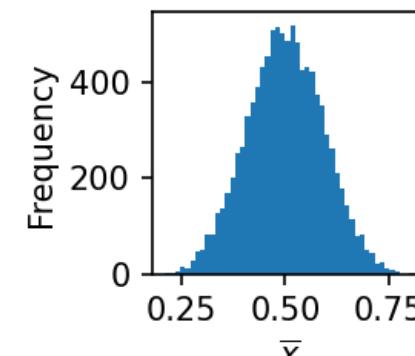
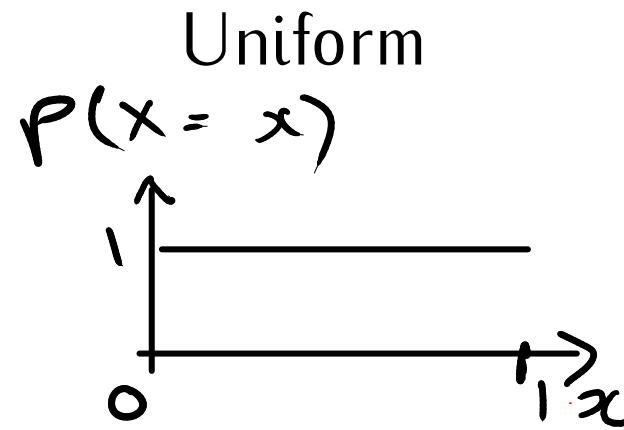
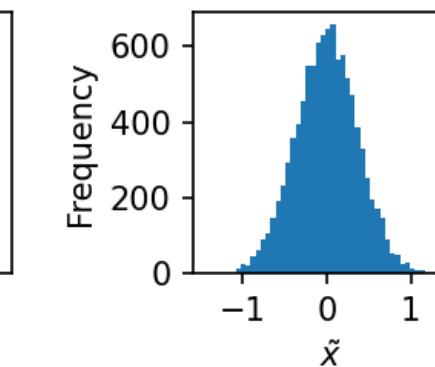
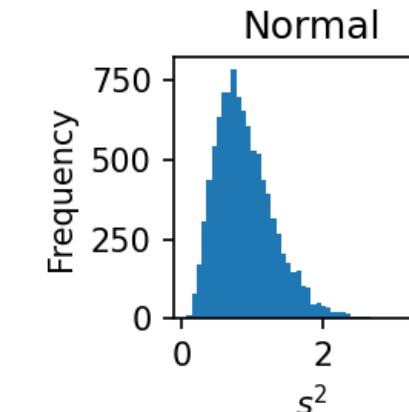
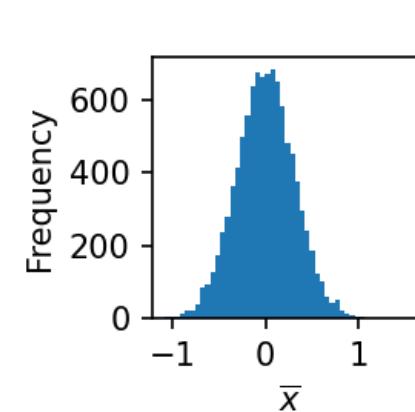
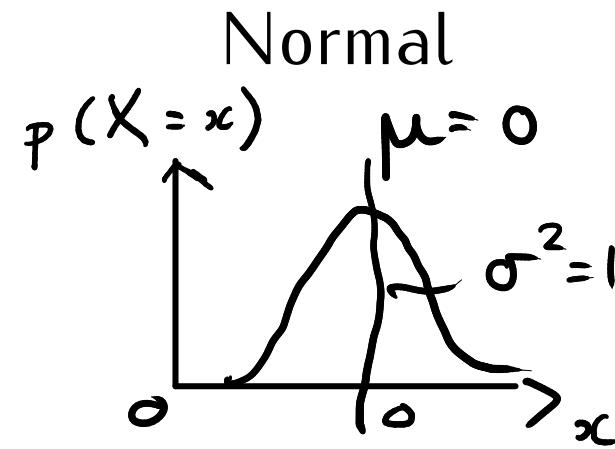


Statistic
mean \bar{X}



Distribution

Mean Variance Median



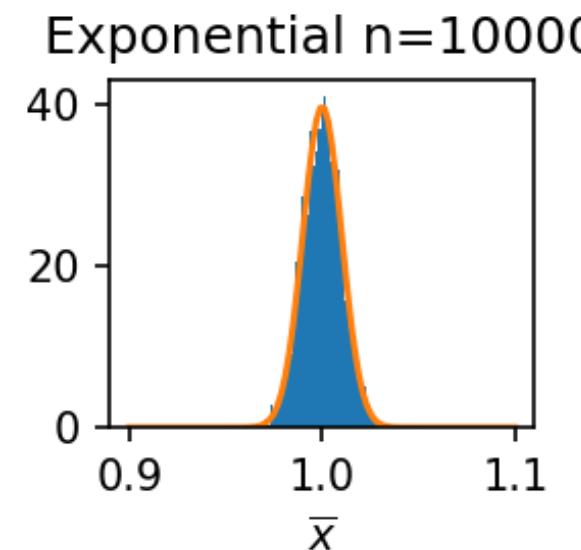
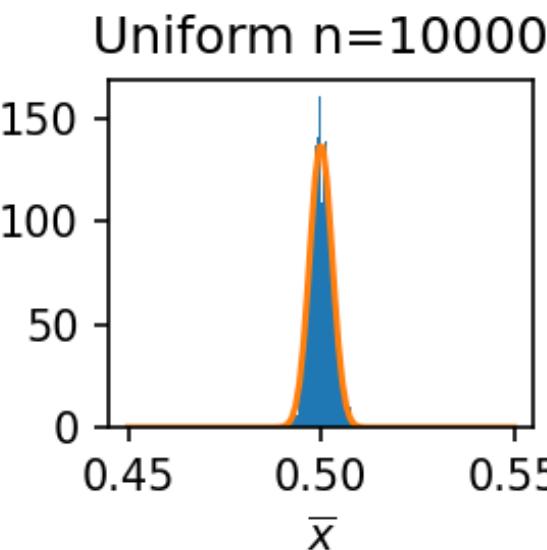
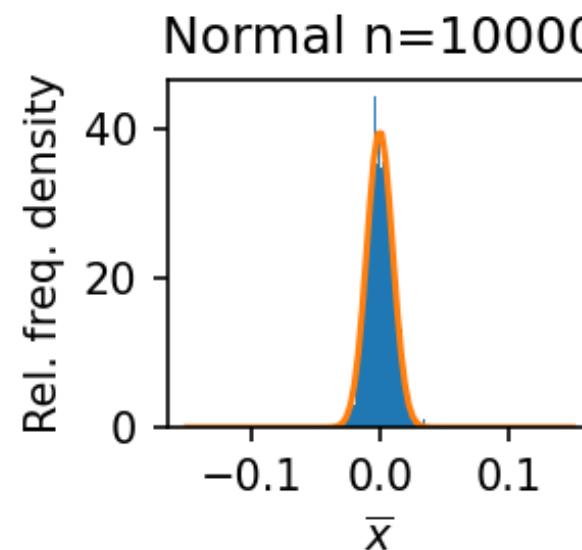
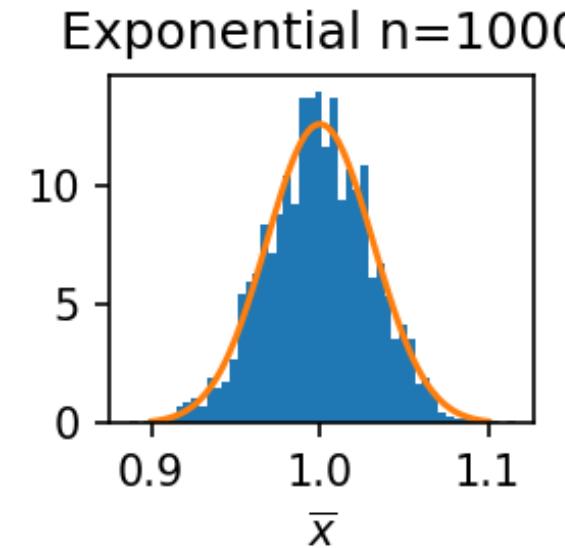
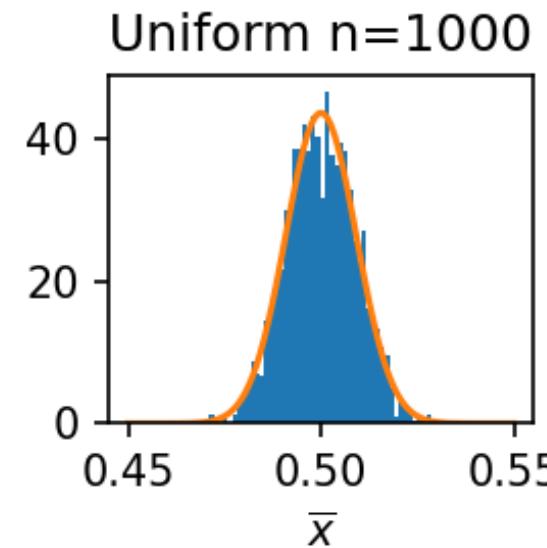
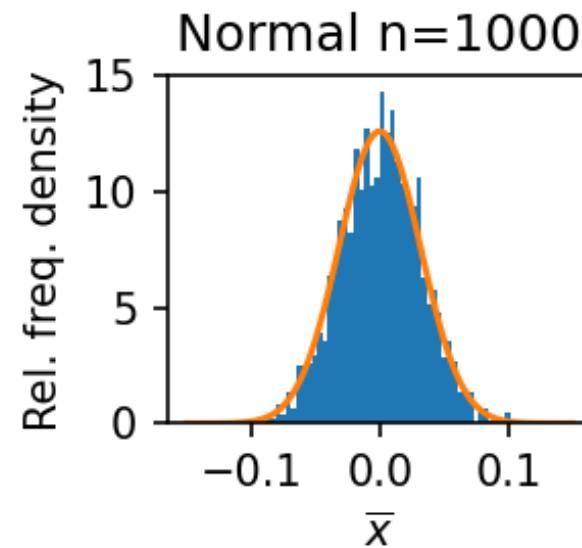
Inf2 - Foundations of Data Science: Randomness, sampling and simulation - Distributions of sample statistics from large samples



THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS
OF
DATA
SCIENCE

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

Expected variance of the mean

Standard error of the mean (SEM)

- Central Limit Theorem applied to tossing an unbiased coin

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 tends to 0.

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 Friday 16 January 2026
to enter a draw to win
one of two £25 vouchers .



<https://edin.ac/47Z4zAU>

