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

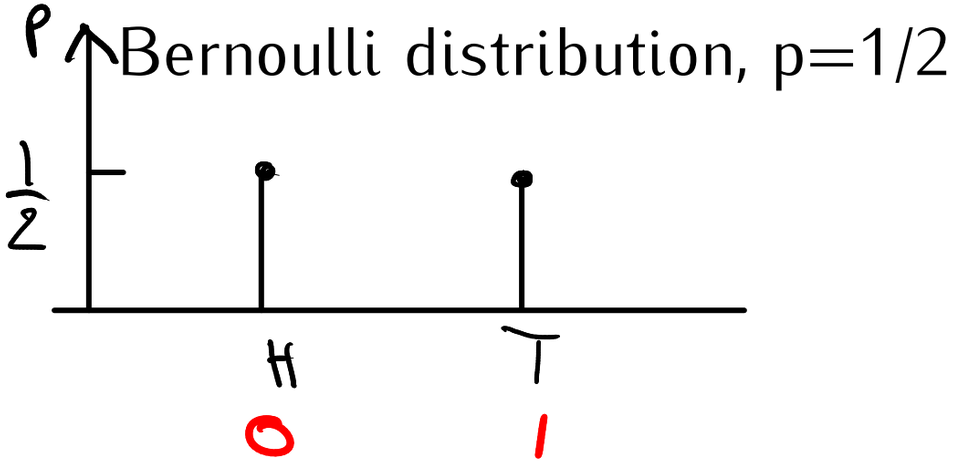
- Steps:
  1. sampling, both random and non-random
  2. definition of a "statistic"
  3. statistical simulation

- 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

Model of unbiased coin



Experiment

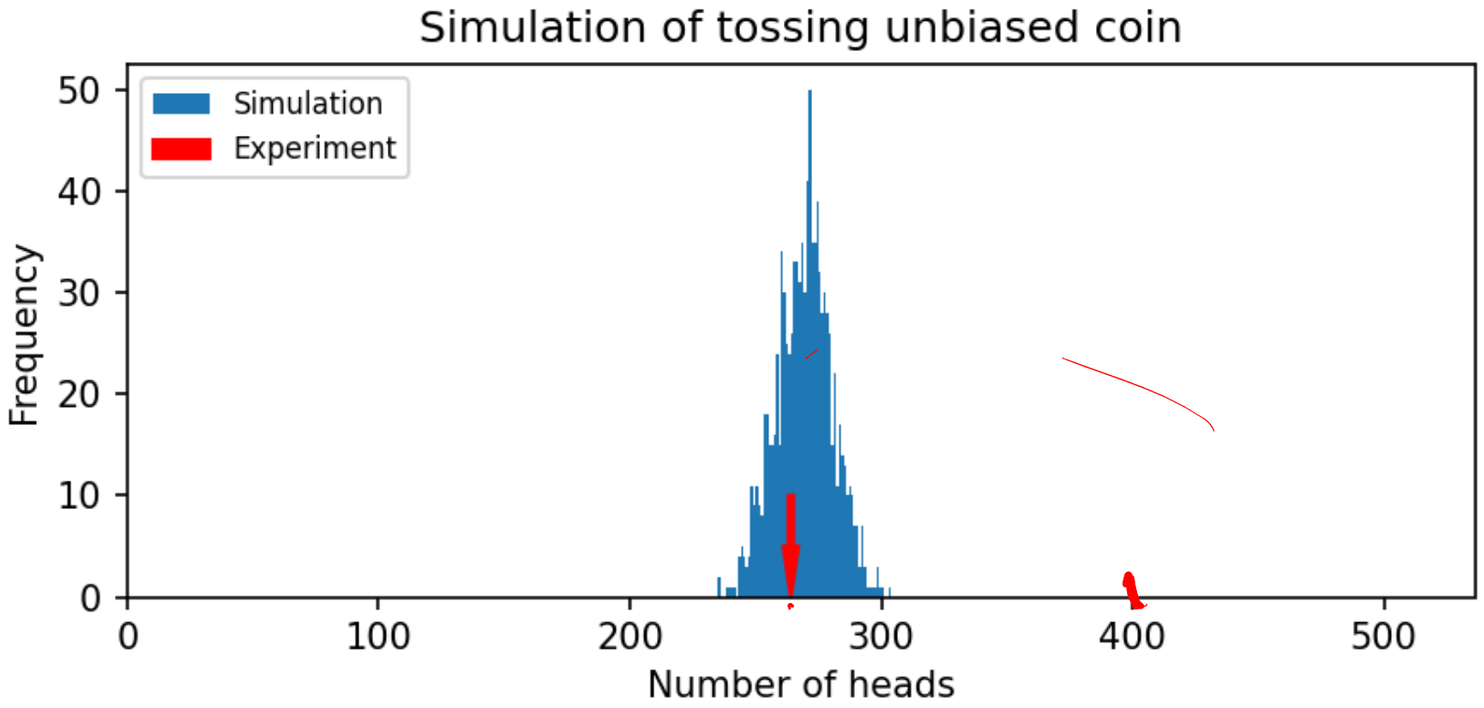
67 students made
536 tosses, of which
264 Heads and 272 Tails

Computational simulation

536 samples, of which
257 Heads and 279 Tails
275
261

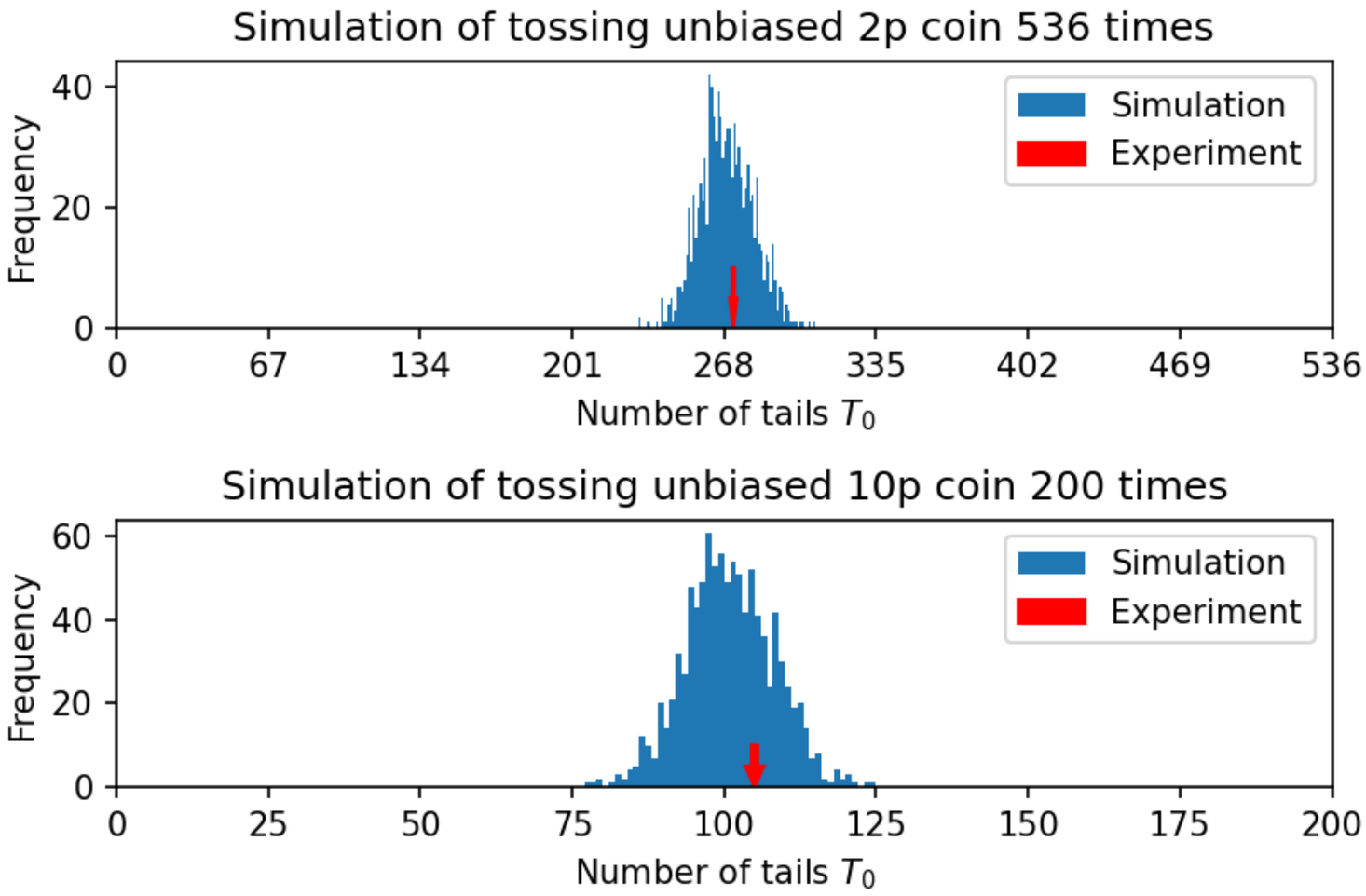
Statistical simulation results

1000 repetitions later... consistent with experiment?

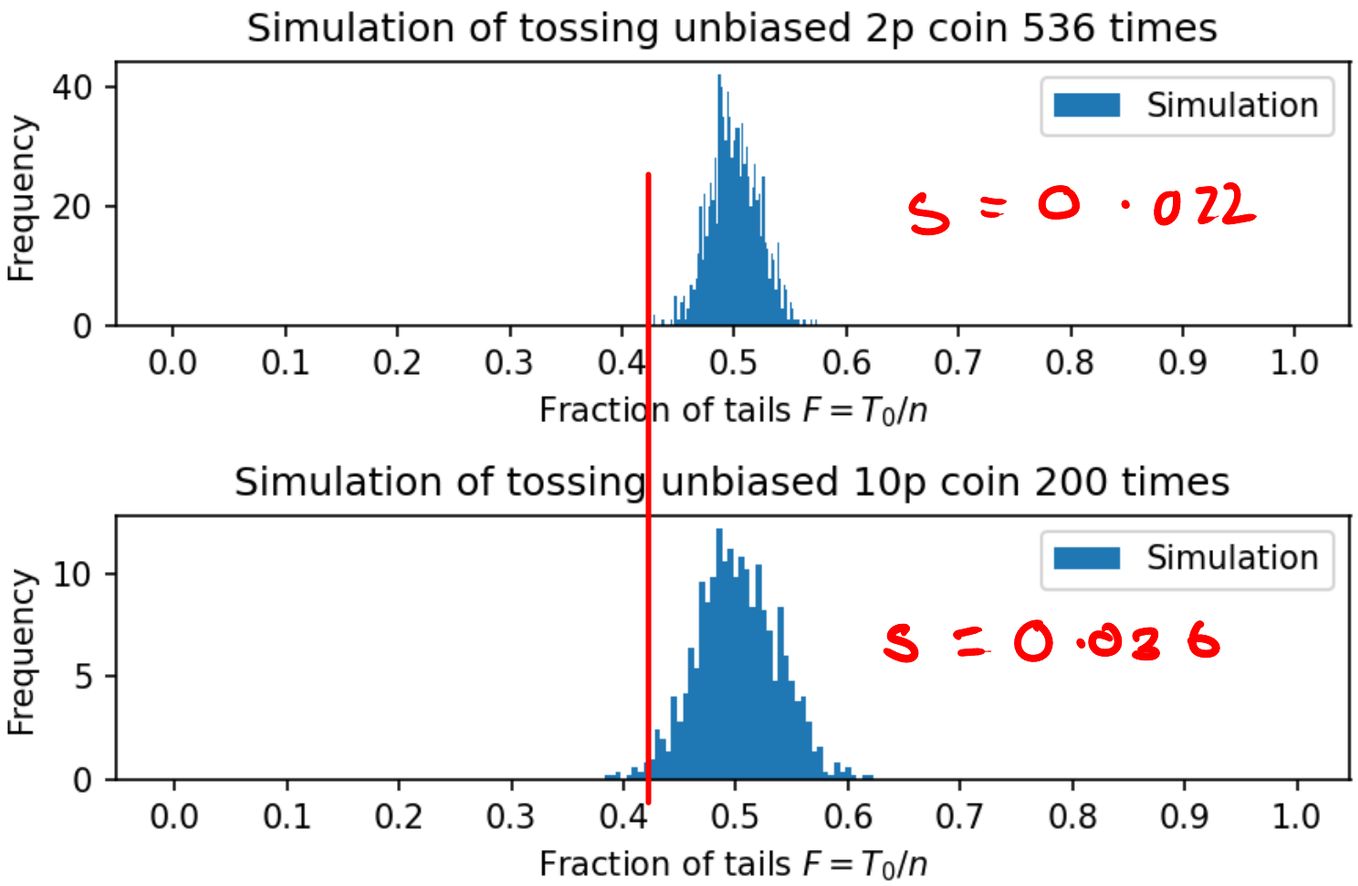


Data is consistent with an unbiased coin

Statistical simulation results – for the 10p coins too



Plot fractions of tosses that are tails to help comparison

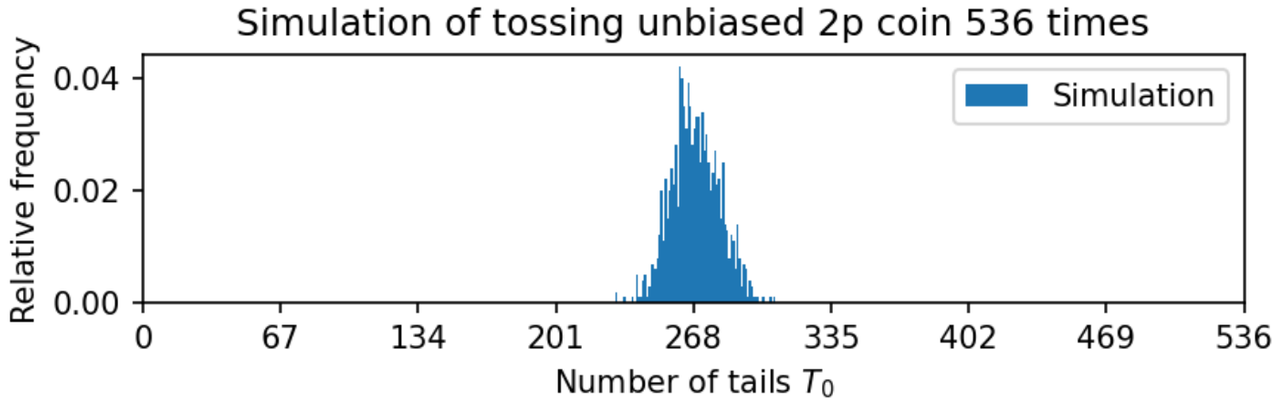


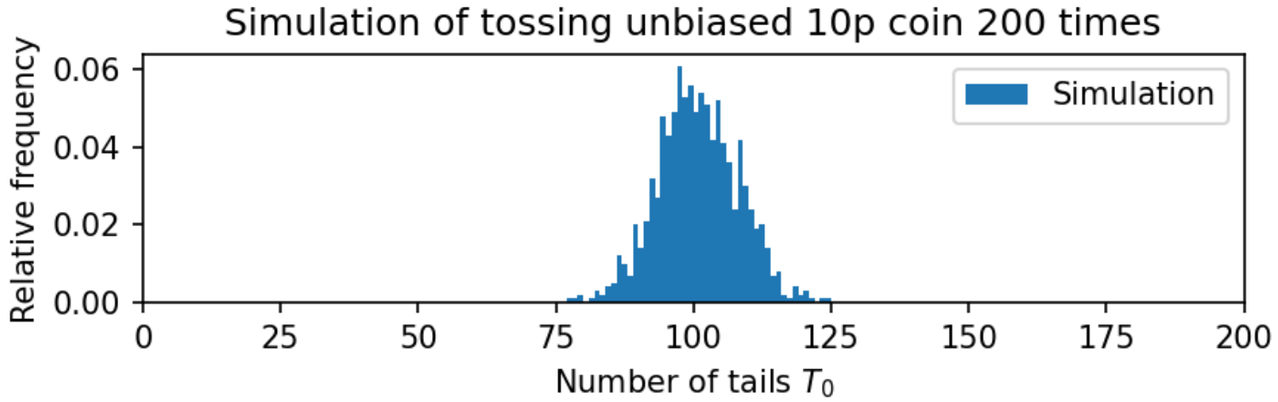
Are the sample means of the fraction of the number of tails in each simulation similar or different?

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

Note on relative frequency

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





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

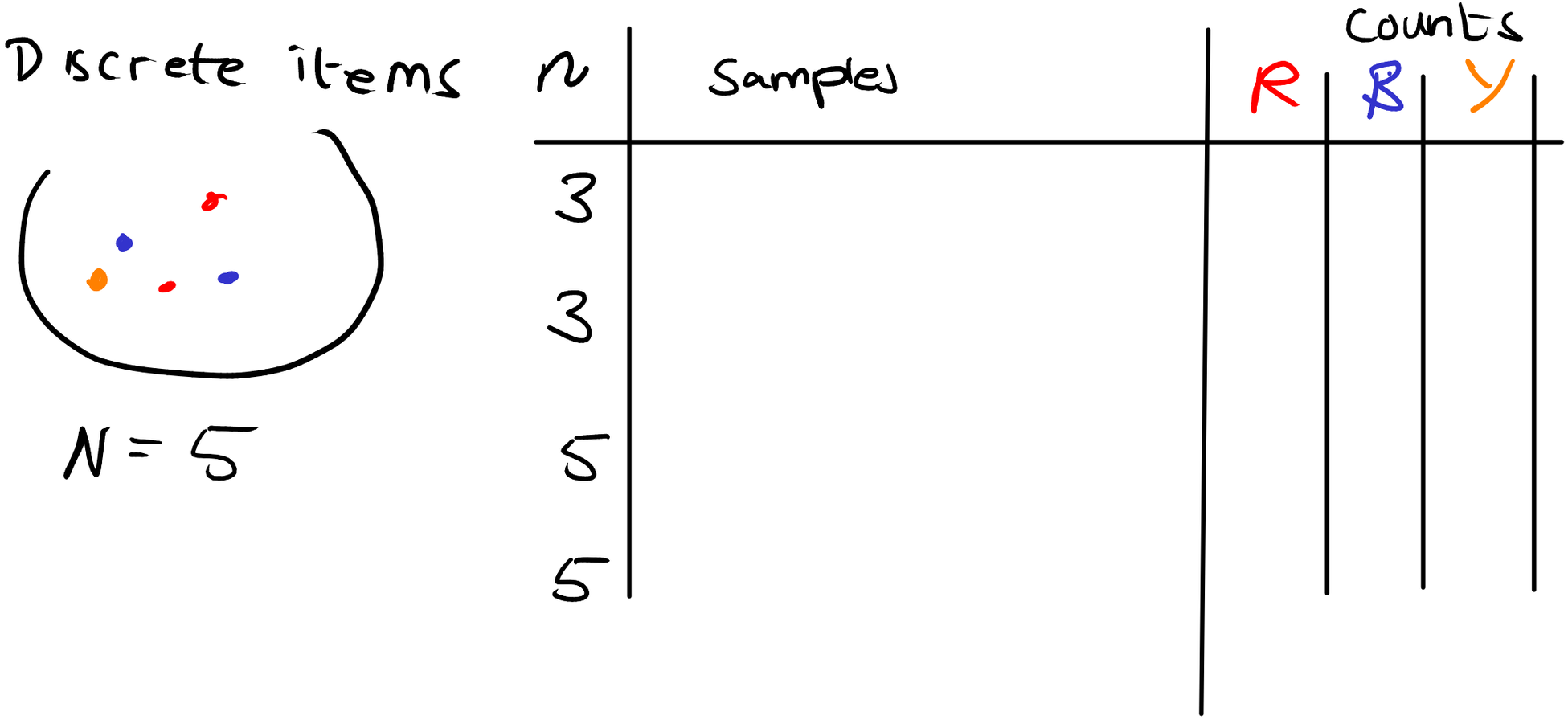
THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

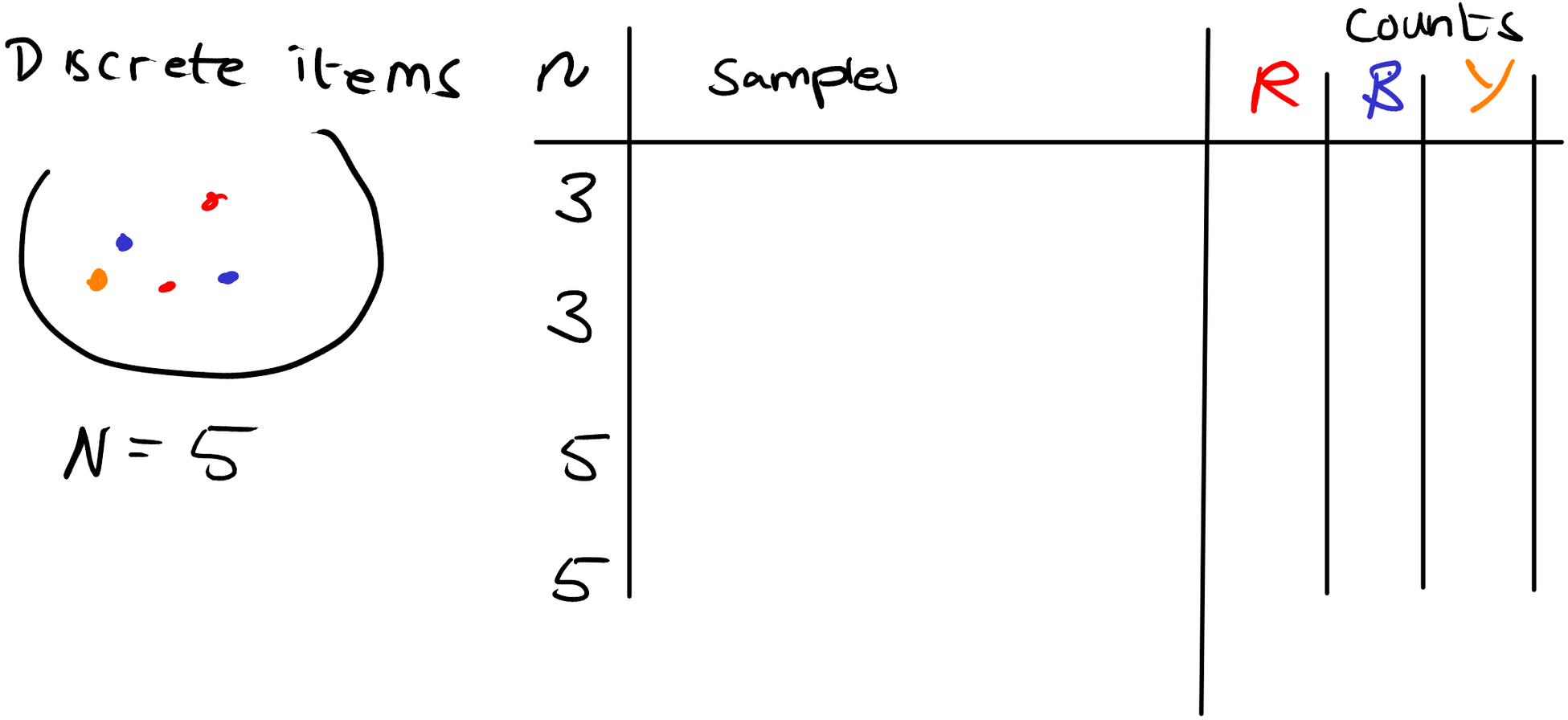
Definition of a random sample (Strictly, an "independent and identically distributed" (iid) random sample)

In a random sample of size n from either
- a probability distribution
- or a finite population of N items
the random variables $X_1, \ldots, X_n$
comprising the sample are all
1. independent and
2. have the same probability distribution

Sampling from a finite population of discrete items without replacement



Sampling from a finite population of discrete items with replacement



Questions

1. Is sampling with replacement an iid random sample?

2. Is sampling without replacement an iid random sample?

If num items sampled n is much less than total number of items N $\Rightarrow$ approx random sample

$$\frac{n}{N} \leq 0.05$$

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 choose to remain anonymous).

Your probability of winning will be proportional 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.)

Why are random samples good?

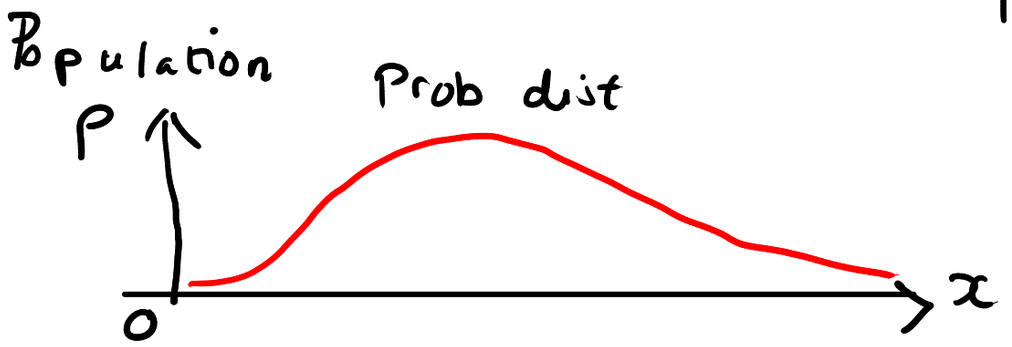
Consider non-random samples

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

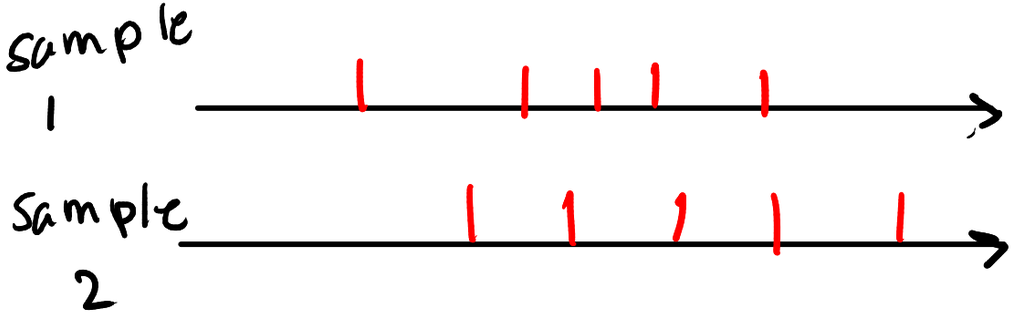
Mstyslav Chernov, Wikimedia Commons, CC BY SA 3.0

Sampling from a probability distribution

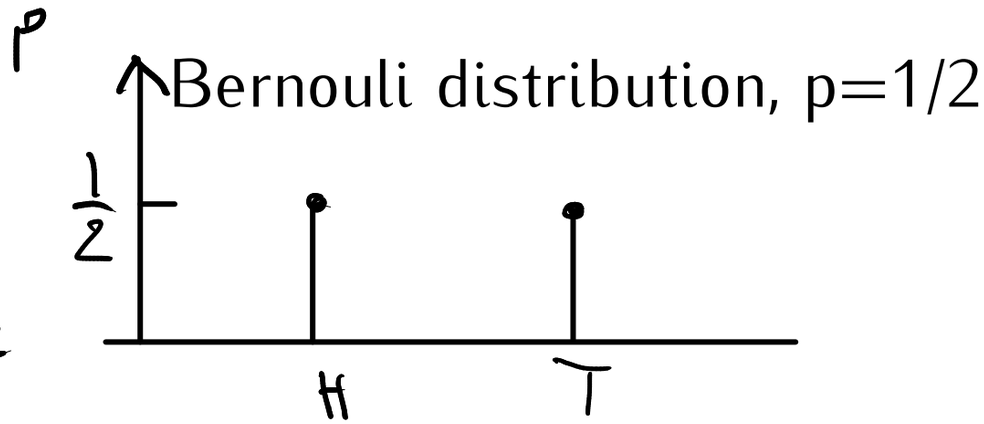
Continuous



sample size n=5



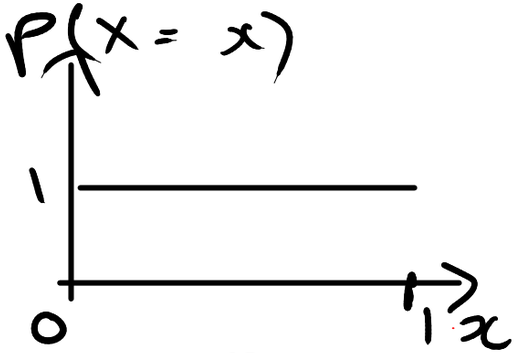
Discrete

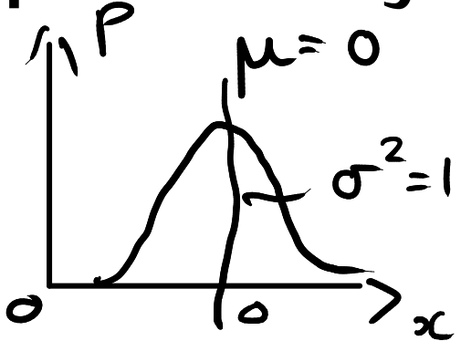


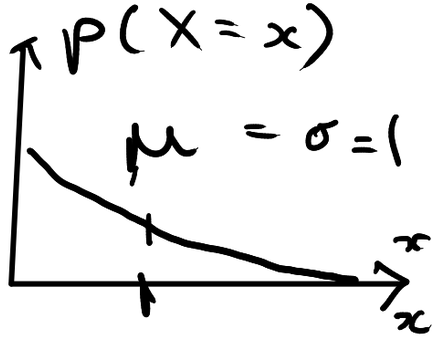
H H T H H

T H T T H

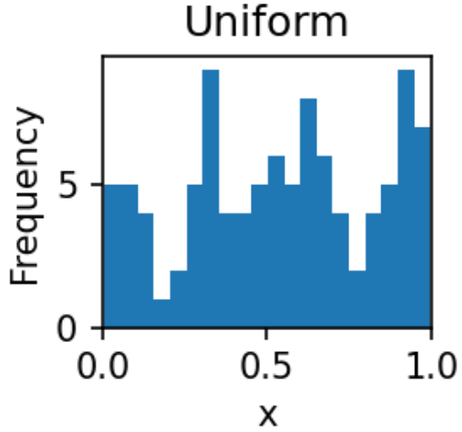
Sampling from continuous probability distributions

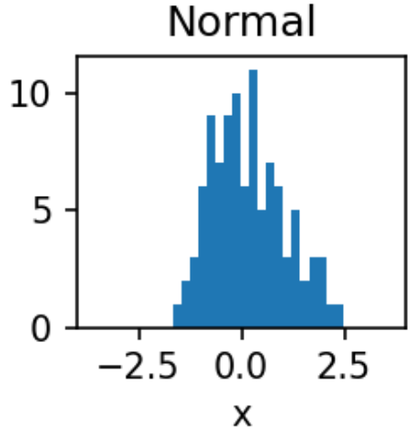


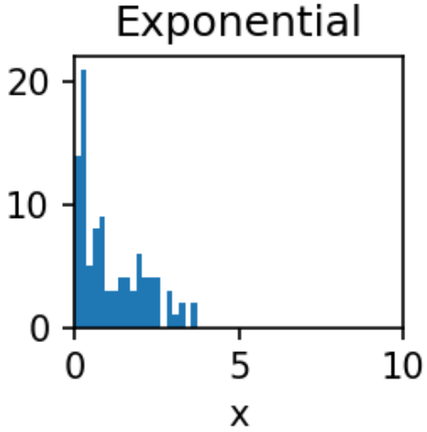




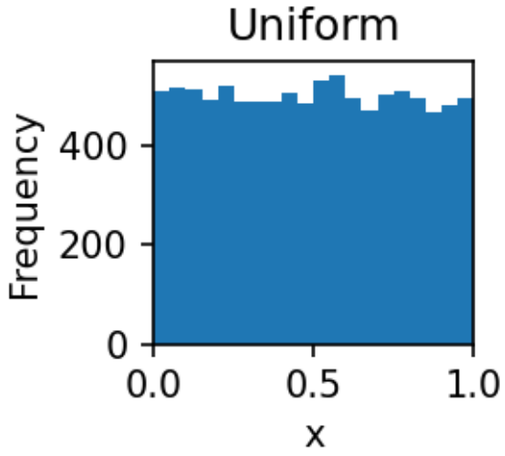
100 samples

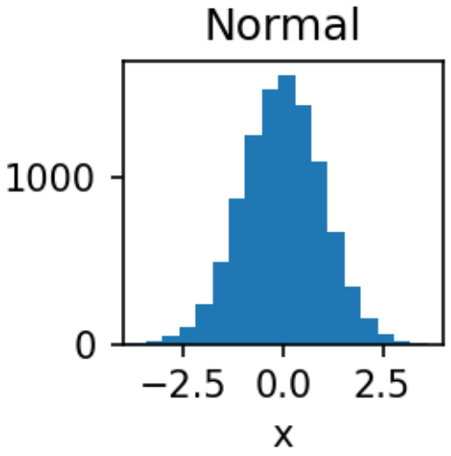


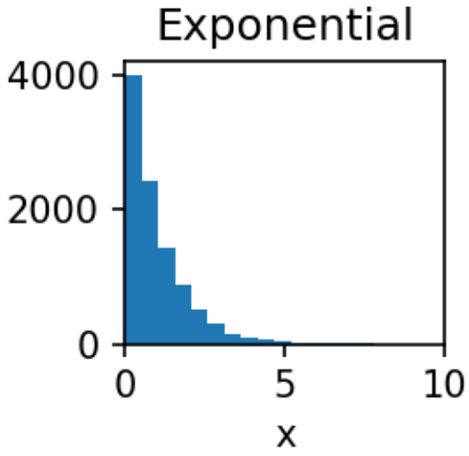




10000 samples







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: T_0 , F , X

Denote observed sample statistic with lower case: t , f , x

Recipe for a statistical simulation

A. Decide on
- Statistic of interest T_0 - num tails
- Population distribution or set of items Bernoulli dist, $p = 0.5$
- Sample size $n = 536$
- 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 $t_0 = 8$
26% of population of Alabama were black
How do we simulate unbiased jury selection?

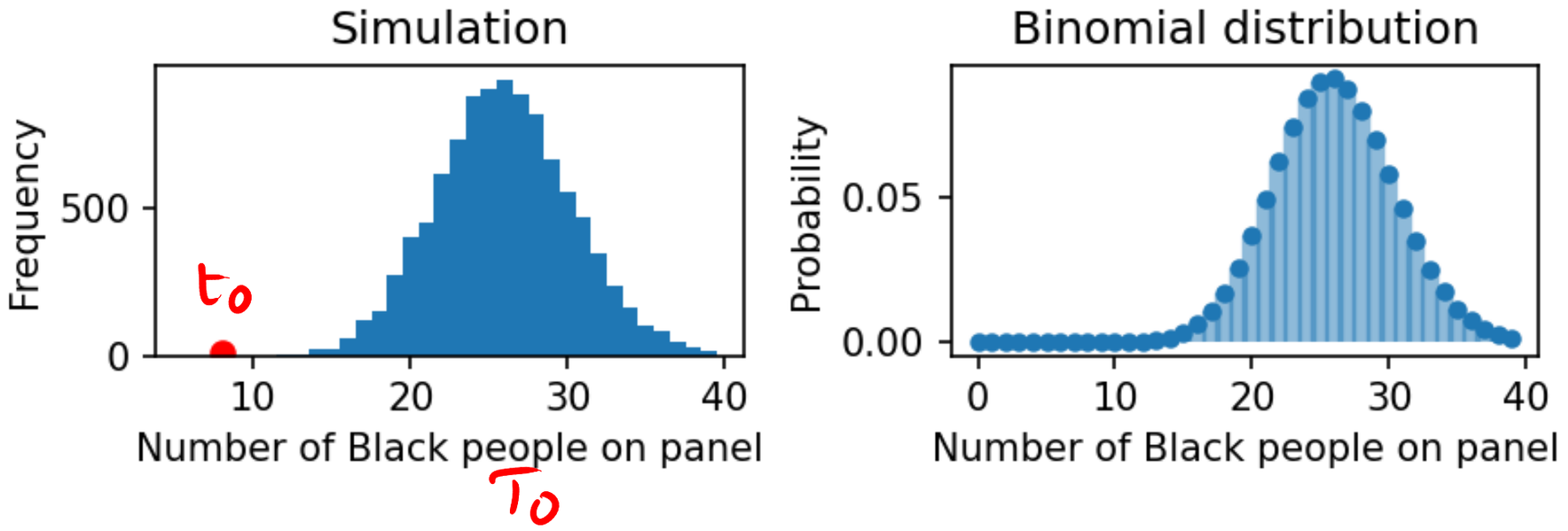
Statistic: T_0 # black on a panel of $n = 100$ members

Pouplation: Bernoulli dist with sample space {Black, White}
$p(\text{Black}) = 0.26$ B W W B B B

Sample size: $n = 100$

Num. repetitions: $k = 10,000$

Swain versus Alabama simulation results

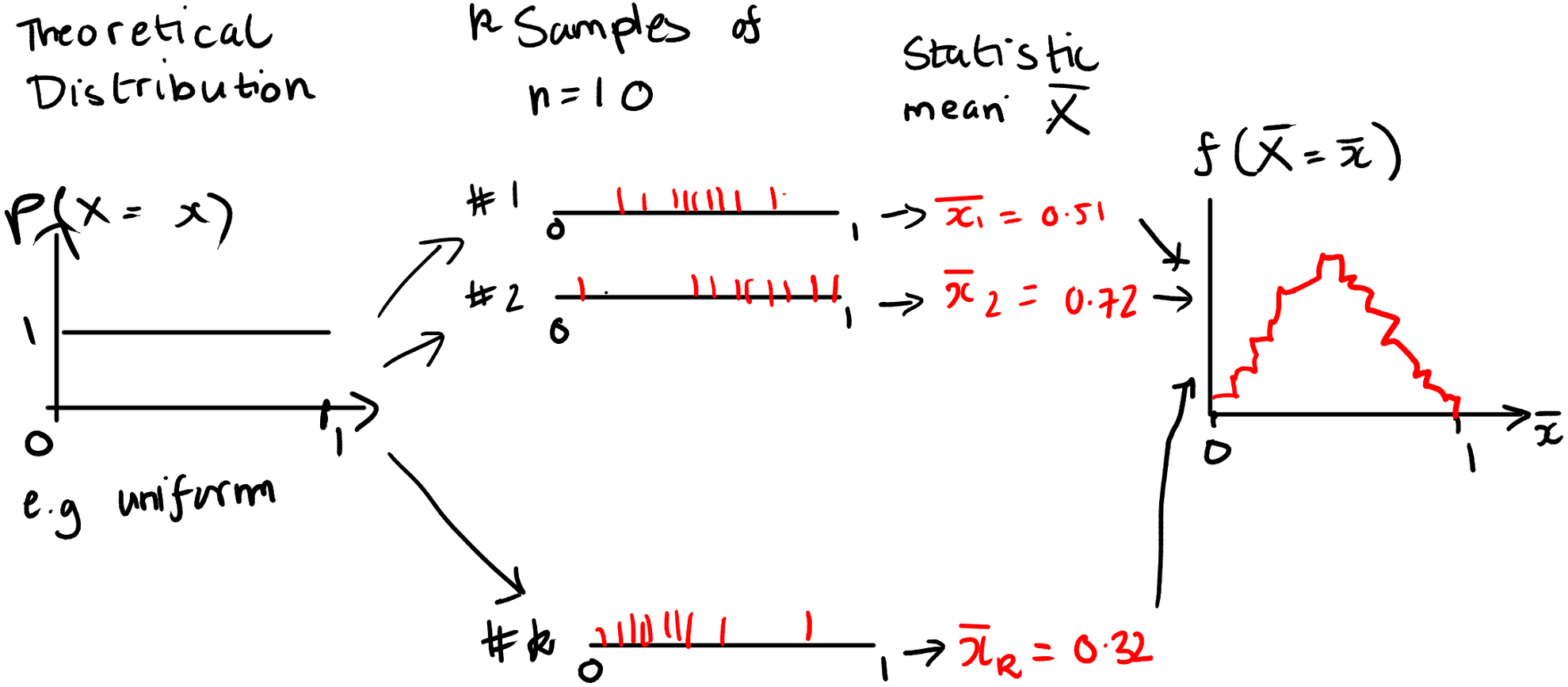


Inf2 - Foundations of Data Science: Randomness, sampling and simulation - Distributions of sample statistics from small samples

THE UNIVERSITY *of* EDINBURGH
informatics

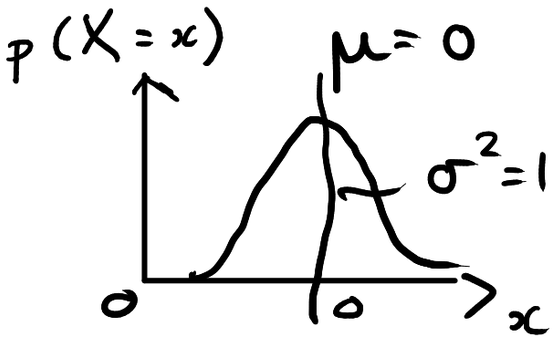
FOUNDATIONS OF DATA SCIENCE

Example: Sampling statistics from continous distributions

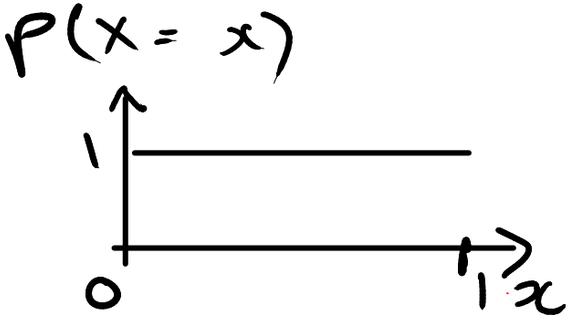


Distribution

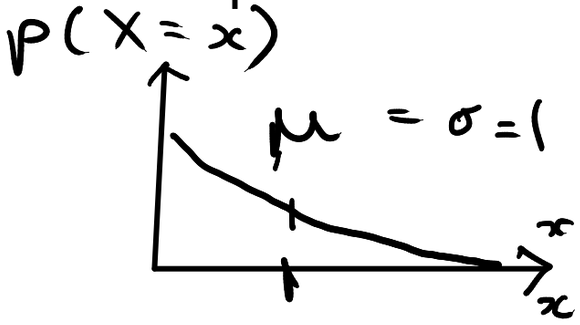
Normal



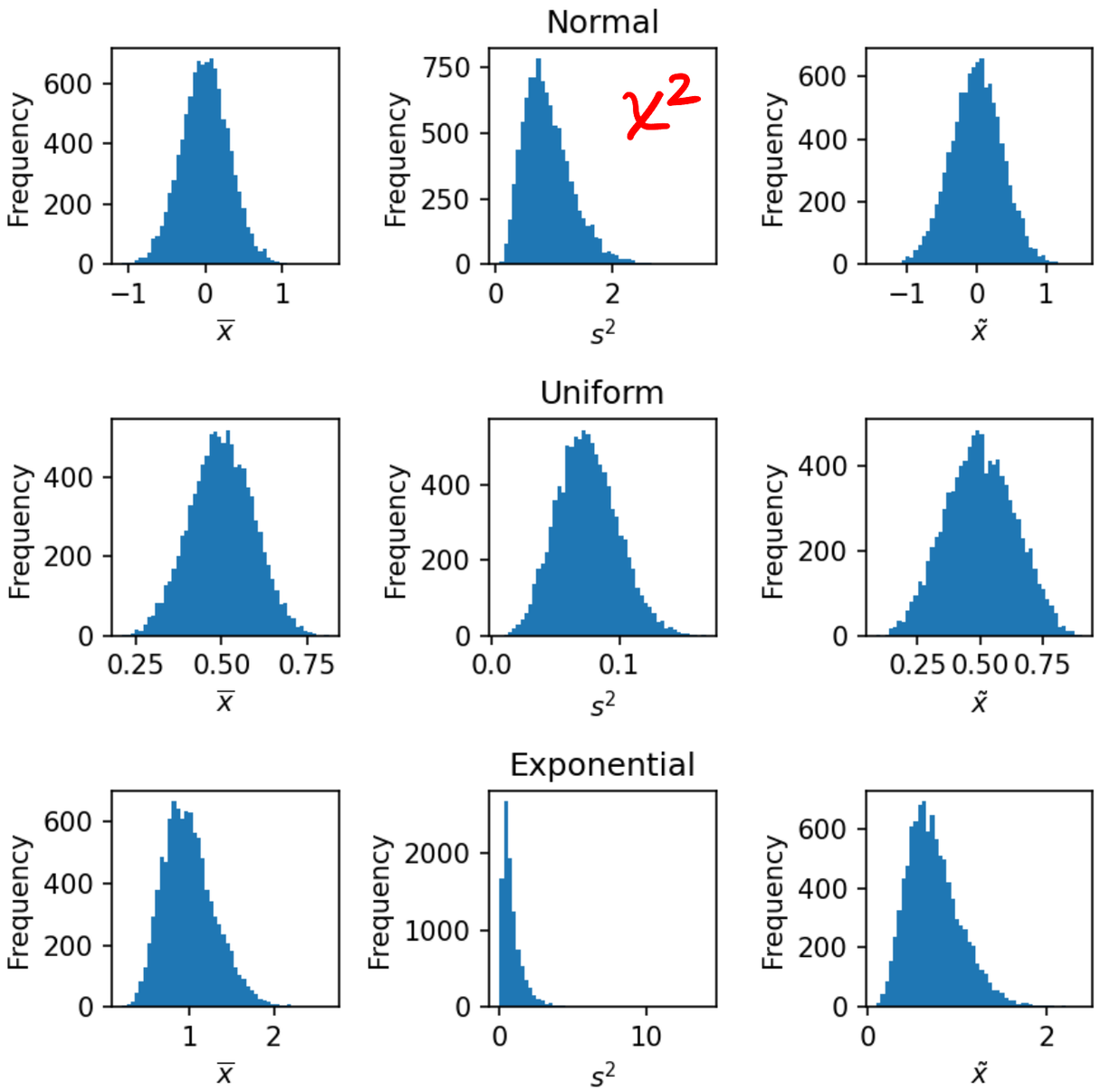
Uniform



Exponential



Mean Variance Median

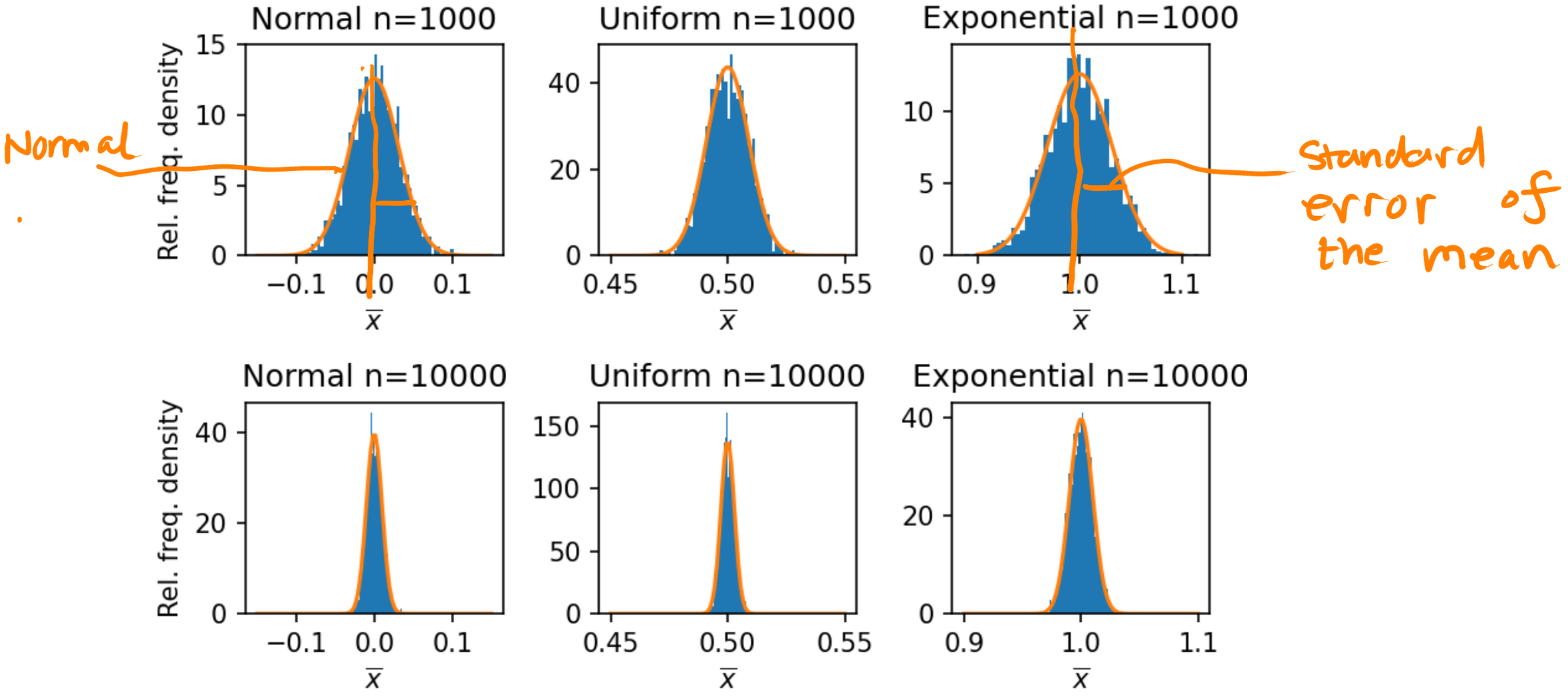


Inf2 - Foundations of Data Science: Randomness, sampling and simulation - Distributions of sample statistics from large samples

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 $\bar{x}$ 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)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem applied to tossing an unbiased coin

Theoretical dist. : Bernoulli dist. with parameter

$$P(X=1) = \frac{1}{2} = p \qquad P(X=0) = \frac{1}{2} = 1-p$$

$$\Rightarrow \mu = \frac{1}{2}$$

$$\sigma^2 = p(1-p) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

CLT $\Rightarrow$ Dist. of mean $\bar{X}$ with n samples will converge on a normal dist. with

$$\mu_{\bar{x}} = \mu - p = \frac{1}{2}$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{1}{4n} \Rightarrow \text{SEM} \quad \sigma_{\bar{x}} = \frac{1}{2\sqrt{n}}$$

COMPARE WITH DISTRIBUTION OF TAILS FROM EARLIER IN LECTURE

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.

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)

- More data (i.e. bigger sample size) -> more certainty

- The Standard Error of the Mean (SEM) is not the same as the Standard Deviation

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