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

Last lecture...

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

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

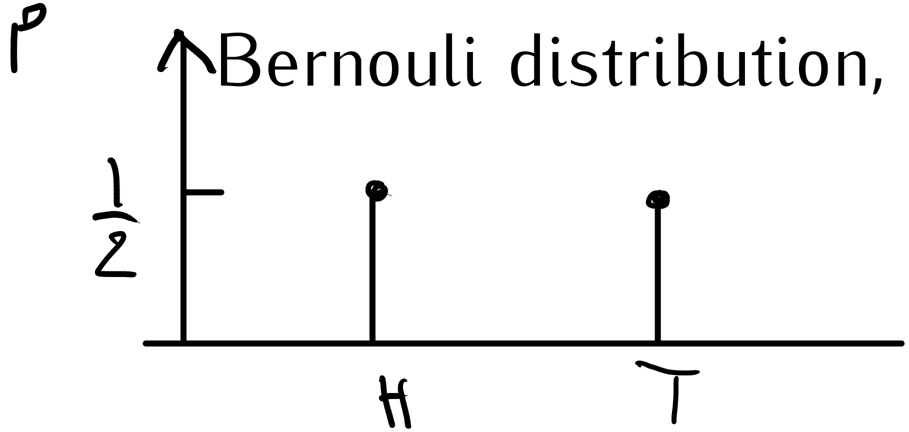
- Then get intuition about what happens as sample size changes
  1. distribution of statistics from small samples
  2. distribution of statistics from large samples

Statistical simulation overview

Reality

Model of unbiased coin

Bernouli distribution, p=1/2



Experiment

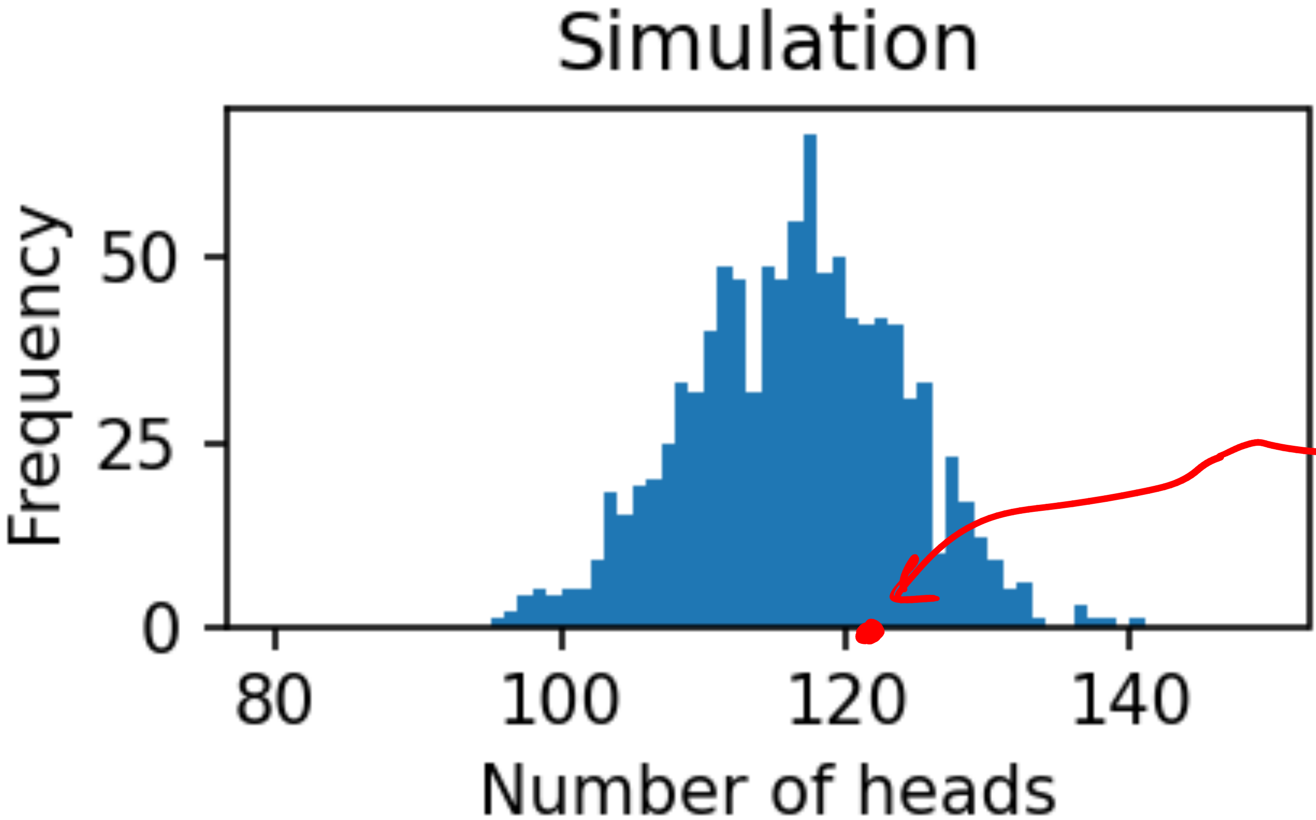
232 tosses, of which
121 Heads and 111 Tails

Computational simulation

232 samples, of which
120 Heads and 112 Tails
106 " " 224 "
⋮ ⋮

Statistical simulation overview

1000 repetitions later... consistent with experiment?



Data/Experiment

consistent with hypothesis of unbiased coin

Definition of a random sample

In a random sample of size n from either
- a probability distribution
- or a finite population of N items

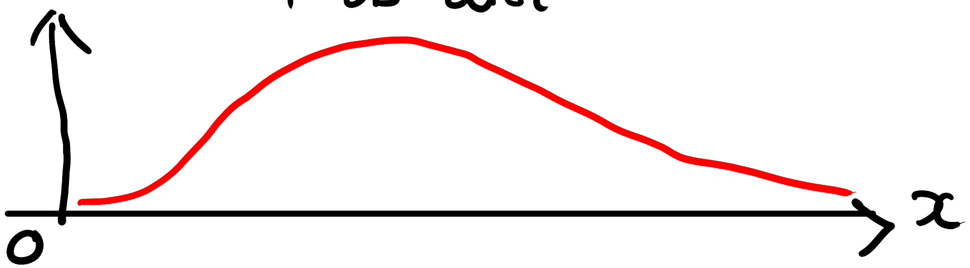
the random variables $X_1, \dots, X_n$
comprising the sample are all
1. independent and
2. have the same probability distribution

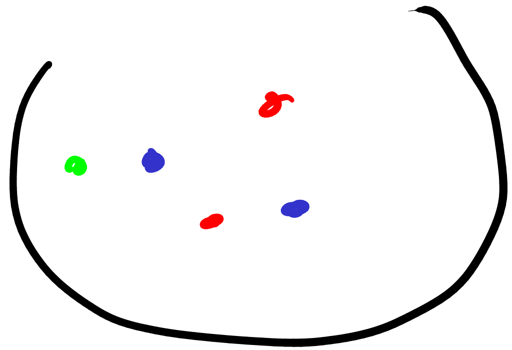
Sampling

Discrete items

Population

Prob dist

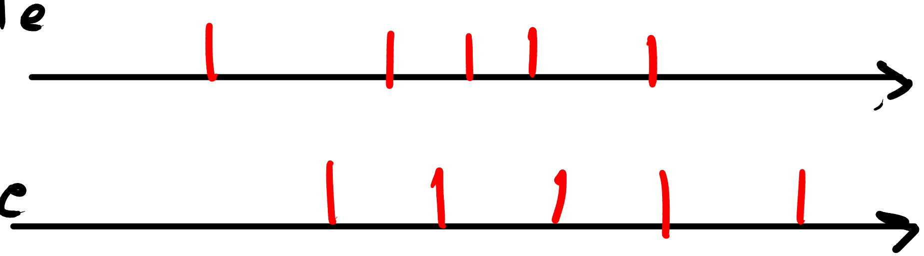


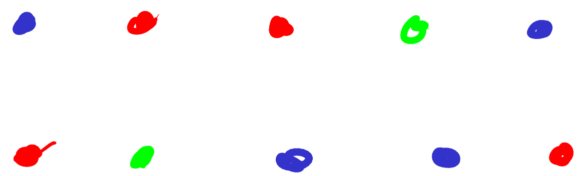


$n = 5$

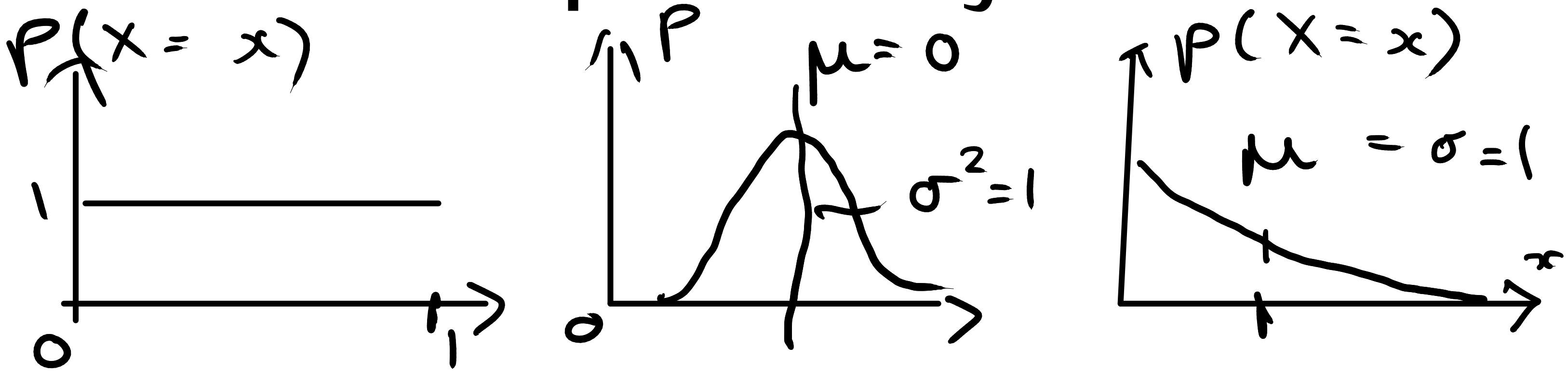
Sample 1

Sample 2



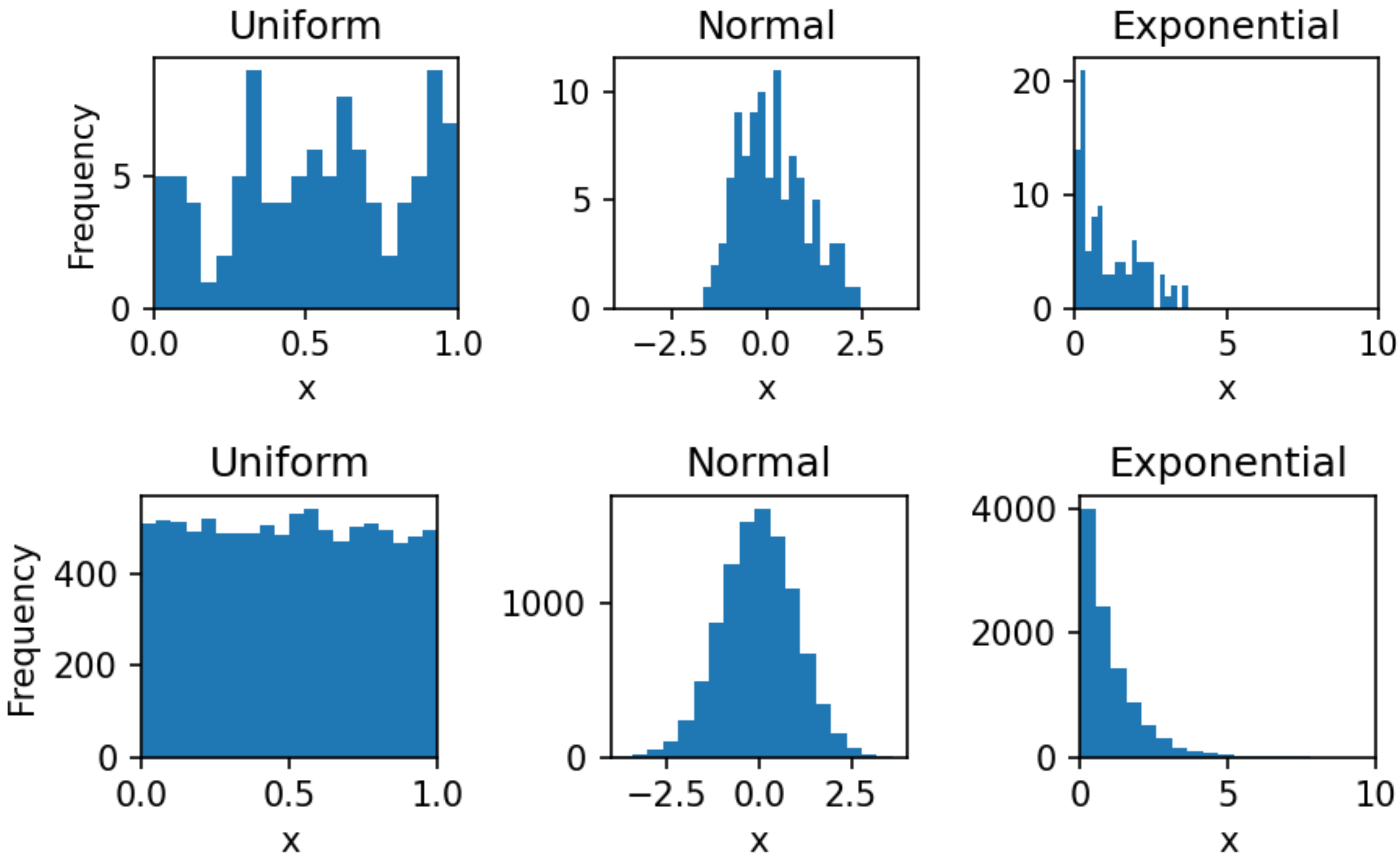


Sampling from continuous probability distributions



100 samples

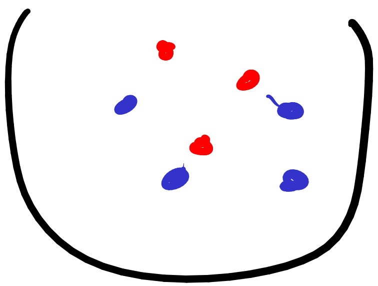
10000 samples



RNGs

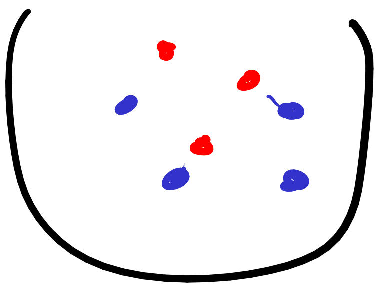
Sampling from a discrete set of items without replacement

$n = 5$



$N = 7$

Sampling from a discrete set of items with replacement



$N = 7$

$n = 5$

Non-random samples

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

Definition of a statistic

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

Examples:

$\underbrace{\text{Number of heads}}_{\text{statistic}}$ from $\underbrace{\text{seq. of coin tosses}}_{\text{Data}}$

Sample	mean	$\bar{x}$	$\bar{X}$
"	variance	s^2	S^2
"	median	$\tilde{x}$	$\tilde{X}$
"	minimum		

Lower case for sample statistic

upper case for statistic of r.v.

Recipe for a statistical simulation

A. Decide on

- Statistic of interest — # heads
- Population distribution or set of items — Bernoulli $p = 0.5$
- Sample size $n = 232$
- 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
Let's simulate unbiased jury selection:

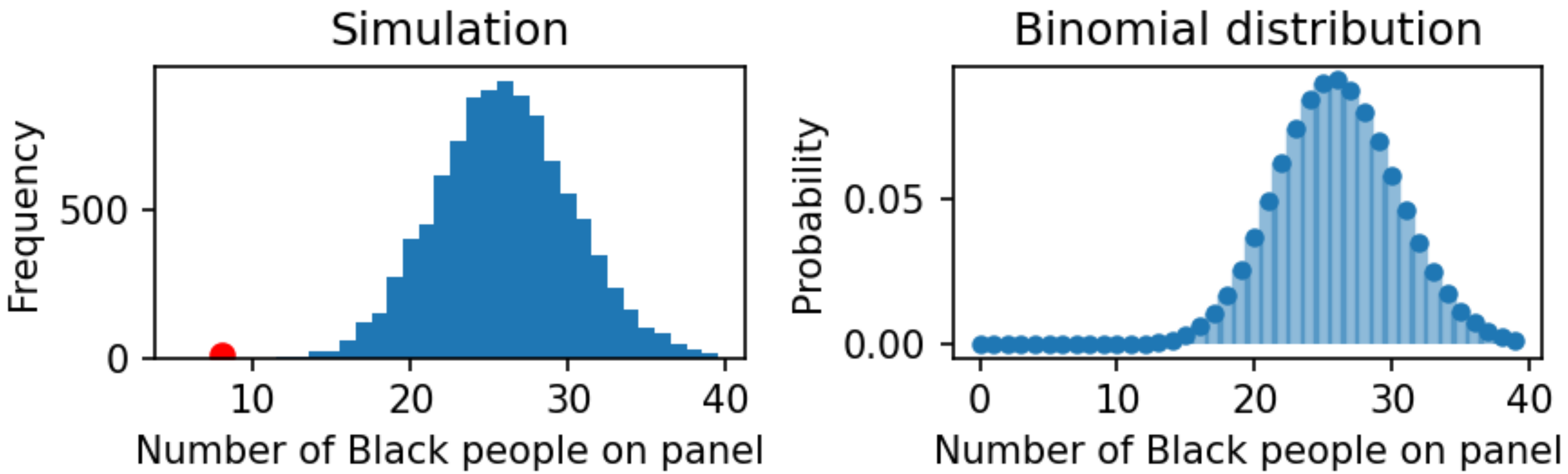
Statistic: T_0 # black people on panel of $n = 100$ members

Population Bernoulli dist with sample space {Black, Non-black}
$P(\text{Black}) = 0.26$

Sample size $n = 100$

Num. repetitions $k = 10,000$

Swain versus Alabama simulation results



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

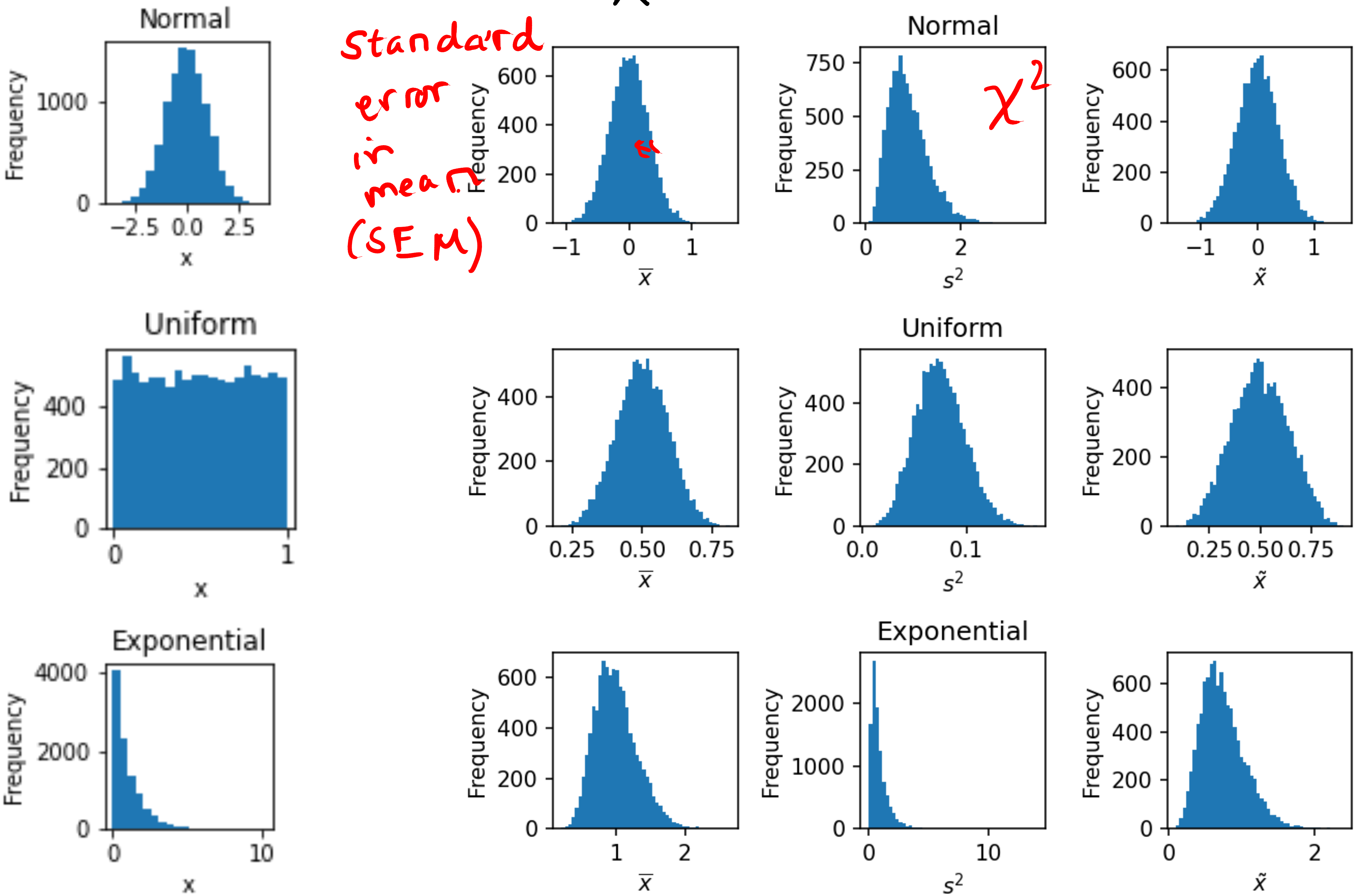
Example: Sampling statistics from continous distributions

- Statistics : $\bar{X}$, S^2 , $\tilde{X}$
- Distributions : Normal, Uniform and exponential
- Sample size : $n=10$
- Num. replications : $R = 10000$

Distribution | Mean $\bar{X}$ | Variance | Median

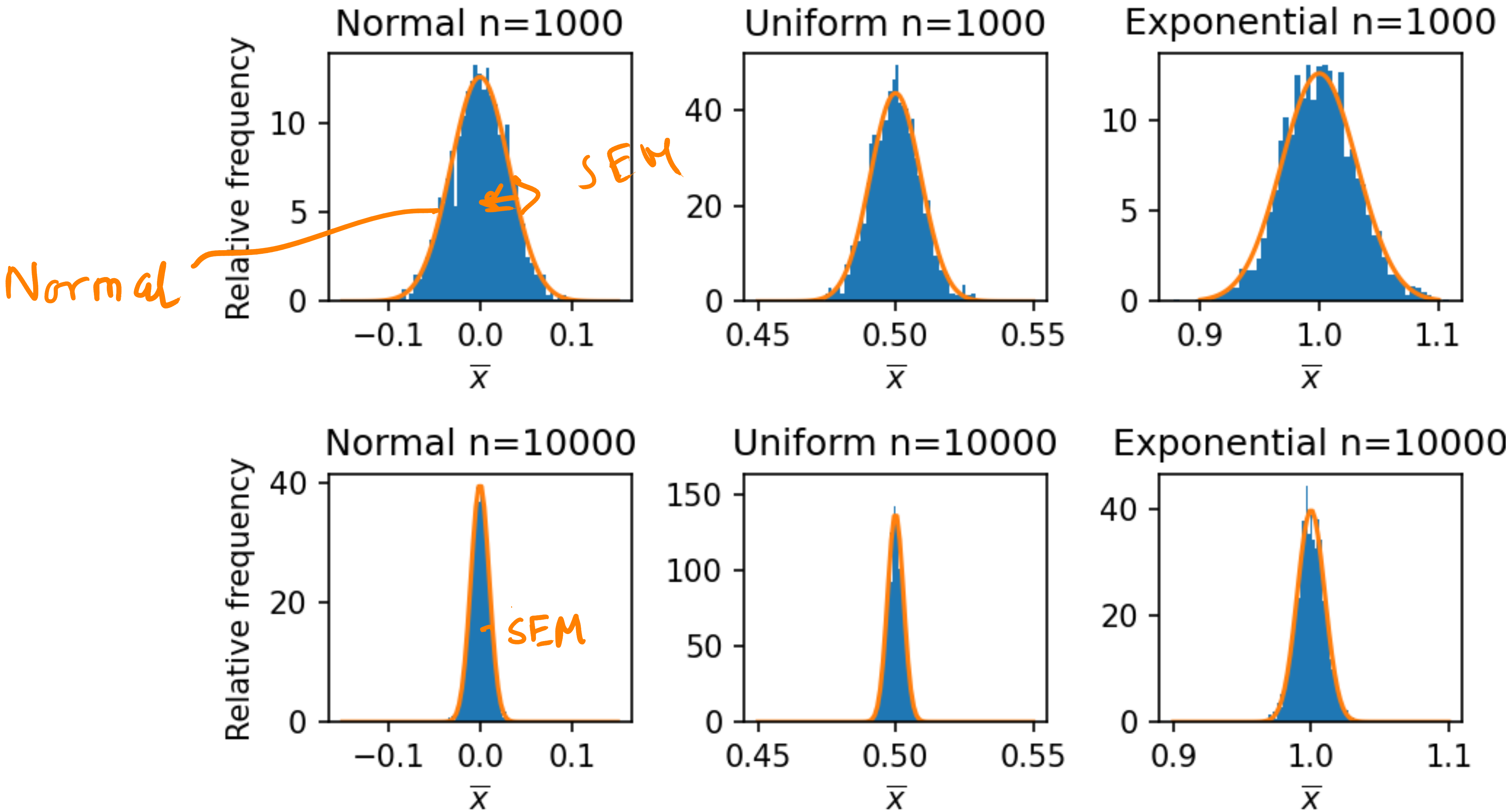
standard error in mean (SEM)

χ^2



Foundations of Data Science: Randomness, sampling and simulation – Distribution of sample mean 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

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

$E(\bar{X}) = \mu$

Sum $E(T_o) = n\mu$

Expected variance of the mean $\sigma^2_{\bar{x}}$

Standard error in 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 tends to 0.

$\Rightarrow$ Law of averages

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)