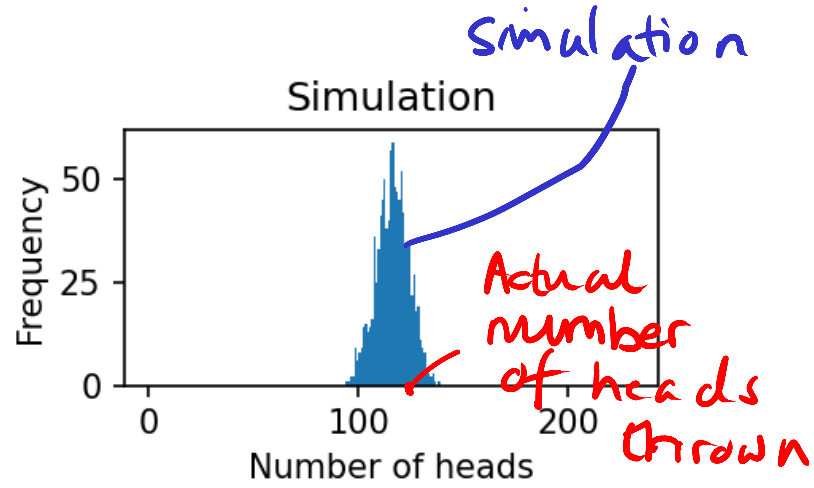
Foundations of Data Science:
Estimation –
Point estimation

Plan for statistical inference

1. Randomness, sampling and simulations (S1 Week 10)
2. Estimation, including confidence intervals (S1 Week 11)
3. Hypothesis testing (S2 Week 1)
4. Logistic regression (S2 Week 1)
5. A/B testing (S2 Week 2)

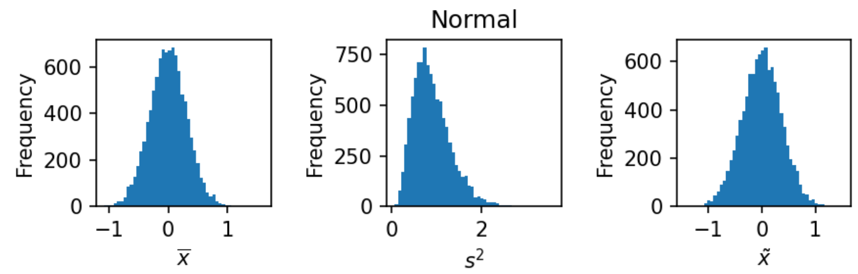
Last lecture...

1. Sampling
   - random
   - non-random

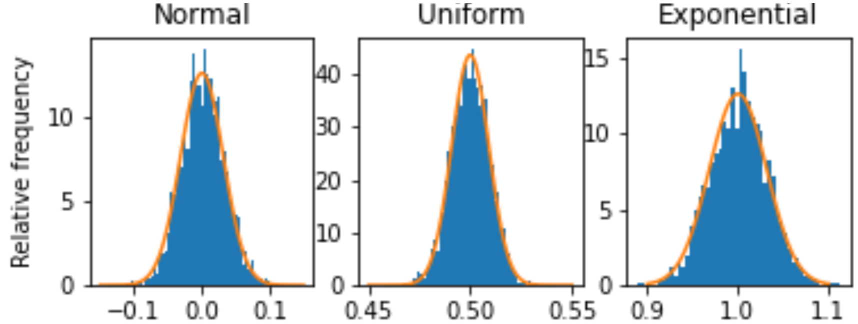


2. Inference on testing the hypothesis that the coin is biased
   - Statistical simulations

3. Sampling distributions of statistics
   - mean, variance, median



4. Sampling distribution of the mean in large samples
   - Central Limit Theorem



Today

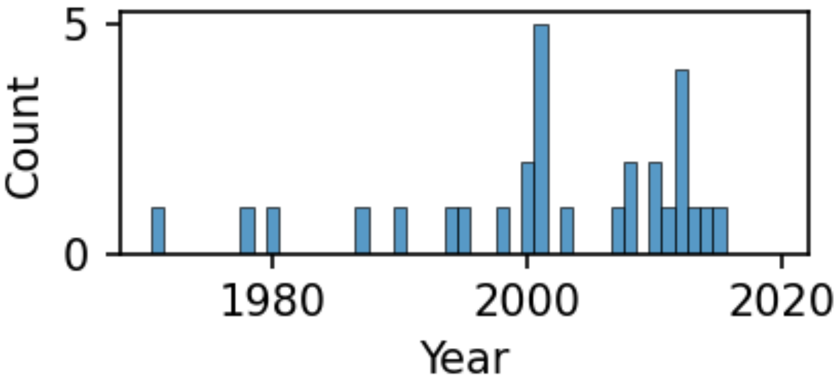
- Big idea: method to determine how precise our estimate of the average age of 2p coin is
  - Confidence interval

- Steps:
  1. Concept of estimator
  2. Sampling distribution of the estimator gives indication of uncertainty in estimate
  3. Confidence interval

Overview

Sample

$n = 29$



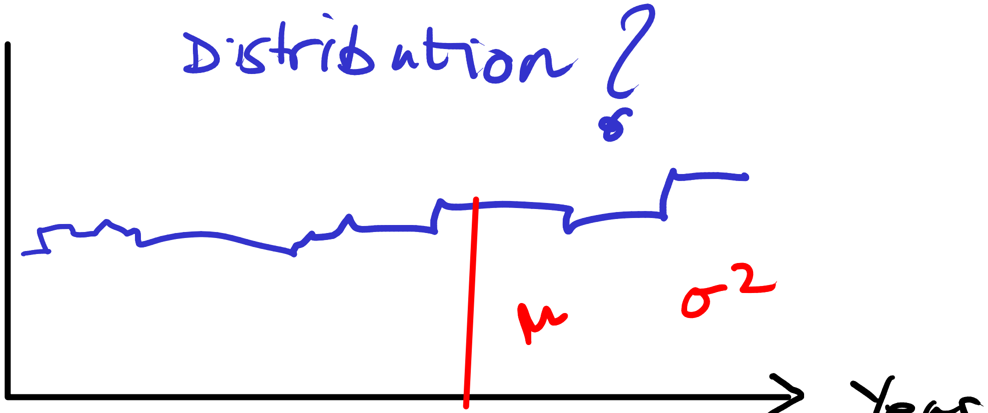
$\bar{x} = 2001.6$ years

$s = 11.4$ years

$\frac{s}{\sqrt{n}} = 2.1$ years

Population

$N \sim 1 \times 10^9$

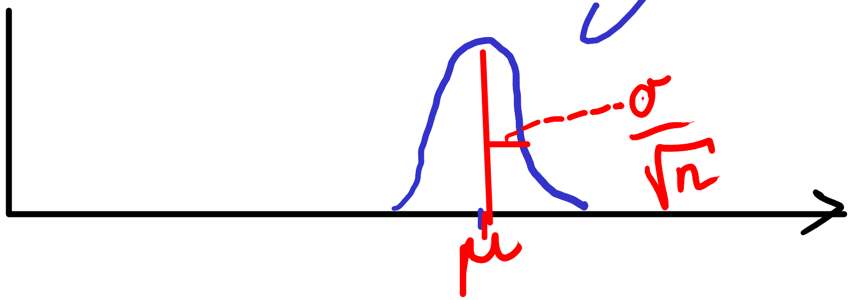


sample 1

$n = 29 \times$ 

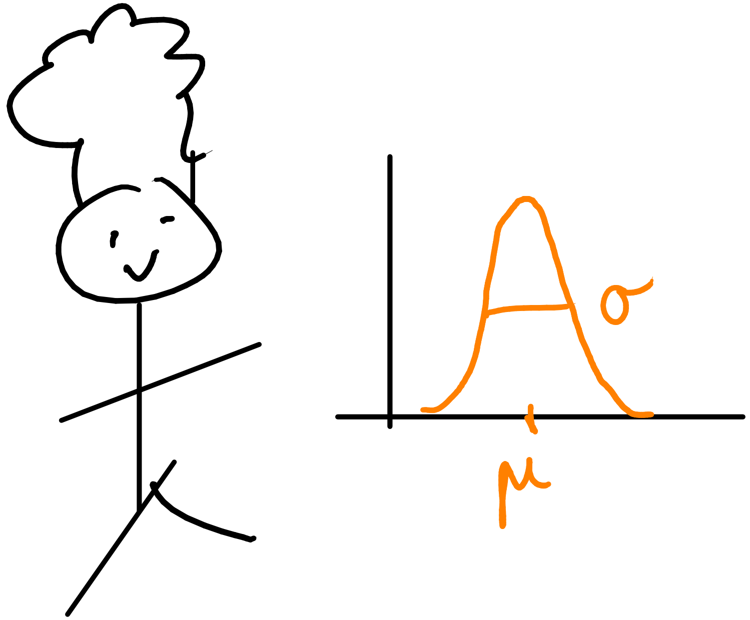
$n = 29 \times$ 

mean $\bar{x}_1 = 2000$ $\bar{x}_2 = 2002$



Sampling dist of mean

A population that's not countable



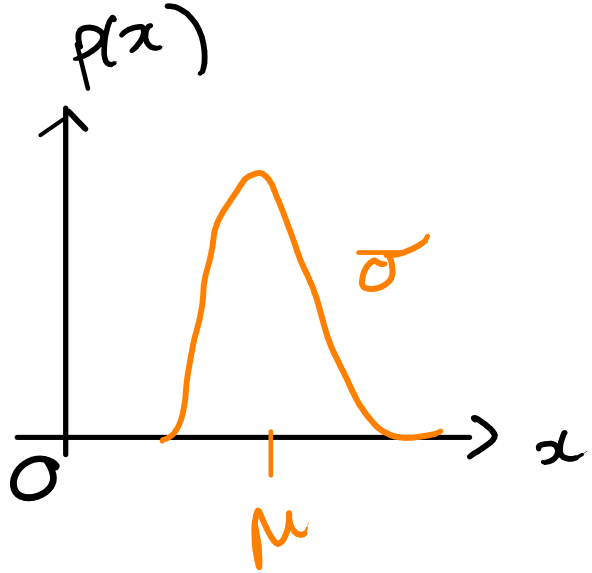
Parameters

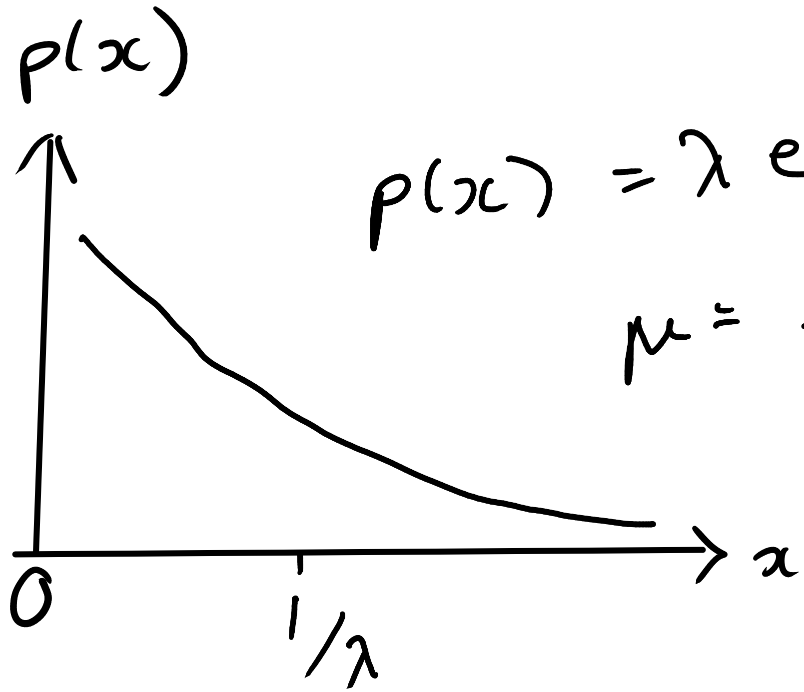
Of a finite population

Mean μ

Variance σ^2

Of a distribution





$$p(x) = \lambda e^{-\lambda x}$$

$$\mu = \frac{1}{\lambda} \qquad \sigma = \frac{1}{\lambda}$$

Problems

1. Construct a point estimator for parameters

2. Determine how accurate that estimate is using confidence intervals

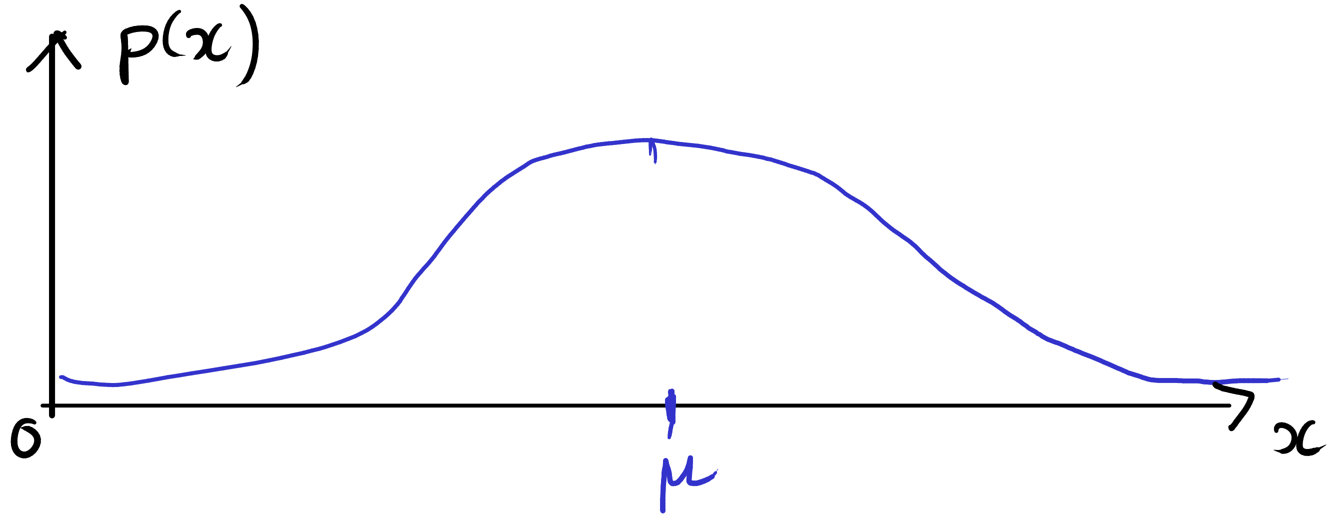
Notation: Generic parameter

Its point estimator

E.g.

There can be more than one estimator for a parameter

E.g. Symmetric distribution:



$\hat{\mu} = \ ?$

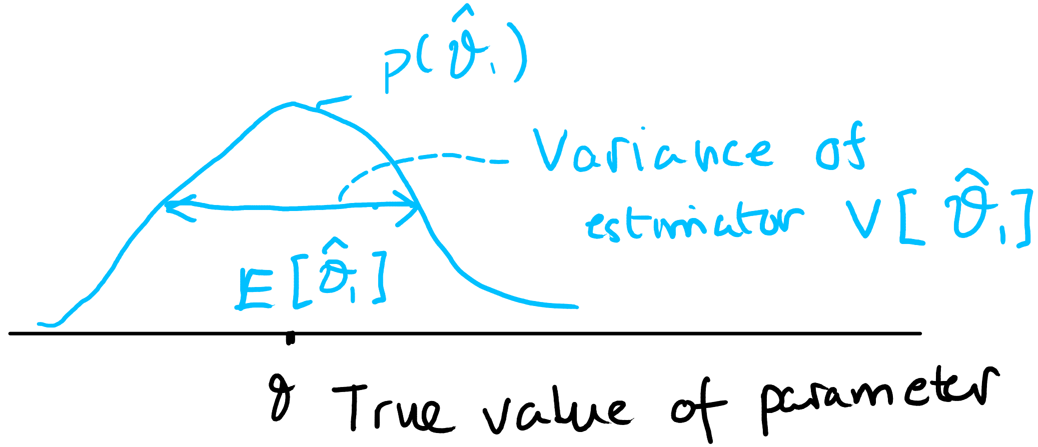
Can you think of something else we could estimate about the population of 2p coins?

- What would you estimate?
- How would you estimate it?
- Would there be any problems with your estimate?

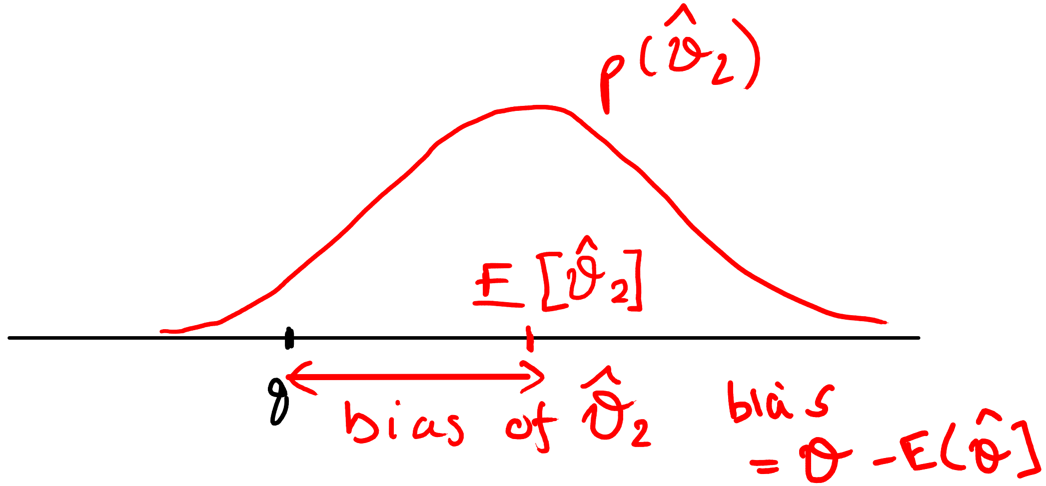
Foundations of Data Science: Estimation – Bias and variance

Estimation bias and variance

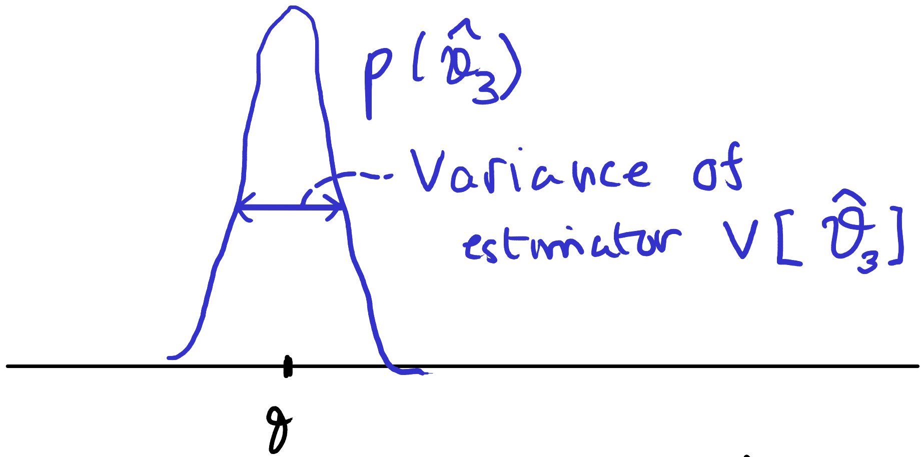
Unbiased estimator



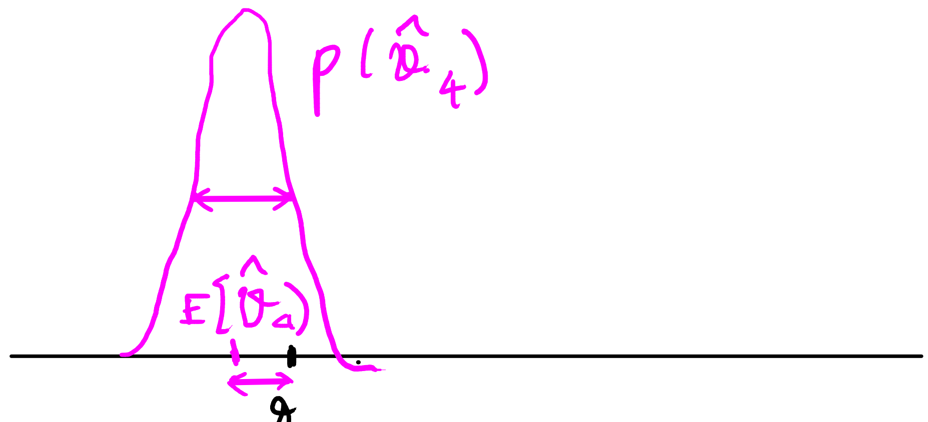
Biased estimator



Unbiased estimator with low variance



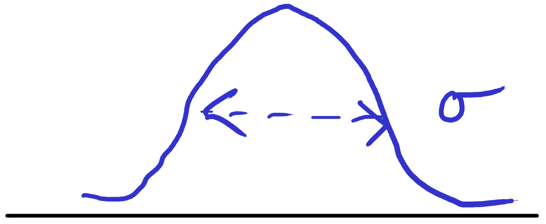
Biased estimator with low variance



$$MSE = E[(\vartheta - \hat{\vartheta})^2] = V[\hat{\vartheta}] + (\vartheta - E[\hat{\vartheta}])^2$$

Example: estimator of mean of normal distribution with known variance

Normal distribution

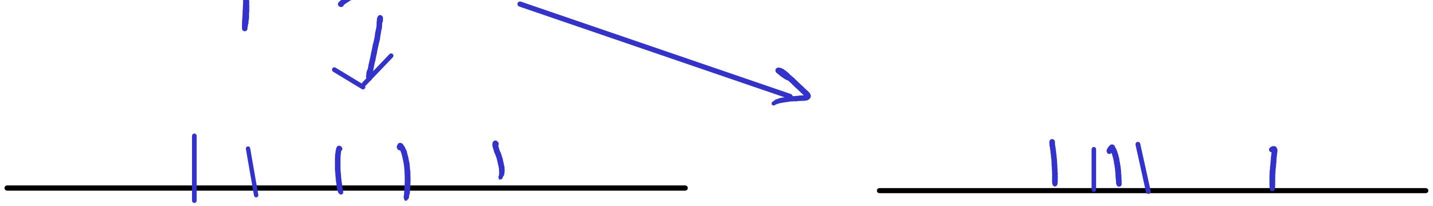


σ – Known

Estimator : $\bar{X} = \sum_{i=1}^{n} X_i$

μ ?

$n = 5$



Standardised variable $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ standardised normal distribution – the z-distribution

$E[\bar{X}] = \mu \Rightarrow E[\bar{X}] - \mu = 0 = \text{bias}$

$\text{MSE} = E[(\bar{X} - \mu)^2] = V[\bar{X}] = \dfrac{\sigma^2}{n}$

Example: a contrived estimator with bias

Estimator : $\hat{\mu} = \bar{X} + 1$

$$\text{bias} = E[\bar{X} + 1] - \mu = E[\bar{X}] + 1 - \mu$$
$$= 1$$

Example from machine learning

Suppose

1. We've used cross-validation to choose the hyperparamters in k-Nearest Neighbours
2. We've estimated the accuracy on the the testing folds the in cross-validation

Identify ϑ and $\hat{\vartheta}$

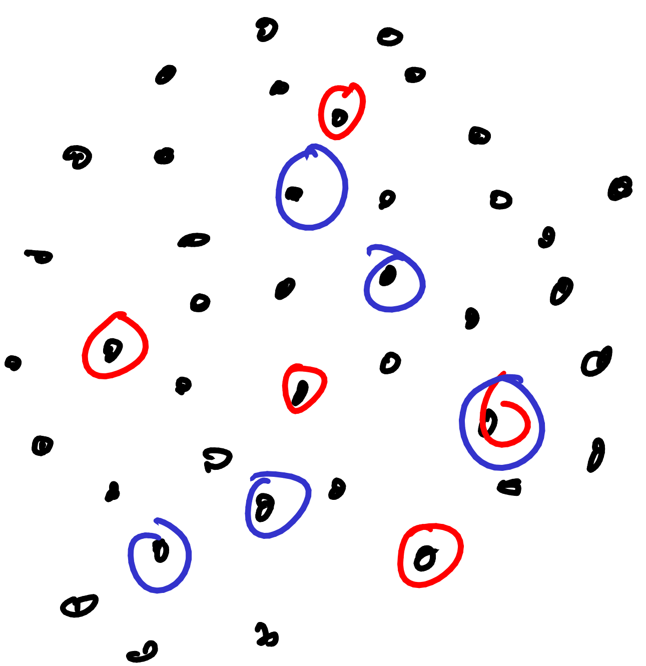
Is $\hat{\vartheta}$ an unbiassed estimator of ϑ ?

Foundations of Data Science: Estimation – Standard error

How far is $\hat{\vartheta}$ from ϑ ?

Ideal world : resample $\hat{\vartheta}$ from the population

E.g. $\hat{\vartheta} = \hat{\mu} = \bar{x} = \sum_{i=1}^{5} x_i$

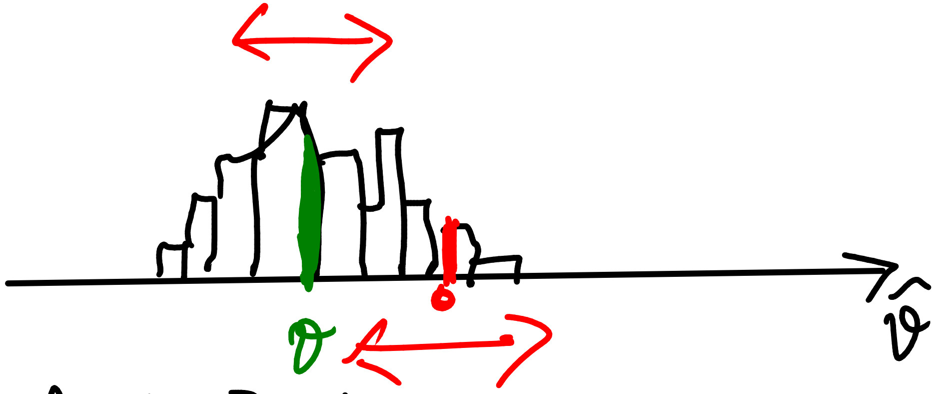


$\hat{\vartheta}_1$

$\hat{\vartheta}_2$

$\vdots$

$\hat{\vartheta}_R$



As $R \to \infty$

Variance tends to $V[\hat{\vartheta}]$

<u>Standard error of</u> an estimator $\sigma_{\hat{\vartheta}} = \sqrt{V[\hat{\vartheta}]}$

Real world

We have only one sample.

We can't resample from the population to estimate $V[\hat{\vartheta}]$

1. For the mean, we can estimate the standard error of the mean using the sample variance of the sample

2. For all estimators, we can use the boostrap method to estimate the distribution of the estimator, and thus the standard error of the estimator (next lecture)

Standard error of an estimator

How far is $\hat{\vartheta}$ from ϑ?

Standard error of an estimator $\sigma_{\hat{\vartheta}} = \sqrt{V[\hat{\vartheta}]}$

$$\simeq \sqrt{MSE(\hat{\vartheta})}$$

Standard error of the mean (SEM) $\sigma_{\hat{\mu}} = \dfrac{\sigma}{\sqrt{n}}$

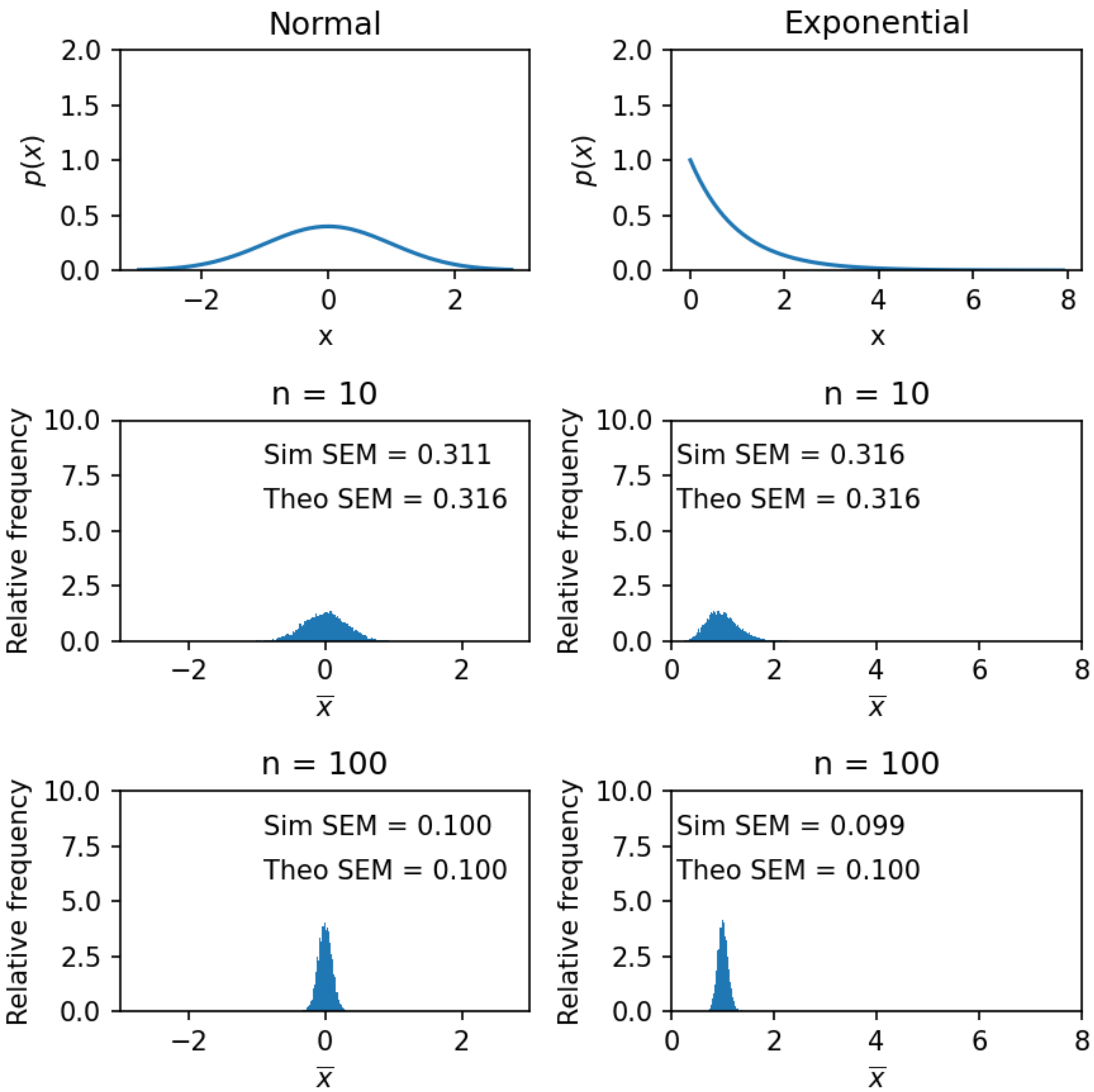
Standard error of mean for known distribution variance σ

Standard deviation

SEM

$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{10}} = 0.316$

$\sigma_{\hat{\mu}} = \frac{1}{\sqrt{100}}$



Estimated standard error for distribution with unknown variance σ

What if we don't know σ ?

Estimated S.E.

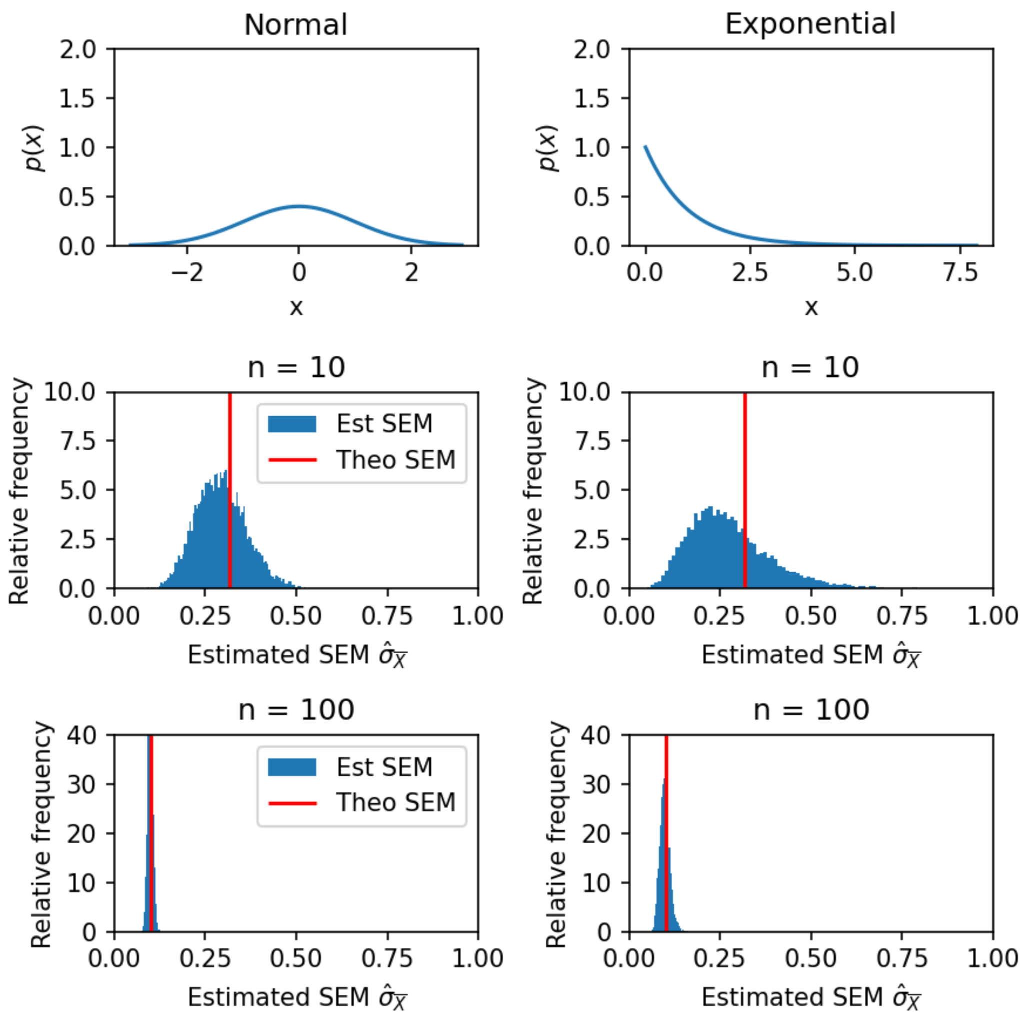
of estimator $\hat{\sigma}_{\hat{\theta}}$

Estimated SEM

$\hat{\sigma}_{\hat{\mu}} = \frac{S}{\sqrt{n}} \leftarrow$ r.v.

n large $\Rightarrow$

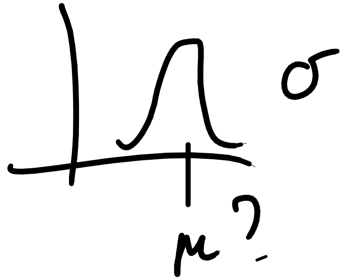
$\hat{\sigma}_{\hat{\mu}} \simeq \sigma_{\hat{\mu}}$



Problem with estimated SEM

Know σ : $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$

$\hat{\mu} = \bar{x}$



$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ normally distributed

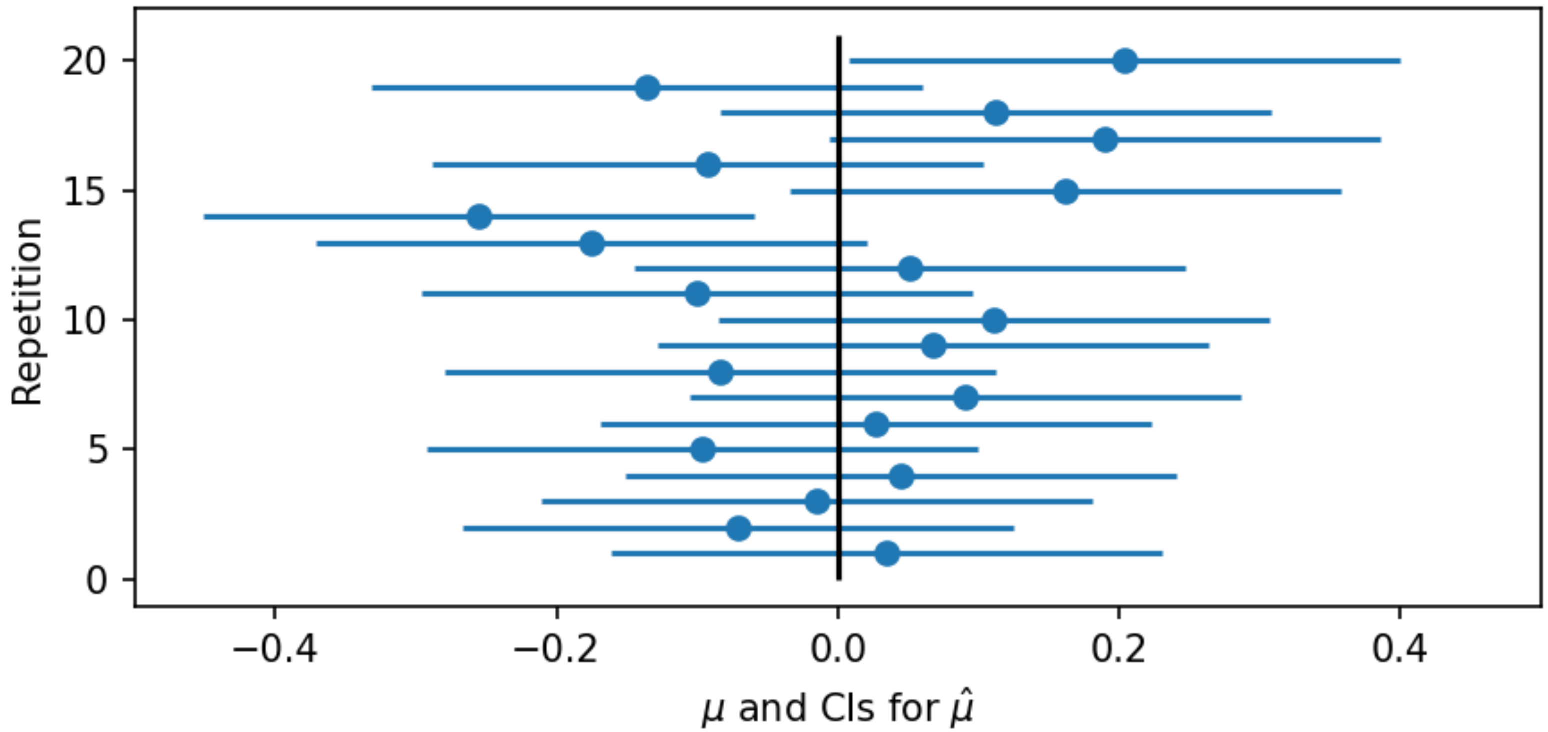
Random variable

Don't know σ : $\frac{\bar{X} - \mu}{S / \sqrt{n}} \Rightarrow$ Not normally dist.

Random variable

Distribution of estimator -> Confidence intervals

Given one point estimate $\hat{\vartheta}$ we want to be able to say that an interval has a specified chance of containing the true parameter ϑ



Summary

1. Progress on estimating the uncertainty in the estimate of the average year of a 2p coin
2. Estimators and parameters
3. Bias and variance of estimators
4. The estimator distribution and standard error
5. The distribution of the mean estimator for a distribution with known variance
5. The distribution of the mean estimator for a distribution with unknown variance
6. Introduction to the confidence interval