Inf2 - Foundations of Data Science: Estimation – Point estimation and confidence intervals

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Announcements

- Updated version of lecture notes
  - Causal inference material from lectures
  - K-NN regression and evaluation from lectures
  - Fixed/improved PCA errors (thanks for student input)
  - Fixed more errors reported by students and us

- Comprehension Questions prize draw
  - Please do 2 sets comprehension questions by 4pm on Tuesday to be entered in draw announced on Wednesday

Week 2 workshops

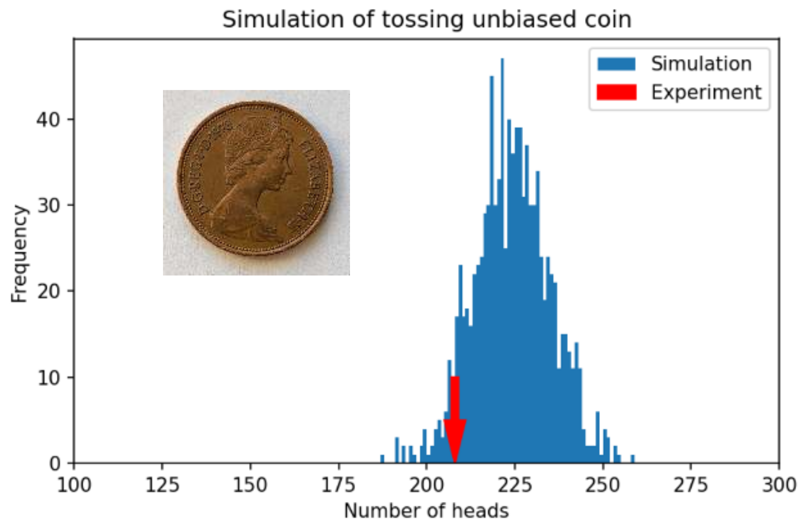
- Paper discussed in this week's workshops will appear on exam
- Not all concepts yet covered, but:
  - "prospective study", "confounders": see Ch 6 of Lecture notes and related lecture
  - "t-tests" - a statistical test for if the difference in two means is significant (big t, low p-value): see Ch 18 of lecture notes
  - "chi squared" χ^2 test: are the two quantities indepdent of each other (ch 17 off LNs). Big χ^2 low p-value -> not independent
  - "odds ratio" - how much more likely is one outcome than another 1 = the same ; >1 more likely - see Ch 20 of LNs

Plan for statistical inference

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

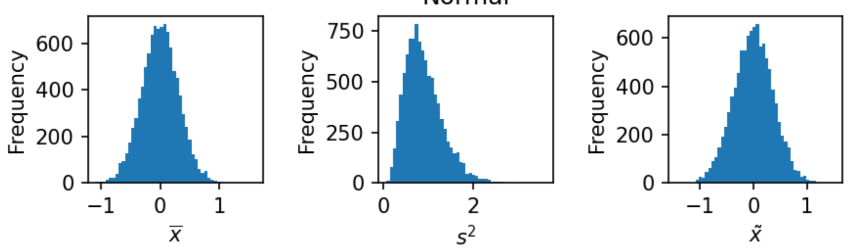
Last lecture...

1. Sampling
   - random
   - non-random

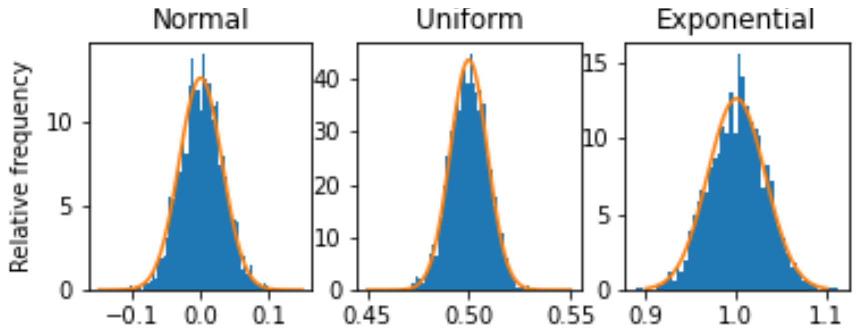


2. Inference on testing the hypothesis that the coin is biased
   - Statistical simulations

Mean Variance Median



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

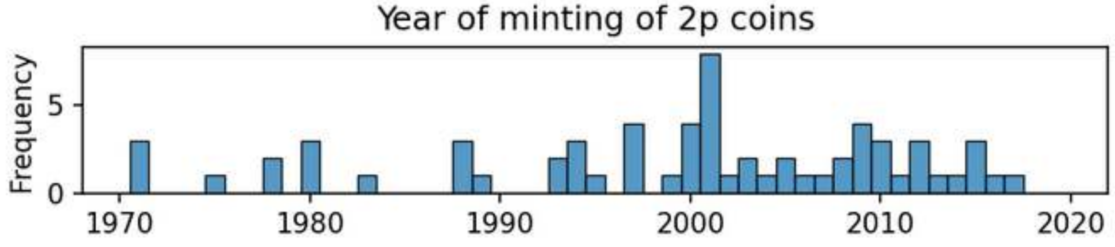
- Steps:
  1. Concepts of parameters and estimators
  2. Sampling distribution of the estimator gives indication of uncertainty in estimate
  3. First attempt at theoretical method of getting confidence interval
     - worked example
  4. Theory on bias and variance of estimators

Overview

Sample

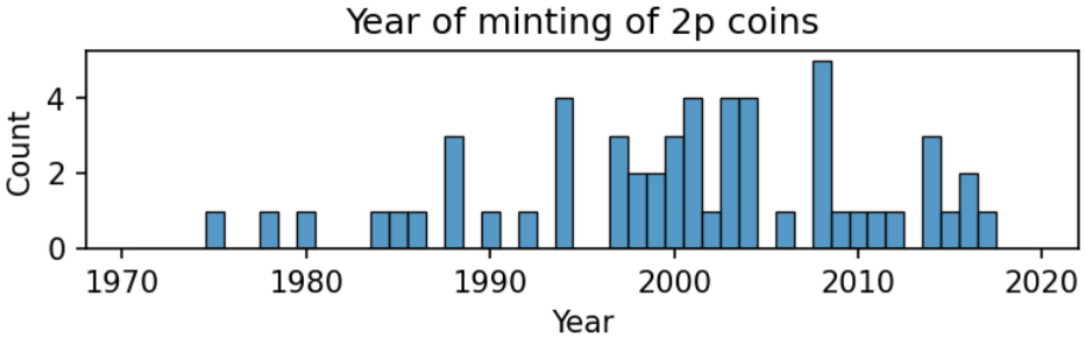
$n = 67$

This year's experiment



$\bar{x} = 1999.8$ yrs
$s = 12.5$ yrs
$n = 67$

The same experiment in 2025 ...

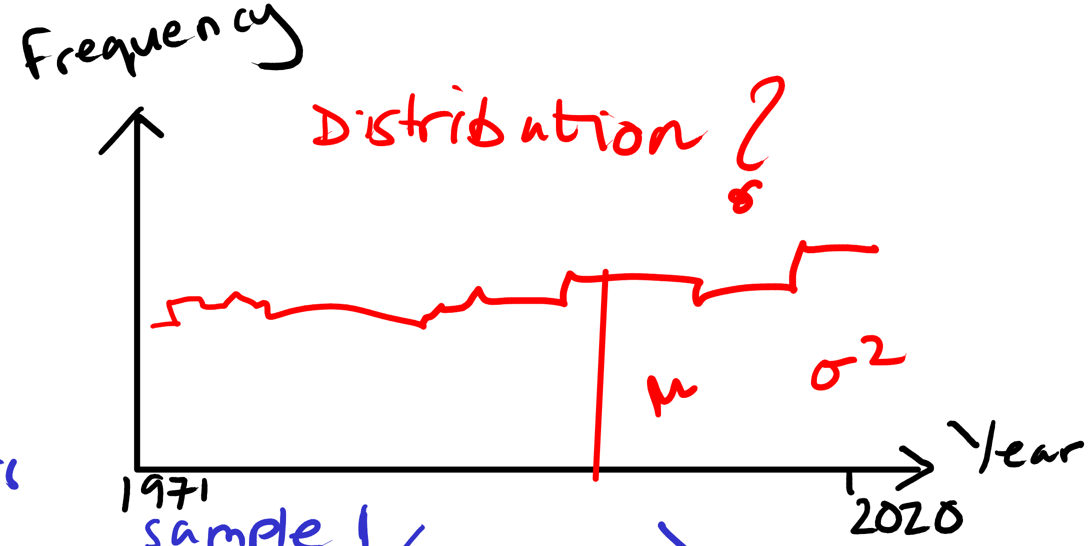


$\bar{x} = 2000.8$ yrs
$s = 10.4$ yrs
$n = 56$

In 2024: $\bar{x} = 2001.6$ yrs
$s = 11.4$ yrs
$n = 29$

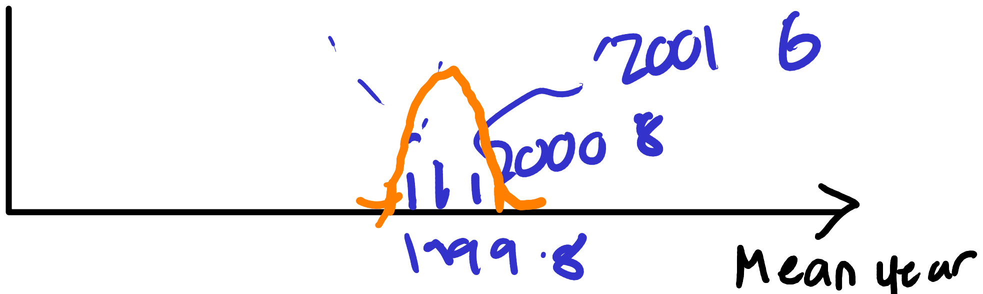
Population

$N \sim 1 \times 10^9$



sample 1 → $n = 67\times$

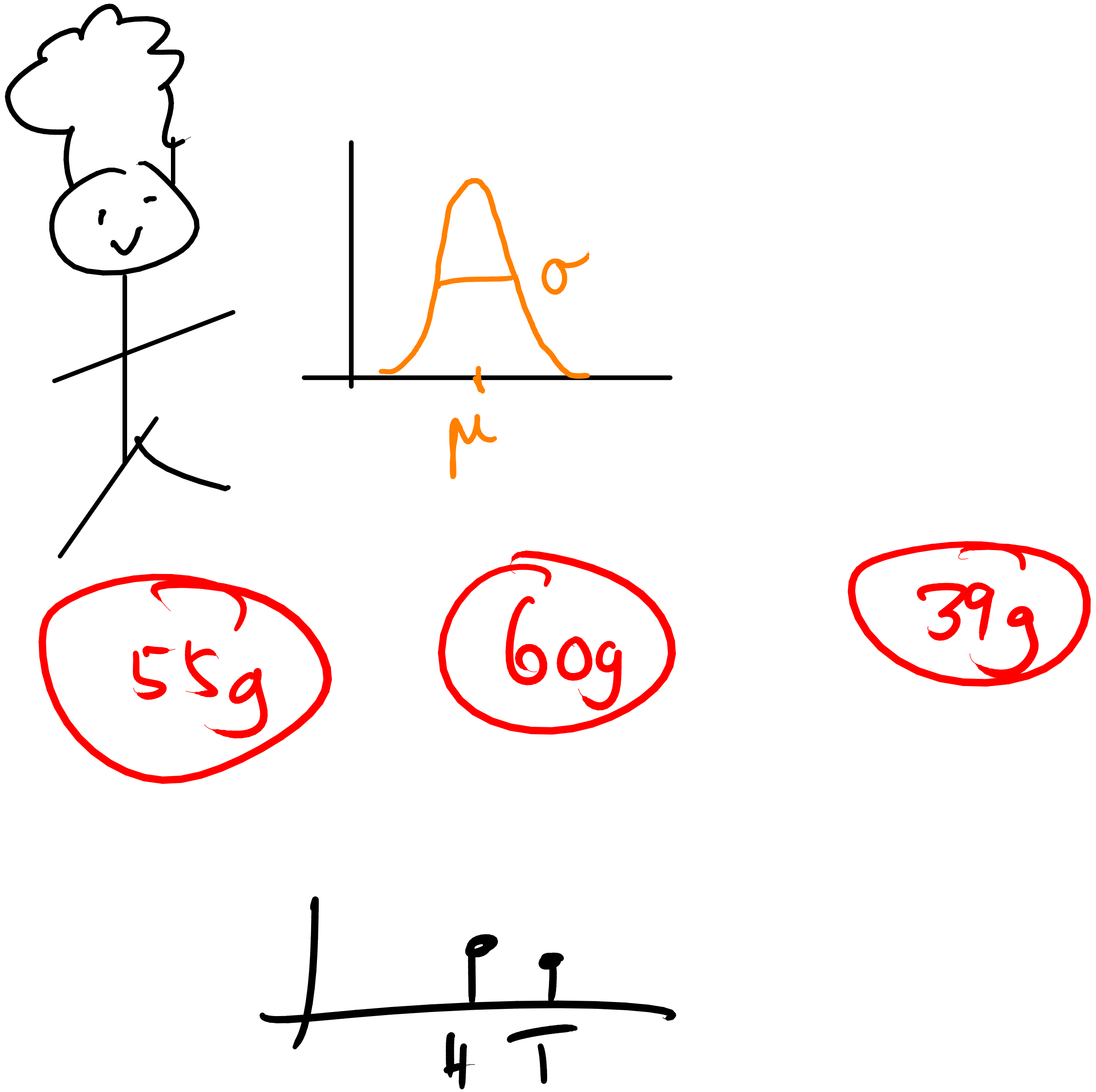
sample 2 ... → $n = 67\times$



Sampling dist of mean

Non-countable populations

Coin tossing



Parameters

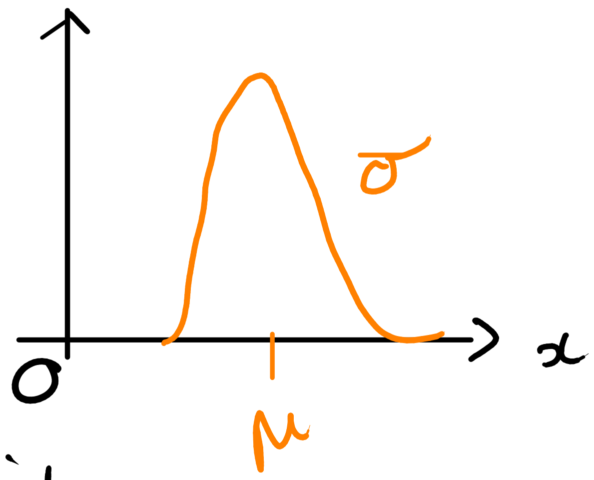
Of a finite population

Mean μ

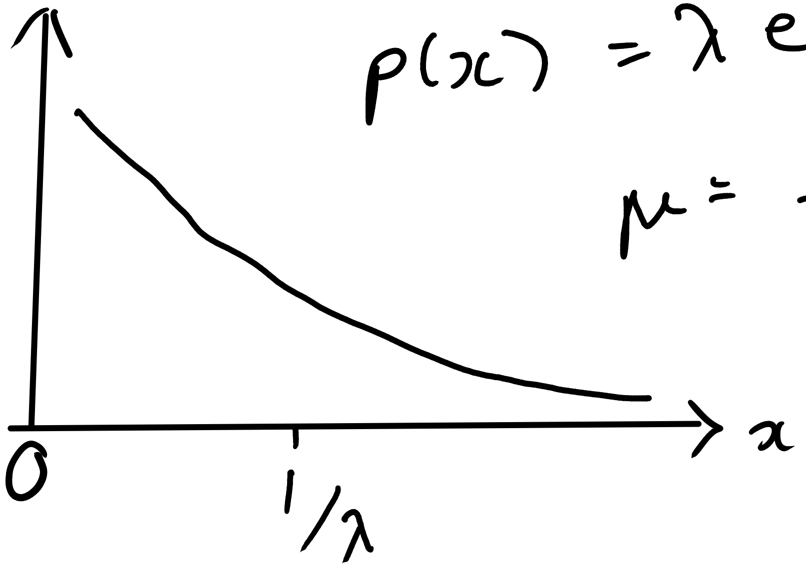
Variance σ^2

Of a distribution

$p(x)$ Normal



$p(x)$ Exponential



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

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

Estimation problems

1. Construct a point estimator for parameter of population

Generic parameter: θ — μ, σ^2, $\tilde{\mu}$

Corresponding estimator: $\hat{\theta}$ — $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\hat{\sigma}^2$, $\hat{\tilde{\mu}}$

Estimator is func. of data

2. Determine how accurate that estimate is using confidence intervals

Find $(u(x), v(X))$ s.t.

$$P(u(x) < \theta < v(X)) = \underbrace{1 - \alpha}_{95\%} \quad \alpha = 0.05$$

Two methods to estimate confidence intervals

1. Theoretical method (start this lecture, finish next)
   - works for some parameters and estimators

2. Statistical sampling (next lecture)
   - works for some parameters and estimators that the theoretical method doesn't work for

Inf2 - Foundations of Data Science: Estimation – First attempt at determing confidence intervals for the mean theoretically

THE UNIVERSITY of EDINBURGH informatics

FOUNDATIONS OF DATA SCIENCE

Coin year point estimation and confidence interval example

What is the point estimator of the mean age of the coins?

Parameter: mean age of coins in the pop. of coins

$\theta = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ where x_i is age of i^{th} coin.

Estimator: mean of sample

$\hat{\theta} = \hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ X_i " " " "

Knowing what we learned about sampling in the last lecture,
what might a good confidence interval of the mean be?

$(\hat{\mu} - a \times SEM, \hat{\mu} + a \times SEM)$

Variance

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$SEM = \hat{\sigma}_{\hat{\mu}}$ Estimated $\hat{\sigma}_{\hat{\mu}} = \frac{S}{\sqrt{n}}$

~~NOT $\hat{\mu} - aS, \hat{\mu} + bS$~~

Exercise: Suppose we had only 5 samples:
1975, 1984, 1998, 2008, 2014
What is the point estimate of the mean and the SEM?

CI: 1995.8 ± 7.3

(1988.5, 2003.1)

Point estimate of mean $\hat{\mu} = \bar{x}$

$= \frac{1}{5}(1974 + 1984 + 1998 + 2008 + 2014) = 1995.8$

Point est of variance $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{5}(x_i^2 - \bar{x}^2)$

$s^2 = \frac{1}{5-1}\left[(1975^2 + 1984^2 + 1998^2 + 2008^2 + 2014^2) - 5 \times 1995.8^2\right]$

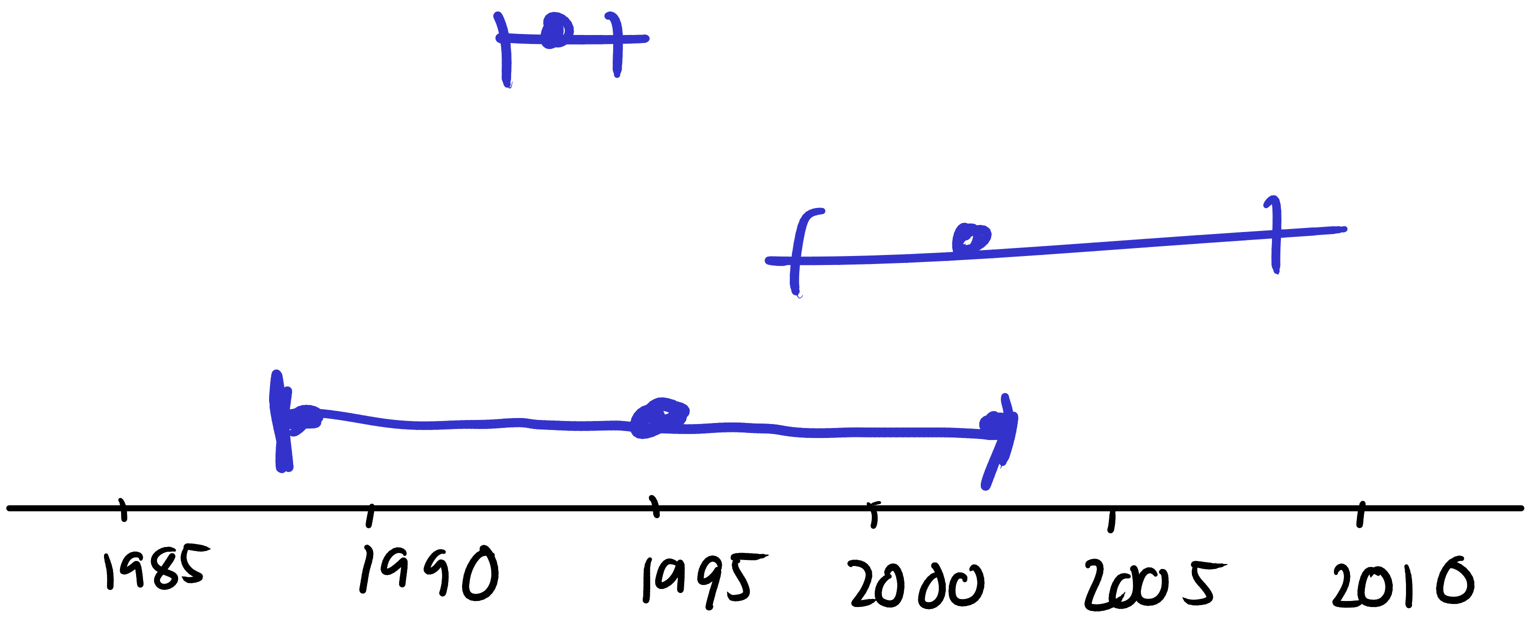
$= 264.200$

sample st. dev. $s = \sqrt{264.000} = 16.254$

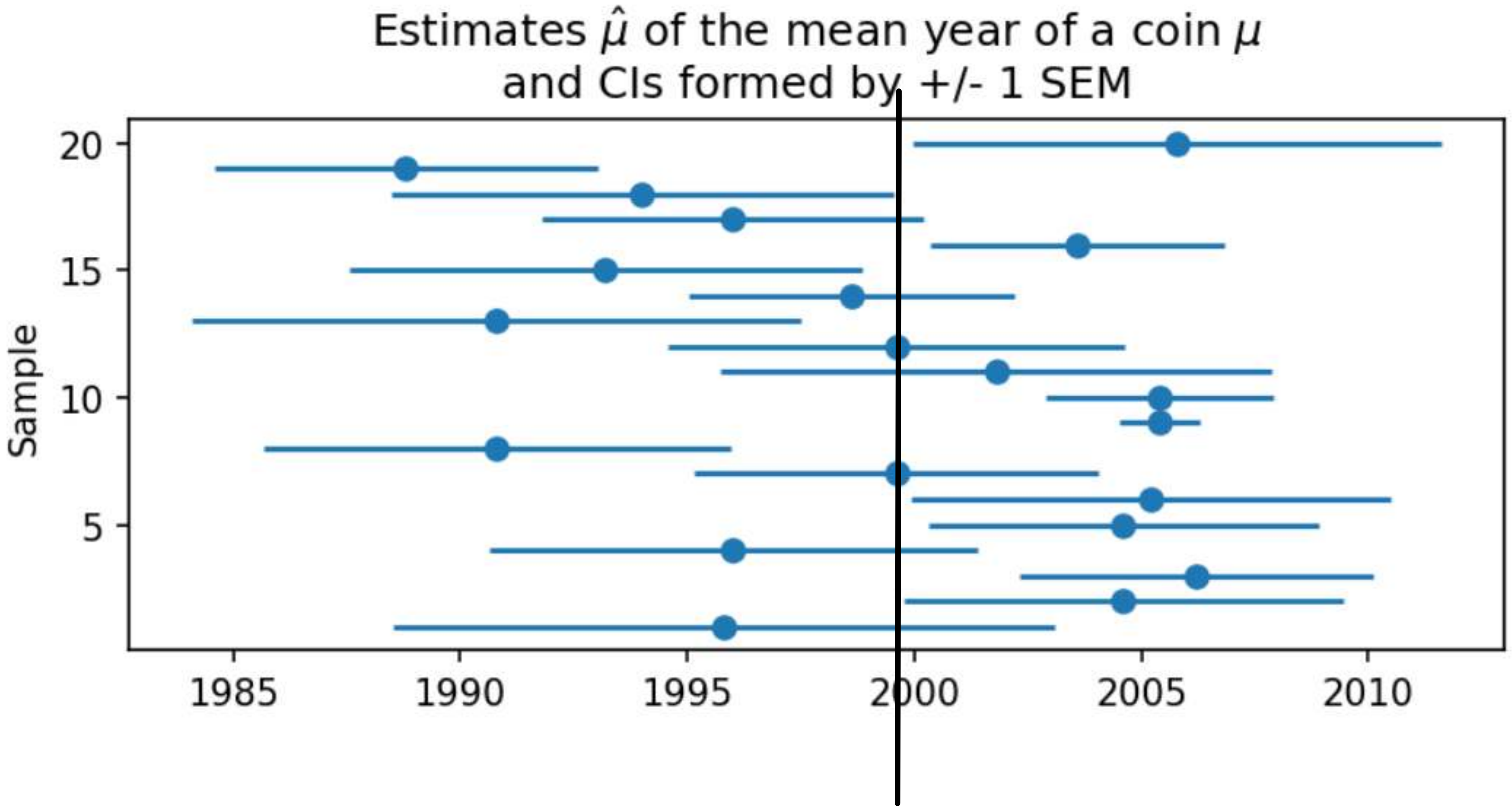
$SEM = \frac{s}{\sqrt{n}} = \frac{s}{\sqrt{5}} = \frac{16.254}{\sqrt{5}} = 7.267 \simeq 7.3$

If you know how to do the above, try computing the same for:
2003, 2004, 2006, 2006, 2008

Graphical representation of confidence intervals



Confidence intervals from different samples



Remaining question for next lecture

How do we calibrate the width of the confidence interval so that there is a specified chance that it encloses the true value?

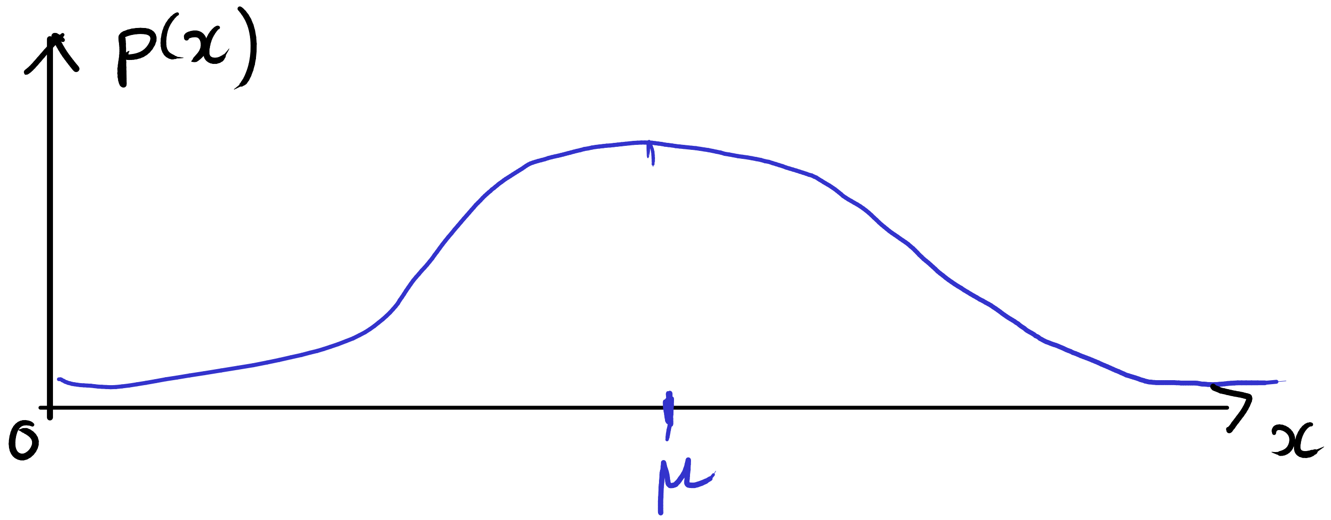
Inf2 - Foundations of Data Science: Estimation – Bias and variance

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

There can be more than one estimator for a parameter

E.g. Symmetric distribution:



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

$= \tilde{X}$

How would we estimate the first year 2p coins were minted (made) from the population of years of 2p coins?

- Would there be any problems with your estimate?

$\hat{X}_{min} = \min(X)$ – estimator

$E(X_{min}) > x_{min}$ ⇒ biased estmator

NOTE AFTER LECTURE

True $\vartheta = x_{min} = 1971$

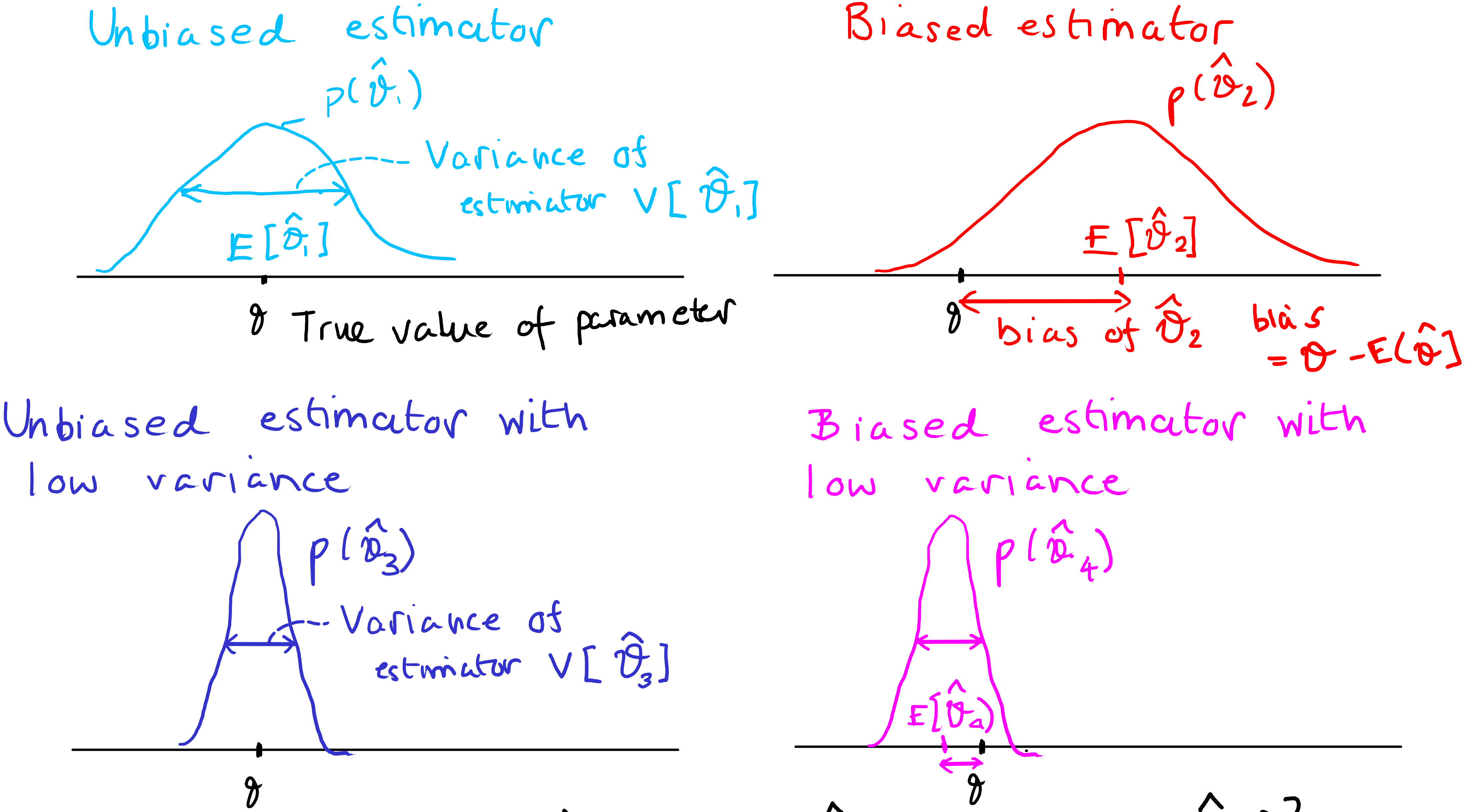
Some repeats of experiment earliest year in sample has been after 1971.

NOTE AFTER LECTURE:

$\vartheta = x_{min}$ ∀ 1, …, N coins

$\hat{\vartheta} = \hat{X}_{min}$

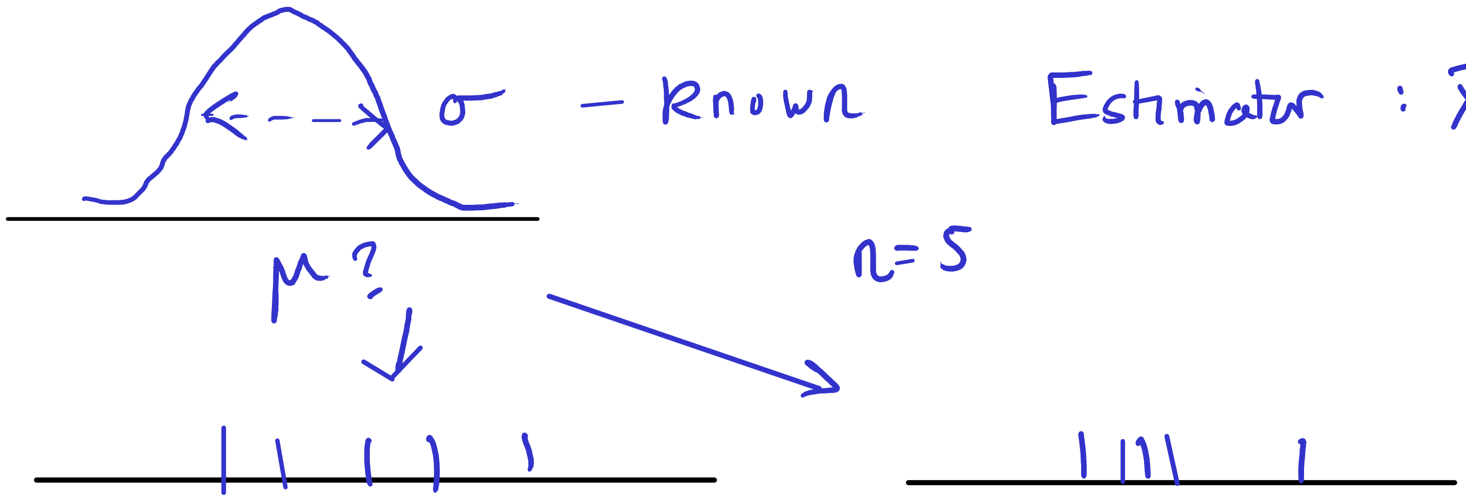
Estimation bias and variance



$$MSE = E[(\vartheta - \hat{\vartheta})^2] = \underbrace{V[\hat{\vartheta}]}_{\text{variance}} + \underbrace{(\vartheta - E[\hat{\vartheta}])^2}_{\text{bias}}$$

Example: estimator of mean of normal distribution with known variance

Normal distribution



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

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

$\Rightarrow$ $\bar{X}$ is an unbiased estimator

Mean Square Error $E[(\bar{X}-\mu)^2] = V[\bar{X}] = \frac{\sigma^2}{n}$

Examples: estimator with bias

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

Estimator of variance (see lecture notes):

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$ – biased

See lecture notes

Bias can be helpful! See comprehension questions on bias and variance

Example from machine learning

Suppose
1. We train k-Nearest Neighbours on some training data
2. We measure the accuracy achieved on the training set $\hat{\vartheta}$
3. We use this accuracy as an estimate of accuracy on unseen data ϑ

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

NOTE AFTER LECTURE:

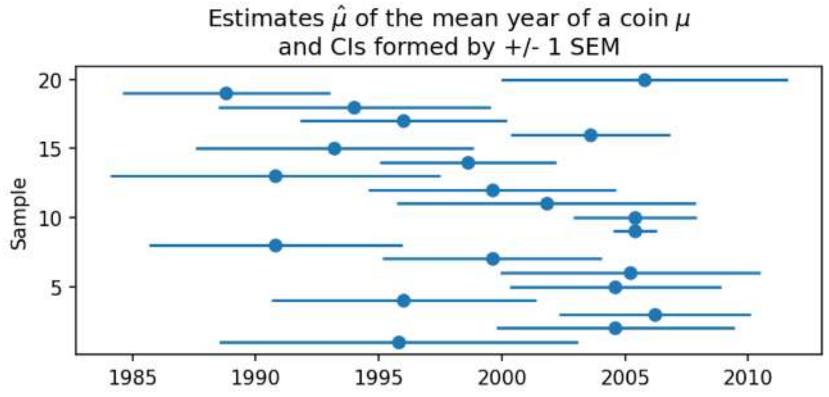
No, $\hat{\vartheta}$ is biased, because it is an estimate of performance on data that the model has been trained on, not an estimate of performance on unseen data.

Remember 1-NN example of perfect performance on training data.

Next lecture

We have only one sample.

We can't resample from the population to estimate $V[\hat{\vartheta}]$
and hence get a confidence interval



1. For the mean, we can estimate the standard error of the mean using the sample variance of the sample (see above) and adjust to get desired level of confidence

2. For other estimators, we can use the boostrap method to estimate the distribution of the estimator, and thus the standard error of the estimator

Summary

1. Progress on estimating the uncertainty in the estimate of the average year of a 2p coin
2. Estimators and parameters
3. Introduction to the confidence interval - theoretical method
4. Bias and variance of estimators