Introduction to Probability

Lecture 11: Estimators (Part II)

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

Easter 2023

UNIVERSITY OF CAMBRIDGE

Outline

Recap: Unbiased Estimators and Bias

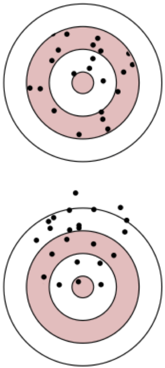
Definition

An estimator T is called an unbiased estimator for a parameter θ if

$$\mathbf{E}[T] = \theta,$$

irrespective of the value θ. The bias is defined as

$$\mathbf{E}[T] - \theta = \mathbf{E}[T - \theta].$$



Source: Edwin Leuven (Point Estimation)

Recap: Unbiased Estimators and Bias

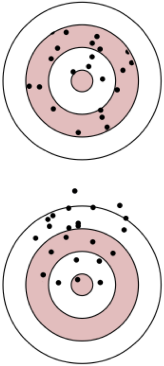
Definition

An estimator T is called an unbiased estimator for a parameter θ if

$$\mathbf{E}[\,T\,] = \theta,$$

irrespective of the value θ. The bias is defined as

$$\mathbf{E}[\,T\,] - \theta = \mathbf{E}[\,T - \theta\,].$$



Source: Edwin Leuven (Point Estimation)

- How can we measure the accuracy of an estimator?
  $\rightsquigarrow$ bias and mean-squared error
- If there are several unbiased estimators, which one to choose? $\rightsquigarrow$ mean-squared error (or variance)

Outline

Estimating Population Sizes (First Version)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an unknown parameter N (so $N = \theta$)
- Each of the IDs is drawn without replacement from the discrete uniform distribution over $\{1, 2, \dots, N\}$

Estimating Population Sizes (First Version)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an unknown parameter N (so $N = \theta$)
- Each of the IDs is drawn without replacement from the discrete uniform distribution over $\{1, 2, \ldots, N\}$
- This is also known as Tank Estimation Problem or (Discrete) Taxi Problem

Estimating Population Sizes (First Version)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an unknown parameter N (so $N = \theta$)
- Each of the IDs is drawn without replacement from the discrete uniform distribution over $\{1, 2, \ldots, N\}$
- This is also known as Tank Estimation Problem or (Discrete) Taxi Problem

$$7, 3, 10, 46, 14$$

Estimating Population Sizes (First Version)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an unknown parameter N (so $N = \theta$)
- Each of the IDs is drawn without replacement from the discrete uniform distribution over $\{1, 2, \ldots, N\}$
- This is also known as Tank Estimation Problem or (Discrete) Taxi Problem

$$7, 3, 10, 46, 14$$

Warning

- As before, we denote the samples $X_1, X_2, \ldots, X_n$

Estimating Population Sizes (First Version)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an unknown parameter N (so $N = \theta$)
- Each of the IDs is drawn without replacement from the discrete uniform distribution over $\{1, 2, \ldots, N\}$
- This is also known as Tank Estimation Problem or (Discrete) Taxi Problem

$$7, 3, 10, 46, 14$$

Warning

- As before, we denote the samples $X_1, X_2, \ldots, X_n$
- Since sampling is without replacement, these are:

Estimating Population Sizes (First Version)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an unknown parameter N (so $N = \theta$)
- Each of the IDs is drawn without replacement from the discrete uniform distribution over $\{1, 2, \ldots, N\}$
- This is also known as Tank Estimation Problem or (Discrete) Taxi Problem

$$7, 3, 10, 46, 14$$

Warning

- As before, we denote the samples $X_1, X_2, \ldots, X_n$
- Since sampling is without replacement, these are:
  - they are not independent! (but identically distributed)

Estimating Population Sizes (First Version)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an unknown parameter N (so $N = \theta$)
- Each of the IDs is drawn without replacement from the discrete uniform distribution over $\{1, 2, \ldots, N\}$
- This is also known as Tank Estimation Problem or (Discrete) Taxi Problem

$$7, 3, 10, 46, 14$$

Warning

- As before, we denote the samples $X_1, X_2, \ldots, X_n$
- Since sampling is without replacement, these are:
  - they are not independent! (but identically distributed)
  - their number must satisfy $n \leq N$

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

- The sample mean is

$$\overline{X}_n = \qquad .$$

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

- The sample mean is

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

- The sample mean is

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- Linearity of expectation applies (even for dependent random var.!):

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

- The sample mean is

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- Linearity of expectation applies (even for dependent random var.!):

$$\mathbf{E}\left[\,\overline{X}_n\,\right] = \frac{n \cdot \mathbf{E}\left[\,X_1\,\right]}{n} = \mathbf{E}\left[\,X_1\,\right]$$

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

- The sample mean is

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- Linearity of expectation applies (even for dependent random var.!):

$$\mathbf{E}\left[\,\overline{X}_n\,\right] = \frac{n \cdot \mathbf{E}\left[\,X_1\,\right]}{n} = \mathbf{E}\left[\,X_1\,\right]$$
$$= \sum_{i=1}^{N} i \cdot \frac{1}{N}$$

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

- The sample mean is

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- Linearity of expectation applies (even for dependent random var.!):

$$\mathbf{E}\left[\,\overline{X}_n\,\right] = \frac{n \cdot \mathbf{E}\left[\,X_1\,\right]}{n} = \mathbf{E}\left[\,X_1\,\right]$$

$$= \sum_{i=1}^{N} i \cdot \frac{1}{N} = \frac{N+1}{2}.$$

First Estimator Based on Sample Mean

Example 1

Construct an unbiased estimator using the sample mean.

Answer

- The sample mean is

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- Linearity of expectation applies (even for dependent random var.!):

$$\mathbf{E}\left[\overline{X}_n\right] = \frac{n \cdot \mathbf{E}\left[X_1\right]}{n} = \mathbf{E}\left[X_1\right] = \sum_{i=1}^{N} i \cdot \frac{1}{N} = \frac{N+1}{2}.$$

- Thus we obtain an unbiased estimator by

$$T_1 := 2 \cdot \overline{X}_n - 1.$$

Example: Odd Behaviour of T_1

- Suppose $n = 5$

Example: Odd Behaviour of T_1

- Suppose $n = 5$
- Let the sample be

$$7, 3, 10, 46, 14$$

Example: Odd Behaviour of T_1

- Suppose $n = 5$
- Let the sample be

$$7, 3, 10, 46, 14$$

- The estimator returns:

$$T_1 = 2 \cdot \overline{X}_n - 1 =$$

Example: Odd Behaviour of T_1

- Suppose $n = 5$
- Let the sample be

$$7, 3, 10, 46, 14$$

- The estimator returns:

$$T_1 = 2 \cdot \overline{X}_n - 1 = 2 \cdot \frac{80}{5} - 1 =$$

Example: Odd Behaviour of T_1

- Suppose $n = 5$
- Let the sample be

$$7, 3, 10, 46, 14$$

- The estimator returns:

$$T_1 = 2 \cdot \overline{X}_n - 1 = 2 \cdot \frac{80}{5} - 1 = 31 \quad ☹$$

Example: Odd Behaviour of T_1

- Suppose $n = 5$
- Let the sample be

$$7, 3, 10, 46, 14$$

- The estimator returns:

$$T_1 = 2 \cdot \overline{X}_n - 1 = 2 \cdot \frac{80}{5} - 1 = 31 \quad ☹$$

This estimator will often unnecessarily underestimate the true value N.

Example: Odd Behaviour of T_1

- Suppose $n = 5$
- Let the sample be

$$7, 3, 10, 46, 14$$

- The estimator returns:

$$T_1 = 2 \cdot \overline{X}_n - 1 = 2 \cdot \frac{80}{5} - 1 = 31 \quad ☹$$

This estimator will often unnecessarily underestimate the true value N.

It is possible (but difficult!) to prove $\mathbf{P}\left[\, T_1 < \max(X_1, X_2, \ldots, X_n) \,\right] \approx 0.5$

Example: Odd Behaviour of T_1

- Suppose $n = 5$
- Let the sample be

$$7, 3, 10, 46, 14$$

- The estimator returns:

$$T_1 = 2 \cdot \overline{X}_n - 1 = 2 \cdot \frac{80}{5} - 1 = 31 \quad ☹$$

This estimator will often unnecessarily underestimate the true value N.

It is possible (but difficult!) to prove $\mathbf{P}\left[\, T_1 < \max(X_1, X_2, \ldots, X_n) \,\right] \approx 0.5$

- Achieving unbiasedness alone is not a good strategy
- Improvement: find an estimator which always returns a value at least $\max(X_1, X_2, \ldots, X_n)$

Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$

Intuition: Constructing an Estimator based on Maximum

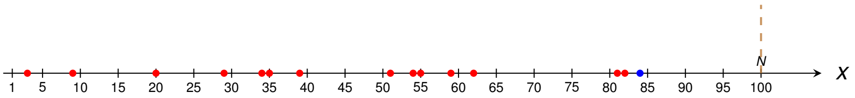
- Suppose $N = 100$ and $n = 15$
- Our samples are:

$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, \color{blue}{84}, 59, 3, 34, 55$$

Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

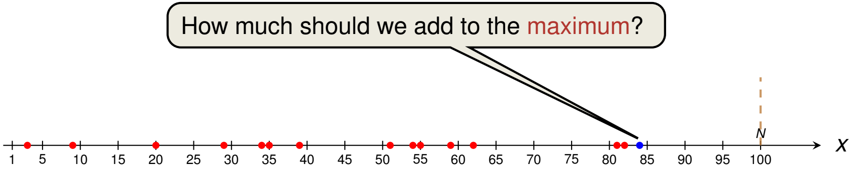
$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, 84, 59, 3, 34, 55$$



Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

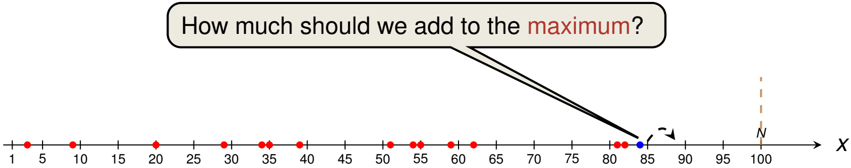
$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, {\color{blue}84}, 59, 3, 34, 55$$



Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

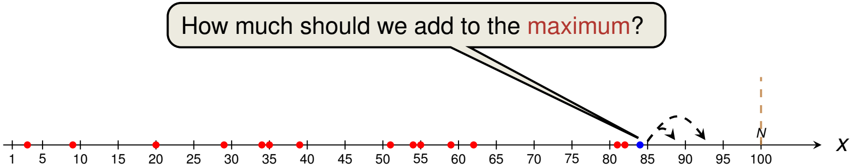
$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, \textcolor{blue}{84}, 59, 3, 34, 55$$



Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

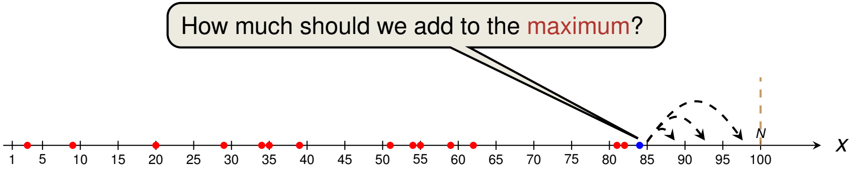
$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, {\color{blue}84}, 59, 3, 34, 55$$



Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

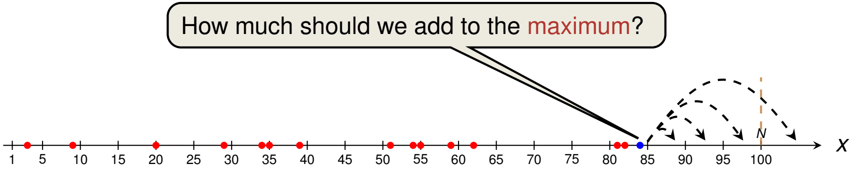
$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, \color{blue}{84}, 59, 3, 34, 55$$



Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, \textcolor{blue}{84}, 59, 3, 34, 55$$

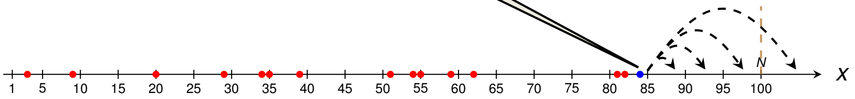


Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, \textcolor{blue}{84}, 59, 3, 34, 55$$

How much should we add to the maximum?

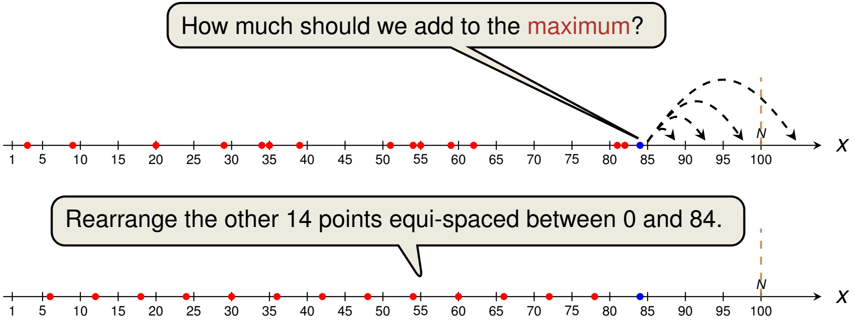


Rearrange the other 14 points equi-spaced between 0 and 84.

Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

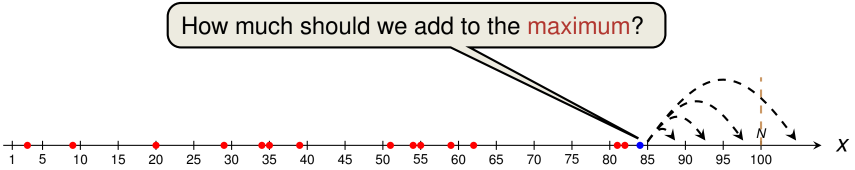
$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, \color{blue}{84}, 59, 3, 34, 55$$

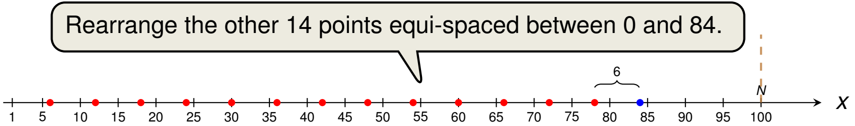


Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, {\color{blue}84}, 59, 3, 34, 55$$

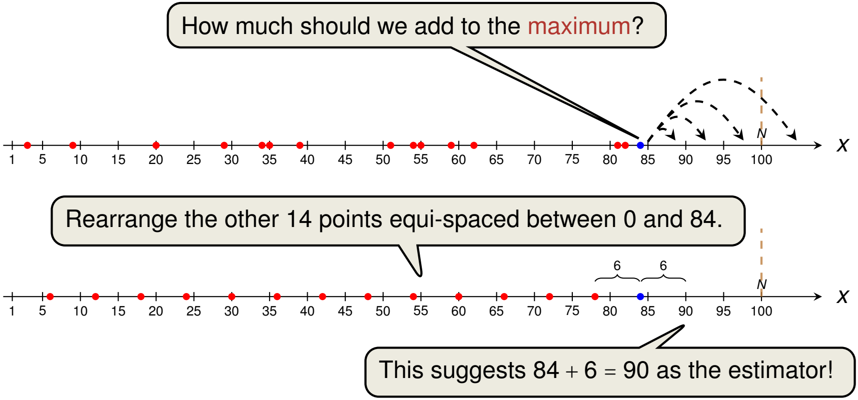




Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

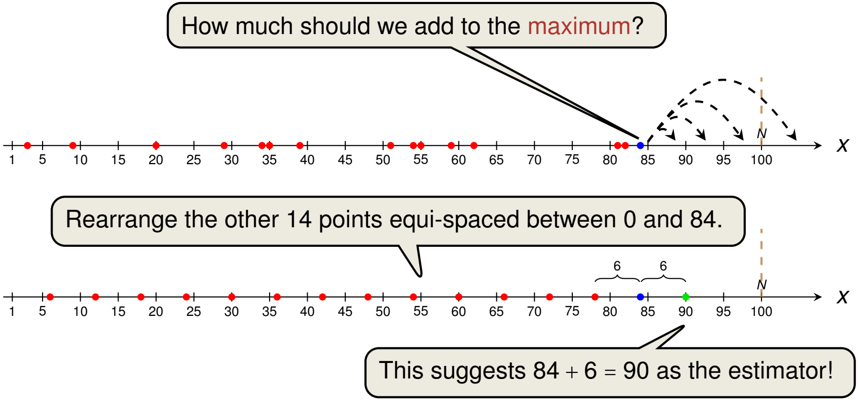
$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, {\color{blue}84}, 59, 3, 34, 55$$



Intuition: Constructing an Estimator based on Maximum

- Suppose $N = 100$ and $n = 15$
- Our samples are:

$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, {\color{blue}84}, 59, 3, 34, 55$$

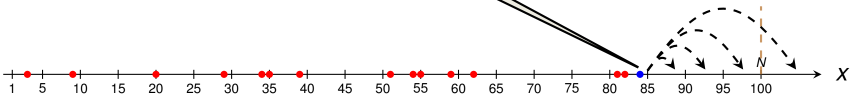


Intuition: Constructing an Estimator based on Maximum

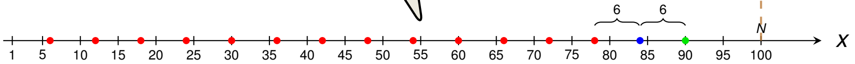
- Suppose $N = 100$ and $n = 15$
- Our samples are:

$$9, 82, 39, 35, 20, 51, 54, 62, 81, 29, \color{blue}{84}, 59, 3, 34, 55$$

How much should we add to the maximum?



Rearrange the other 14 points equi-spaced between 0 and 84.



$\max(X_1, \ldots, X_n) + \frac{\max(X_1,\ldots,X_n)}{n-1}$

This suggests $84 + 6 = 90$ as the estimator!

Deriving the Estimator Based on Maximum

Example 2

Construct an unbiased estimator using $\max(X_1, \ldots, X_n)$

Answer

Deriving the Estimator Based on Maximum

Example 2

Construct an unbiased estimator using $\max(X_1, \ldots, X_n)$

Answer

- Calculate expectation of the maximum (for details see Dekking et al.)

$$\mathbf{E}\left[\, \max(X_1, \ldots, X_n) \,\right] =$$

Deriving the Estimator Based on Maximum

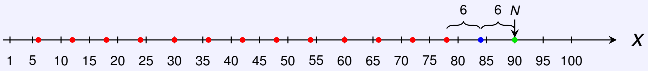
Example 2

Construct an unbiased estimator using $\max(X_1, \ldots, X_n)$

Answer

- Calculate expectation of the maximum (for details see Dekking et al.)

$$\mathbf{E}\left[\max(X_1, \ldots, X_n)\right] =$$



Deriving the Estimator Based on Maximum

Example 2

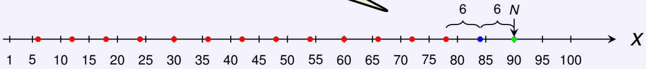
Construct an unbiased estimator using $\max(X_1, \ldots, X_n)$

Answer

- Calculate expectation of the maximum (for details see Dekking et al.)

$$\mathbf{E}\left[\,\max(X_1, \ldots, X_n)\,\right] =$$

Equi-spaced configuration would suggest $\max(X_1, \ldots, X_n) \approx \frac{n-1}{n} \cdot N$



Deriving the Estimator Based on Maximum

Example 2

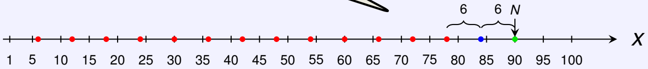
Construct an unbiased estimator using $\max(X_1, \ldots, X_n)$

Answer

- Calculate expectation of the maximum (for details see Dekking et al.)

$$\mathbf{E}\left[\, \max(X_1, \ldots, X_n) \,\right] = \ \ldots = \frac{n}{n+1} \cdot N + \frac{n}{n+1} = \frac{n}{n+1} \cdot (N+1).$$

Equi-spaced configuration would suggest $\max(X_1, \ldots, X_n) \approx \frac{n-1}{n} \cdot N$



Deriving the Estimator Based on Maximum

Example 2

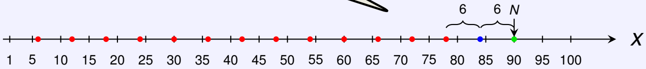
Construct an unbiased estimator using $\max(X_1, \ldots, X_n)$

Answer

- Calculate expectation of the maximum (for details see Dekking et al.)

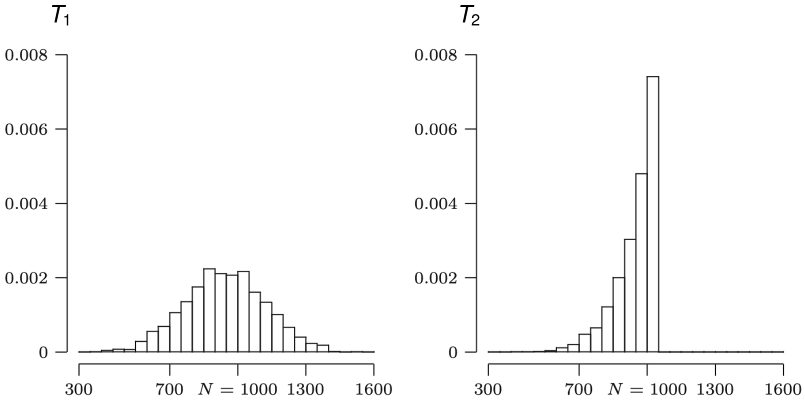
$$\mathbf{E}\left[\,\max(X_1, \ldots, X_n)\,\right] = \ldots = \frac{n}{n+1} \cdot N + \frac{n}{n+1} = \frac{n}{n+1} \cdot (N+1).$$

Equi-spaced configuration would suggest $\max(X_1, \ldots, X_n) \approx \frac{n-1}{n} \cdot N$



- Hence we obtain an unbiased estimator by

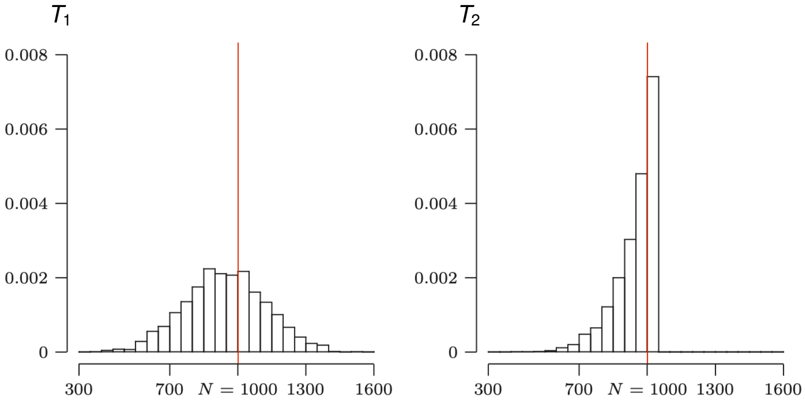
$$T_2 := \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1.$$



Source: Modern Introduction to Statistics

Figure: Histogram of 2000 values for T_1 and T_2, when $N = 1000$ and $n = 10$.

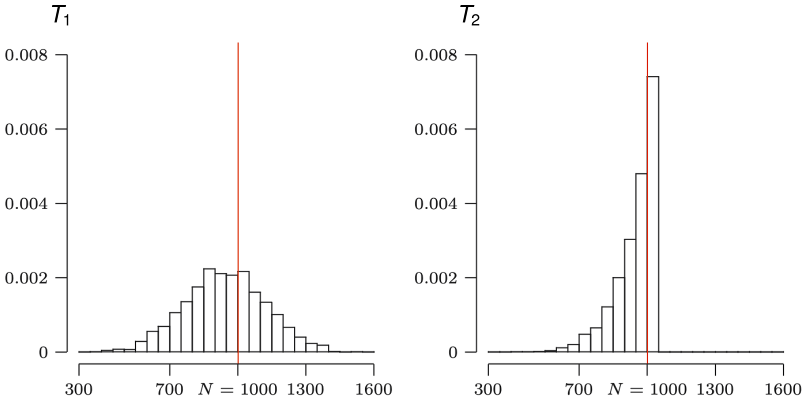
Empirical Analysis of the two Estimators



Source: Modern Introduction to Statistics

Figure: Histogram of 2000 values for T_1 and T_2, when $N = 1000$ and $n = 10$.

Empirical Analysis of the two Estimators



Source: Modern Introduction to Statistics

Figure: Histogram of 2000 values for T_1 and T_2, when $N = 1000$ and $n = 10$.

Can we find a quantity that captures the superiority of T_2 over T_1?

Outline

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}\left[\,T\,\right] = \mathbf{E}\left[\,(T - \theta)^2\,\right].$$

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}\left[\,T\,\right] = \mathbf{E}\left[\,(T-\theta)^2\,\right].$$

- According to this, estimator T_1 better than T_2 if $\mathbf{MSE}\left[\,T_1\,\right] < \mathbf{MSE}\left[\,T_2\,\right]$.

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}[\,T\,] = \mathbf{E}\left[\,(T-\theta)^2\,\right].$$

- According to this, estimator T_1 better than T_2 if $\mathbf{MSE}[\,T_1\,] < \mathbf{MSE}[\,T_2\,]$.

Bias-Variance Decomposition

The mean squared error can be decomposed into:

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}[\,T\,] = \mathbf{E}\left[\,(T-\theta)^2\,\right].$$

- According to this, estimator T_1 better than T_2 if $\mathbf{MSE}[\,T_1\,] < \mathbf{MSE}[\,T_2\,]$.

Bias-Variance Decomposition

The mean squared error can be decomposed into:

$$\mathbf{MSE}[\,T\,] = (\mathbf{E}[\,T\,]-\theta)^2 + \mathbf{V}[\,T\,]$$

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}[\,T\,] = \mathbf{E}\left[\,(T-\theta)^2\,\right].$$

- According to this, estimator T_1 better than T_2 if $\mathbf{MSE}[\,T_1\,] < \mathbf{MSE}[\,T_2\,]$.

Bias-Variance Decomposition

The mean squared error can be decomposed into:

$$\mathbf{MSE}[\,T\,] = \underbrace{(\mathbf{E}[\,T\,]-\theta)^2}_{=\,\mathrm{Bias}^2} + \ \mathbf{V}[\,T\,]$$

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}[\,T\,] = \mathbf{E}\left[(T-\theta)^2\right].$$

- According to this, estimator T_1 better than T_2 if $\mathbf{MSE}[\,T_1\,] < \mathbf{MSE}[\,T_2\,]$.

Bias-Variance Decomposition

The mean squared error can be decomposed into:

$$\mathbf{MSE}[\,T\,] = \underbrace{(\mathbf{E}[\,T\,]-\theta)^2}_{=\,\text{Bias}^2} + \underbrace{\mathbf{V}[\,T\,]}_{=\,\text{Variance}}$$

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}[\,T\,] = \mathbf{E}\left[\,(T-\theta)^2\,\right].$$

- According to this, estimator T_1 better than T_2 if $\mathbf{MSE}[\,T_1\,] < \mathbf{MSE}[\,T_2\,]$.

Bias-Variance Decomposition

The mean squared error can be decomposed into:

$$\mathbf{MSE}[\,T\,] = \underbrace{(\mathbf{E}[\,T\,]-\theta)^2}_{=\,\text{Bias}^2} + \underbrace{\mathbf{V}[\,T\,]}_{=\,\text{Variance}}$$

- If T_1 and T_2 are both unbiased, T_1 is better than T_2 iff $\mathbf{V}[\,T_1\,] < \mathbf{V}[\,T_2\,]$.

Mean Squared Error

Mean Squared Error Definition

Let T be an estimator for a parameter θ. The mean squared error of T is

$$\mathbf{MSE}[\,T\,] = \mathbf{E}\left[\,(T-\theta)^2\,\right].$$

- According to this, estimator T_1 better than T_2 if $\mathbf{MSE}[\,T_1\,] < \mathbf{MSE}[\,T_2\,]$.

Bias-Variance Decomposition

The mean squared error can be decomposed into:

$$\mathbf{MSE}[\,T\,] = \underbrace{(\mathbf{E}[\,T\,]-\theta)^2}_{=\,\text{Bias}^2} + \underbrace{\mathbf{V}[\,T\,]}_{=\,\text{Variance}}$$

- If T_1 and T_2 are both unbiased, T_1 is better than T_2 iff $\mathbf{V}[\,T_1\,] < \mathbf{V}[\,T_2\,]$.

$\rightsquigarrow$ Minimum-Variance Unbiased Estimator (MVUE) (the unbiased estimator with the smallest variance).

Bias-Variance Decomposition: Derivation

Example 3

We need to prove: $\mathbf{MSE}[T] = (\mathbf{E}[T] - \theta)^2 + \mathbf{V}[T]$.

Answer

Bias-Variance Decomposition: Derivation

Example 3

We need to prove: $\mathbf{MSE}[\,T\,] = (\mathbf{E}[\,T\,] - \theta)^2 + \mathbf{V}[\,T\,]$.

Answer

$$\mathbf{MSE}[\,T\,] = \mathbf{E}\left[\,(T - \theta)^2\,\right]$$

Bias-Variance Decomposition: Derivation

Example 3

We need to prove: $\mathbf{MSE}[\,T\,] = (\mathbf{E}[\,T\,] - \theta)^2 + \mathbf{V}[\,T\,]$.

Answer

$$\begin{aligned} \mathbf{MSE}[\,T\,] &= \mathbf{E}\left[\,(T-\theta)^2\,\right] \\ &= \mathbf{E}\left[\,T^2 - 2T\theta + \theta^2\,\right] \end{aligned}$$

Bias-Variance Decomposition: Derivation

Example 3

We need to prove: $\mathbf{MSE}[\,T\,] = (\mathbf{E}[\,T\,] - \theta)^2 + \mathbf{V}[\,T\,]$.

Answer

$$
\begin{aligned}
\mathbf{MSE}[\,T\,] &= \mathbf{E}\left[\,(T-\theta)^2\,\right] \\
&= \mathbf{E}\left[\,T^2 - 2T\theta + \theta^2\,\right] \\
&= \mathbf{E}[\,T\,]^2 - 2\cdot\mathbf{E}[\,T\,]\cdot\theta + \theta^2 + \mathbf{E}\left[\,T^2\,\right] - \mathbf{E}[\,T\,]^2
\end{aligned}
$$

Bias-Variance Decomposition: Derivation

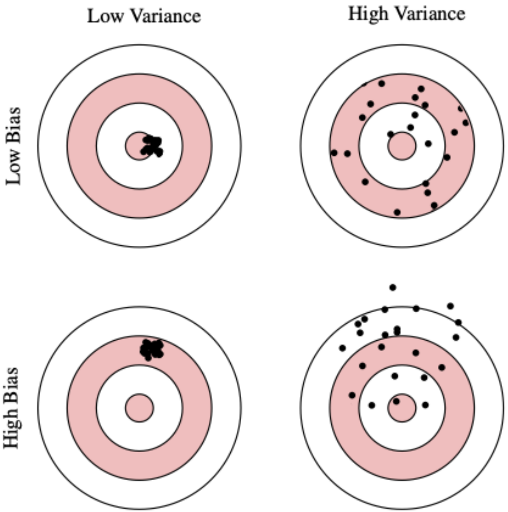
Example 3

We need to prove: $\mathbf{MSE}[\,T\,] = (\mathbf{E}[\,T\,] - \theta)^2 + \mathbf{V}[\,T\,]$.

Answer

$$
\begin{aligned}
\mathbf{MSE}[\,T\,] &= \mathbf{E}\left[\,(T-\theta)^2\,\right] \\
&= \mathbf{E}\left[\,T^2 - 2T\theta + \theta^2\,\right] \\
&= \mathbf{E}[\,T\,]^2 - 2\cdot\mathbf{E}[\,T\,]\cdot\theta + \theta^2 + \mathbf{E}\left[\,T^2\,\right] - \mathbf{E}[\,T\,]^2 \\
&= (\mathbf{E}[\,T\,] - \theta)^2 + \mathbf{V}[\,T\,].
\end{aligned}
$$

Bias-Variance Decomposition: Illustration



Source: Edwin Leuven (Point Estimation)

Example 4

It holds that $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2 \cdot \overline{X}_n - 1$.

Answer

Example 4

It holds that $\mathbf{MSE}[\, T_1 \,] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2 \cdot \overline{X}_n - 1$.

Answer

- Since T_1 is unbiased, $\mathbf{MSE}[\, T_1 \,] = (\mathbf{E}[\, T_1 \,] - \theta)^2 + \mathbf{V}[\, T_1 \,] = \mathbf{V}[\, T_1 \,]$, and

Example 4

It holds that $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2 \cdot \overline{X}_n - 1$.

Answer

- Since T_1 is unbiased, $\mathbf{MSE}[\,T_1\,] = (\mathbf{E}[\,T_1\,] - \theta)^2 + \mathbf{V}[\,T_1\,] = \mathbf{V}[\,T_1\,]$, and

$$\mathbf{V}[\,T_1\,] = \mathbf{V}\left[2 \cdot \overline{X}_n - 1\right] = 4 \cdot \mathbf{V}\left[\overline{X}_n\right] = \frac{4}{n^2} \cdot \mathbf{V}[\,X_1 + \cdots + X_n\,]$$

Example 4

It holds that $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2\cdot\overline{X}_n - 1$.

Answer

- Since T_1 is unbiased, $\mathbf{MSE}[\,T_1\,] = (\mathbf{E}[\,T_1\,] - \theta)^2 + \mathbf{V}[\,T_1\,] = \mathbf{V}[\,T_1\,]$, and

$$\mathbf{V}[\,T_1\,] = \mathbf{V}\left[2\cdot\overline{X}_n - 1\right] = 4\cdot\mathbf{V}\left[\overline{X}_n\right] = \frac{4}{n^2}\cdot\mathbf{V}[\,X_1 + \cdots + X_n\,]$$

- Note: The X_i's are not independent!

Example 4

It holds that $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2\cdot\overline{X}_n - 1$.

Answer

- Since T_1 is unbiased, $\mathbf{MSE}[\,T_1\,] = (\mathbf{E}[\,T_1\,] - \theta)^2 + \mathbf{V}[\,T_1\,] = \mathbf{V}[\,T_1\,]$, and

$$\mathbf{V}[\,T_1\,] = \mathbf{V}\left[2\cdot\overline{X}_n - 1\right] = 4\cdot\mathbf{V}\left[\overline{X}_n\right] = \frac{4}{n^2}\cdot\mathbf{V}[\,X_1 + \cdots + X_n\,]$$

- Note: The X_i's are not independent!
- Use generalisation of $\mathbf{V}[\,X_1 + X_2\,] = \mathbf{V}[\,X_1\,] + \mathbf{V}[\,X_2\,] + 2\cdot\mathbf{Cov}[\,X_1, X_2\,]$ (Exercise Sheet) to n r.v.'s, and then that the X_i's are identically distributed, and also the (X_i, X_j), $i \neq j$:

Example 4

It holds that $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2\cdot\overline{X}_n - 1$.

Answer

- Since T_1 is unbiased, $\mathbf{MSE}[\,T_1\,] = (\mathbf{E}[\,T_1\,]-\theta)^2 + \mathbf{V}[\,T_1\,] = \mathbf{V}[\,T_1\,]$, and

$$\mathbf{V}[\,T_1\,] = \mathbf{V}\left[2\cdot\overline{X}_n - 1\right] = 4\cdot\mathbf{V}\left[\overline{X}_n\right] = \frac{4}{n^2}\cdot\mathbf{V}[\,X_1+\cdots+X_n\,]$$

- Note: The X_i's are not independent!
- Use generalisation of $\mathbf{V}[\,X_1+X_2\,] = \mathbf{V}[\,X_1\,] + \mathbf{V}[\,X_2\,] + 2\cdot\mathbf{Cov}[\,X_1,X_2\,]$ (Exercise Sheet) to n r.v.'s, and then that the X_i's are identically distributed, and also the (X_i,X_j), $i\neq j$:

$$\begin{aligned}\mathbf{V}[\,X_1+\cdots+X_n\,] &= \sum_{i=1}^{n}\mathbf{V}[\,X_i\,] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbf{Cov}\left[X_i,X_j\right]\\ &= n\cdot\mathbf{V}[\,X_1\,] + 2\binom{n}{2}\cdot\mathbf{Cov}[\,X_1,X_2\,].\end{aligned}$$

Example 4

It holds that $\mathbf{MSE}[T_1] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2 \cdot \overline{X}_n - 1$.

Answer

- Since T_1 is unbiased, $\mathbf{MSE}[T_1] = (\mathbf{E}[T_1] - \theta)^2 + \mathbf{V}[T_1] = \mathbf{V}[T_1]$, and

$$\mathbf{V}[T_1] = \mathbf{V}\left[2 \cdot \overline{X}_n - 1\right] = 4 \cdot \mathbf{V}\left[\overline{X}_n\right] = \frac{4}{n^2} \cdot \mathbf{V}[X_1 + \cdots + X_n]$$

- Note: The X_i's are not independent!
- Use generalisation of $\mathbf{V}[X_1 + X_2] = \mathbf{V}[X_1] + \mathbf{V}[X_2] + 2 \cdot \mathbf{Cov}[X_1, X_2]$ (Exercise Sheet) to n r.v.'s, and then that the X_i's are identically distributed, and also the (X_i, X_j), $i \neq j$:

$$\begin{aligned}\mathbf{V}[X_1 + \cdots + X_n] &= \sum_{i=1}^{n} \mathbf{V}[X_i] + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{Cov}\left[X_i, X_j\right] \\ &= n \cdot \mathbf{V}[X_1] + 2\binom{n}{2} \cdot \mathbf{Cov}[X_1, X_2].\end{aligned}$$

- $\mathbf{V}[X_1] = \frac{(N+1)(N-1)}{12}$, and with "more effort" (see Dekking et al.)

$$\mathbf{Cov}[X_1, X_2] = -\frac{1}{12}(N+1).$$

Example 4

It holds that $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, where $T_1 = 2\cdot\overline{X}_n - 1$.

Answer

- Since T_1 is unbiased, $\mathbf{MSE}[\,T_1\,] = (\mathbf{E}[\,T_1\,] - \theta)^2 + \mathbf{V}[\,T_1\,] = \mathbf{V}[\,T_1\,]$, and

$$\mathbf{V}[\,T_1\,] = \mathbf{V}\left[2\cdot\overline{X}_n - 1\right] = 4\cdot\mathbf{V}\left[\overline{X}_n\right] = \frac{4}{n^2}\cdot\mathbf{V}[\,X_1+\cdots+X_n\,]$$

- Note: The X_i's are not independent!
- Use generalisation of $\mathbf{V}[\,X_1+X_2\,] = \mathbf{V}[\,X_1\,] + \mathbf{V}[\,X_2\,] + 2\cdot\mathbf{Cov}[\,X_1,X_2\,]$ (Exercise Sheet) to n r.v.'s, and then that the X_i's are identically distributed, and also the (X_i,X_j), $i\neq j$:

$$\begin{aligned}\mathbf{V}[\,X_1+\cdots+X_n\,] &= \sum_{i=1}^{n}\mathbf{V}[\,X_i\,] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbf{Cov}\left[\,X_i,X_j\,\right]\\ &= n\cdot\mathbf{V}[\,X_1\,] + 2\binom{n}{2}\cdot\mathbf{Cov}[\,X_1,X_2\,].\end{aligned}$$

- $\mathbf{V}[\,X_1\,] = \frac{(N+1)(N-1)}{12}$, and with "more effort" (see Dekking et al.)

$$\mathbf{Cov}[\,X_1,X_2\,] = -\frac{1}{12}(N+1).$$

- Rearranging and simplifying gives

$$\mathbf{V}[\,T_1\,] = \frac{(N+1)(N-n)}{3n}.$$

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that **MSE** $[\, T_2 \,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

Answer

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1, \ldots, X_n)\,]$

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}[\, T_2 \,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\, T_2 \,]$ which reduces to $\mathbf{V}[\, \max(X_1, \ldots, X_n) \,]$
- One can prove:

$$\mathbf{V}[\, \max(X_1, \ldots, X_n) \,] =$$

Analysis of the MSE for T_2 (Sketch)

Example 5

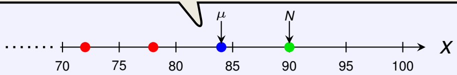
It holds that **MSE** $[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1, \ldots, X_n)\,]$
- One can prove:

$$\mathbf{V}[\,\max(X_1, \ldots, X_n)\,] =$$

Equi-spaced (idealised) configuration



Analysis of the MSE for T_2 (Sketch)

Example 5

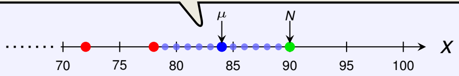
It holds that $\mathbf{MSE}[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1, \ldots, X_n)\,]$
- One can prove:

$$\mathbf{V}[\,\max(X_1, \ldots, X_n)\,] =$$

Equi-spaced (idealised) configuration



Analysis of the MSE for T_2 (Sketch)

Example 5

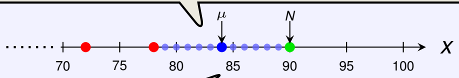
It holds that $\mathbf{MSE}[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1, \ldots, X_n)\,]$
- One can prove:

$$\mathbf{V}[\,\max(X_1, \ldots, X_n)\,] =$$

Equi-spaced (idealised) configuration



Maximum could have equally likely taken any value between 79 and 90

Analysis of the MSE for T_2 (Sketch)

Example 5

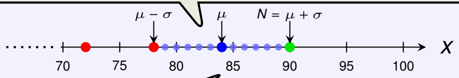
It holds that $\mathbf{MSE}[\, T_2 \,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\, T_2 \,]$ which reduces to $\mathbf{V}[\, \max(X_1, \ldots, X_n) \,]$
- One can prove:

$$\mathbf{V}[\, \max(X_1, \ldots, X_n) \,] =$$

Equi-spaced (idealised) configuration suggests a standard deviation of $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}\left[\, T_2 \,\right] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

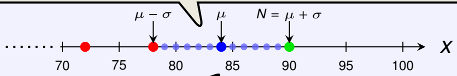
Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}\left[\, T_2 \,\right]$ which reduces to $\mathbf{V}\left[\, \max(X_1, \ldots, X_n) \,\right]$
- One can prove:

For details see Dekking et al.

$$\mathbf{V}\left[\, \max(X_1, \ldots, X_n) \,\right] = \cdots = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2}$$

Equi-spaced (idealised) configuration suggests a standard deviation of $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}[\, T_2 \,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

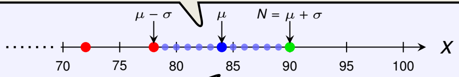
Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\, T_2 \,]$ which reduces to $\mathbf{V}[\, \max(X_1, \ldots, X_n) \,]$
- One can prove:

For details see Dekking et al.

$$\mathbf{V}[\, \max(X_1, \ldots, X_n) \,] = \cdots = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2} = \Theta\left(\frac{N^2}{n^2}\right)$$

Equi-spaced (idealised) configuration suggests a standard deviation of $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

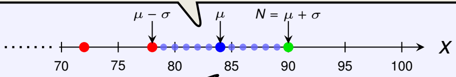
Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1, \ldots, X_n)\,]$
- One can prove:

For details see Dekking et al.

$$\mathbf{V}[\,\max(X_1, \ldots, X_n)\,] = \cdots = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2} = \Theta\left(\frac{N^2}{n^2}\right)$$

Equi-spaced (idealised) configuration suggests a standard deviation of $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

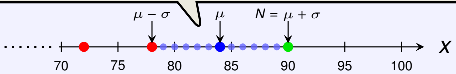
Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1, \ldots, X_n)\,]$
- One can prove:

For details see Dekking et al.

$$\mathbf{V}[\,\max(X_1, \ldots, X_n)\,] = \cdots = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2} = \Theta\left(\frac{N^2}{n^2}\right)$$

Equi-spaced (idealised) configuration suggests a standard deviation of $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

- $\mathbf{MSE}[\,T_2\,]$ is much lower than $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, i.e., $\frac{\mathbf{MSE}[\,T_1\,]}{\mathbf{MSE}[\,T_2\,]} = \frac{n+2}{3}$

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1,\ldots,X_n) - 1$.

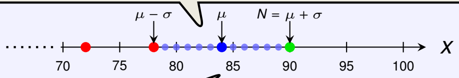
Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1,\ldots,X_n)\,]$
- One can prove:

For details see Dekking et al.

$$\mathbf{V}[\,\max(X_1,\ldots,X_n)\,] = \cdots = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2} = \Theta\left(\frac{N^2}{n^2}\right)$$

Equi-spaced (idealised) configuration suggests a standard deviation of $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

- $\mathbf{MSE}[\,T_2\,]$ is much lower than $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, i.e., $\frac{\mathbf{MSE}[\,T_1\,]}{\mathbf{MSE}[\,T_2\,]} = \frac{n+2}{3}$
- $\Rightarrow$ confirms simulations suggesting that T_2 is better than T_1!

Analysis of the MSE for T_2 (Sketch)

Example 5

It holds that $\mathbf{MSE}[\,T_2\,] = \Theta\left(\frac{N^2}{n^2}\right)$, where $T_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n) - 1$.

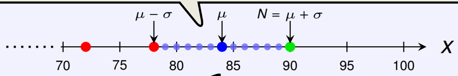
Answer

- T_2 is unbiased $\Rightarrow$ need $\mathbf{V}[\,T_2\,]$ which reduces to $\mathbf{V}[\,\max(X_1, \ldots, X_n)\,]$
- One can prove:

For details see Dekking et al.

$$\mathbf{V}[\,\max(X_1, \ldots, X_n)\,] = \cdots = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2} = \Theta\left(\frac{N^2}{n^2}\right)$$

Equi-spaced (idealised) configuration suggests a standard deviation of $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

- $\mathbf{MSE}[\,T_2\,]$ is much lower than $\mathbf{MSE}[\,T_1\,] = \Theta\left(\frac{N^2}{n}\right)$, i.e., $\frac{\mathbf{MSE}[\,T_1\,]}{\mathbf{MSE}[\,T_2\,]} = \frac{n+2}{3}$
- $\Rightarrow$ confirms simulations suggesting that T_2 is better than T_1!
- can be shown T_2 is the best unbiased estimator, i.e., it minimises MSE.

Outline

Recap

Estimating Population Sizes

Mean Squared Error

Estimating Population Sizes through Collisions

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \ldots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \ldots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

New Model

- Population/ID space of size $|S| = N$
- We take uniform samples from S with replacement
- **Goal:** Find estimator for N

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \dots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

New Model

- Population/ID space of size $|S| = N$
- We take uniform samples from S with replacement
- **Goal:** Find estimator for N

- Suppose $n = 6$, $N = 11$, $S = \{3, 4, 7, 8, 10, 15.83356, 20, 21, 56, 81, 10000\}$

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \ldots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

New Model

- Population/ID space of size $|S| = N$
- We take uniform samples from S with replacement
- **Goal:** Find estimator for N

- Suppose $n = 6$, $N = 11$, $S = \{3, 4, 7, 8, 10, 15.83356, 20, 21, 56, 81, 10000\}$
- Let the sample be

$$10, 81, 20, 3, 81, 10000$$

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \ldots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

New Model

- Population/ID space of size $|S| = N$
- We take uniform samples from S with replacement
- **Goal:** Find estimator for N

- Suppose $n = 6$, $N = 11$, $S = \{3, 4, 7, 8, 10, 15.83356, 20, 21, 56, 81, 10000\}$
- Let the sample be

$$10, 81, 20, 3, 81, 10000$$

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \ldots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

New Model

- Population/ID space of size $|S| = N$
- We take uniform samples from S with replacement
- **Goal:** Find estimator for N

- Suppose $n = 6$, $N = 11$, $S = \{3, 4, 7, 8, 10, 15.83356, 20, 21, 56, 81, 10000\}$
- Let the sample be

$$10, 81, 20, 3, 81, 10000$$

As we do not know S, our only clue are elements that **were sampled twice.**

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \ldots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

New Model

- Population/ID space of size $|S| = N$
- We take uniform samples from S with replacement
- **Goal:** Find estimator for N

- Suppose $n = 6$, $N = 11$, $S = \{3, 4, 7, 8, 10, 15.83356, 20, 21, 56, 81, 10000\}$
- Let the sample be

$$10, 81, 20, 3, 81, 10000$$

Let us call this a **collision**

As we do not know S, our only clue are elements that **were sampled twice.**

A New Estimation Problem

Previous Model

- Population/ID space $S = \{1, 2, \ldots, N\}$
- We take uniform samples from S without replacement
- **Goal:** Find estimator for N

This also applies to situations where elements are not labelled before we see them first time (e.g., Mark & Recapture Method)

New Model

- Population/ID space of size $|S| = N$
- We take uniform samples from S with replacement
- **Goal:** Find estimator for N

- Suppose $n = 6$, $N = 11$, $S = \{3, 4, 7, 8, 10, 15.83356, 20, 21, 56, 81, 10000\}$
- Let the sample be

$$10, 81, 20, 3, 81, 10000$$

Let us call this a **collision**

As we do not know S, our only clue are elements that **were sampled twice.**

Birthday Problem

Birthday Problem: Given a set of i people

Birthday Problem

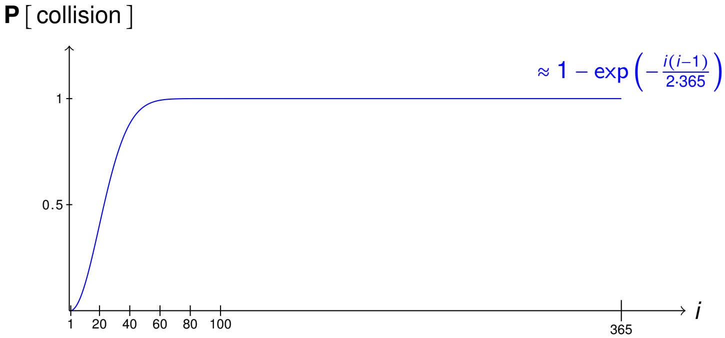
Birthday Problem: Given a set of i people

- What is the probability of having two with the same birthday (i.e., having at least one collision)?

Birthday Problem

Birthday Problem: Given a set of i people

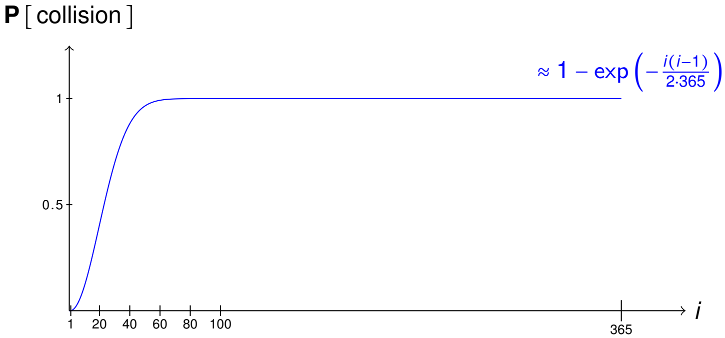
- What is the probability of having two with the same birthday (i.e., having at least one collision)?



Birthday Problem

Birthday Problem: Given a set of i people

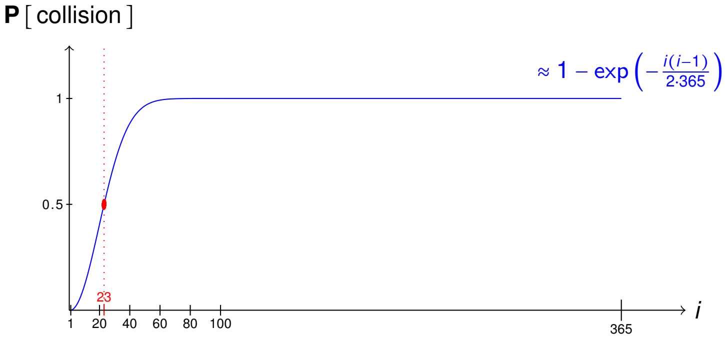
- What is the probability of having two with the same birthday (i.e., having at least one collision)?
- What is the expected number of people one needs to ask until the first collision occurs?



Birthday Problem

Birthday Problem: Given a set of i people

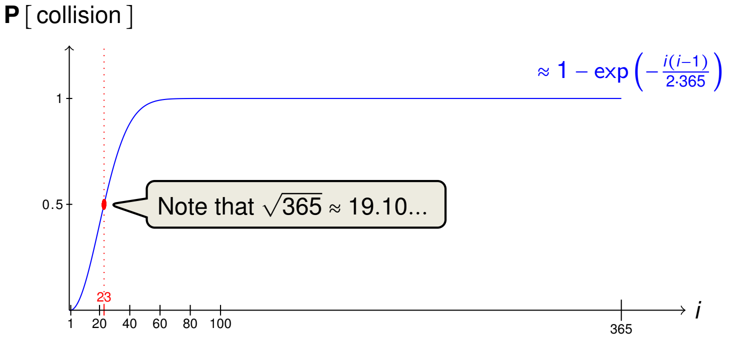
- What is the probability of having two with the same birthday (i.e., having at least one collision)?
- What is the expected number of people one needs to ask until the first collision occurs?



Birthday Problem

Birthday Problem: Given a set of i people

- What is the probability of having two with the same birthday (i.e., having at least one collision)?
- What is the expected number of people one needs to ask until the first collision occurs?



Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

```
FIND-FIRST-COLLISION(S)
C = ∅
For i = 1, 2, ...
    Take next i.i.d. sample X_i from S
    If X_i ∉ C then C ← C ∪ {X_i}
    else return T(i)
End For
```

Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

```
FIND-FIRST-COLLISION(S)
C = ∅
For i = 1, 2, ...
    Take next i.i.d. sample X_i from S
    If X_i ∉ C then C ← C ∪ {X_i}
    else return T(i)
End For
```

$T(i)$ will be the value of the estimator if algo returns after i rounds. (We want T unbiased)

Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

FIND-FIRST-COLLISION(S)

```
C = ∅
For i = 1, 2, ...
    Take next i.i.d. sample X_i from S
    If X_i ∉ C then C ← C ∪ {X_i}
    else return T(i)
End For
```

$T(i)$ will be the value of the estimator if algo returns after i rounds. (We want T unbiased)

- Running Time: The expected time until the algorithm stops is:

Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

FIND-FIRST-COLLISION(S)

$C = \emptyset$
For $i = 1, 2, \ldots$
  Take next i.i.d. sample X_i from S
  If $X_i \notin C$ **then** $C \leftarrow C \cup \{X_i\}$
  else return $T(i)$
End For

$T(i)$ will be the value of the estimator if algo returns after i rounds. (We want T unbiased)

- Running Time: The expected time until the algorithm stops is:
  = the expected number of samples until a collision...

Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

FIND-FIRST-COLLISION(S)

```
C = ∅
For i = 1, 2, ...
    Take next i.i.d. sample X_i from S
    If X_i ∉ C then C ← C ∪ {X_i}
    else return T(i)
End For
```

$T(i)$ will be the value of the estimator if algo returns after i rounds. (We want T unbiased)

- Running Time: The expected time until the algorithm stops is:
  = the expected number of samples until a collision...

Same as the birthday problem, but now with $|S| = N$ days... ☺

Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

FIND-FIRST-COLLISION(S)

$C = \emptyset$
For $i = 1, 2, \ldots$
    Take next i.i.d. sample X_i from S
    If $X_i \notin C$ **then** $C \leftarrow C \cup \{X_i\}$
    else return $T(i)$
End For

$T(i)$ will be the value of the estimator if algo returns after i rounds. (We want T unbiased)

- Running Time: The expected time until the algorithm stops is:
  $=$ the expected number of samples until a collision...

Same as the birthday problem, but now with $|S| = N$ days... ☺

Expected Running Time (Knuth, Ramanujan)

$$\sqrt{\frac{\pi N}{2}} - \frac{1}{3} + O\left(\frac{1}{\sqrt{N}}\right).$$

Estimation via Collision: The Algorithm

Recall: As we do not know S, our only information are **collisions**.

FIND-FIRST-COLLISION(S)

$C = \varnothing$

For $i = 1, 2, \ldots$

    Take next i.i.d. sample X_i from S

    If $X_i \notin C$ **then** $C \leftarrow C \cup \{X_i\}$

    else return $T(i)$

End For

$T(i)$ will be the value of the estimator if algo returns after i rounds. (We want T unbiased)

- Running Time: The expected time until the algorithm stops is:
  = the expected number of samples until a collision...

Same as the birthday problem, but now with $|S| = N$ days... ☺

Expected Running Time (Knuth, Ramanujan)

$$\sqrt{\frac{\pi N}{2}} - \frac{1}{3} + O\left(\frac{1}{\sqrt{N}}\right).$$

Exercise: Prove a bound of $\leq 2 \cdot \sqrt{N}$

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$.

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$. We want

$$1 = \mathbf{E}[\,T\,] = T(2)$$

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$. We want

$$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$. We want

$$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$

- Case $|S| = 2$: Algo stops after 2 or 3 rounds (w.p. $1/2$ each).

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$. We want
$$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$
- Case $|S| = 2$: Algo stops after 2 or 3 rounds (w.p. 1/2 each). We want
$$2 = \mathbf{E}[\,T\,] = \frac{1}{2} \cdot T(2) + \frac{1}{2} \cdot T(3)$$

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$. We want
$$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$
- Case $|S| = 2$: Algo stops after 2 or 3 rounds (w.p. 1/2 each). We want
$$2 = \mathbf{E}[\,T\,] = \frac{1}{2} \cdot T(2) + \frac{1}{2} \cdot T(3) \qquad \Rightarrow \qquad T(3) = 3.$$

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$. We want

$$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$

- Case $|S| = 2$: Algo stops after 2 or 3 rounds (w.p. 1/2 each). We want

$$2 = \mathbf{E}[\,T\,] = \frac{1}{2} \cdot T(2) + \frac{1}{2} \cdot T(3) \qquad \Rightarrow \qquad T(3) = 3.$$

- Case $|S| = 3$: gives $3 = \mathbf{E}[\,T\,] = \frac{1}{3} \cdot T(2) + \frac{4}{9} \cdot T(3) + \frac{2}{9} \cdot T(4)$

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$.
  We want
  $$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$
- Case $|S| = 2$: Algo stops after 2 or 3 rounds (w.p. 1/2 each).
  We want
  $$2 = \mathbf{E}[\,T\,] = \frac{1}{2} \cdot T(2) + \frac{1}{2} \cdot T(3) \qquad \Rightarrow \qquad T(3) = 3.$$
- Case $|S| = 3$: gives $3 = \mathbf{E}[\,T\,] = \frac{1}{3} \cdot T(2) + \frac{4}{9} \cdot T(3) + \frac{2}{9} \cdot T(4)$
  $\Rightarrow T(4) = 6$, similarly, $T(5) = 10$ etc.

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$.
  We want
$$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$
- Case $|S| = 2$: Algo stops after 2 or 3 rounds (w.p. 1/2 each).
  We want
$$2 = \mathbf{E}[\,T\,] = \frac{1}{2} \cdot T(2) + \frac{1}{2} \cdot T(3) \qquad \Rightarrow \qquad T(3) = 3.$$
- Case $|S| = 3$: gives $3 = \mathbf{E}[\,T\,] = \frac{1}{3} \cdot T(2) + \frac{4}{9} \cdot T(3) + \frac{2}{9} \cdot T(4)$
  $\Rightarrow T(4) = 6$, similarly, $T(5) = 10$ etc.
- can continue to define $T(i)$ inductively in this way (note T is unique)

Estimation via Collision: Getting the Estimator Unbiased

Example 6

It is possible to define $T(i)$, $i \in \mathbb{N}$, such that $\mathbf{E}[\,T\,] = |S|$ for any set S.

Answer

- We outline a construction by induction.
- Case $|S| = 1$: Algo always stops after $i = 2$ rounds and returns $T(2)$. We want

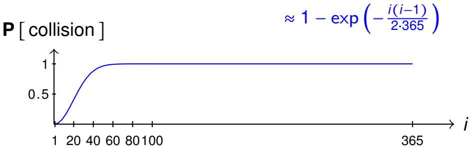
$$1 = \mathbf{E}[\,T\,] = T(2) \qquad \Rightarrow \qquad T(2) = 1.$$

- Case $|S| = 2$: Algo stops after 2 or 3 rounds (w.p. 1/2 each). We want

$$2 = \mathbf{E}[\,T\,] = \frac{1}{2} \cdot T(2) + \frac{1}{2} \cdot T(3) \qquad \Rightarrow \qquad T(3) = 3.$$

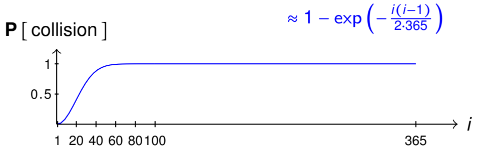
- Case $|S| = 3$: gives $3 = \mathbf{E}[\,T\,] = \frac{1}{3} \cdot T(2) + \frac{4}{9} \cdot T(3) + \frac{2}{9} \cdot T(4)$
$\Rightarrow T(4) = 6$, similarly, $T(5) = 10$ etc.
- can continue to define $T(i)$ inductively in this way (note T is unique)
(proof that $T(i) = \binom{i}{2}$ is harder)

Mark & Recapture Method (non-examinable)



Source: Wikipedia

Mark & Recapture Method (non-examinable)

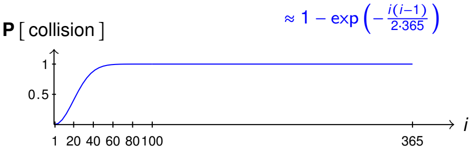


Source: Wikipedia

Mark & Recapture Method:

- **First phase:** A portion of the population is captured, marked and released
- **Second phase:** Another portion is captured and the number of marked individuals is counted

Mark & Recapture Method (non-examinable)



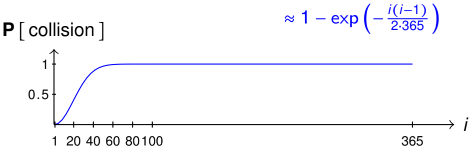
Source: Wikipedia

Mark & Recapture Method:

- **First phase:** A portion of the population is captured, marked and released
- **Second phase:** Another portion is captured and the number of marked individuals is counted

A similar method making use of collisions again!

Mark & Recapture Method (non-examinable)



Source: Wikipedia

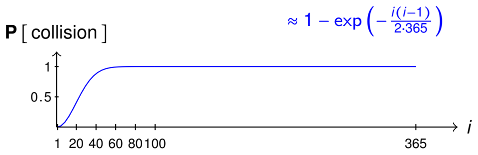
Mark & Recapture Method:

- **First phase:** A portion of the population is captured, marked and released
- **Second phase:** Another portion is captured and the number of marked individuals is counted

A similar method making use of collisions again!

- Let n be the number of marked animals, and N be the (unknown) size of population
- Let k be the number of caught marked animals (in the second visit), and K be the number of caught animals (in the second visit)

Mark & Recapture Method (non-examinable)



Source: Wikipedia

Mark & Recapture Method:

- **First phase:** A portion of the population is captured, marked and released
- **Second phase:** Another portion is captured and the number of marked individuals is counted

A similar method making use of collisions again!

- Let n be the number of marked animals, and N be the (unknown) size of population
- Let k be the number of caught marked animals (in the second visit), and K be the number of caught animals (in the second visit)

$$\frac{k}{K} \approx \frac{n}{N} \qquad \Rightarrow \qquad N \approx n \cdot \frac{K}{k}.$$