# The Lazy Select Algorithm

K. Subramani[1]

[1] Lane Department of Computer Science and Electrical Engineering
West Virginia University

21 February, 2012

# Outline

# Outline

1. The median statistic

2. Deterministic Approaches

# Outline

# Median

# Median

**Definition**

The median of an array **A** of $n$ distinct elements, where $n$ is an odd number is $\mathbf{A}[\frac{n+1}{2}]$, after **A** has been sorted.

# Median

**Definition**

The median of an array **A** of $n$ distinct elements, where $n$ is an odd number is $\mathbf{A}[\frac{n+1}{2}]$, after **A** has been sorted. This definition differs from the book definition.

# Median

**Definition**

The median of an array **A** of $n$ distinct elements, where $n$ is an odd number is $\mathbf{A}[\frac{n+1}{2}]$, after **A** has been sorted. This definition differs from the book definition.

*Note*

*The median is an important clustering statistic that is used as a reference point in a number of practical applications.*

# Median

**Definition**

The median of an array **A** of $n$ distinct elements, where $n$ is an odd number is $\mathbf{A}[\frac{n+1}{2}]$, after **A** has been sorted. This definition differs from the book definition.

***Note***

*The median is an important clustering statistic that is used as a reference point in a number of practical applications.*

***Note***

*The oddness of n and the distinctness of the elements of* **A** *are simplifying assumptions.*

# Median

**Definition**

The median of an array **A** of $n$ distinct elements, where $n$ is an odd number is $\mathbf{A}[\frac{n+1}{2}]$, after **A** has been sorted. This definition differs from the book definition.

***Note***

*The median is an important clustering statistic that is used as a reference point in a number of practical applications.*

***Note***

*The oddness of n and the distinctness of the elements of* **A** *are simplifying assumptions. Our algorithm will work even if these assumptions are relaxed.*

# Deterministic Approaches

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

(i) Break the array into $\frac{n}{5}$ groups of 5 elements each.

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

(i) Break the array into $\frac{n}{5}$ groups of 5 elements each. Call the buckets $b_1, b_2, \ldots b_{\frac{n}{5}}$.

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

(i) Break the array into $\frac{n}{5}$ groups of 5 elements each. Call the buckets $b_1, b_2, \ldots b_{\frac{n}{5}}$.

(ii) Find the median of each bucket.

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

(i) Break the array into $\frac{n}{5}$ groups of 5 elements each. Call the buckets $b_1, b_2, \ldots b_{\frac{n}{5}}$.

(ii) Find the median of each bucket.

(iii) Recursively find the median of medians.

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

(i) Break the array into $\frac{n}{5}$ groups of 5 elements each. Call the buckets $b_1, b_2, \ldots b_{\frac{n}{5}}$.
(ii) Find the median of each bucket.
(iii) Recursively find the median of medians. Call the median of medians $p$.

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

(i) Break the array into $\frac{n}{5}$ groups of 5 elements each. Call the buckets $b_1, b_2, \ldots b_{\frac{n}{5}}$.

(ii) Find the median of each bucket.

(iii) Recursively find the median of medians. Call the median of medians $p$.

(iv) Use $p$ as a pivot to split the array into two sub-arrays $L$ and $G$, which respectively contain the elements of **A** that are smaller than $p$ and larger thatn $p$.

# Deterministic Approaches

## Sorting

Simply sort the array **A** in $O(n \cdot \log n)$ time and then return $\mathbf{A}[\frac{n+1}{2}]$.

## The bucketing approach

(i) Break the array into $\frac{n}{5}$ groups of 5 elements each. Call the buckets $b_1, b_2, \ldots b_{\frac{n}{5}}$.

(ii) Find the median of each bucket.

(iii) Recursively find the median of medians. Call the median of medians $p$.

(iv) Use $p$ as a pivot to split the array into two sub-arrays $L$ and $G$, which respectively contain the elements of **A** that are smaller than $p$ and larger thatn $p$.

(v) Recurse on the appropriate piece.

# Analysis

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.
(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.
(iii) Step (iii) takes $T(\frac{n}{5})$ time.
(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.
(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time.

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.
(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.
(iii) Step (iii) takes $T(\frac{n}{5})$ time.
(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.
(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time. Why?

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.

(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time. Why?

Accordingly, we have,

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.

(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time. Why?

Accordingly, we have,

$$T(n) \leq c \cdot n + T(\frac{n}{5}) + T(\frac{7 \cdot n}{10})$$

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.

(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time. Why?

Accordingly, we have,

$$\begin{aligned} T(n) &\leq c \cdot n + T(\frac{n}{5}) + T(\frac{7 \cdot n}{10}) \\ &\in O(n). \end{aligned}$$

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.

(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time. Why?

Accordingly, we have,

$$\begin{aligned} T(n) &\leq c \cdot n + T(\frac{n}{5}) + T(\frac{7 \cdot n}{10}) \\ &\in O(n). \end{aligned}$$

## Note

*Will the above analysis work when the bucket size is 3?*

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.

(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time. Why?

Accordingly, we have,

$$\begin{aligned} T(n) &\leq c \cdot n + T(\frac{n}{5}) + T(\frac{7 \cdot n}{10}) \\ &\in O(n). \end{aligned}$$

## *Note*

*Will the above analysis work when the bucket size is* 3*?* 7*?*

# Analysis

## Analysis

Let $T(n)$ denote the time taken by the median finding algorithm.

(i) Step (i) takes $c_1 \cdot n$ time, for some constant $c_1$.

(ii) Step (ii) takes $c_2 \cdot n$ time, for some constant $c_2$.

(iii) Step (iii) takes $T(\frac{n}{5})$ time.

(iv) Step (iv) takes $c_3 \cdot n$ time, for some constant $c_3$.

(v) Step (v) takes $T(\frac{7 \cdot n}{10})$ time. Why?

Accordingly, we have,

$$\begin{aligned} T(n) &\leq c \cdot n + T(\frac{n}{5}) + T(\frac{7 \cdot n}{10}) \\ &\in O(n). \end{aligned}$$

## *Note*

*Will the above analysis work when the bucket size is* $3$*?* $7$*? The key observation is that recursive calls need to be performed on pieces whose cumulative size is at most* $(1-\varepsilon) \cdot n$*, for some* $\varepsilon > 0$.

# Randomized Selection

# Randomized Selection

## Principal Ideas

# Randomized Selection

## Principal Ideas

# Randomized Selection

## Principal Ideas

- A random sample **R** of **A** reveals sufficient information about its median $m$, w.h.p.

# Randomized Selection

## Principal Ideas

- A random sample **R** of **A** reveals sufficient information about its median $m$, w.h.p.
- **R** contains markers $d$ and $u$, such that $d \leq m \leq u$, w.h.p.

# Randomized Selection

## Principal Ideas

- A random sample **R** of **A** reveals sufficient information about its median $m$, w.h.p.
- **R** contains markers $d$ and $u$, such that $d \leq m \leq u$, w.h.p.
- From the random sample and the markers, we can construct a set **C**, such that **C** is of small size and contains the median w.h.p.

## Randomized Selection

### Principal Ideas

- A random sample **R** of **A** reveals sufficient information about its median $m$, w.h.p.
- **R** contains markers $d$ and $u$, such that $d \leq m \leq u$, w.h.p.
- From the random sample and the markers, we can construct a set **C**, such that **C** is of small size and contains the median w.h.p.
- The set **C** can be sorted to determine the median.

# The Lazy Select Algorithm

# The Lazy Select Algorithm

**Function** LAZY-SELECT-MEDIAN(**A**)

Construct a set **R**, by choosing $n^{\frac{3}{4}}$ elements from **A**, uniformly and at random.
Sort **R** using some optimal sorting algorithm.
Let $d$ be the $(\frac{1}{2}n^{\frac{3}{4}} - \sqrt{n})^{th}$ smallest element of **R**.
Let $u$ be the $(\frac{1}{2}n^{\frac{3}{4}} + \sqrt{n})^{th}$ smallest element of **R**.
Construct $\mathbf{C} = \{x : x \in \mathbf{A}, d \leq x \leq u\}$. Also compute $l_d = |\{x \in \mathbf{A} : x < d\}|$ and $l_u = |\{x \in \mathbf{A} : x > u\}|$.
**if** $(l_d > \frac{n}{2})$ **then**
    **return** (“**Fail.**”)
**else**
    **if** $(l_u > \frac{n}{2})$ **then**
        **return** (“**Fail.**”)
    **end if**
**end if**
**if** $(|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}})$ **then**
    **return** (“**Fail.**”)
**end if**
Sort **C** in ascending order.
**return** $(C[\frac{n}{2} - l_d + 1])$

# Correctness

# Correctness

**Theorem**

*The* LAZY-SELECT-MEDIAN() *algorithm runs in linear time and it if it does not output **Fail**, it returns the median of* **A***.*

# Correctness

## Theorem

*The* LAZY-SELECT-MEDIAN() *algorithm runs in linear time and it if it does not output **Fail**, it returns the median of* **A**.

## Modes of failure

# Correctness

## Theorem

*The* LAZY-SELECT-MEDIAN() *algorithm runs in linear time and it if it does not output* ***Fail****, it returns the median of* **A***.*

## Modes of failure

(i) Set **C** does not contain the median.

# Correctness

## Theorem

*The* LAZY-SELECT-MEDIAN() *algorithm runs in linear time and it if it does not output* ***Fail****, it returns the median of* **A**.

## Modes of failure

(i) Set **C** does not contain the median. (Exclusion error).

# Correctness

**Theorem**

*The* Lazy-Select-Median() *algorithm runs in linear time and it if it does not output* ***Fail****, it returns the median of* **A***.*

**Modes of failure**

(i) Set **C** does not contain the median. (Exclusion error).

(ii) Set **C** is too large.

# Correctness

## Theorem

*The* LAZY-SELECT-MEDIAN() *algorithm runs in linear time and it if it does not output **Fail**, it returns the median of **A**.*

## Modes of failure

(i) Set **C** does not contain the median. (Exclusion error).

(ii) Set **C** is too large. (Size error).

# Correctness

**Theorem**

*The* LAZY-SELECT-MEDIAN() *algorithm runs in linear time and it if it does not output **Fail**, it returns the median of* **A**.

**Modes of failure**

(i) Set **C** does not contain the median. (Exclusion error).

(ii) Set **C** is too large. (Size error).

We will argue that the probability of each of the above events occurring in a run, is at most $O(n^{-\frac{1}{4}})$.

# Correctness

## Theorem

*The* LAZY-SELECT-MEDIAN() *algorithm runs in linear time and it if it does not output* ***Fail****, it returns the median of* **A**.

## Modes of failure

(i) Set **C** does not contain the median. (Exclusion error).

(ii) Set **C** is too large. (Size error).

We will argue that the probability of each of the above events occurring in a run, is at most $O(n^{-\frac{1}{4}})$. Hence with probability at least $1 - O(n^{-\frac{1}{4}})$, the algorithm correctly outputs the median, on a given run.

# Exclusion error

# Exclusion error

## Observations

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$,

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$.

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$.

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?)

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event.

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

Let

$$X_i =$$

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ element of } \mathbf{R} \text{ is less than or equal to the median} \end{cases}$$

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ element of } \mathbf{R} \text{ is less than or equal to the median} \\ 0, & \text{otherwise} \end{cases}$$

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ element of } \mathbf{R} \text{ is less than or equal to the median} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) \quad =$$

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ element of } \mathbf{R} \text{ is less than or equal to the median} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) \quad = \quad \frac{\frac{n+1}{2}}{n}$$

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ element of } \mathbf{R} \text{ is less than or equal to the median} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$\begin{aligned} P(X_i = 1) &= \frac{\frac{n+1}{2}}{n} \\ &= \frac{1}{2} + \frac{1}{2 \cdot n} \end{aligned}$$

# Exclusion error

## Observations

- Exclusion error can occur in one of two ways, viz., (a) $l_d > \frac{n}{2}$, and (b) $l_u > \frac{n}{2}$. We will focus on deriving bounds for (a); the bounds for (b) can be derived in identical fashion.
- $l_d > \frac{n}{2} \Rightarrow |\{x \in \mathbf{R} : x \leq m\}| < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})$. (Why?) If not, the median would be between $d$ and $u$!
- Let $E_1$ denote the above event. We will show that $P(E_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Analysis

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ element of } \mathbf{R} \text{ is less than or equal to the median} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$\begin{aligned} P(X_i = 1) &= \frac{\frac{n+1}{2}}{n} \\ &= \frac{1}{2} + \frac{1}{2 \cdot n} \end{aligned}$$

We thus see that each $X_i$ is a Bernoulli random variable with parameter $p = \frac{1}{2} + \frac{1}{2 \cdot n}$.

# Exclusion error (contd.)

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$.

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$.

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$. It follows that,

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$. It follows that,

$$E[Y_1] =$$

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$. It follows that,

$$E[Y_1] = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$. It follows that,

$$\begin{aligned} E[Y_1] &= \sum_{i=1}^{n^{\frac{3}{4}}} X_i \\ &= \frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} \end{aligned}$$

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$. It follows that,

$$\begin{aligned} E[Y_1] &= \sum_{i=1}^{n^{\frac{3}{4}}} X_i \\ &= \frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} \\ \mathbf{Var}[Y_1] &= \end{aligned}$$

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$. It follows that,

$$\begin{aligned} E[Y_1] &= \sum_{i=1}^{n^{\frac{3}{4}}} X_i \\ &= \frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} \\ \mathbf{Var}[Y_1] &= n^{\frac{3}{4}} \cdot \mathbf{Var}[X_i] \end{aligned}$$

# Exclusion error (contd.)

## Analysis (contd.)

Let $Y_1 = |\{x \in \mathbf{R} : x \leq m\}|$. Clearly,

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

It follows that $Y_1$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $\frac{1}{2} + \frac{1}{2 \cdot n}$. It follows that,

$$\begin{aligned} E[Y_1] &= \sum_{i=1}^{n^{\frac{3}{4}}} X_i \\ &= \frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} \\ \mathbf{Var}[Y_1] &= n^{\frac{3}{4}} \cdot \mathbf{Var}[X_i] \\ &< \frac{1}{4} n^{\frac{3}{4}} \end{aligned}$$

# Exclusion error (contd.)

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$P(E_1) \quad = \quad P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n}))$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$\begin{aligned} P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\ &= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \end{aligned}$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$\begin{aligned} P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\ &= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\ &\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \end{aligned}$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$\begin{aligned}
P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\
&= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\
&\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \\
&= P((E[Y_1] - Y_1) > \sqrt{n})
\end{aligned}$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$
\begin{aligned}
P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\
&= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\
&\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \\
&= P((E[Y_1] - Y_1) > \sqrt{n}) \\
&\leq P(|E[Y_1] - Y_1| > \sqrt{n})
\end{aligned}
$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$
\begin{aligned}
P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\
&= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\
&\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \\
&= P((E[Y_1] - Y_1) > \sqrt{n}) \\
&\leq P(|E[Y_1] - Y_1| > \sqrt{n}) \\
&\leq \frac{\mathbf{Var}[Y_1]}{(\sqrt{n})^2}
\end{aligned}
$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$\begin{aligned}
P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\
&= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\
&\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \\
&= P((E[Y_1] - Y_1) > \sqrt{n}) \\
&\leq P(|E[Y_1] - Y_1| > \sqrt{n}) \\
&\leq \frac{\mathbf{Var}[Y_1]}{(\sqrt{n})^2} \\
&\leq \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n}
\end{aligned}$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$
\begin{aligned}
P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\
&= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\
&\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \\
&= P((E[Y_1] - Y_1) > \sqrt{n}) \\
&\leq P(|E[Y_1] - Y_1| > \sqrt{n}) \\
&\leq \frac{\mathbf{Var}[Y_1]}{(\sqrt{n})^2} \\
&\leq \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\
&\leq \frac{1}{4} \cdot n^{-\frac{1}{4}}
\end{aligned}
$$

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$\begin{aligned}
P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\
&= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\
&\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \\
&= P((E[Y_1] - Y_1) > \sqrt{n}) \\
&\leq P(|E[Y_1] - Y_1| > \sqrt{n}) \\
&\leq \frac{\mathbf{Var}[Y_1]}{(\sqrt{n})^2} \\
&\leq \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\
&\leq \frac{1}{4} \cdot n^{-\frac{1}{4}}
\end{aligned}$$

Let $E_2$ denote the event that $l_u > \frac{n}{2}$.

# Exclusion error (contd.)

## Analysis (contd.)

Applying Chebyshev's inequality, we have,

$$\begin{aligned}
P(E_1) &= P(Y_1 < (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\
&= P((\frac{1}{2} \cdot n^{\frac{3}{4}} - Y_1) > \sqrt{n}) \\
&\leq P((\frac{1}{2} \cdot n^{\frac{3}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} - Y_1) > \sqrt{n}) \\
&= P((E[Y_1] - Y_1) > \sqrt{n}) \\
&\leq P(|E[Y_1] - Y_1| > \sqrt{n}) \\
&\leq \frac{\mathbf{Var}[Y_1]}{(\sqrt{n})^2} \\
&\leq \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\
&\leq \frac{1}{4} \cdot n^{-\frac{1}{4}}
\end{aligned}$$

Let $E_2$ denote the event that $l_u > \frac{n}{2}$. Arguing in identical fashion as above, we can show that $P(E_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

## Size error

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are greater than the median $m$ (call this event $Z_1$),

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are smaller than the median $m$ (call this event $Z_2$).

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

Assume that $Z_1$ has occurred. This means that, $u$ has rank at least $\frac{1}{2} \cdot n + 2 \cdot n^{\frac{3}{4}}$ in **A**.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

Assume that $Z_1$ has occurred. This means that, $u$ has rank at least $\frac{1}{2} \cdot n + 2 \cdot n^{\frac{3}{4}}$ in **A**. Recall that $u$ is the element of rank $\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}$ in **R**.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

Assume that $Z_1$ has occurred. This means that, $u$ has rank at least $\frac{1}{2} \cdot n + 2 \cdot n^{\frac{3}{4}}$ in $\mathbf{A}$. Recall that $u$ is the element of rank $\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}$ in $\mathbf{R}$.
Consider the set $\mathbf{L}$ of the largest $\frac{1}{2} \cdot n - 2 \cdot n^{\frac{3}{4}}$ elements of $\mathbf{A}$.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

Assume that $Z_1$ has occurred. This means that, $u$ has rank at least $\frac{1}{2} \cdot n + 2 \cdot n^{\frac{3}{4}}$ in **A**. Recall that $u$ is the element of rank $\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}$ in **R**.

Consider the set **L** of the largest $\frac{1}{2} \cdot n - 2 \cdot n^{\frac{3}{4}}$ elements of **A**. How many samples of **R** are there in **L**?

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

Assume that $Z_1$ has occurred. This means that, $u$ has rank at least $\frac{1}{2} \cdot n + 2 \cdot n^{\frac{3}{4}}$ in **A**. Recall that $u$ is the element of rank $\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}$ in **R**.
Consider the set **L** of the largest $\frac{1}{2} \cdot n - 2 \cdot n^{\frac{3}{4}}$ elements of **A**. How many samples of **R** are there in **L**? At least $n^{\frac{3}{4}} - (\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n})$

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of $\mathbf{C}$ are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

Assume that $Z_1$ has occurred. This means that, $u$ has rank at least $\frac{1}{2} \cdot n + 2 \cdot n^{\frac{3}{4}}$ in $\mathbf{A}$. Recall that $u$ is the element of rank $\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}$ in $\mathbf{R}$.
Consider the set $\mathbf{L}$ of the largest $\frac{1}{2} \cdot n - 2 \cdot n^{\frac{3}{4}}$ elements of $\mathbf{A}$. How many samples of $\mathbf{R}$ are there in $\mathbf{L}$? At least $n^{\frac{3}{4}} - (\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}) = \frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n}$.

# Size error

## Analysis

Let $Z$ denote the event that $|\mathbf{C}| > 4 \cdot n^{\frac{3}{4}}$. Observe that one of the following two events must occur:

(i) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are greater than the median $m$ (call this event $Z_1$),

(ii) At least $2 \cdot n^{\frac{3}{4}}$ of the elements of **C** are smaller than the median $m$ (call this event $Z_2$).

We shall show that $P(Z_1) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$. It follows that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$, by symmetry. Hence, $P(Z) \leq P(Z_1) + P(Z_2) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

Assume that $Z_1$ has occurred. This means that, $u$ has rank at least $\frac{1}{2} \cdot n + 2 \cdot n^{\frac{3}{4}}$ in **A**. Recall that $u$ is the element of rank $\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}$ in **R**.
Consider the set **L** of the largest $\frac{1}{2} \cdot n - 2 \cdot n^{\frac{3}{4}}$ elements of **A**. How many samples of **R** are there in **L**? At least $n^{\frac{3}{4}} - (\frac{1}{2} \cdot n^{\frac{3}{4}} + \sqrt{n}) = \frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n}$.
Thus, $P(Z_1)$ is precisely the probability that the number of samples of **R** in **L** is at least $\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n}$.

# Size error (contd.)

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i =$$

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \end{cases}$$

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) \quad =$$

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) = \frac{|\mathbf{L}|}{n}$$

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) = \frac{|\mathbf{L}|}{n}$$

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) = \frac{|\mathbf{L}|}{n}$$

We thus see that each $X_i$ is a Bernoulli random variable with parameter $p = \frac{1}{2} - 2 \cdot n^{-\frac{1}{4}}$.

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) = \frac{|\mathbf{L}|}{n}$$

We thus see that each $X_i$ is a Bernoulli random variable with parameter $p = \frac{1}{2} - 2 \cdot n^{-\frac{1}{4}}$.

Observe that the total number of samples of **R** which are in **L** is given by:

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) = \frac{|\mathbf{L}|}{n}$$

We thus see that each $X_i$ is a Bernoulli random variable with parameter $p = \frac{1}{2} - 2 \cdot n^{-\frac{1}{4}}$.

Observe that the total number of samples of **R** which are in **L** is given by: $X = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$.

# Size error (contd.)

## Analysis (contd.)

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample of } \mathbf{R} \text{ is in } \mathbf{L} \\ 0, & \text{otherwise} \end{cases}$$

Clearly,

$$P(X_i = 1) = \frac{|\mathbf{L}|}{n}$$

We thus see that each $X_i$ is a Bernoulli random variable with parameter $p = \frac{1}{2} - 2 \cdot n^{-\frac{1}{4}}$.

Observe that the total number of samples of **R** which are in **L** is given by: $X = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$. Clearly $X$ is a binomial random variable with parameters $n^{\frac{3}{4}}$ and $p$.

# Size error (contd.)

# Size error (contd.)

## Analysis (contd.)

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$E[X] \quad =$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$E[X] = \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n}$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$
\begin{aligned}
E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\
\mathbf{Var}[X] &=
\end{aligned}
$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \textbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1 - p) \end{aligned}$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \textbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\ &< \end{aligned}$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \mathbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\ &< \frac{1}{4} \cdot n^{\frac{3}{4}} \end{aligned}$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$
\begin{aligned}
E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\
\textbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\
&< \frac{1}{4} \cdot n^{\frac{3}{4}}
\end{aligned}
$$

Applying Chebyshev's inequality, we conclude that,

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \mathbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\ &< \frac{1}{4} \cdot n^{\frac{3}{4}} \end{aligned}$$

Applying Chebyshev's inequality, we conclude that,

$$P(Z_1) = P(X \geq (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n}))$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \mathbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\ &< \frac{1}{4} \cdot n^{\frac{3}{4}} \end{aligned}$$

Applying Chebyshev's inequality, we conclude that,

$$\begin{aligned} P(Z_1) &= P(X \geq (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\ &= P((X - \frac{1}{2} \cdot n^{\frac{3}{4}}) \geq -\sqrt{n}) \end{aligned}$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \mathbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\ &< \frac{1}{4} \cdot n^{\frac{3}{4}} \end{aligned}$$

Applying Chebyshev's inequality, we conclude that,

$$\begin{aligned} P(Z_1) &= P(X \geq (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\ &= P((X - \frac{1}{2} \cdot n^{\frac{3}{4}}) \geq -\sqrt{n}) \\ &= P((X - (\frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n})) \geq \sqrt{n}) \end{aligned}$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \mathbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\ &< \frac{1}{4} \cdot n^{\frac{3}{4}} \end{aligned}$$

Applying Chebyshev's inequality, we conclude that,

$$\begin{aligned} P(Z_1) &= P(X \geq (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\ &= P((X - \frac{1}{2} \cdot n^{\frac{3}{4}}) \geq -\sqrt{n}) \\ &= P((X - (\frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n})) \geq \sqrt{n}) \\ &= P((X - E[X]) \geq \sqrt{n}) \end{aligned}$$

# Size error (contd.)

## Analysis (contd.)

It follows that,

$$\begin{aligned} E[X] &= \frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n} \\ \mathbf{Var}[X] &= n^{\frac{3}{4}} \cdot p \cdot (1-p) \\ &< \frac{1}{4} \cdot n^{\frac{3}{4}} \end{aligned}$$

Applying Chebyshev's inequality, we conclude that,

$$\begin{aligned} P(Z_1) &= P(X \geq (\frac{1}{2} \cdot n^{\frac{3}{4}} - \sqrt{n})) \\ &= P((X - \frac{1}{2} \cdot n^{\frac{3}{4}}) \geq -\sqrt{n}) \\ &= P((X - (\frac{1}{2} \cdot n^{\frac{3}{4}} - 2 \cdot \sqrt{n})) \geq \sqrt{n}) \\ &= P((X - E[X]) \geq \sqrt{n}) \\ &\leq P(|X - E[X]| \geq \sqrt{n}) \end{aligned}$$

# Size error (contd.)

# Size error (contd.)

## Analysis (contd.)

$$P(Z_1) \leq$$

# Size error (contd.)

## Analysis (contd.)

$$P(Z_1) \leq \frac{\mathbf{Var}[X]}{(\sqrt{n})^2}$$

# Size error (contd.)

## Analysis (contd.)

$$
\begin{aligned}
P(Z_1) &\leq \frac{\mathbf{Var}[X]}{(\sqrt{n})^2} \\
&< \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\
&= \frac{1}{4} \cdot n^{-\frac{1}{4}}
\end{aligned}
$$

# Size error (contd.)

## Analysis (contd.)

$$\begin{aligned} P(Z_1) &\leq \frac{\textbf{Var}[X]}{(\sqrt{n})^2} \\ &< \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\ &= \frac{1}{4} \cdot n^{-\frac{1}{4}} \end{aligned}$$

Using a symmetric argument, we conclude that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$ as well,

# Size error (contd.)

## Analysis (contd.)

$$
\begin{aligned}
P(Z_1) &\leq \frac{\mathbf{Var}[X]}{(\sqrt{n})^2} \\
&< \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\
&= \frac{1}{4} \cdot n^{-\frac{1}{4}}
\end{aligned}
$$

Using a symmetric argument, we conclude that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$ as well, and hence, $P(Z) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.

# Size error (contd.)

## Analysis (contd.)

$$
\begin{aligned}
P(Z_1) &\leq \frac{\mathbf{Var}[X]}{(\sqrt{n})^2} \\
&< \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\
&= \frac{1}{4} \cdot n^{-\frac{1}{4}}
\end{aligned}
$$

Using a symmetric argument, we conclude that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$ as well, and hence, $P(Z) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.
Therefore, the probability that the LAZY-SELECT-MEDIAN() algorithm fails is at most, $\frac{1}{4} \cdot n^{-\frac{1}{4}} + \frac{1}{4} \cdot n^{-\frac{1}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}}$

# Size error (contd.)

## Analysis (contd.)

$$
\begin{aligned}
P(Z_1) &\leq \frac{\mathbf{Var}[X]}{(\sqrt{n})^2} \\
&< \frac{\frac{1}{4} \cdot n^{\frac{3}{4}}}{n} \\
&= \frac{1}{4} \cdot n^{-\frac{1}{4}}
\end{aligned}
$$

Using a symmetric argument, we conclude that $P(Z_2) \leq \frac{1}{4} \cdot n^{-\frac{1}{4}}$ as well, and hence, $P(Z) \leq \frac{1}{2} \cdot n^{-\frac{1}{4}}$.
Therefore, the probability that the LAZY-SELECT-MEDIAN() algorithm fails is at most, $\frac{1}{4} \cdot n^{-\frac{1}{4}} + \frac{1}{4} \cdot n^{-\frac{1}{4}} + \frac{1}{2} \cdot n^{-\frac{1}{4}} = n^{-\frac{1}{4}}$.

# Size error (contd.)

## Analysis (contd.)

$$\begin{aligned} P(Z_1) &\leq \frac{\mathbf{Var}[X]}{(\sqrt{n})^2} \\ &< \frac{\frac{1}{4}\cdot n^{\frac{3}{4}}}{n} \\ &= \frac{1}{4}\cdot n^{-\frac{1}{4}} \end{aligned}$$

Using a symmetric argument, we conclude that $P(Z_2) \leq \frac{1}{4}\cdot n^{-\frac{1}{4}}$ as well, and hence, $P(Z) \leq \frac{1}{2}\cdot n^{-\frac{1}{4}}$.
Therefore, the probability that the LAZY-SELECT-MEDIAN() algorithm fails is at most, $\frac{1}{4}\cdot n^{-\frac{1}{4}} + \frac{1}{4}\cdot n^{-\frac{1}{4}} + \frac{1}{2}\cdot n^{-\frac{1}{4}} = n^{-\frac{1}{4}}$. It follows that with probability $(1 - \frac{1}{n^{\frac{1}{4}}})$, the LAZY-SELECT-MEDIAN() algorithm finds the median of **A** in one round.