Sets: Basics and Vocabulary

Video companion

1 Set theory basics

- What is a set?
- Cardinality (size)
- Intersections
- Unions

2 What is a set?

Vocab: A *set* is made up of *elements*.

Example: $A = \{1, 2, -3, 7\}$ and $E = \{\text{apple, monkey, Daniel Egger}\}$

- $2 \in A$: "2 is an element of A"
- $8 \notin A$: "8 is NOT an element of A"

3 Cardinality

Vocab: The *cardinality* (size) of a set is the number of elements in it.

- $|A| = 4$ (there are 4 elements in A, so the cardinality is 4)
- $|E| = 3$ (there are 3 elements in E, so the cardinality is 3)

4 Intersections

The *intersection* is defined as elements that are in both sets.

Symbol $\cap$: "intersects" (and)

Example: $A = \{1, 2, -3, 7\}$ and $B = \{2, -3, 8, 10\}$ and $D = \{5, 10\}$

- $A \cap B = \{2, -3\}$
- $B \cap D = \{10\}$

In general, $A \cap B = \{x \in A \text{ and } x \in B\}$

If there are no elements in common, the answer is the empty set $\emptyset$. The cardinality of the empty set $|\emptyset| = 0$.

- $A \cap D = \{\emptyset\}$

5 Unions

The *union* is defined as elements that are in either set.

Symbol $\cup$: "union" (or)

Example: $A = \{1, 2, -3, 7\}$ and $B = \{2, -3, 8, 10\}$ and $D = \{5, 10\}$

- $A \cup B = \{1, 2, -3, 7, 8, 10\}$
- $A \cup D = \{1, 2, -3, 7, 5, 10\}$

In general, $A \cup B = \{x \in A \text{ or } x \in B\}$.

Sets: Venn Diagrams

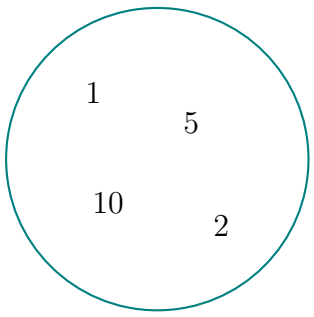
Video companion

1 Visualizing sets

- Venn diagrams
- Inclusion-exclusion formula
- Medical testing example, re-visited

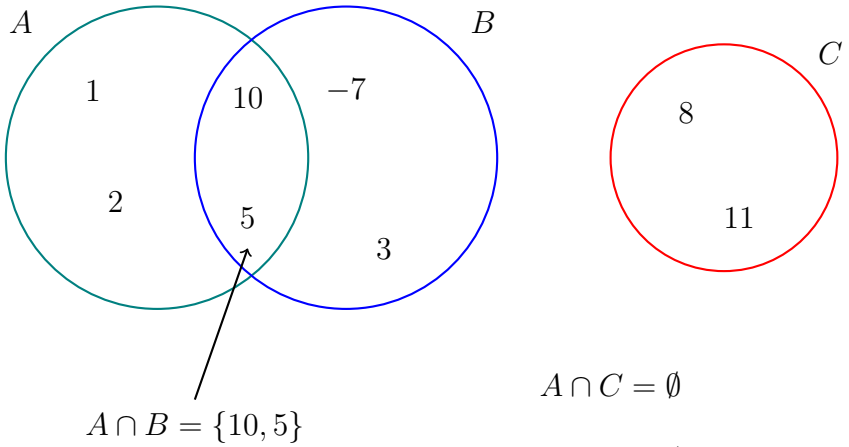
2 Single set

$A = \{1, 5, 10, 2\}$ $\quad |A| = 4$

$A =$ 

3 Multiple sets

$A = \{1, 5, 10, 2\}$
$B = \{5, -7, 10, 3\}$
$C = \{8, 11\}$

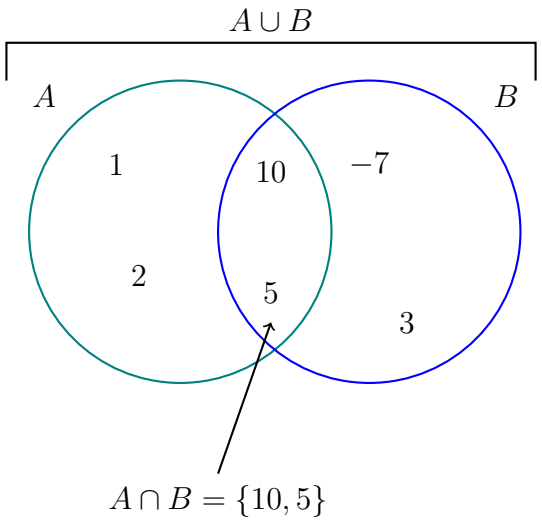


$A \cap B = \{10, 5\}$

$A \cap C = \emptyset$

$B \cap C = \emptyset$

4 Inclusion-exclusion formula



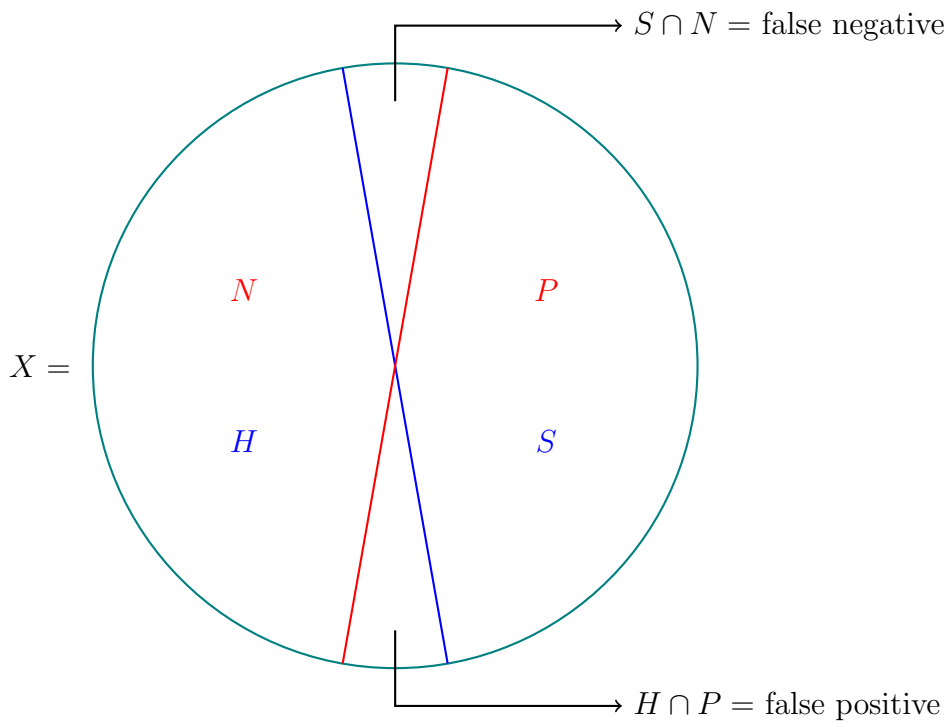
Inclusion-exclusion formula:

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Check with this example:

$$6 \stackrel{?}{=} 4 + 4 - 2$$
$$6 = 6 \quad \checkmark$$

5 Medical testing example



$X = H \cup S \qquad H \cap S = \emptyset$
$S = N \cup P \qquad N \cap P = \emptyset$

Data Science Math Skills
Paul Bendich and Daniel Egger
Duke University

Sets: Medical Testing Example

Video companion

1 Example using set theory

VBS: "very bad syndrome"

X = set of people in a clinical trial

$S = \{x \in X : x \text{ has VBS}\}$
$H = \{x \in X : x \text{ does not have VBS}\}$

$X = S \cup H$ (you either have VBS or you don't)
$S \cap H = \emptyset$ (no one both has and doesn't have it)

Point of medial testing to figure out whether a person is in S or in H

2 Test

$P = \{x \in X : x \text{ tests positive for VBS}\}$
$N = \{x \in X : x \text{ tests negative for VBS}\}$

$P \cup N = X$ (you either test positive or negative)
$P \cap N = \emptyset$ (no one tests both positive and negative)

In a perfect world, S would equal P—the sick people would always test positive, and H would equal N—the healthy people would always test negative.

...but this is not always the case.

$S \cap P$	$H \cap N$	$S \cap N$	$H \cap P$
true positive	true negative	false negative	false positive

3 Cardinality

$\frac{|S|}{|X|}$ = proportion of people in the study who do genuinely have VBS
$\frac{|H|}{|X|}$ = proportion of people in the study without VBS
$\frac{|S|}{|X|} + \frac{|H|}{|X|} = 1$

$\frac{\lvert S \cap P \rvert}{\lvert S \rvert}$	true positive rate	would like to be close to 1
$\frac{\lvert H \cap P \rvert}{\lvert H \rvert}$	false positive rate	would like to be as small as possible
$\frac{\lvert S \cap N \rvert}{\lvert S \rvert}$	false negative rate	would like to be as small as possible
$\frac{\lvert H \cap N \rvert}{\lvert H \rvert}$	true negative rate	would like to be close to 1

Numbers: The Real Number Line

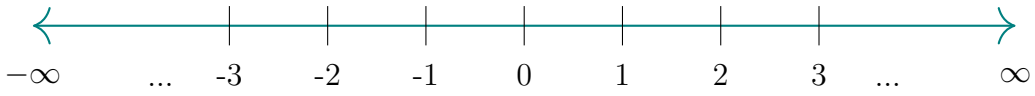
Video companion

1 Introduction

- What is $\mathbb{R}$?
- Positive, negative
- Absolute value

2 Integers and rational numbers

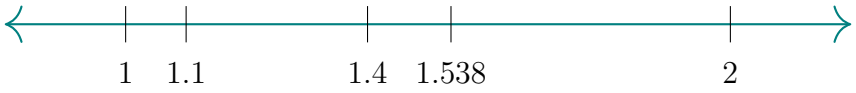
Graph of $\mathbb{R}$, the real numbers:



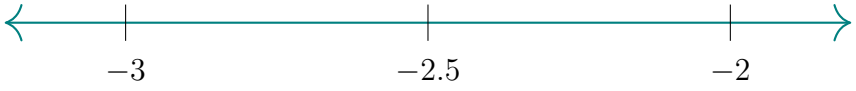
Subset of real numbers, integers:

$$\mathbb{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}$$

Segment between 1 and 2:

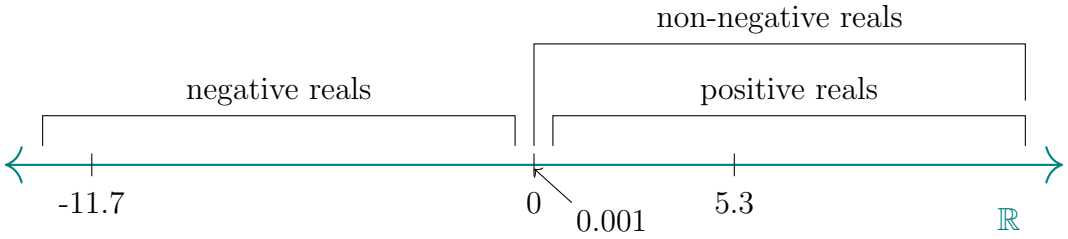


Segment between -3 and -2:



Some real numbers terminate, and some do not.
The number $\pi = 3.14159...$ is *irrational*, i.e. it does not repeat after the decimal point.

3 Sets of real numbers



4 Absolute value

The absolute value of a number x, $|x|$, is the distance from x to 0.

Example:

$$|7.1| = 7.1$$
$$|-7.1| = 7.1 = -(-7.1)$$

General rule:
For any $x \in \mathbb{R}$,

$$|x| = \begin{cases} x, & \text{if } x \text{ is non-negative} \\ -x, & \text{if } x \text{ is negative} \end{cases}$$

Check:

$$|8.7| = 8.7$$
$$|-10| = -(-10) = 10$$

Data Science Math Skills
Paul Bendich and Daniel Egger
Duke University

Numbers: Greater-than and Less-than

Video companion

1 Inequalities, basic idea

Introduction to symbols:

$a < b$	"a is less than b"
$x > y$	"x is greater than y"
$c \leq d$	"c is less than or equal to d"
$z \geq w$	"z is greater than or equal to w"
$e \ll f$	"e is much, much less than f"

2 Inequality on the real number line

$2 < 3.1$	"2 is to the left of 3.1 on the real number line"
$-11.78 < 3.1$	"-11.78 is to the left of 3.1 on the real number line"

For any $a < b$, a must be to the left of b on the real number line.

$3.1 > 2$	"3.1 is to the right of 2 on the real number line"

In general, a is less than b, if, and only if, b is greater than a:

$$\boxed{a < b \iff b > a}$$

3 Much, much less than

$x \ll y$ "x is much, much less than y"
(Not proper math, but used frequently in data science)

For example, $1 \ll 1,000,000$, which is reasonable but not possible to prove "true"

4 Less than or equal to

$a \leq b$ means $a < b$ or $a = b$

Examples:

Is $2 \leq 3.1$ true?

$$\left.\begin{array}{ll} 2 < 3.1 & \checkmark \\ 2 = 3.1 & \times \end{array}\right] \checkmark$$

Is $2 \leq 2$ true?

$$\left.\begin{array}{ll} 2 < 2 & \times \\ 2 = 2 & \checkmark \end{array}\right] \checkmark$$

Is $2 \leq 0.8$ true?

$$\left.\begin{array}{ll} 2 < 0.8 & \times \\ 2 = 0.8 & \times \end{array}\right] \times$$

Paul Bendich and Daniel Egger
Duke University

Numbers: Algebra with Inequalities

Video companion

1 Introduction

- Review algebra with equalities ($=$)
  - how?
  - why?
- Learn algebra with inequalities ($<, >, \leq, \geq$)
  - what works
  - A BIG WARNING

2 Algebra with equalities

$$\begin{aligned} 4 &= 4 \\ 4+3 &= 4+3 \\ 7 &= 7 \quad \checkmark \end{aligned}$$

Rule:
If $a = b$, then $a + c = b + c$.

Example:

$$\begin{aligned} x+3 &= 10 \\ (x+3)-3 &= 10-3 \\ x &= 7 \end{aligned}$$

Similarly with multiplication,

$$\begin{aligned} 4 &= 4 \\ 2 \cdot 4 &= 2 \cdot 4 \\ 8 &= 8 \quad \checkmark \end{aligned}$$

$$\begin{aligned} 4 &= 4 \\ (-3)\cdot 4 &= (-3)\cdot 4 \\ -12 &= -12 \quad \checkmark \end{aligned}$$

Rule:
If a, b, and c are numbers, and $c \neq 0$, and $a = b$, then $c \cdot a = c \cdot b$.

Example:

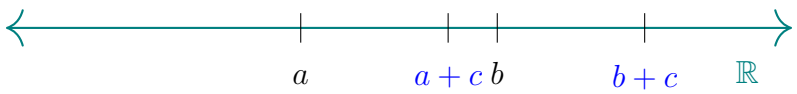
$$\begin{aligned} -5x &= 15 \\ \left(-\frac{1}{5}\right)\cdot(-5x) &= \left(-\frac{1}{5}\right)\cdot 15 \\ x &= -3 \end{aligned}$$

3 Algebra with inequalities

$$\begin{aligned} 4 &< 7 \\ 4+2 &\overset{?}{<} 7+2 \\ 6 &\overset{?}{<} 9 \quad \checkmark \end{aligned}$$

$$\begin{aligned} 4 &< 7 \\ 4-1 &\overset{?}{<} 7-1 \\ 3 &\overset{?}{<} 6 \quad \checkmark \end{aligned}$$

Rule:
If $a < b$, then $a + c < b + c$.



Example:

$$\begin{aligned} x + 3 &< 10 \\ (x+3) - 3 &< 10 - 3 \\ x &< 7 \end{aligned}$$

$x \in (-\infty, 7)$

Test cases with multiplication:

$$\begin{aligned} 5 &< 8 \\ 3 \cdot 5 &\overset{?}{<} 3 \cdot 8 \\ 15 &\overset{?}{<} 40 \quad \checkmark \end{aligned}$$

$$\begin{aligned} 5 &< 8 \\ (-1) \cdot 5 &\overset{?}{<} (-1) \cdot 8 \\ -5 &\overset{?}{<} -8 \quad \times \\ -5 &> -8 \quad ! \end{aligned}$$

Rule:

Suppose $a < b$.

If $c > 0$, then $a \cdot c < b \cdot c$.

If $c < 0$, then $a \cdot c > b \cdot c$.

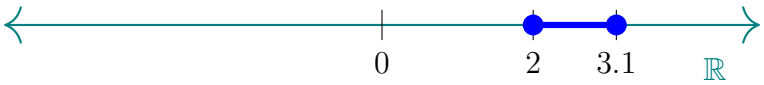
Example:

$$\begin{aligned} -2x &< 10 \\ \left(-\frac{1}{2}\right) \cdot (-2x) &> \left(-\frac{1}{2}\right) \cdot 10 \\ x &> -5 \end{aligned}$$

Numbers: Intervals and Interval Notation

Video companion

1 Closed intervals



Real number line is an infinite set. There are also infinite subsets.

$[2, 3.1] = \{x \in \mathbb{R} : 2 \leq x \leq 3.1\}$

$2.3 \in [2, 3.1]$ because $2 \leq 2.3 \leq 3.1$
$3 \in [2, 3.1]$
$3.1 \in [2, 3.1]$
$1 \notin [2, 3.1]$ because $2 \not\leq 1 \leq 3.1$

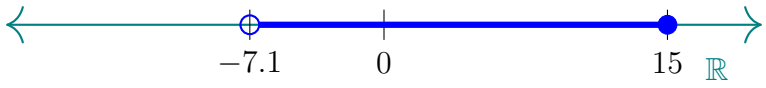
2 Open intervals

$(5, 8) = \{x \in \mathbb{R} : 5 < x < 8\}$

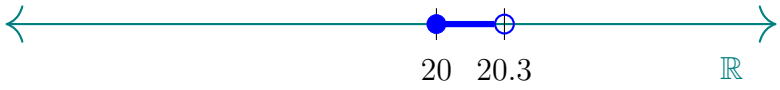
$5.5 \in (5, 8)$ because $5 < 5.5 < 8$
$5.0001 \in (5, 8)$
$5 \notin (5, 8)$ because $5 \not< 5 < 8$

The intervals $[5, 8]$ and $(5, 8)$ differ at exactly two numbers: 5 and 8.

3 Half-open intervals



$(-7.1, 15] = \{x \in \mathbb{R} : -7.1 < x \leq 15\}$

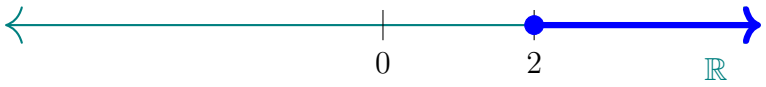


$[20, 20.3) = \{x \in \mathbb{R} : 20 \leq x < 20.3\}$

4 Recap vocabulary

- Closed intervals $[2, 3.1]$
- Open intervals $(5, 8)$
- Half-open intervals $(2, 3]$, $[20, 20.3)$

5 Rays



$[2, \infty) = \{x \in \mathbb{R} : x \geq 2\}$

Another example:
$(-\infty, 7.1) = \{x \in \mathbb{R} : x < 7.1\}$

6 What does an "answer" mean?

Solving an equality gives you a number:

$$\begin{aligned} x + 5 &= 10 \\ x &= 5 \end{aligned}$$

Solving an inequality give you an interval:

$$\begin{aligned} 1 \leq x + 5 &< 10 \\ -4 \leq x &< 5 \\ x &\in [-4, 5) \end{aligned}$$

Paul Bendich and Daniel Egger
Duke University

Sigma Notation: Introduction to Summation

Video companion

1 Sigma notation (Σ)

Examples that will be seen in this video:

$$\sum_{i=1}^{4} i^2 = 30$$

$$\sum_{i=1}^{5} (2i+3) = 45$$

$$\sum_{j=3}^{7} \frac{j}{2} = \frac{25}{2}$$

2 First example

Example:

$$\begin{aligned}\sum_{i=1}^{4} i^2 &= 1^2 + 2^2 + 3^2 + 4^2 \\ &= 30\end{aligned}$$

$i = 1$ on bottom tells us to *start* with $i = 1$.
4 on top tells us to *finish* with $i = 4$.
Implicitly know that you increment by 1.

For each number i that you count,

$i = 1 : i^2 = 1^2$
$i = 2 : i^2 = 2^2$
$i = 3 : i^2 = 3^2$
$i = 4 : i^2 = 4^2$

then the Σ tells you to *sum* the results.

3 Second example

Example:

$$\sum_{i=1}^{5}(2i+3) = (2(1)+3)+(2(2)+3)+(2(3)+3)+(2(4)+3)+(2(5)+3)$$
$$= 45$$

Work for problem:

$$i=1 : 2i+3 = 2(1)+3$$
$$i=2 : 2i+3 = 2(2)+3$$
$$i=3 : 2i+3 = 2(3)+3$$
$$i=4 : 2i+3 = 2(4)+3$$
$$i=5 : 2i+3 = 2(5)+3$$

4 Third example

Example:

$$\sum_{j=3}^{7}\frac{j}{2} = \frac{3}{2}+\frac{4}{2}+\frac{5}{2}+\frac{6}{2}+\frac{7}{2} = \frac{25}{2}$$
$$\sum_{r=3}^{7}\frac{r}{2} = \frac{25}{2}$$

j and r are "dummy indices," symbols for counters.

$$\sum_{☺=3}^{7}\frac{☺}{2} = \frac{25}{2}$$

Common choices for indices:
i, j, k, l, r, m, n

Paul Bendich and Daniel Egger
Duke University

Sigma Notation: Simplification Rules

Video companion

1 Distributive property

Examples:

$$\sum_{i=1}^{4} i^2 = 30$$

$$\begin{aligned}\sum_{i=1}^{4} 3i^2 &= 3(1)^2 + 3(2)^2 + 3(3)^2 + 3(4)^2 \\ &= 3[1^2 + 2^2 + 3^2 + 4^2] \\ &= 3\left[\sum_{i=1}^{4} i^2\right]\end{aligned}$$

$$\sum_{r=4}^{25} 18r^3 = 18\left[\sum_{r=4}^{25} r^3\right]$$

This is due to the *distributive property*:

$$a(b+c) = ab + ac$$

In other words, constants inside the summed expression can be pulled outside.

2 Commutative property

$$\begin{aligned}\sum_{i=1}^{4}(i^2+2i) &= \left(1^2+2(1)\right)+\left(2^2+2(2)\right)+\left(3^2+2(3)\right)+\left(4^2+2(4)\right)\\ &= \left(1^2+2^2+3^2+4^2\right)+\left(2(1)+2(2)+2(3)+2(4)\right)\\ &= \left(\sum_{i=1}^{4} i^2\right)+\left(\sum_{i=1}^{4} 2i\right)\end{aligned}$$

This is due to the *commutative property*:

$$a+b=b+a$$

In other words, we can add the terms in any order.

3 Summation of constants

Examples:

$$\begin{aligned}\sum_{k=1}^{10} 5 &= 5+5+5+5+5+5+5+5+5+5\\ &= 10\cdot 5\\ &= 50\end{aligned}$$

$$\begin{aligned}\sum_{r=1}^{7} 8 &= 8+8+8+8+8+8+8\\ &= 7\cdot 8\\ &= 56\end{aligned}$$

When summing constants, you can multiply the constant by the number of indices you count.

Data Science Math Skills
Paul Bendich and Daniel Egger
Duke University

Sigma Notation: Mean and Variance

Video companion

1 Introduction

Important equations for this video:

$$X = \{x_1, ..., x_n\}$$
$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$${\sigma_x}^2 = \frac{1}{n}\left[\sum_{i=1}^{n} (x_i - \mu_x)^2\right]$$

The symbol μ_x is the "mean of x," and ${\sigma_x}^2$ is the "variance of x." The standard deviation is denoted σ_x.

2 Mean

Example:

$$Z = \{1, 5, 12\}$$
$$|Z| = 3$$
$$\mu_z = \frac{1+5+12}{3} = \frac{18}{3} = 6$$

The mean μ_z is also denoted $\mu(z)$ or simply μ.

Symbolic example:

$$Y = \{y_1, y_2, y_3, y_4\}$$
$$\mu_y = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$$
$$= \frac{1}{4}\left(\sum_{i=1}^{4} y_i\right)$$

In general, suppose you have a set

$$X = \{x_1, x_2, ..., x_n\},$$

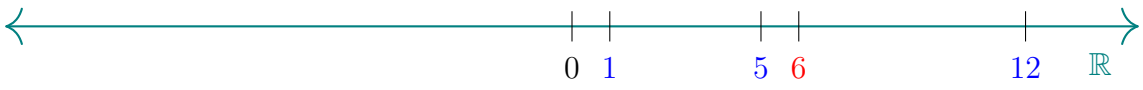
then the mean of X is

$$\mu_x = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right).$$

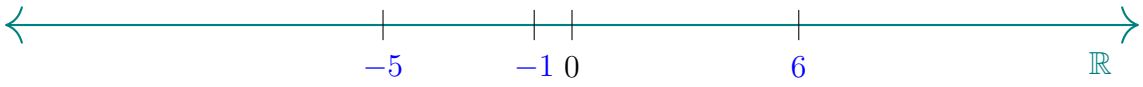
The variable i is a counter. The variable n is a number, which tells you when to stop counting.

3 Mean centering

$$Z = \{1, 5, 12\}$$
$$\mu_z = 6$$



$$Z' = \{1-6, 5-6, 12-6\}$$
$$= \{-5, -1, 6\}$$
$$\mu_{z'} = 0$$



Mean centering data produces a new data set, which has the same relationships, but the mean is zero.

4 Variance

$$Z = \{1, 5, 12\}$$
$$\mu_z = 6$$

$$W = \{5, 6, 7\}$$
$$\mu_w = 6$$

Set Z (blue) is more "spread out" than set W (olive).

If $X = \{x_1, ..., x_n\}$, the variance of X is

$$\sigma_x{}^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \mu_x)^2\right].$$

The standard deviation is given by

$$\sigma_x = \sqrt{\sigma_x{}^2}.$$

Z and W have the same mean, but Z is more spread out, so σ_z should be greater than σ_w.

$$\begin{aligned}
\sigma_w^2 &= \frac{1}{3}\left[\sum_{i=1}^{3}(w_i - \mu_w)^2\right] \\
&= \frac{1}{3}\left[(5-6)^2 + (6-6)^2 + (7-6)^2\right] \\
&= \frac{1}{3}\left[(-1)^2 + 0^2 + 1^2\right] \\
&= \frac{2}{3} \\
\sigma_w &= \sqrt{\frac{2}{3}}
\end{aligned}$$

$$\begin{aligned}
\sigma_z^2 &= \frac{1}{3}\left[\sum_{i=1}^{3}(z_i - \mu_z)^2\right] \\
&= \frac{1}{3}\left[(1-6)^2 + (5-6)^2 + (12-6)^2\right] \\
&= \frac{1}{3}\left[(-5)^2 + (-1)^2 + 6^2\right] \\
&= \frac{62}{3} \\
\sigma_w &= \sqrt{\frac{62}{3}}
\end{aligned}$$

$\sigma_z^2 \gg \sigma_w^2$, so Z is much more spread out than W.