

Convexity, Infimum and Sums

Last modified: 2026-10-27 20:53:54 (PDT)

Convex functions

Exercise 0.1 (A value compared with an average of values). Let $f(x) = (x - 2)^2$, and take the points $x = 0$ and $y = 4$ with $t = \frac{1}{2}$.

1. Compute $f(tx + (1 - t)y)$.
2. Compute $tf(x) + (1 - t)f(y)$.
3. Is $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$?

Solution

Solution 0.1.

1. With $t = \frac{1}{2}$,

$$\begin{aligned} tx + (1 - t)y &= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 4 \\ &= 2, \end{aligned}$$

so

$$\begin{aligned} f(2) &= (2 - 2)^2 \\ &= 0. \end{aligned}$$

- 2.

$$\begin{aligned} f(0) &= (0 - 2)^2 \\ &= 4 \end{aligned}$$

and

$$\begin{aligned} f(4) &= (4 - 2)^2 \\ &= 4, \end{aligned}$$

so $\frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 4 = 4$.

3. Yes: $0 \leq 4$. The value of f at the midpoint of 0 and 4 is below the average of its values at the two endpoints.

Definition 0.1 (Line segment). Let $x, y \in \mathbb{R}^p$. The **line segment** from x to y [1] is the set of points

$$\{tx + (1 - t)y : t \in [0, 1]\}.$$

The value $t = 1$ gives x , the value $t = 0$ gives y , and $t = \frac{1}{2}$ gives the midpoint $\frac{1}{2}(x + y)$.

Example 0.1 (A line segment in $\mathbb{R}$). For $p = 1$, $x = 0$ and $y = 4$, the line segment from x to y is the **interval** $[0, 4]$. The value $t = \frac{1}{4}$ gives the point $\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 4 = 3$.

Definition 0.2 (Chord). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be a **function**, and let $x, y \in \mathbb{R}^p$. The **chord** of f from x to y [2] is the line segment (Definition 0.1) in $\mathbb{R}^{p+1}$ from the point $(x, f(x))$ to the point $(y, f(y))$ of the **graph** of f :

$$\{(tx + (1 - t)y, tf(x) + (1 - t)f(y)) : t \in [0, 1]\}.$$

Above the point $tx + (1 - t)y$, the chord has height $tf(x) + (1 - t)f(y)$.

Example 0.2 (A chord of x^2). For $f(x) = x^2$, $x = -1$ and $y = 3$, the chord runs from $(-1, f(-1)) = (-1, 1)$ to $(3, f(3)) = (3, 9)$. With $t = \frac{1}{2}$, it passes through $(\frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 3, \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 9) = (1, 5)$, so above the point 1 it has height 5.

Definition 0.3 (Convex function). Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be a **function**. f is **convex** [3] if

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

for all $x, y \in \mathbb{R}^p$ and all $t \in [0, 1]$.

i Remark (Chords lie on or above the graph)

Remark 0.1 (Chords lie on or above the graph). The point $tx + (1 - t)y$ lies on the **line segment** from x to y , and the right-hand side is the height above that point of the **chord** of f from x to y . So f is convex when every chord lies on or above the **graph**.

For example, take $f(x) = x^2$, $x = -1$, $y = 3$, and $t = \frac{1}{2}$. The point is $\frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 3 = 1$,

where the graph has height $f(1) = 1$ and the chord has height

$$\begin{aligned}\frac{1}{2}f(-1) + \frac{1}{2}f(3) &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 9 \\ &= 5.\end{aligned}$$

The chord is above the graph: $1 \leq 5$.

Example 0.3 (A function that is not convex). The function in [Exercise 2 in Equalities, Inequalities and Minimizers](#), $f(x) = x^3 - 3x$, is not convex. Take $x = -2$, $y = 0$, $t = \frac{1}{2}$. The point is $\frac{1}{2} \cdot (-2) + \frac{1}{2} \cdot 0 = -1$, where the graph has height

$$\begin{aligned}f(-1) &= -1 + 3 \\ &= 2,\end{aligned}$$

but the chord has height

$$\begin{aligned}\frac{1}{2}f(-2) + \frac{1}{2}f(0) &= \frac{1}{2}(-8 + 6) + \frac{1}{2} \cdot 0 \\ &= -1,\end{aligned}$$

and $2 \leq -1$ is false.

Example 0.4 (A convex and a non-convex function). $f(x) = (x - 2)^2$ is convex (Definition 0.3): by expanding the square, $f(tx + (1-t)y) - tf(x) - (1-t)f(y) = -t(1-t)(x-y)^2 \leq 0$. Its local minimizer $x^* = 2$ is also a global minimizer, since

$$\begin{aligned}f(x) &= (x - 2)^2 \\ &\geq 0 \\ &= f(2)\end{aligned}$$

for every x .

Without convexity, a local minimizer need not be global: $f(x) = x^3 - 3x$ is not convex (Example 0.3), and it has a local minimizer at $x^* = 1$ that is not global ([Exercise 2 in Equalities, Inequalities and Minimizers](#)).

Theorem 0.1 (Local minimizers of convex functions are global). *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be a convex function. Every local minimizer of f is a global minimizer of f over $\mathbb{R}^p$.*

i Proof

Proof. Let x^* be a local minimizer of f , so there is a $\delta > 0$ with $f(x^*) \leq f(x)$ whenever $\|x - x^*\| < \delta$ (Definition 20 in [Equalities, Inequalities and Minimizers](#)). Suppose x^* is not

a global minimizer (Definition 19 in Equalities, Inequalities and Minimizers). Then some $y \in \mathbb{R}^p$ has $f(y) < f(x^*)$, and in particular $y \neq x^*$.

Let $t = \min \left\{ \frac{1}{2}, \frac{\delta}{2\|y - x^*\|} \right\}$, so $t \in (0, 1)$, and let $z = ty + (1-t)x^*$. Then $z - x^* = t(y - x^*)$, so $\|z - x^*\| = t\|y - x^*\| \leq \delta/2 < \delta$.

By convexity,

$$f(z) \leq tf(y) + (1-t)f(x^*) < tf(x^*) + (1-t)f(x^*) = f(x^*),$$

where the strict inequality uses $t > 0$ and $f(y) < f(x^*)$. So $f(z) < f(x^*)$ with $\|z - x^*\| < \delta$, which contradicts x^* being a local minimizer. Hence x^* is a global minimizer. $\square$

Theorem 0.2 (Jensen's inequality for a weighted average). *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be a [convex function](#) [4]. Let $x_1, \dots, x_n \in \mathbb{R}^p$, and let $w_1, \dots, w_n \geq 0$ be weights with $\sum_{i=1}^n w_i = 1$. Then*

$$f\left(\sum_{i=1}^n w_i x_i\right) \leq \sum_{i=1}^n w_i f(x_i).$$

Proof

Proof. The proof is by induction on n .

For $n = 1$ the weight is $w_1 = 1$, and both sides equal $f(x_1)$.

Now let $n \geq 2$, and assume the inequality holds for $n - 1$ points. If $w_n = 1$, the other weights are all 0, and both sides equal $f(x_n)$. Otherwise let

$$\begin{aligned} s &= 1 - w_n \\ &= \sum_{i=1}^{n-1} w_i \\ &> 0 \end{aligned}$$

and

$$y = \sum_{i=1}^{n-1} \frac{w_i}{s} x_i.$$

The weights w_i/s are non-negative and sum to 1, so the induction assumption gives

$$f(y) \leq \sum_{i=1}^{n-1} \frac{w_i}{s} f(x_i).$$

Since $\sum_{i=1}^n w_i x_i = sy + (1-s)x_n$ and $s \in (0, 1]$, the definition of a convex function

(Definition 0.3) gives

$$f\left(\sum_{i=1}^n w_i x_i\right) = f(sy + (1-s)x_n) \leq sf(y) + (1-s)f(x_n).$$

Substituting the bound on $f(y)$, and using $1-s = w_n$,

$$sf(y) + (1-s)f(x_n) \leq \sum_{i=1}^{n-1} w_i f(x_i) + w_n f(x_n) = \sum_{i=1}^n w_i f(x_i).$$

□

Example 0.5 (The average of the squares is at least the square of the average). Take $f(x) = x^2$. It is convex by Definition 0.3: for $t \in [0, 1]$,

$$tx^2 + (1-t)y^2 - (tx + (1-t)y)^2 = t(1-t)(x-y)^2 \geq 0.$$

Take the three points $x_1 = 1$, $x_2 = 2$, $x_3 = 6$, and the equal weights

$$\begin{aligned} w_1 &= w_2 \\ &= w_3 \\ &= \frac{1}{3}. \end{aligned}$$

The weighted average of the points is their ordinary average, $\frac{1}{3}(1+2+6) = 3$, so the left side of Theorem 0.2 is $f(3) = 9$. The right side is the average of the squares, $\frac{1}{3}(1+4+36) = \frac{41}{3}$. The inequality holds: $9 \leq \frac{41}{3}$.

The gap is $\frac{41}{3} - 9 = \frac{14}{3}$. That is the average squared distance of the points from their average, $\frac{1}{3}((1-3)^2 + (2-3)^2 + (6-3)^2) = \frac{14}{3}$, which statistics calls the variance of the three points (about their [mean](#), defined later). For $f(x) = x^2$, Jensen's inequality says this average squared distance is never negative.

i Remark (The inequality reverses for concave functions)

Remark 0.2 (The inequality reverses for concave functions). If f is [concave](#), then $-f$ is convex, and applying Theorem 0.2 to $-f$ reverses the inequality: $f\left(\sum_i w_i x_i\right) \geq \sum_i w_i f(x_i)$. For $f(x) = -x^2$ and the points in Example 0.5, $f(3) = -9 \geq -\frac{41}{3}$.

Infimum and supremum

Definition 0.4 (Upper bound). Let $A \subseteq \mathbb{R}$. A real number u is an **upper bound** for A [5] if $a \leq u$ for all $a \in A$.

Definition 0.5 (Lower bound). Let $A \subseteq \mathbb{R}$. A real number ℓ is a **lower bound** for A [5] if $\ell \leq a$ for all $a \in A$.

Definition 0.6 (Bounded above). A set $A \subseteq \mathbb{R}$ is **bounded above** if it has an upper bound (Definition 0.4). A real-valued **function** is bounded above if its **image** is.

Definition 0.7 (Bounded below). A set $A \subseteq \mathbb{R}$ is **bounded below** if it has a lower bound (Definition 0.5). A real-valued **function** is bounded below if its **image** is.

Definition 0.8 (Bounded). A set $A \subseteq \mathbb{R}$ is **bounded** if it is both bounded above (Definition 0.6) and bounded below (Definition 0.7). A real-valued **function** is bounded if its **image** is.

Example 0.6 (Bounded and unbounded sets).

- For $A = (0, 1]$, 1, 2, and 100 are upper bounds, and 0 and -5 are lower bounds, so A is bounded. 0.5 is not an upper bound, because $0.6 \in A$ and $0.6 > 0.5$.
- The natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ are bounded below, by 1, but not bounded above: for any real number u , some natural number n satisfies $n > u$.

Theorem 0.3 (Completeness of the real numbers). *Let $A \subseteq \mathbb{R}$ be nonempty [6].*

- *If A is **bounded above**, the set of upper bounds of A has a **minimum**.*
- *If A is **bounded below**, the set of lower bounds of A has a **maximum**.*

Proof

Proof. The first statement is the least-upper-bound property of $\mathbb{R}$ [7], Definition 1.8, p. 4, and Theorem 1.19, p. 8. For the second, let $-A = \{-a : a \in A\}$. Negating both sides of an inequality reverses it (**Theorem 4 in Equalities, Inequalities and Minimizers**), so ℓ is a lower bound of A exactly when $-\ell$ is an upper bound of $-A$. The set $-A$ is nonempty and bounded above, so by the first statement its upper bounds have a minimum u , and then $-u$ is the largest lower bound of A . $\square$

Example 0.7 (Completeness for an interval). For $A = (1, 2]$, the lower bounds of A are the numbers $\ell \leq 1$, and the largest of them is 1. The upper bounds of A are the numbers $u \geq 2$, and the smallest of them is 2.

Definition 0.9 (Infimum, greatest lower bound). Let $A \subseteq \mathbb{R}$ be nonempty and bounded below (Definition 0.7). The **infimum** of A [8], written $\inf A$, is the greatest real number t satisfying $t \leq a$ for all $a \in A$:

$$\inf A \stackrel{\text{def}}{=} \max\{t \in \mathbb{R} : \forall a \in A, t \leq a\}$$

If A is nonempty but not bounded below, we write $\inf A = -\infty$ by convention.

i Remark (Existence of the infimum, and when it is a minimum)

Remark 0.3 (Existence of the infimum, and when it is a minimum). The maximum in Definition 0.9 always exists, by the completeness of the real numbers (Theorem 0.3). For example, for $A = (1, 2]$, the numbers t with $t \leq a$ for all $a \in A$ are those with $t \leq 1$, and the largest of them is 1, so $\inf A = 1$.

If the infimum belongs to A , it equals the minimum: $\inf A = \min A$. For example,

$$\begin{aligned}\inf[1, 2] &= 1 \\ &= \min[1, 2].\end{aligned}$$

For $A = (1, 2]$, the infimum 1 is not in A , and A has no minimum.

Example 0.8 (Numerical examples of infimum).

- $\inf\{1, 2, 3\} = 1$, since 1 is the smallest element.

-

$$\begin{aligned}\inf(0.5, 1] &= 0.5 \\ &= \min[0.5, 1] : \end{aligned}$$

for intervals open below, the infimum equals the minimum of the corresponding closed-below interval, even though $0.5 \notin (0.5, 1]$. More generally,

$$\begin{aligned}\inf(c, b] &= \min[c, b] \\ &= c\end{aligned}$$

for any $c < b$.

- $\inf\{t \geq 0 : t > 0.5\} = 0.5$, even though 0.5 itself is not in the set.

- $\inf\{-1, -2, -3, \dots\} = -\infty$, because no real number is less than or equal to every element of that set.

Definition 0.10 (Supremum, least upper bound). Let $A \subseteq \mathbb{R}$ be nonempty and bounded above (Definition 0.6). The **supremum** of A [8], written $\sup A$, is the smallest real number t satisfying $a \leq t$ for all $a \in A$:

$$\sup A \stackrel{\text{def}}{=} \min\{t \in \mathbb{R} : \forall a \in A, a \leq t\}$$

If A is nonempty but not bounded above, we write $\sup A = +\infty$ by convention.

i Remark (Existence of the supremum, and when it is a maximum)

Remark 0.4 (Existence of the supremum, and when it is a maximum). The minimum in Definition 0.10 always exists, by the completeness of the real numbers (Theorem 0.3). For example, for $A = [1, 2)$, the numbers t with $a \leq t$ for all $a \in A$ are those with $t \geq 2$, and the smallest of them is 2, so $\sup A = 2$.

If the supremum belongs to A , it equals the maximum: $\sup A = \max A$. For example,

$$\begin{aligned} \sup[1, 2] &= 2 \\ &= \max[1, 2]. \end{aligned}$$

For $A = [1, 2)$, the supremum 2 is not in A , and A has no maximum.

Example 0.9 (Numerical examples of supremum).

- $\sup\{1, 2, 3\} = 3$, since 3 is the largest element.
- $\sup\{t \geq 0 : t < 0.5\} = 0.5$, even though 0.5 itself is not in the set.
- $\sup\{1, 2, 3, \dots\} = +\infty$, because no real number is greater than or equal to every element of that set.

Sums

Definition 0.11 (Term of a sum). In a sum $a_1 + a_2 + \dots + a_n$, each of the numbers or expressions $a_1, a_2, \dots, a_n$ being added is a **term** of the sum.

Example 0.10 (Terms of a sum).

- The sum $3 + 5 + 9$ has three terms: 3, 5 and 9.
- The sum $x^2 - 3x + 7 = x^2 + (-3x) + 7$ has three terms: x^2 , $-3x$ and 7. A subtracted

expression counts as a term with a minus sign.

Definition 0.12 (Identity element). Let S be a set, such as the real numbers $\mathbb{R}$, and let $\star$ be an operation that combines two elements a and b of S into an element $a \star b$ of S , such as addition ($a + b$) or multiplication ($a \times b$) of real numbers. An element u of S is an **identity element** for $\star$ [9] if $a \star u = a$ and $u \star a = a$ for every $a \in S$.

Example 0.11 (Identity elements for addition and multiplication).

- 0 is the identity element for addition: for example, $5 + 0 = 5$ and $0 + 5 = 5$ (Theorem 0.4).
- 1 is the identity element for multiplication: for example, $5 \times 1 = 5$ and $1 \times 5 = 5$ (Theorem 0.7).
- 0 is not an identity element for subtraction: $5 - 0 = 5$, but $0 - 5 = -5 \neq 5$.

Definition 0.13 (Commutative operation). An operation $\star$ on a set S (Definition 0.12) is **commutative** [10] if $a \star b = b \star a$ for all $a, b \in S$: the order of the two inputs does not matter. Some sources call a commutative operation **symmetric**, as in the theorem titles “Sums are symmetric” and “Products are symmetric”.

Example 0.12 (A commutative operation, and one that is not).

- Addition is commutative: for example,

$$\begin{aligned} 2 + 5 &= 7 \\ &= 5 + 2. \end{aligned}$$

- Subtraction is not commutative: $5 - 3 = 2$, but $3 - 5 = -2$.

Definition 0.14 (Associative operation). An operation $\star$ on a set S (Definition 0.12) is **associative** [11] if $(a \star b) \star c = a \star (b \star c)$ for all $a, b, c \in S$: which pair is combined first does not matter.

Example 0.13 (An associative operation, and one that is not).

- Multiplication is associative: for example,

$$\begin{aligned}(2 \times 3) \times 4 &= 6 \times 4 \\ &= 24\end{aligned}$$

and

$$\begin{aligned}2 \times (3 \times 4) &= 2 \times 12 \\ &= 24.\end{aligned}$$

- Subtraction is not associative:

$$\begin{aligned}(8 - 4) - 2 &= 4 - 2 \\ &= 2,\end{aligned}$$

but

$$\begin{aligned}8 - (4 - 2) &= 8 - 2 \\ &= 6.\end{aligned}$$

Theorem 0.4 (Adding zero changes nothing).

$$a + 0 = a$$

Theorem 0.5 (Sums are symmetric).

$$a + b = b + a$$

Theorem 0.6 (Sums are associative). *When adding three numbers, it does not matter which pair you add first:*

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

Example 0.14 (Grouping a sum two ways).

$$\begin{aligned}(2 + 3) + 4 &= 5 + 4 \\ &= 9,\end{aligned}$$

and

$$\begin{aligned}2 + (3 + 4) &= 2 + 7 \\ &= 9.\end{aligned}$$

Products

Theorem 0.7 (Multiplying by 1 changes nothing).

$$a \times 1 = a$$

Theorem 0.8 (Products are symmetric).

$$a \times b = b \times a$$

Theorem 0.9 (Products are associative).

$$(a \times b) \times c = a \times (b \times c)$$

Division

Theorem 0.10 (Division can be written as a product). *If $b \neq 0$, then*

$$\frac{a}{b} = a \times \frac{1}{b}$$

Sums and products together

Definition 0.15 (Distributes over). An operation $\star$ on a set S **distributes over** an operation $\diamond$ on S if

$$a \star (b \diamond c) = (a \star b) \diamond (a \star c)$$

for all $a, b, c \in S$.

Example 0.15 (Addition does not distribute over multiplication).

$$\begin{aligned} 2 + (3 \times 4) &= 2 + 12 \\ &= 14, \end{aligned}$$

but

$$\begin{aligned} (2 + 3) \times (2 + 4) &= 5 \times 6 \\ &= 30. \end{aligned}$$

So Definition 0.15 fails with $\star = +$ and $\diamond = \times$.

Theorem 0.11 (Multiplication is distributive).

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

Definition 0.16 (Distributive law). The **distributive law** [12] is the statement that multiplication distributes over addition (Definition 0.15): $a \times (b + c) = (a \times b) + (a \times c)$ (Theorem 0.11). We then say multiplication is **distributive**.

Example 0.16 (Multiplication distributes over addition).

$$\begin{aligned} 3 \times (4 + 5) &= 3 \times 9 \\ &= 27, \end{aligned}$$

and

$$\begin{aligned} (3 \times 4) + (3 \times 5) &= 12 + 15 \\ &= 27. \end{aligned}$$

Definition 0.17 (Like terms). Two **terms** of a sum are **like terms** [13] if they are the same product of variables, each to the same power, possibly multiplied by different constants: c_1m and c_2m , where c_1 and c_2 are constants and m is that product.

Definition 0.18 (Collecting like terms). To **collect** like terms (Definition 0.17) is to replace their sum by one term, using the distributive law (Theorem 0.11): $c_1m + c_2m = (c_1 + c_2)m$.

Example 0.17 (Like and unlike terms).

- $3ab$ and $-5ab$ are like terms; collecting them gives

$$\begin{aligned} 3ab + (-5ab) &= (3 - 5)ab \\ &= -2ab. \end{aligned}$$

- $2x$ and $2x^2$ are not like terms: x appears to different powers.

Exercise 0.2 (Expand a squared sum). Is $(3+4)^2$ equal to 3^2+4^2 ? Then expand $(a+b)^2$ for any numbers a and b , using only the distributive law and the rules above.

Solution

Solution 0.2. No:

$$\begin{aligned}(3+4)^2 &= 7^2 \\ &= 49,\end{aligned}$$

while

$$\begin{aligned}3^2+4^2 &= 9+16 \\ &= 25.\end{aligned}$$

The difference, $49 - 25 = 24$, is $2 \cdot 3 \cdot 4$.

To see where that extra term comes from, write the square as a product and apply the distributive law (Theorem 0.11) twice:

$$\begin{aligned}(a+b)^2 &= (a+b)(a+b) && \text{(definition of a square)} \\ &= (a+b)a + (a+b)b && \text{(distributive law)} \\ &= (a^2+ba) + (ab+b^2) && \text{(distributive law, twice)} \\ &= a^2+ab+ab+b^2 && \text{(commutative and associative laws)} \\ &= a^2+2ab+b^2 && \text{(collect like terms, Definition 0.18)}\end{aligned}$$

The step “commutative and associative laws” uses Theorem 0.8 to write ba as ab , and Theorem 0.6 to drop the parentheses.

Theorem 0.12 (Product of two sums (FOIL)). *For any numbers a , b , c and d ,*

$$(a+b)(c+d) = ac + ad + bc + bd.$$

*The four products pair the **F**irst terms (ac), the **O**uter terms (ad), the **I**nnner terms (bc) and the **L**ast terms (bd), which gives this rule its name, **FOIL**.*

i Proof

Proof.

$$\begin{aligned}(a+b)(c+d) &= (a+b)c + (a+b)d && \text{(distributive law, Theorem 0.11)} \\ &= (ac+bc) + (ad+bd) && \text{(distributive law, twice)} \\ &= ac+ad+bc+bd && \text{(commutative and associative laws)}\end{aligned}$$

□

Example 0.18 (Checking FOIL). With $a = 1$, $b = 2$, $c = 3$ and $d = 4$,

$$\begin{aligned}(1 + 2)(3 + 4) &= 3 \cdot 7 \\ &= 21,\end{aligned}$$

and

$$\begin{aligned}1 \cdot 3 + 1 \cdot 4 + 2 \cdot 3 + 2 \cdot 4 &= 3 + 4 + 6 + 8 \\ &= 21.\end{aligned}$$

Setting $c = a$ and $d = b$ gives $(a + b)^2 = a^2 + ab + ba + b^2$, which is the [square of a sum](#) once ab and ba are combined.

Theorem 0.13 (Square of a sum). *For any numbers a and b ,*

$$(a + b)^2 = a^2 + 2ab + b^2$$

i Proof

Proof. By [Solution 0.2](#). □

i Remark (Square of a difference)

Remark 0.5 (Square of a difference). Replacing b by $-b$ in [Theorem 0.13](#) gives $(a - b)^2 = a^2 - 2ab + b^2$. For example, the square $(y - \hat{y})^2$ of the difference between an observed value y and a prediction $\hat{y}$ of it expands this way. With $y = 5$ and $\hat{y} = 3$,

$$\begin{aligned}(5 - 3)^2 &= 2^2 \\ &= 4,\end{aligned}$$

and

$$\begin{aligned}5^2 - 2 \cdot 5 \cdot 3 + 3^2 &= 25 - 30 + 9 \\ &= 4.\end{aligned}$$

Summation notation

Exercise 0.3 (Expand a sum). The expression $\sum_{i=1}^4 i^2$ is shorthand for a sum of four terms.

1. Guess which four terms, and add them up.
2. Write $\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$ for $N = 3$ without the $\sum$ symbol.

Solution

Solution 0.3.

1. Replace i by each of 1, 2, 3, 4 in turn, and add the results:

$$\begin{aligned}\sum_{i=1}^4 i^2 &= 1^2 + 2^2 + 3^2 + 4^2 \\ &= 1 + 4 + 9 + 16 \\ &= 30\end{aligned}$$

2. Replace i by each of 1, 2, 3 in turn inside the parentheses, add the three squares, and multiply the total by $\frac{1}{3}$:

$$\frac{1}{3} \sum_{i=1}^3 (y_i - \hat{y}_i)^2 = \frac{1}{3} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2]$$

Definition 0.19 (Summation notation). Let m and n be integers with $m \leq n$, and let $a_m, a_{m+1}, \dots, a_n$ be numbers. The **sum** of a_m through a_n is

$$\sum_{i=m}^n a_i \stackrel{\text{def}}{=} a_m + a_{m+1} + \dots + a_n$$

The numbers m and n are its lower and upper limits: the first and last values of the index.

Definition 0.20 (Index of a sum). The variable i in $\sum_{i=m}^n a_i$ (Definition 0.19) is the **index** of the sum.

i Remark (The index is a placeholder)

Remark 0.6 (The index is a placeholder). The name of the index does not change the sum: $\sum_{i=1}^n a_i$ and $\sum_{j=1}^n a_j$ are the same number. For example,

$$\begin{aligned}\sum_{i=1}^3 i &= 1 + 2 + 3 \\ &= 6\end{aligned}$$

and

$$\begin{aligned}\sum_{j=1}^3 j &= 1 + 2 + 3 \\ &= 6.\end{aligned}$$

Definition 0.21 (Empty sum). When the upper limit is less than the lower limit ($n < m$), the sum $\sum_{i=m}^n a_i$ has no terms. By convention, such an **empty sum** equals 0. For example, $\sum_{i=1}^0 a_i = 0$.

Exercise 0.4 (Rearrange a sum). Let c be a number, and let a_1, a_2, a_3 and b_1, b_2, b_3 be numbers. Using Definition 0.19 and the rules of algebra above, show that:

1. $\sum_{i=1}^3 c a_i = c \sum_{i=1}^3 a_i$;
2. $\sum_{i=1}^3 (a_i + b_i) = \sum_{i=1}^3 a_i + \sum_{i=1}^3 b_i$.

Does either argument depend on there being exactly three terms?

Solution

Solution 0.4. Each step below applies Definition 0.19, the distributive law (Theorem 0.11), or the commutative and associative laws of addition (Theorem 0.5 and Theorem 0.6).

1. Expand the sum, then factor out c :

$$\begin{aligned}\sum_{i=1}^3 c a_i &= c a_1 + c a_2 + c a_3 && \text{(expand the sum)} \\ &= c(a_1 + a_2 + a_3) && \text{(distributive law)} \\ &= c \sum_{i=1}^3 a_i && \text{(collect the sum)}\end{aligned}$$

2. Expand the sum, then regroup the terms:

$$\begin{aligned}\sum_{i=1}^3 (a_i + b_i) &= (a_1 + b_1) + (a_2 + b_2) + (a_3 + b_3) && \text{(expand the sum)} \\ &= (a_1 + a_2 + a_3) + (b_1 + b_2 + b_3) && \text{(commutative and associative laws)} \\ &= \sum_{i=1}^3 a_i + \sum_{i=1}^3 b_i && \text{(collect the sum)}\end{aligned}$$

Neither argument uses the number of terms: the same steps work for any lower and upper limits.

Theorem 0.14 (A constant factor comes out of a sum). *For any number c and numbers $a_m, \dots, a_n$,*

$$\sum_{i=m}^n c a_i = c \sum_{i=m}^n a_i$$

i Proof

Proof. By Solution 0.4, part 1. □

Theorem 0.15 (A sum of sums splits). *For any numbers $a_m, \dots, a_n$ and $b_m, \dots, b_n$,*

$$\sum_{i=m}^n (a_i + b_i) = \sum_{i=m}^n a_i + \sum_{i=m}^n b_i$$

i Proof

Proof. By Solution 0.4, part 2. □

Exercise 0.5 (Sum over a set). Let $A = \{-1, 2, 5\}$ and let $f(x) = x^2$. The expression $\sum_{x \in A} f(x)$ means: add up $f(x)$ for each element x of A . Write $\sum_{x \in A} f(x)$ without the $\sum$ symbol.

Solution

Solution 0.5. Replace x by each element of A in turn, and add:

$$\sum_{x \in A} f(x) = f(-1) + f(2) + f(5)$$

Exercise 0.6 (Value of a sum over a set). In the setting of Exercise 0.5, find the value of $\sum_{x \in A} f(x)$.

Solution

Solution 0.6. Evaluate each term and add:

$$\begin{aligned} f(-1) + f(2) + f(5) &= (-1)^2 + 2^2 + 5^2 \\ &= 1 + 4 + 25 \\ &= 30 \end{aligned}$$

Exercise 0.7 (List the terms in a different order). In the setting of Exercise 0.5, list the elements of A in a different order, and write the term for each element in that order.

Solution

Solution 0.7. In the order $5, -1, 2$, the terms are 25, 1 and 4.

Exercise 0.8 (Add the terms in the new order). In the setting of Exercise 0.7, add up the terms.

Solution

Solution 0.8. The total is $25 + 1 + 4 = 30$.

Exercise 0.9 (Does the order change the total?). Say whether the total in Exercise 0.8 is the same as the total in Exercise 0.6.

Solution

Solution 0.9. The total is the same, because addition does not depend on the order of the terms.

Definition 0.22 (Sum over a finite set). Let $A = \{x_1, x_2, \dots, x_k\}$ be a **finite set** with $k \geq 1$ different elements. Let f be a **function** that gives a number $f(x)$ for each element x of A . The **sum of f over A** is

$$\sum_{x \in A} f(x) \stackrel{\text{def}}{=} f(x_1) + f(x_2) + \dots + f(x_k)$$

If A has no elements, the sum has no terms, and by convention it equals 0.

i Remark (The order of the terms does not matter)

Remark 0.7 (The order of the terms does not matter). The order in which we list the elements of A does not change the sum. The reason is that addition is commutative and associative: we can reorder and regroup the terms of a finite sum without changing the total. For example, if $A = \{1, 2, 3\}$ and $f(x) = x^2$, listing A as 1, 2, 3 gives $1 + 4 + 9 = 14$, and listing A as 3, 1, 2 gives $9 + 1 + 4 = 14$. Exercise 0.6 and Exercise 0.8 show another example.

When $A = \{m, m+1, \dots, n\}$, this sum is the same as $\sum_{i=m}^n f(i)$ from Definition 0.19.

i Remark (Leaving the set out)

Remark 0.8 (Leaving the set out). Some authors leave the set out and write $\sum_x f(x)$. This shorthand means the sum over every value x can take, $\sum_{x \in \mathcal{R}(x)} f(x)$, where $\mathcal{R}(x)$ is the **range** of x . For example, if x is the outcome of one roll of a six-sided die, $\sum_x f(x)$ means $\sum_{x \in \{1, 2, 3, 4, 5, 6\}} f(x)$. In these notes, we write the set out in full (see **Limits of sums and notational shorthands**).

Example 0.19 (A sum over the empty set). For example, $\sum_{x \in \{\}} x^2 = 0$.

Quotients

Definition 0.23 (Quotient). For real numbers a and b with $b \neq 0$, the **quotient** of a by b is the result of dividing a by b :

$$\frac{a}{b}$$

Definition 0.24 (Fraction). A quotient $\frac{a}{b}$ (Definition 0.23) is also called a **fraction**.

Definition 0.25 (Numerator). In a fraction $\frac{a}{b}$ (Definition 0.24), a is its **numerator**.

Definition 0.26 (Denominator). In a fraction $\frac{a}{b}$ (Definition 0.24), b is its **denominator**.

Example 0.20 (Numerator and denominator of a fraction). The fraction $\frac{6}{4}$ has numerator 6 and denominator 4.

Example 0.21 (A quotient). The quotient of 6 by 4 is $\frac{6}{4} = 1.5$. The quotient of 6 by 0 is undefined, because Definition 0.23 requires a nonzero denominator.

Definition 0.27 (Follow-up). In a study that observes participants over time, a participant's **follow-up** is the length of time the study observes that participant.

Definition 0.28 (Person-years). In a study that observes participants over time, the study's **person-years** are the sum of the follow-up times (Definition 0.27) of all participants, measured in years.

Example 0.22 (Counting person-years). If 10 people are each followed for 2 years, the study has $10 \cdot 2 = 20$ person-years. If one person is followed for 3 years and another for $\frac{1}{2}$ year, the study has $3 + \frac{1}{2} = 3.5$ person-years.

Definition 0.29 (Rate). A **rate** is a quotient of two quantities, usually with a denominator that measures time, such as weeks or person-years of follow-up (Definition 0.28). For example, 12 new cases in 4 weeks is a rate of $\frac{12}{4} = 3$ new cases per week, and 30 cases over 10,000 person-years of follow-up is a rate of $\frac{30}{10,000} = 0.003$ cases per person-year.

cf. [https://en.wikipedia.org/wiki/Rate_\(mathematics\)](https://en.wikipedia.org/wiki/Rate_(mathematics))

Definition 0.30 (Ratios). A **ratio** is a quotient in which the numerator and denominator are measured using the same unit scales.

cf. <https://en.wikipedia.org/wiki/Ratio>

Example 0.23 (A ratio, and a quotient that is not one).

- A board 150 cm long and one 75 cm long have length ratio $\frac{150 \text{ cm}}{75 \text{ cm}} = 2$: both lengths are in centimeters, so the units cancel and the ratio has none.

- A sample of mass 300 g and volume 150 cm³ gives the quotient $\frac{300 \text{ g}}{150 \text{ cm}^3} = 2 \text{ g per cm}^3$, its density. The numerator and denominator are in different units, so this quotient is not a ratio.

Definition 0.31 (Proportion). In statistics, a **proportion** typically means a ratio where the numerator represents a subset of the denominator.

See https://en.wikipedia.org/wiki/Population_proportion.

See also [https://en.wikipedia.org/wiki/Proportion_\(mathematics\)](https://en.wikipedia.org/wiki/Proportion_(mathematics)) for other meanings.

Example 0.24 (A proportion, and a ratio that is not one). In a clinic with 120 patients, 30 of whom smoke:

- the proportion of patients who smoke is $\frac{30}{120} = 0.25$: the 30 smokers are a subset of the 120 patients;
- the ratio of smokers to non-smokers is $\frac{30}{90} = \frac{1}{3}$. Both counts are of patients, so this quotient is a ratio, but it is not a proportion: the 30 smokers are not part of the 90 non-smokers.

Definition 0.32 (Proportional). Two functions $f(x)$ and $g(x)$ are **proportional** if their ratio $\frac{f(x)}{g(x)}$ does not depend on x . (cf. [https://en.wikipedia.org/wiki/Proportionality_\(mathematics\)](https://en.wikipedia.org/wiki/Proportionality_(mathematics)))

Example 0.25 (Proportional and non-proportional functions).

- $f(x) = 6x^2$ and $g(x) = 2x^2$ are proportional: for $x \neq 0$,

$$\begin{aligned}\frac{f(x)}{g(x)} &= \frac{6x^2}{2x^2} \\ &= 3,\end{aligned}$$

which does not depend on x .

- $f(x) = x + 1$ and $g(x) = x$ are not proportional: for $x \neq 0$,

$$\begin{aligned}\frac{f(x)}{g(x)} &= \frac{x+1}{x} \\ &= \frac{x}{x} + \frac{1}{x} \\ &= 1 + \frac{1}{x},\end{aligned}$$

which is 2 at $x = 1$ and $\frac{3}{2}$ at $x = 2$.

Additional reference for elementary algebra: https://en.wikipedia.org/wiki/Population_proportion#Mathematical_definition

Limits of sums and notational shorthands

Definition 0.33 (Lower limit of a sum or integral). In a **sum** $\sum_{i=m}^n a_i$, the **lower limit** m is the first value of the index i . In an **integral** $\int_a^b f(x) dx$, the **lower limit** a is the left end of the interval of integration. In both, the lower limit is written below the symbol.

Definition 0.34 (Upper limit of a sum or integral). In a **sum** $\sum_{i=m}^n a_i$, the **upper limit** n is the last value of the index i . In an **integral** $\int_a^b f(x) dx$, the **upper limit** b is the right end of the interval of integration. In both, the upper limit is written above the symbol.

Definition 0.35 (Limits of integration). In an **integral** $\int_a^b f(x) dx$, the lower limit (Definition 0.33) a and the upper limit (Definition 0.34) b together are the **limits of integration**.

Example 0.26 (Reading off limits).

- In $\sum_{i=2}^5 i^2$, the lower limit is 2 and the upper limit is 5, so the sum is $2^2 + 3^2 + 4^2 + 5^2 = 54$.
- In $\int_0^3 x dx$, the lower limit is 0 and the upper limit is 3.

i Remark (Two meanings of “limit”)

Remark 0.9 (Two meanings of “limit”). The lower and upper limits of a sum or integral (Definition 0.33, Definition 0.34) are not limits in the calculus sense, such as the **limit of a function** or the **limit of a sequence**. They are the ends of the values the index or variable runs over. The two meanings meet in an integral such as $\int_0^\infty f(x) dx$, which means $\lim_{b \rightarrow \infty} \int_0^b f(x) dx$: the upper limit b is sent to infinity by a calculus limit.

Definition 0.36 (Notational shorthand). A **notational shorthand** is a way of writing an expression that leaves part of the expression out, such as the set a sum runs over or the limits of an integral, and relies on the reader to supply the missing part from context.

Definition 0.37 (Full form). The **full form** of a notational shorthand (Definition 0.36) is the expression with every part written out.

i These notes write the full form

In these notes, we avoid notational shorthands in permanent writing and write the full form of each expression, even where a shorthand would be shorter. A reader who supplies the missing part of a shorthand from a different context than the writer intended reads a different expression than the one the writer meant. Handwritten board work is often less complete, but it should still write out ranges and limits where it can.

We make one exception: writing two factors side by side means multiplying them, as in $2x$ for $2 \cdot x$ and ab for $a \cdot b$. We still write the multiplication sign when the side-by-side form could be misread:

- $2 \cdot 3$, not 23, which reads as twenty-three;
- $a \cdot (b + c)$ when a could be a function, since $a(b + c)$ could mean a evaluated at $b + c$.

Definition 0.38 (Range of a variable). The **range** of a variable x , written $\mathcal{R}(x)$, is the set of values that x can take.

For example, if x is the number of heads in two tosses of a coin, then $\mathcal{R}(x) = \{0, 1, 2\}$. A **random variable** X (defined in the Morrison Lab’s probability notes) is a **function**, and the set of values X can take, $\mathcal{R}(X)$, is the **image** of X . These notes use “range” only for variables: for a function, “range” can mean either the image or the codomain (**Range**). In statistics, “the range” of a dataset can also mean its largest value minus its smallest value; $\mathcal{R}(x)$ is a set, not that difference.

Table 1 lists shorthands you may encounter in other sources, the full form each one abbreviates, and the part each one leaves out. In the table, $x_1, \dots, x_n$ and $a_1, \dots, a_n$ are n numbers, b_{ij} is the entry in row i and column j of an $m \times n$ array, k is an integer with $1 \leq k \leq n$, f is a **function** from $\mathbb{R}$ to $\mathbb{R}$, and, in the first row, $\mathcal{R}(x)$ is a finite set.

Table 1: Common notational shorthands and their full forms

Shorthand	Full form	Part left out
$\sum_x f(x)$	$\sum_{x \in \mathcal{R}(x)} f(x)$	the set the sum runs over (sum over a finite set)
$\sum_i x_i$	$\sum_{i=1}^n x_i$	the lower and upper limits of the index (summation notation)
$\sum x_i$	$\sum_{i=1}^n x_i$	the index and its lower and upper limits

Shorthand	Full form	Part left out
$\prod_i x_i$	$\prod_{i=1}^n x_i$	the lower and upper limits of the index
$\prod x_i$	$\prod_{i=1}^n x_i$	the index and its lower and upper limits
$\sum_{i,j} b_{ij}$	$\sum_{i=1}^m \sum_{j=1}^n b_{ij}$	the limits of both indices
$\sum_{i \neq k} a_i$	$\sum_{i \in \{1, \dots, n\} \setminus \{k\}} a_i$	the set $\{1, \dots, n\}$ of indices that k is removed from
$\int f(x) dx$, used for a number	$\int_{-\infty}^{\infty} f(x) dx$	the limits of integration
$\int f$	$\int_{-\infty}^{\infty} f(x) dx$	the limits, the variable of integration, and the differential dx
$x_{1:n}$	$(x_1, x_2, \dots, x_n)$	the explicit list of entries
“the function x^2 ”	“the function $f : \mathbb{R} \rightarrow \mathbb{R}$ with $f(x) = x^2$ ”	the function’s name, domain, and codomain

i Remark ($\int f(x) dx$ has two readings)

Remark 0.10 ($\int f(x) dx$ has two readings). In calculus, $\int f(x) dx$ with no limits is not a shorthand: it is the **indefinite integral** of f , a family of functions rather than a number. Some sources, especially in probability, also write $\int f(x) dx$ for the number $\int_{-\infty}^{\infty} f(x) dx$. Writing the limits out tells the reader which of the two is meant.

Example 0.27 (Writing out an index shorthand). Let $n = 3$ and $(x_1, x_2, x_3) = (2, 5, 1)$. A source that writes $\sum_i x_i$ means $\sum_{i=1}^3 x_i$:

$$\begin{aligned}
\sum_{i=1}^3 x_i &= x_1 + x_2 + x_3 && \text{(definition of summation notation)} \\
&= 2 + 5 + 1 && \text{(substitute the values)} \\
&= 8 && \text{(add)}
\end{aligned}$$

- [1] Wikipedia contributors, “Line segment — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Line_segment
- [2] Wikipedia contributors, “Chord (geometry) — Wikipedia, the free encyclopedia.” 2026. Available: [https://en.wikipedia.org/wiki/Chord_\(geometry\)](https://en.wikipedia.org/wiki/Chord_(geometry))
- [3] Wikipedia contributors, “Convex function — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Convex_function
- [4] Wikipedia contributors, “Jensen’s inequality — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Jensen%27s_inequality

- [5] Wikipedia contributors, “Upper and lower bounds — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Upper_and_lower_bounds
- [6] Wikipedia contributors, “Least-upper-bound property — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Least-upper-bound_property
- [7] W. Rudin, *Principles of mathematical analysis*, 3rd ed. in International series in pure and applied mathematics. New York: McGraw-Hill, 1976.
- [8] Wikipedia contributors, “Infimum and supremum — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Infimum_and_supremum
- [9] Wikipedia contributors, “Identity element — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Identity_element
- [10] Wikipedia contributors, “Commutative property — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Commutative_property
- [11] Wikipedia contributors, “Associative property — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Associative_property
- [12] Wikipedia contributors, “Distributive property — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Distributive_property
- [13] Wikipedia contributors, “Like terms — Wikipedia, the free encyclopedia.” 2026. Available: https://en.wikipedia.org/wiki/Like_terms