Defining terms clearly

Last modified: 2026-11-07 21:34:25 (PST)

Clear definitions are essential to effective scientific writing. Every specialized term should have an explicit, concise definition before its first use. Readers should not need to search for the meaning of a term or infer it from context alone. For example:

Definition 0.1 (Term)**.** A **term** is a word or phrase with a specific, technical meaning within a particular field or context.

Example 0.1 ("Significant" used as a term)**.** In "the difference was statistically significant at the 5% level", "significant" is a term (Definition 0.1): it means that a hypothesis test rejected its null hypothesis at that level.

Example 0.2 ("Significant" used as an ordinary word)**.** In "staffing is a significant problem for the clinic", "significant" is not a term (Definition 0.1). It carries its everyday meaning, "large" or "important", and no hypothesis test is involved.

Definition 0.2 (Population)**.** In statistics, a **population** is the complete set of all items or individuals of interest.

Example 0.3 (A population of patients)**.** For a study of blood-sugar control, the population (Definition 0.2) might be all adults in California with type 2 diabetes in 2024.

Example 0.4 (A sample is not the population)**.** The 500 patients who enrolled in that study are not the population (Definition 0.2). They are a sample drawn from the population, and the study's estimates describe the population only as far as the sample represents it.

Remark 0.1 (Population is not only geographic). In everyday use, "population" usually means the people living in a place. The statistical sense in Definition 0.2 is broader: a population can be all the patients with a condition, all the batches a factory produces, or any other set of items of interest.

Guidelines for defining terms

Follow these principles when introducing new terms:

- **Provide explicit definitions:** State clearly what a term means, rather than assuming readers will understand from context.
- **Be concise:** Definitions should be brief and focused, using only the words necessary to convey the meaning. A definition block holds the defining statement and nothing else: no examples, special cases, motivation or caveats.
- **State technical definitions and results in both prose and a display equation:** Give a quantitative term, or a theorem or corollary, a one-sentence prose statement and the formula it stands for as a display equation, in the same definition or theorem block; together the sentence and the formula are the statement. A formula left inline in the sentence is easy to miss and hard for later derivations to point to. The prose tells the reader what the quantity means; the formula makes it exact and lets later derivations cite it. For example, define the ordinary least squares estimate of a parameter vector $\underset{\sim}{\theta}$ as "the value of $\underset{\sim}{\theta}$ that minimizes the residual sum of squares" *and* as

  $$\underset{\sim}{\hat{\theta}} := \arg\min_{\underset{\sim}{\theta}} \mathrm{RSS}(\underset{\sim}{\theta}),$$

  not with only one of the two. Build the formula from terms already defined, rather than expanding them again: once the residual and total sums of squares are defined, write

  $$R^2 := 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}},$$

  not the two sums that RSS and TSS stand for.
- **Define a statistical model by its distribution, and derive its other forms:** Define a regression model by the distribution of the outcome conditional on the covariates, centered on a named mean function. For example, define simple linear regression as $Y_i \mid X_i = x_i \sim \mathrm{N}(\mu_i, \sigma^2)$, independently, with $\mu_i := \mu(x_i)$ and $\mu(x) := \beta_0 + \beta_x x$. Then define the deviation $\varepsilon_i := Y_i - \mu(x_i)$, and state $Y_i = \mu_i + \varepsilon_i$ with $\varepsilon_i \mid X_i = x_i \sim \mathrm{N}(0, \sigma^2)$ as a result proved from those definitions. Writing the model as "$Y_i = \beta_0 + \beta_x x_i + \varepsilon_i$ with

Gaussian errors" makes the definition rest on a quantity that has not been defined yet, and hides that the model is a statement about the outcome's distribution.

- **Define a new symbol by its most direct, literal meaning, and derive its other forms:** Define a shorthand symbol directly from the objects it summarizes, then prove any equivalent form as a result. For example, with deviations $\delta_i := y_i - \mu$ from the population mean μ, define the sample mean of the deviations as

$$\bar{\delta} := \frac{1}{n} \sum_{i=1}^{n} \delta_i,$$

  then show that $\bar{\delta} = \bar{y} - \mu$. Defining $\bar{\delta}$ as $\bar{y} - \mu$ states a result in place of the meaning, and hides that the symbol is a mean of the deviations.

- **Put only the defining equation in a definition, and state equivalences as results:** A definition div displays one equation, the one that links the symbol of the defined term to its definitional expression. It does not display a chain of equal expressions, and it does not display an equivalent form. It does not state that the defined concept is the same as, or equivalent to, another concept either. State each equality between expressions, and each equivalence between concepts, in its own result (a theorem, lemma, corollary or proposition) after the definition. For example, define the residual variance at x as $\sigma^2 \stackrel{\text{def}}{=} \mathrm{Var}(\varepsilon \mid X = x)$, then state $\sigma^2 = \mathrm{Var}(Y \mid X = x)$ as a result. A definition that also shows $= \mathrm{Var}(Y \mid X = x)$ asserts a result without a statement or a proof, and the reader cannot tell which equality is the definition. The same holds for a sentence: a definition that adds "equivalently, the variance of the outcome given the covariates" states a result, so it belongs in one.

- **Write both sides of a definition in the same notation:** The symbol on the left names an object, and the right side expands that object's own definition. It does not describe the object in another way, because the step from one description to the other is a result. For example, with the conditional expectation function μ and its prediction error $\varepsilon_{\text{pred}}(\mu)$, define the residual risk as $R(\mu) = \mathbb{E}\left[\left(\varepsilon_{\text{pred}}(\mu)\right)^2\right]$, not as $R(\mu) = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right]$. The second form spells the function μ out on the right but names it on the left, so a reader cannot tell whether the two sides describe the same object. Derive the second form from the first in a proof.

- **Reuse the notation and the concept names of the result that introduces an object:** When an earlier theorem names an object, every later definition about it uses the same symbols and the same concept. Do not coin an alias symbol or a synonym term for it. For example, a theorem that decomposes the squared-error risk and names the minimizing function μ fixes the residual risk as the risk $R(\mu)$ of that function. A definition that names the residual risk through a second symbol for the same function leaves the reader to check that the two symbols denote one object.

- **Make a definition as concise as possible:** State the term with the fewest words and symbols that fix it exactly. Build on terms that are already defined, and link them, instead of spelling out again what they mean. For example, define a mode of a distribution as an element of the argmax of its density or mass function, not by writing out the inequality that the argmax stands for. The inequality belongs in the definition of the argmax, which the reader can follow the link to. A definition that restates a linked term is longer to read and drifts from the original if either changes.
- **Write a definition with an expectation operator where one fits:** When a quantity is a probability-weighted sum, or a probability-weighted integral, of a function of a random variable, define it as the expectation of that function, not as the sum, or the integral, written out. The expectation operator already stands for the sum when the variable is discrete and for the integral when it is continuous, so one definition covers both and need not repeat the support. For example, define the entropy of a discrete random variable X with mass function p as

$$H(X) \stackrel{\text{def}}{=} \mathbb{E}[-\log p(X)],$$

  not as $-\sum_x p(x) \log p(x)$ with a note that the sum runs over the x with $p(x) > 0$. The expectation sums over the possible values of X, which are the values x with $p(x) > 0$, so $-\log p(x)$ is defined at each of them. Likewise, define the variance of any random variable X as

$$\operatorname{Var}(X) \stackrel{\text{def}}{=} \mathbb{E}\big[(X - \mathbb{E}[X])^2\big],$$

  not as a sum for one case and an integral for the other. The expanded sum or integral belongs in a result that follows from the definition of expectation. Keep the sum or integral when no expectation expresses the quantity, or when the definition is the one that defines expectation itself.
- **Write the defining expression with the highest-level macros available:** Write each quantity in the defining expression with the semantic macro of the highest-level concept that names it, such as `\entropyf{X}` for an entropy, not with the lower-level expression that the macro stands for. Among equivalent defining expressions, choose the one that makes the most important property visible. For example, define the mutual information of discrete random variables X and Y as

$$I(X;Y) \stackrel{\text{def}}{=} H(X) + H(Y) - H(X,Y),$$

  which is symmetric in X and Y at a glance, and not as $H(Y) - H(Y \mid X)$ or as a sum over the joint mass function. The other forms are results that follow from the definition.

- **Define a probability distribution by its density or mass function:** Define a named distribution directly, by its probability density function (pdf) or probability mass function (pmf) and its support. State any relationship to other distributions as a theorem proved from those definitions. For example, define the chi-square distribution with k degrees of freedom by its pdf

$$f(x) := \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\,\Gamma(k/2)}, \quad x \in (0, \infty),$$

  and state as a theorem that, for a positive integer k, the sum of the squares of k independent standard normal random variables has this distribution. The density is defined for every real $k > 0$; the theorem covers only integer k. Defining the chi-square distribution as that sum turns a result into a definition, and leaves the reader without the density that later calculations need.

- **Define a term directly, not as a modification of another term:** State what the term is, in its own words and symbols. Do not make the reader recall a different term and then apply a list of changes to it. For example, define the expected prediction error as the population mean of the squared prediction error over training sets and new observations, rather than as "the sample mean squared error of predictions, with the average over observations replaced by an expectation". The second wording makes the reader build a term that the definition then discards. Link a related term when it helps, and relate the two in an example or a remark after the definition.

- **Define in the most general form the document needs:** State the concept at the level of generality the rest of the document relies on, not just for the first case the reader encounters. For example, define the residual sum of squares as the sum of squared residuals of any fitted model, not as a formula for a straight line. A definition tied to one special case has to be restated, or silently stretched, the first time the document needs the general case.

- **Follow the definition with examples, from general to specific:** Put each special case in its own example block right after the definition, ordered from the most general to the most specific, each linking back to the definition or to the example it specializes (see Order: general definition, then examples).

- **Define terms at first use:** Place the definition of a term before the first definition, theorem or passage that relies on it, so that every link to a definition points backward rather than forward (see Avoid forward references). When a definition or theorem uses a term whose definition comes later in the document, reorder the content so the definition comes first; do not remove the hyperlink to hide the forward reference.

- **Give every key term a formal definition in a `#def-` div:** A term that the document relies on needs its own definition div, even if an exercise, a solution, an example or a sentence of prose has already explained it. An explanation inside an exercise or solution

cannot be linked to as a definition, and a student who skips the exercise never sees it. Find the gaps by listing the terms the document uses, such as the bolded phrases and the terms in section titles, and checking that each one has a `#def-` div, or a link to the page that has one. This extends *Give each term its own definition block* below from terms defined in passing to terms explained only in exercises. For example, if "loss function" first appears in an exercise that asks students to compute one, add a `#def-loss-function` div before the exercise, and link the exercise's mentions to it.

- **Pair each technical term with its mathematical symbol:** Whenever a definition, an example or a list names a quantity in words, write its symbol beside the name. For example, write "the slope β_x" and "the residual sum of squares RSS", not "the slope" and "the residual sum of squares" alone. The words tell the reader what the quantity means, and the symbol tells the reader which letters stand for it in the equations that follow.
- **Give a definition a symbol only when the object recurs in formulas:** Introduce a new symbol or macro for a defined term when the term appears in at least two formulas. A term that appears only in words, such as "pairwise disjoint", needs no symbol. A term that recurs in formulas, such as a level set, needs one, so the reader sees the same object written the same way each time. Define the symbol by the term's most literal meaning, such as the preimage of a single value for a level set. No check can decide this rule, because whether a term "recurs" depends on how the notes use it, so a reviewer judges it.
- **Define and link every technical term at its first use in each div:** Inside a definition, theorem, example or other div, link a technical term to its formal definition the first time it appears in that div. Outside divs, link it the first time it appears in each prose passage. A link in an earlier div does not carry over, because readers jump into the middle of a page. Link once per div or passage, not at every mention. If the term has no formal definition yet, write the definition in its own definition div, then link to it. Never define a term inline inside another div, because an inline definition cannot be linked to from anywhere else. In a series of related documents, such as course notes, first look for a definition in an earlier document and link to it if one exists. Otherwise write the definition in the earliest document that needs the term. A passage is the prose outside divs between one heading and the next, and each div counts separately from the passage around it and from every other div. For example, if a div defines noise as the outcome minus the "true function", link the term as `[true function](#def-true-function)`, and first add a definition div with that ID if the document lacks one. A counterexample: the definition of noise links "mean". The example block after it also uses "mean", but without a link, wrongly assuming the earlier link carries over. The example block should link "mean" at its own first use.
- **Link the longest defined phrase:** When a longer phrase has its own definition, link the whole phrase to that definition, not a shorter term inside it. For example, link `[logistic regression](#def-logistic-regression)`, not "logistic `[regression](#def-regression)`". The shorter term names a different concept, so linking it sends the

reader to the wrong definition. Link the shorter term only where it stands alone. The same holds when the definition is on another page or site: write `[likelihood ratio test](https://morrison-lab.github.io/sds/intro-MLEs.html#def-lrt-stat)`, not "`[likelihood](https://morrison-lab.github.io/sds/intro-MLEs.html#def-lik)` ratio test", which sends the reader to the definition of likelihood, a different concept. If no definition titled with the whole phrase exists, link the phrase to the closest definition that covers it, here the likelihood ratio statistic, and the check then learns the phrase. The `check-term-links` check in gha flags a partial link.

- **Link the term where the sentence uses it, not in a parenthetical list:** Write the name as a link inside the sentence that needs it. Do not name the concepts in the sentence and then list their definitions in a parenthetical after it. For example, write `the expected squared prediction error is the sum of the squared [bias](#def-bias), the [variance](#def-variance), and the [residual variance](#def-residual-variance)`, not `the expected squared prediction error is the sum of the squared bias, the variance, and the residual variance (definitions: [bias](#def-bias), [variance](#def-variance), [residual variance](#def-residual-variance))`. The same holds for a source: link the term's words to the source's page instead of adding a numbered citation after them, as Link the words, not a trailing citation describes. The parenthetical makes the reader read each term twice, and it separates the link from the place where the term does its work. If a term sits inside math, where a link cannot go, name the term in words in the sentence and link those words. The `check-quarto-prose` check in gha flags a parenthetical that opens with `(definition:` or `(definitions:`.

- **Do not presuppose what the reader has not encountered:** A definite noun phrase, one that starts with "the", presupposes that the reader already knows what it refers to. "Fit a line to the synthetic data" is unjustified if no earlier sentence has said which data these are, or where they came from. Check each "the …" phrase for a referent the reader has already seen. If there is none, introduce the referent first, or write "a" or "some" in place of "the". Then link each definite noun phrase to what it refers to, with a cross-reference such as `[the synthetic data](#exm-synthetic-data)`, so that a reader who has forgotten or skipped the referent can find it. Skip the link when the referent is introduced in the same div, or in the same prose passage, as the phrase. Inside a div, the antecedent has to be in that div: an earlier div does not count, because readers jump into the middle of a page. For example, an exercise that says "Explain the labels" with no labels named in it has no antecedent, even if an earlier exercise names some. Name the labels in the exercise, or link them to where they are introduced.

- **Provide examples and counterexamples:** Every definition should include at least one concrete example that illustrates how the term is used, and at least one counterexample: a case that looks as if it fits the definition but does not. The example shows what the term covers; the counterexample shows where its boundary is, which is often where readers misapply it. Put each in its own example block after the definition.

- **Give each term its own definition block:** In a Quarto document, put every definition in its own `#def-` div (see Using theorem environments in Quarto), one term per div, with commentary after the div rather than inside it, in a remark block (`#rem-`) when it is about the concept. Never nest one theorem-type block (any of the prefixes listed in Using theorem environments in Quarto, such as `#def-`, `#exm-`, `#exr-`, `#sol-` or `#rem-`) inside another; give each its own div, in order. A bolded term in running prose is usually an inline definition that belongs in a div, including a term defined in passing. Give closely related terms separate divs too: only full synonyms share one (see *Name the synonyms* below). For example, the left and right distributive laws are two properties, each with its own identifier, example and exercise, so define them in two divs. A reader who needs one of them can then link to exactly that one.

  Example 0.1 (Splitting two related definitions)**.**

  ✗ **Definition (distributive laws).** Multiplication *distributes over addition on the left* if $a(b+c) = ab + ac$ and *on the right* if $(a+b)c = ac + bc$.

  ✓ **Definition (left distributivity).** Multiplication *distributes over addition on the left* if $a(b+c) = ab + ac$ for all a, b and c.

  ✓ **Definition (right distributivity).** Multiplication *distributes over addition on the right* if $(a+b)c = ac + bc$ for all a, b and c.

  The first block cannot be linked to as "left distributivity" without also pointing at the other law, and one example cannot illustrate both. For numbers, right distributivity follows from left distributivity because multiplication commutes, so one example checks both. Matrix multiplication does not commute, so establishing one law does not establish the other, and each needs its own example.

- **Name the synonyms:** When a field uses several words for one concept, such as *learning*, *training* and *fitting* a model, every word that names the concept is an alternative name, and each is treated exactly as the primary name: put all of them in the div title as a comma-separated list, with the primary name first, as in "Postulate, axiom", and bold each where it first appears in the definition. Do not use parentheses or a phrase such as "also called" in the title. This covers the names other sources use, such as a textbook's, which belong in the title and in bold too, as in "Risk, expected loss, population risk, true risk, generalization error, expected prediction error". Say which one the document uses. Name near-synonyms in the definition block or in a callout beside it, and state how they differ. A reader who encounters the other words in another source can then connect them to the concept already learned (see Minimize unnecessary jargon).

- **Tell synonyms from hyponyms:** A *synonym* names the same concept as the primary term, so it belongs in the div title. A *hyponym* names a special case of another term,

such as "linear regression" for "regression model", so it is a different concept. Give each hyponym its own `#def-` div, and never put it in the title of the general term or introduce it with "also called" (Ezra, 2026-10-23). When unsure, check both directions. Is every instance of the new term an instance of the old term? Is every instance of the old term an instance of the new term? If only the first holds, the new term is a hyponym.

- **Tell synonyms from equivalent concepts:** Only true synonyms share a div title as aliases: names for one concept, which has one definition. Two concepts that happen to be equivalent are different concepts, however often they coincide. Give each its own definition, an example after each, and a theorem showing that they agree (see *Put only the defining equation in a definition, and state equivalences as results* above). A test: if each name has its own natural definition, they are two concepts. For example, a loss defined as the negative log of the predicted probability of the observed class and a loss defined as the cross-entropy between two distributions each have a natural definition, so give each its own div and state that they agree as a theorem, instead of titling one div with both names.

- **Keep an entrenched misnomer, and clarify it once:** Some names the literature has fixed are technically wrong, such as "linear regression" for a model with an intercept, whose mean is an affine function of the predictors, not a linear one (Ezra, 2026-10-28). Wherever the wording is yours, use the precise word: write "affine in the predictors" for a function with an intercept, and "linear" only when the intercept is zero. Keep a name the reader will meet in other sources, such as linear regression, linear decay or linear approximation, because renaming it cuts the reader off from that literature. Clarify it once, in parentheses, at its definition or its first use on a page, as in "linear (strictly, affine) decay", and use the established name unqualified after that.

- **Distinguish estimands from estimators:** An *estimand* is the quantity a study is after, defined on the population or the data-generating distribution; an *estimator* is the rule that computes an estimate of it from a sample. Define a quantity first as its estimand, then give the estimator that targets it, and keep the two in separate notation (for example, θ and $\hat{\theta}$). A definition that gives only the sample formula hides what the formula estimates, and invites the reader to treat one sample's value as the quantity itself. Keep that order across the whole document, not only within one definition: the estimand's definition, with its exercise and example, comes before the estimator's, and the estimator's definition links back to it. For example, define the risk of a model, its expected loss on new data, before the empirical risk, its average loss on the training data; define the population mean before the sample mean, and a parameter before its maximum likelihood estimate. Reading the estimator first leaves the reader with a formula and no answer to the question of what it is for. When several related quantities are defined together, give a table with one row per quantity and a column each for its name, its estimand and its estimator. For a binary classifier with prediction $\hat{Y}$ of the true class Y, and counts TP, FP, FN and TN from a test set of size n:

Name	Estimand	Estimator
Accuracy	$\Pr(\hat{Y} = Y)$	$(\mathrm{TP} + \mathrm{TN})/n$
Precision	$\Pr(Y = 1 \mid \hat{Y} = 1)$	$\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$
Recall	$\Pr(\hat{Y} = 1 \mid Y = 1)$	$\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$

The precision of a classifier is a property of the classifier and the population it is used on; $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ is one test set's estimate of it.

- **Name the target of every estimator and predictor:** Every estimator has one target estimand, and the definition says which. A prediction is not an estimator of a mean. Its estimand is the observed outcome value of a new data point, not the population mean of that outcome. For example, the fitted model at x is an estimator of the mean function value $\mu(x)$, so its bias is measured against that mean, while the prediction at x has the new outcome Y_* as its estimand, so its prediction error is measured against Y_*. Write "the fitted model at x as an estimator of the mean function value", or "the prediction of the new outcome", and never leave the target implicit.
- **Say which error:** A bare "error" does not say what is measured, on which data, or whether it is a value computed from one sample or an expectation. Every use of "error" (and of "loss") names three things: the quantity measured, such as the squared prediction error; the data that it is computed on, such as the training set, the validation set, the test set or the population; and whether it is a sample value or an expectation. For example, write "the prediction error on the validation set", or "the expected squared prediction error", and never "the error" or "the smallest error". A term that already carries all three, such as the generalization error, can stand alone after its definition.
- **Make the fitted model the subject, not the fitting procedure:** A prediction, an error, a bias and a variance are all properties of fitted models, so a sentence about them has a fitted model as its subject, and a choice among candidates is a choice among fitted models. Write "the fitted model predicts", "the prediction of the fitted model at x", and "choose among fitted models", and never "the procedure predicts" or "choose among procedures". Name the fitting procedure only when the sentence is about how a fitted model is built: the inputs that the procedure needs, what it costs, or an expectation over the training sets to which it is applied. In math, write $\hat{y}_{\mathcal{D}_{\text{train}}}$ for the fitted model, the model that the procedure produces from the training set, and g only for the procedure itself. Where the training set is not salient, write the fitted model as $\hat{y}$ instead, the symbol that the predicted-value bullet below builds on. For example, write "the bias of the prediction $\hat{y}$", not "the bias of g".
- **Use "prediction rule" only for a fitted model in its role as a predictor:** Not every fitted model is used to predict outcomes, so "prediction rule" is a narrower term, not another name for "fitted model". Define it in its own `#def-prediction-rule` div as a fitted model that is used to predict outcomes at a new covariate value, and link

it to that definition at its first use. Write "the fitted model" for a sentence about the model itself, such as its form, its coefficients or its fit to the training set. Write "the prediction rule" only where the sentence needs the role of predicting a new outcome to tell it from other uses of the fitted model. The verb does not change the term: keep writing "the fitted model predicts", as the bullet above requires. For example, write "the fitted model has slope 2" and "a clinic uses the prediction rule for new patients", and not "the prediction rule has slope 2". If the notes once defined "prediction rule" as a span inside the fitted model's definition, keep that anchor on the new div, so that links to it still land there.

- **Write a predicted value with the predicted-value macro:** A predicted value is the fitted model applied to a covariate value, so write it with a predicted-value macro and not as the procedure applied to a covariate value and a training set. The macro `\predval` takes no argument and prints the predicted value of the outcome at the covariate value that the sentence is about. The macro `\predvalf{x}` takes a covariate value x and prints the prediction $\hat{y}(x)$, a value of the outcome's type. The macro `\candpredvalf{l}{x}` takes a candidate index l and a covariate value x, and prints the prediction of candidate l's fitted model at x. Suppress the covariate value where it is not salient: write `\predval` when the sentence is not about which covariate value the prediction is at, and `\predvalf` or `\candpredvalf` only when the covariate value matters, such as when comparing predictions at two covariate values or averaging a prediction over them. For example, write $\hat{y}$ in "the bias of the prediction $\hat{y}$" when one covariate value is fixed throughout, write $\hat{y}(x)$ and $\hat{y}_l(x)$ when the covariate value is salient, and never $g(x; \mathcal{D}_{\text{train}})$ or $g_l(x; \mathcal{D}_{\text{train}})$.
- **Apply a function to its argument with the macro's f form:** Where a macro names a function, such as a fitting procedure, the macro with an `f` ending applies it to its argument, so write the application with that macro and not with a hand-typed `\paren`. For example, write $g(\mathcal{D}_{\text{train}})$ for the fitted model that g gives on the training set $\mathcal{D}_{\text{train}}$, and never $g(\mathcal{D}_{\text{train}})$. The same holds for an operator and every delimiter macro (`\paren`, `\sb`, `\cb`, `\set`, `\abs`, `\norm` and `\angb`), and for an alias of the operator: write $\mathbb{E}[y]$, and never $\mathbb{E}[y]$. Before adding an f form, check that no macro already prints the notation: `\candfitted{l}` already prints the fitted model of candidate l, and a second macro for the same notation is not an f form. When no macro prints the notation, create the f form in the macros repository first, in the same session, then bump the pin and use it, as for any new macro. A missing f form is never a reason to type the operator and the delimiter separately.
- **Qualify "mean", "average" and their shorthands as population or sample:** A bare "mean" or "average" does not say whether it is a quantity of the population or a value computed from one sample. Write "population mean" for the expected value of a distribution, and "sample mean" for the average of the observed values; write "population average" and "sample average" in the same way. Prefer "mean" with the qualifier to "average". The same holds for shorthands: write pMSE and sMSE, pRMSE and sRMSE, and pMAE and sMAE, with a leading p for the population version and s for the sample

version, and never a bare MSE, RMSE or MAE. Spelled out, the same words apply: "population mean squared error", "sample root mean squared error", "population mean absolute error". Define each shorthand once, at first use, as "population mean squared error (pMSE)". A term that already says which, such as "mean function" or "running mean", needs no extra word, and neither does expected value notation such as $\mathbb{E}[Y]$. For example, write "the sample mean of the training outcomes estimates the population mean of the outcome", not "the mean estimates the mean". Drop the qualifier only when it is completely clear from context, such as inside a definition that has just said which one it concerns. When in doubt, keep it.

Examples of term definitions

Here are examples of well-defined terms:

Definition 0.1 (Confidence interval)**.** A **confidence interval** with confidence level $1-\alpha$ for a parameter θ is an interval $[L, U]$ computed from a sample that contains θ with probability $1-\alpha$ over repeated samples:

$$\Pr(\theta \in [L, U]) = 1 - \alpha.$$

Example 0.1 (A confidence interval for mean height)**.** If we calculate a 95% confidence interval (Definition 0.1) for mean height as [165 cm, 175 cm], we are 95% confident that the true population mean height falls within this range.

Example 0.2 (A range that is not a confidence interval for the mean)**.** The interval from the 2.5th to the 97.5th percentile of the observed heights, [150 cm, 190 cm], is not a 95% confidence interval (Definition 0.1) for mean height. The percentile interval describes where individual heights fall, not where the population mean is likely to be, and it does not shrink as the sample grows.

Definition 0.2 (Active voice)**. Active voice** is a sentence structure where the subject performs the action expressed by the verb.

Example 0.3 (A sentence in active voice)**.** "The researcher conducted the experiment" is in active voice (Definition 0.2): the subject, "the researcher", performs the action, "conducted".

Example 0.4 (A sentence that is not in active voice)**.** "The experiment was conducted by the researcher" is not in active voice (Definition 0.2). Its subject, "the experiment", receives the action instead of performing it, so the sentence is in passive voice.

Order: general definition, then examples

When a concept has special cases, write the general definition first and compactly, then one example block per special case, ordered from the most general to the most specific. Put any commentary in a remark block after them. When the special cases have their own definitions, define the general concept first from prerequisite terms defined before it, and then define the special cases after it. Do not define the general concept by listing the special-case definitions that follow it, or move the general definition after its special cases: define the prerequisite terms first (for example, define lower and upper bounds first, then define an interval as a subset of the real numbers between two bounds, possibly including one or both bounds, and then define closed, open and half-open intervals). For example, a statistics text that defines the residual sum of squares would give these blocks, in this order:

Definition 0.1 (Residual sum of squares)**.** The **residual sum of squares** of a model fitted to data is the sum of its squared residuals:

$$\mathrm{RSS} = \sum_{i=1}^{n} r_i^2.$$

Example 0.1 (Residual sum of squares of a straight line)**.** For a straight line with intercept β_0 and slope β_x, the fitted value of observation i is $\beta_0 + \beta_x x_i$, so the residual sum of squares is a function of the two coefficients:

$$\mathrm{RSS}(\beta_0, \beta_x) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_x x_i)^2.$$

Example 0.2 (Residual sum of squares of a line through three points)**.** For the points $(0, 1)$, $(1, 2)$, $(2, 2)$ and the line with $\beta_0 = 1$ and $\beta_x = 0.5$ (Example 0.1), the residuals are 0, 0.5 and 0, so $\mathrm{RSS}(1, 0.5) = 0.25$.

Example 0.3 (A sum of squares that is not a residual sum of squares)**.** For the same three points, the sum of squared deviations of y_i from their mean, 5/3, is 2/3. That quantity is the total sum of squares, not the residual sum of squares (Definition 0.1) of the line in Example 0.2, because it measures spread around the mean rather than around

the line's fitted values.

The definition stays one sentence and one equation, so it still applies when the document later fits a model that is not a line. The line is a special case, so it is an example, and the numerical case is a special case of the line, so it follows the line. The counterexample comes after the examples, once the reader knows what the definition covers.

Give a defined quantity its expression and its prose reading

When a term names a quantity with a mathematical form, the definition states the expression and then reads it aloud in a prose sentence that says the same thing. The expression makes the definition exact. The prose reading lets the reader check that they understood it. The two must match (see Make the prose say what the formula says). A term with no mathematical form, such as active voice, keeps its prose definition alone.

Definition 0.1 (Variance, with its expression and its prose reading)**.** The **variance** of a random variable X is

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big].$$

In words, it is the population mean of the squared deviation of X from its population mean.

Example 0.1 (A definition that gives only one of the two)**.** "The variance is how spread out X is" gives no expression, so a reader cannot compute it. "The variance is $\mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$" gives no reading, so a reader cannot check their understanding. Definition 0.1 gives both.

Define an activity separately from doing it well

A term that names an activity, such as fitting, estimating or generalizing, does not also name how well the activity is done. Give the activity one definition, and give the quality its own definition and its own term, such as "a good fit" or "a consistent estimator". Three reasons:

- **The activity can be done badly.** A definition that builds the quality in, such as "an algorithm generalizes when it does well on new data", makes "generalizes poorly" a contradiction.
- **The quality needs a criterion.** "Does well" and "good" are comparisons without a standard. State the quantity that measures the quality and the threshold or ordering applied to it.

- **Sources use the name both ways.** Name the sense that a document uses, so that a reader who knows the other sense is not misled.

Example 0.1 (Estimation and a consistent estimator)**.** *Estimation* is the computation of a value for a parameter from a sample. An estimator is *consistent* when its estimates converge in probability to the parameter as the sample size grows. An estimator can be used to estimate, and still be inconsistent.

Example 0.2 (Generalization and generalizing well)**.** *Generalization* is the use of a model fitted to a training set to predict the outputs of inputs outside it. A model *generalizes well* when its average prediction error on new inputs and their outputs, from the target population, is at most a stated tolerance. A model can generalize, in the first sense, and generalize poorly, in the second.

Example 0.3 (Fitting and a good fit)**.** *Fitting* is choosing the parameter values of a model from data, for example by minimizing the residual sum of squares. A *good fit* is a fitted model whose residual sum of squares is at most a stated tolerance. A line can be fitted to any data, including data that it fits badly.

Why examples matter

Examples serve several crucial purposes:

- **Concrete understanding:** Examples transform abstract concepts into tangible instances that readers can visualize and understand.
- **Verification:** Readers can work through examples themselves to verify they understand the definition or theorem.
- **Application:** Examples show how to apply definitions and theorems to solve real problems.
- **Memory:** Concrete examples are easier to remember than abstract definitions alone.
- **Boundaries:** Counterexamples show where a definition stops applying, which is where readers most often misapply it.

Common pitfalls to avoid

- **Circular definitions:** Do not define a term using the term itself.

Incorrect: "Conciseness is the quality of being concise."

Correct: "Conciseness is the quality of expressing ideas using only the words necessary to communicate meaning clearly."

Qualifying the term does not make the definition less circular. "A new observation relative to a training set is a new observation when …" still uses the term in its own definition. Define it directly instead: "A new observation relative to a training set is a covariate value and an outcome drawn from the same population as the training observations and independent of them."

- **Vague definitions:** Avoid definitions that rely on imprecise language.

  Incorrect: "A variable is something that can change."

  Correct: "A variable is a symbol representing a value that can take on different values."

- **A characteristic without its object:** A mode, kernel, support, maximum, or median belongs to one object, and the sentence names that object. Name the object whose definition the characteristic is stated for, such as the density, the distribution, the likelihood, or the kernel, and use "of the distribution" only when the definition is for distributions. Where a text moves from a distribution to its density, say so once rather than switching silently. A bare "the mode" leaves the reader to guess the object, and a wrong guess misattributes the characteristic: a mode is a mode of one function, and the modes of a kernel and of its density coincide only by a theorem, not by definition.

  Incorrect: "The gamma distribution has the kernel $x^{a-1}e^{-x}$, so maximizing it finds the mode."

  Correct: "The gamma density has the kernel $x^{a-1}e^{-x}$, so maximizing the kernel finds the mode of the density." A kernel is defined for a density, not for a distribution.

- **Missing examples:** Never leave a definition without at least one example.

- **Examples without context:** Ensure examples clearly illustrate the specific aspect of the definition they are meant to demonstrate.

Mathematical statements require examples

Mathematical theorems, lemmas, corollaries, and postulates should always include examples that demonstrate their application. Abstract mathematical statements become more accessible when readers can see concrete instances.

Using theorem environments in Quarto

When writing mathematical or technical content in Quarto, use theorem environments to demarcate and highlight the structure of your content. These environments provide consistent formatting, automatic numbering, and cross-reference capabilities.

Quarto provides built-in theorem environments including:

- `#thm-` for theorems
- `#lem-` for lemmas
- `#cor-` for corollaries
- `#prp-` for propositions
- `#cnj-` for conjectures
- `#def-` for definitions
- `#exm-` for examples
- `#exr-` for exercises
- `#sol-` for solutions, named after their exercise (`#exr-foo` and `#sol-foo`)
- `#rem-` for remarks

See the Quarto documentation on theorems and proofs for complete details.

Example syntax:

```
::: {#thm-pythagorean}

## Pythagorean theorem

In a right triangle,
the square of the hypotenuse equals the sum of squares of the other two sides:

$$a^2 + b^2 = c^2.$$

:::
```

This produces automatically numbered output like "Theorem 2.1 (Pythagorean theorem)" and can be referenced elsewhere in your document using `@thm-pythagorean`.

Example: Pythagorean theorem

Theorem 0.1 (Pythagorean theorem)**.** *In a right triangle, the square of the hypotenuse equals the sum of squares of the other two sides:*

$$a^2 + b^2 = c^2.$$

Example 0.1 (Pythagorean theorem example)**.** For a triangle with sides 3, 4, and 5:

$$\begin{aligned} 3^2 + 4^2 &= 9 + 16 \\ &= 25 \\ &= 5^2. \end{aligned}$$

Definition 0.1 (Postulate (axiom))**.** A **postulate** (or **axiom**) is a statement accepted as true without proof, serving as a starting point for further reasoning.

Example 0.2 (Euclid's parallel postulate)**.** Through a point not on a line, exactly one line can be drawn parallel to the given line.
If we have line L and point P not on L, only one line through P will never intersect L.

Presenting a derivation: exercise, solution, theorem, proof

In teaching material, present a derivation as one or more exercises, each followed by its solution, then the theorem that records the result, with a short proof that cites the exercises. The reader encounters the question before the answer, can try each step before reading it, and can find the result itself without searching through the working.

Follow these principles:

- **One exercise per step.** Split a long derivation into exercises a reader can attempt one at a time, such as each partial derivative, then solving the resulting equations, then checking the second derivative.
- **The solution follows its exercise.** Put each solution directly after its exercise, and give it an id named after the exercise (`#exr-foo` and `#sol-foo`), so a reader who wants to check one step can find its working.
- **The theorem states the result; the proof cites results.** The theorem gives the result in prose and as a display equation. Its proof is a few sentences that cite the results it rests on, not a second copy of their working. A proof that does no derivation of its own is one line citing the solution of the exercise that asks for the result, and every later argument cites the theorem, never the solution (see Cite results, not solutions).

- **Define notation first.** Introduce any notation the exercises use in its own definition before them, not inside the theorem that follows them.
- **One operation per line.** Inside each solution, write every displayed line with a single operation and its justification (see One operation per step).
- **One result per theorem.** Give each theorem, corollary or lemma block a single result, with its own exercise and proof. Two results joined by a semicolon, or set side by side in one display equation, usually belong in two blocks: each can then be cited on its own, and each proof cites only the result it rests on.
- **Start from the side that simplifies.** When a derivation would add and subtract a term to turn one side into the other, start from the other side instead. Simplifying an expression needs only definitions and algebra, while adding and subtracting a term asks the reader to accept a step whose purpose shows only later. Another route is to solve a definition for the term you want: add μ_i to both sides of $\varepsilon_i = Y_i - \mu_i$, then simplify the right-hand side to Y_i, one operation per line. Adding a term to both sides of an equation is an ordinary step, not the pattern to avoid.

Definition 0.1 (Sample mean)**.** The **sample mean** of numbers $x_1, \dots, x_n$ is their average:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Exercise 0.1 (Deviations from the mean)**.** Show that

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0,$$

where $\bar{x}$ is the sample mean (Definition 0.1).

Solution

Solution 0.1.

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} && \text{(distribute the sum over the terms)} \\
&= \sum_{i=1}^{n} x_i - n\bar{x} && \text{(sum of a constant)} \\
&= n\bar{x} - n\bar{x} && (\textstyle\sum_i x_i = n\bar{x}, \text{ by the definition of } \bar{x}) \\
&= 0 && \text{(subtract)}
\end{aligned}$$

Theorem 0.1 (Deviations from the mean sum to zero)**.** *The deviations of numbers $x_1, \dots, x_n$ from their sample mean $\bar{x}$ (Definition 0.1) sum to zero:*

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

Proof

Proof. Exercise 0.1 derives this result. □

Example 0.1 (Deriving an equality from the side that simplifies)**.** Suppose an outcome Y_i has mean μ_i and deviation $\varepsilon_i := Y_i - \mu_i$, and a derivation needs $Y_i = \mu_i + \varepsilon_i$.

✗

$$\begin{aligned} Y_i &= Y_i - \mu_i + \mu_i && \text{(add and subtract } \mu_i\text{)} \\ &= \varepsilon_i + \mu_i && \text{(definition of } \varepsilon_i\text{)} \\ &= \mu_i + \varepsilon_i && \text{(reorder the terms)} \end{aligned}$$

✓

$$\begin{aligned} \mu_i + \varepsilon_i &= \mu_i + (Y_i - \mu_i) && \text{(definition of } \varepsilon_i\text{)} \\ &= \mu_i + Y_i - \mu_i && \text{(remove the parentheses)} \\ &= Y_i + \mu_i - \mu_i && \text{(reorder the terms)} \\ &= Y_i + (\mu_i - \mu_i) && \text{(group the last two terms)} \\ &= Y_i + 0 && (a - a = 0) \\ &= Y_i && (a + 0 = a) \end{aligned}$$

The ✗ version's first line adds a term the reader has no reason to expect. The ✓ version starts from $\mu_i + \varepsilon_i$, which a definition expands, and every line after that rewrites or simplifies what is already there. The ✓ version is longer because it gives a line to each step that a single "cancel" justification would hide (see One operation per step).

Example 0.2 (Splitting a theorem that states two results)**.**

✗ **Theorem.** Each outcome is its mean plus its deviation; given the covariates, each deviation is Gaussian with mean 0:

$$Y_i = \mu_i + \varepsilon_i, \qquad \varepsilon_i \mid X_i = x_i \sim \mathrm{N}(0, \sigma^2).$$

✓ **Theorem 1.** Each outcome is its mean plus its deviation:

$$Y_i = \mu_i + \varepsilon_i.$$

Theorem 2. Given the covariates, each deviation is Gaussian with mean 0:

$$\varepsilon_i \mid X_i = x_i \sim \mathrm{N}(0, \sigma^2).$$

The first result is algebra from the definition of ε_i; the second needs the model's distribution. In separate blocks, each gets the proof it needs, and a later step that uses only the first can cite only the first.