

Probability for Data Science

Last modified: 2026-09-28 23:45:35 (PDT)

Welcome

Probability theory is the branch of mathematics concerned with formalizing and quantifying uncertainty. It is the foundation on which statistical inference is built: before we can reason about what data tell us about the world, we need a precise language for describing random phenomena.

These notes collect the probability that data science courses assume. Some key results are listed here, organized by topic:

- [Notation](#)
- [Probability basics](#): sample spaces and events, probability axioms, conditional probability, and Bayes' theorem
- [Random variables](#): PMF, PDF, CDF, quantile function, and survival, hazard, and cumulative hazard functions
- [Independence](#): independence, conditional independence, and (conditionally) IID random variables
- [Expectation](#): expectation, LOTUS, conditional expectation, Fubini-Tonelli, and the law of iterated expectations
- [Variance and covariance](#): deviation, variance, conditional variance, homoskedasticity, and covariance
- [Key distributions](#): Bernoulli, Poisson, negative binomial, and Weibull distributions
- [Limit theorems](#): the Central Limit Theorem

Most of this material should be review from an introductory probability or mathematical statistics course (e.g., UC Davis's Epi 202). These notes began as the probability chapter of the Morrison Lab's *[Regression Models for Epidemiology](#)*, which applies them to regression and survival models (see [Additional resources](#)).

Using these notes in another site

Course sites include these notes as a git submodule named `pds` at the site's root, and include fragments with paths that start with `pds/`, for example `{{< include pds/_subfiles/_thm-bayes.qmd >}}`. This site includes its own fragments the same way, through a `pds` symlink that points at the repository root.

Quarto resolves `@id` cross-references only within one rendered page, so a host site that links to a result here uses an explicit link, `[text](expectation.qmd#thm-lotus)`.

Additional resources

- Miller (2017)
- Morrison Lab's *Regression Models for Epidemiology*, which applies this material to regression and survival analysis

References

Miller, Steven J. 2017. *The Probability Lifesaver: All the Tools You Need to Understand Chance*. A Princeton Lifesaver Study Guide. Princeton University Press. <https://doi.org/10.1515/9781400885381>.