Probability for Data Science

Last modified: 2026-11-07 21:34:35 (PST)

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, binomial, 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 link to these pages by URL; they do not include this repository as a git submodule. A host site that keeps a copy of this repository at its root, named `pds`, can still 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

Comparable courses

Several universities offer open-access courses with comparable or complementary coverage of probability for data science, computer science, and statistics:

- **Stanford University CS 109**: *Probability for Computer Scientists* (Chris Piech, Mehran Sahami, and David Varodayan) introduces counting, probability axioms, random variables, joint distributions, limit theorems, and machine learning applications.
- **Stanford University STATS 116**: *Theory of Probability* (John Duchi) provides a rigorous mathematical treatment of probability spaces, conditioning, random variables, transform methods, and limit theorems (see also lecture notes by Amir Dembo).
- **UC Berkeley Data 140 / Stat 140**: *Probability for Data Science* (Ani Adhikari and Jim Pitman) blends mathematical probability theory with Python simulations and computational data science applications.
- **Stanford University EE 278**: *Probability and Statistical Inference* (Ayfer Özgür and Tsachy Weissman) builds on probability for hypothesis testing, empirical risk minimization, concentration inequalities (Markov, Chebyshev, Hoeffding), Gaussian random vectors and minimum mean-square error estimation; its previous notes are handwritten lectures on these topics. The prerequisite probability course has open lecture notes by David Tse and Abbas El Gamal.
- **Harvard University Stat 110**: *Introduction to Probability* (Joseph K. Blitzstein; companion textbook Blitzstein and Hwang [1], online at probabilitybook.net) covers conditioning, distributions, transform methods, and Markov chains with an emphasis on intuitive storytelling.
- **MIT 18.05**: *Introduction to Probability and Statistics* (Jeremy Orloff and Jonathan Bloom, MIT OpenCourseWare) bridges probability, discrete and continuous random variables, and both Bayesian and frequentist statistical inference.

- **MIT 6.041 / 6.431**: *Probabilistic Systems Analysis and Applied Probability* (John Tsitsiklis, MIT OpenCourseWare) covers probability spaces, conditioning, independence, transform methods, and Bernoulli and Poisson processes.
- **Cal Poly / Kevin Ross**: *An Introduction to Probability and Simulation* (Ross [2]) develops probability concepts through simulation in R and Python, providing intuitive frameworks including conditioning as "slicing and renormalizing" and "taking out what is known" in conditional expectation.

Textbooks and companion notes

- Ross [3], an undergraduate calculus-based probability text
- Gut [4], a graduate text on the measure-theoretic foundations of probability, as a companion to Billingsley [5]
- Miller [6]
- Blitzstein and Hwang [1]
- Hastie et al. [7]
- James et al. [8]
- Amir Dembo and Kevin Ross, *Stochastic Processes* lecture notes (Dembo and Ross [9], Stanford MATH 136 / STAT 219)
- Amir Dembo, *Probability Theory* lecture notes (Stanford STAT 310 / MATH 230, the measure-theoretic sequel to this material; used in STATS 310B)
- John Duchi, *Lecture Notes on Statistics and Information Theory* (Stanford STATS 311), which develops entropy and divergences beyond the definitions here
- Morrison Lab's *Regression Models for Epidemiology*, which applies this material to regression and survival analysis

Video channels

The DATA 571 (Machine Learning) course notes keep a curated list of online video channels, with levels, for probability and statistics as well as machine learning, AI, computer science and mathematics.

References

[1] J. K. Blitzstein and J. Hwang, *Introduction to probability*, 2nd ed. Chapman; Hall/CRC, 2019. doi: 10.1201/9780429428357.

[2] K. Ross, "An introduction to probability and simulation." Bookdown, 2022. Available: https://bookdown.org/kevin_davisross/probsim-book/

[3] S. M. Ross, *A first course in probability*, 10th ed. Hoboken, NJ: Pearson, 2019.

[4] A. Gut, *Probability: A graduate course*, 2nd ed. in Springer texts in statistics. New York: Springer, 2013. doi: 10.1007/978-1-4614-4708-5.

[5] P. Billingsley, *Probability and measure*, 3rd ed. in Wiley series in probability and mathematical statistics. New York: Wiley, 1995.

[6] S. J. Miller, *The probability lifesaver: All the tools you need to understand chance*. in A princeton lifesaver study guide. Princeton: Princeton University Press, 2017. doi: 10.1515/9781400885381.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in R*, 2nd ed. Springer, 2021. doi: 10.1007/978-1-0716-1418-1.

[9] A. Dembo and K. Ross, "Stochastic processes." Stanford University, 2021. Available: https://adembo.su.domains/math-136/nnotes.pdf