Statistics for Data Science

Last modified: 2026-11-12 22:02:32 (PST)

These notes collect the statistics that data science courses assume.

Topics

The notes are organized in four groups of pages.

Describing data:

- Types of variables: numerical and categorical variables, scales of measurement, the taxonomy of data types, and binary and count random variables;
- Nonparametric models: the empirical CDF, order statistics, and sample quantiles;
- Exploratory data analysis: sample statistics, graphs, correlation, and contingency tables, illustrated with the WCGS data.

Inference:

- Estimation: estimands, estimators, bias, mean squared error, and standard error;
- Statistical inference: hypothesis tests, p-values, the t, chi-square, and F distributions, confidence intervals, and how to interpret negative findings;
- Maximum likelihood inference: likelihood, score, information, the asymptotic distribution of MLEs, Wald and likelihood ratio tests, and Newton-Raphson, with worked examples;
- Properties of estimators: consistency, relative efficiency, and the Cramér–Rao lower bound.
- Sufficiency and the Rao–Blackwell theorem: sufficient statistics, the factorization theorem, and improving an unbiased estimator.

Methods:

- Comparing means: t-tests and one-way ANOVA;
- Comparing proportions: the chi-square test and Fisher's exact test;

- Correlation and simple linear regression;
- The bootstrap;
- Permutation tests: exchangeability, the permutation distribution, and exact and Monte Carlo permutation p-values;
- Multiple testing: the family-wise error rate, the false discovery rate, the Bonferroni, Holm, and Benjamini–Hochberg procedures, and adjusted p-values;
- Validating predictions: overfitting, generalization error, training/validation/test splits, k-fold and leave-one-out cross-validation, and choosing a model by cross-validation;
- Classification and diagnostic tests: sensitivity, specificity, prevalence, and positive and negative predictive values.

Bayesian:

- Bayesian inference: the Bayesian paradigm, priors, posteriors, and hierarchical models;
- Markov chain Monte Carlo: Monte Carlo integration, the Metropolis–Hastings and Gibbs samplers, convergence diagnostics, and the deviance information criterion;
- Fitting models by Bayesian inference with JAGS: a proportion, logistic regression, survival, random effects, and Bayesian model averaging.

Comparable courses

Stanford's statistics department posts the materials of several courses that cover the same inference topics at a more mathematical level:

- **STATS 200**, *Introduction to Statistical Inference* (Zhou Fan, Autumn 2016): 29 lecture PDFs on testing, maximum likelihood, the bootstrap, Bayesian analysis and regression, following Rice's *Mathematical Statistics and Data Analysis*.
- **STATS 300A**, *Theory of Statistics* (Andrea Montanari, Autumn 2018; see also Lester Mackey's 2015 scribed notes): finite-sample optimality, exponential families, and testing theory.
- **STATS 300B**, *Theory of Statistics II* (John Duchi, Winter 2021): slides on asymptotic statistics and concentration.
- **STATS 305B**, *Methods for Applied Statistics* (Jonathan Taylor): contingency tables, logistic regression and generalized linear models.

Origin

These notes began as the statistics appendices of *Regression Models for Epidemiology*, and keep those chapters' file names, except that two long chapters are now split into several pages. Most results keep their rme `#id` anchors on the same page; the exceptions are:

- the maximum likelihood practice exercises (`exr-prac-*`), which moved from the estimation page to the maximum likelihood page;
- the null and alternative hypothesis definitions, which moved from basic statistical methods to statistical inference;
- the sample statistics (mean, median, variance, standard deviation, interquartile range, proportion, correlation) and the contingency table, which rme defined on two pages; they are now defined once, on the exploratory data analysis page, under the ids rme used on its basic statistical methods page (such as `def-sample-mean`), so the exploratory-descriptive ids from rme `def-mean`, `def-median`, `def-eda-variance`, `def-eda-sd`, `def-quantile`, `def-iqr`, and `def-correlation` are retired;
- the basic statistical methods sections on categorical outcomes, correlation and simple linear regression, and the bootstrap (such as `sec-two-group-categorical`, `sec-correlation`, `sec-simple-linear-regression`, and `sec-bootstrap-ci`), which moved to their own pages, with their ids;
- the Bayesian inference sections on computation (`sec-foundations` through `sec-dic`) and the worked examples (`sec-bayes-examples` and the sections after it), which moved to the MCMC and JAGS pages, with their ids.

Where these notes rely on probability or calculus, they link to the lab's notes on probability and mathematics for data science, which are the canonical home for the probability and mathematics appendices of rme.

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 `sds`, can still include fragments with paths that start with `sds/`, for example `{{< include sds/_subfiles/intro-MLEs/_def_mle.qmd >}}`. This site includes its own fragments the same way, through an `sds` 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, such as `[text](estimation.qmd#def-bias)`.