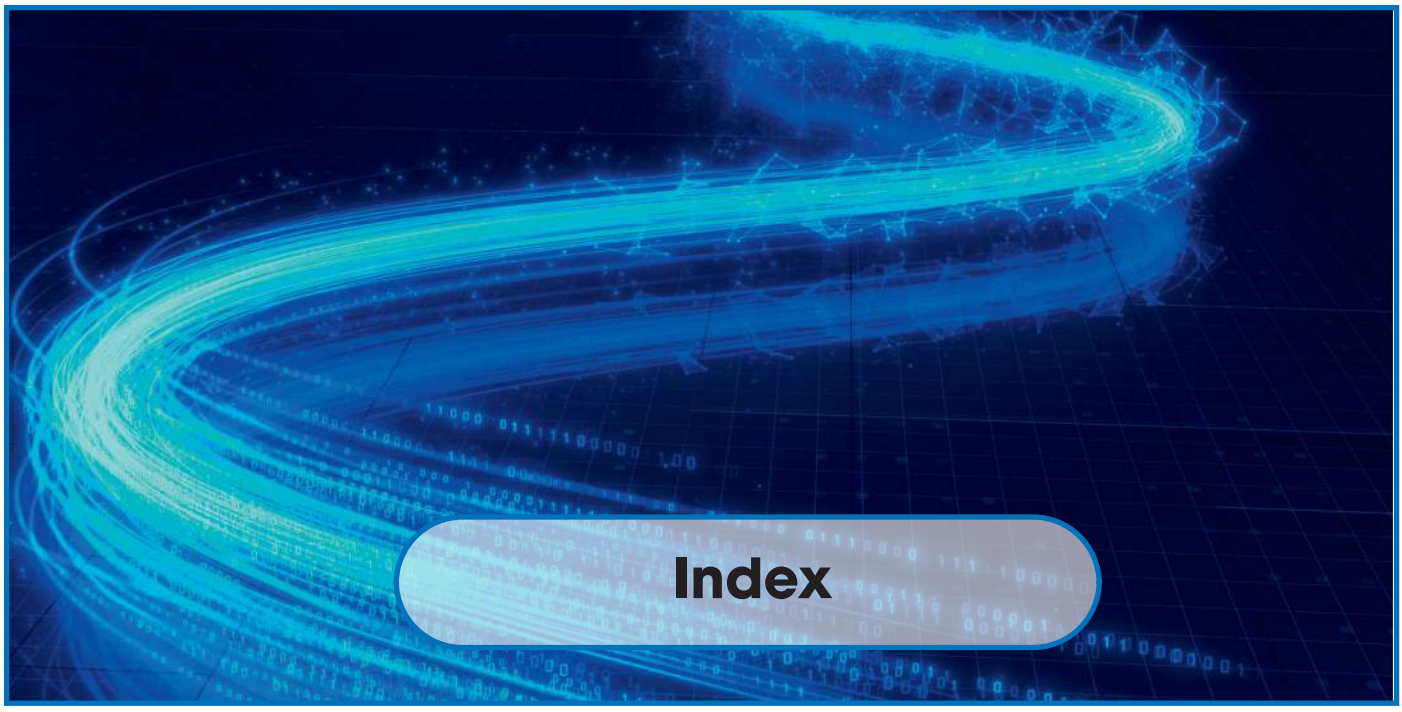
CAMBRIDGE

Cambridge University Press & Assessment
978-1-107-11304-6 — Introduction to Probability and Statistics for Data Science
Steven E. Rigdon, Ronald D. Fricker, Jr, Douglas C. Montgomery
Index
More Information



additive law of probability, 58
adjusted R^2, 561, 563
alternative hypothesis, 350
analysis of variance
    F statistic, 415
    partition of the total sum of squares, multiple regression, 535
    for simple linear regression, 526
    table for simple linear regression, 529
approximate pivotal quantity, 312
approximating maximum likelihood estimators, 289
asymptotic normality of the MLE, 646
autocorrelation, 50
autocorrelation plots, 50
autoregressive integrated moving average (ARIMA) models, 635

bar charts, 32, 33
Bayes' Theorem, 74, 575
Bayesian credible interval, 582
Bayesian methods in statistics, 574
Bayesian paradoxes, 585, 586
Benjamini–Hochberg method, 787
Bernoulli random variable, 85
Bernoulli trials, 132
beta distribution, 202
bias, 235, 237
binomial distribution, 132, 133
bivariate normal distribution, 186
blocking, 439, 443
Bonferroni's inequality, 780, 789
Boole's inequality, 780, 782
bootstrap, 656, 662
    confidence intervals, 667
box plots, 40, 42

categorical data, 9
    analysis, 459
causation and correlation, 21
centered moving average, 621
Central Limit Theorem, 255
chi-square distribution, 196
clinical trial, 231
cluster sampling, 236, 338
combination, 69
comparative box plots, 42
complement of an event, 56
complementary events, 60
conditional distributions, 111, 113, 121
conditional expectation, 122
conditional probability, 62
conditioning on a random variable, 123
confidence interval, 303, 307, 312
    on the intercept, simple linear regression, 547
    for differences of proportions, 317
    for differences in means, 319
    for mean response, simple linear regression, 522
    for paired data, 324
    on the slope, simple linear regression, 512
    from survey data, 330
confounding, 231
conjugate priors, 588
consistent estimator, 278
constant time series process, 628
contingency table, 459
    testing homogeneity, 471
    testing independence, 465
continuity correction, 479
continuous data, 8
continuous random variable, 93
contrasts, 794
convenience sampling, 234
convergence in distribution, 257

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-11304-6 — Introduction to Probability and Statistics for Data Science
Steven E. Rigdon, Ronald D. Fricker, Jr, Douglas C. Montgomery
Index
More Information

convergence in probability, 267, 279
correlation, 109, 551, 553, 555
counting techniques, 66
covariance, 108
covariates, 495
critical region, 352
cross-sectional data, 8
cross-validation, 754, 755, 760, 765
  with classification data, 765
crossing, 443
cumulative distribution function, 86, 97

data science cycle, 230
data science, 2
DeMorgan's Laws, 71
dependent events, 62
descriptive statistics, 5, 7
design of experiments, 430
discrete data, 8
discrete random variable, 82
discrete sample space, 55
discrimination or classification, 714, 715
distribution of the estimator of the intercept, 505
distribution of the estimator of the slope, 504
distributive law, 71
double exponential smoothing, 642

empirical rule, 186
equivalence of least squares and maximum likelihood in linear regression, 502
errors of coverage, 239
estimator, 273
event, 55
expected mean squares, 416
expected value and variance properties, 102
expected value of the error sum of squares, 508
experimental data, 231
experimental design principles, 442
exponential distribution, 173

F distribution, 201
factorials, 67
factorization theorem, 295
false discovery proportion, 785
false discovery rate, 787
family-wise error rate, 776
first-order autoregressive process, 637
Fisher's exact test, 473
fractional factorial designs, 436
frequentist methods in statistics, 574

gamma distribution, 189, 190
geometric distribution, 141
Gibbs sampling, 604
goodness-of-fit tests, 460

Hanning filter, 621
hazard function, 174
hierarchical Bayes models, 606
hinges, 25
histogram, 35, 36
Holm's method, 783
hypergeometric distribution, 149
hyperparameters, 607
hypothesis, 350
hypothesis testing, 348
  comparing two means, 386
  comparing two proportions, 398
  comparing two variances, 402
  on the mean, variance known, 357
  on the mean, variance unknown, 364
  with paired samples, 404
  on a proportion, 370
  on the variance, 373

independent events, 59
independent random variables, 109
indicator variable, 9
indicator variables in regression, 539
inference on a linear combination of regression model parameters, 546
inferential statistics, 5
intercept estimator in simple linear regression, 499
invariance property of maximum likelihood estimators, 289

jackknife, 656
Jeffreys prior for the binomial, 597
Jeffreys prior for the normal distribution, 598
joint distributions, 105

k-fold cross-validation, 762

lag *k* autocorrelation coefficient, 624
large sample confidence interval, 307
lattice (trellis) plots, 38
law of total probability, 71
least absolute deviations criterion, 501
least squares estimator, multiple regression, 533
least squares function, 497
  multiple regression, 533
least squares normal equations, 498
leave-one-out cross-validation, 760
likelihood principle, 297
likelihood ratio tests, 374
linear combinations of random variables, 244, 245, 246
linear data smoother, 621
linear discriminant function, 727
linear trend time series process, 629
link functions, 707
Ljung–Box goodness-of-fit statistic, 625
logistic function, 685
logistic regression model, 684, 740
longitudinal (time series) data, 8
longitudinal data graphics, 46

marginal distributions, 106
matrix expression of the regression model, 517
matrix plot, 538

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-11304-6 — Introduction to Probability and Statistics for Data Science
Steven E. Rigdon, Ronald D. Fricker, Jr, Douglas C. Montgomery
Index
More Information

maximum Likelihood, 283
McNemar's test, 481
mean, 9
measurement error, 5, 240
measures of central tendency, 9, 10, 13, 15
measures of location, 9
median, 13
memoryless property of exponential random variables, 178
method of moments, 280
Metropolis–Hastings algorithm, 601, 603
mixed autoregressive moving average (ARMA) process, 638
mode, 15
model adequacy checking in regression, 565–567
model matrix, 517
model matrix, multiple regression, 532
moment-generating functions, 209, 211, 212, 214
Monte Carlo Markov chain (MCMC), 601
moving average (MA) processes, 636
moving median of span N, 621
multinomial distribution, 160
multinomial logistic regression, 698
multiple linear regression model, 531
multiplicative law of probability, 63, 64
multiplicative seasonal exponential smoothing time series model, 633
multivariate normal distribution, 186, 188
mutually exclusive events, 56

negative binomial distribution, 146
nominal qualitative data, 8
noncentral chi-square distribution, 444
noncentral F distribution, 445
noncentral t distribution, 444
noninformative priors, 596
nonresponse error, 240
normal approximation to the binomial, 259
normal distribution, 180
normal probability plot of effects, 435
normal probability plots of residuals, 566
N-span moving average, 619
null hypothesis, 350

objective probability, 64
observational data, 232
observational study, 430
observed Fisher information matrix, 652
odds ratio (OR), 695
one-way table, 25
ordinal qualitative data, 8
overfitting, 751

parameter, 85, 226
parametric bootstrap, 670
partial autocorrelation function, 625
patterns on residual plots, 568
percentiles, 23
permutation, 67
pie charts, 35
pivotal quantity, 312
Poisson distribution, 154
Poisson regression, 703
polynomial regression models, 543
  building, 544
pooled variance estimate, 388
population, 4, 226
  mean, 9
  variance, 17
posterior distribution, 577, 581, 719
posterior mean, 582
power, 354, 360
power for one-factor ANOVA, 448
power for the two-sample t-test, 445
predicted value from a regression model, 497
prediction interval for the response, simple linear regression, 525
prediction vs. explanation in regression, 493
prior distribution, 577, 581, 719
probability, 54
  axioms, 55, 56
  measure, 56
probability density function, 93
probability mass function, 82, 83
p-value, 357, 359

quadratic discriminant analysis, 737
qualitative data, 8
quantile–quantile (QQ) plots, 215, 435
quantiles, 24, 176
quantitative data, 8
quartile, 24

R statistical software, 6
R^2, 561
random sample, 4, 227
random variable, 82
randomization, 236
  in designed experiments, 442
range, 18
regression trees, 707
regressor or predictor variables, 495
rejection region, 352
relative risk, 697
repeated cross-sectional plots, 48
replication, 230, 443
residual, 497
residual plots, 566–568
response variable, 495

sample, 4, 226
  autocorrelation function, 625
  correlation, 20
  covariance, 19
  mean, 10
  median, 13
  point method for determining probabilities, 64
  random, 4
  size for confidence intervals, 326
  space, 54

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-11304-6 — Introduction to Probability and Statistics for Data Science
Steven E. Rigdon, Ronald D. Fricker, Jr, Douglas C. Montgomery
Index
More Information

sample (*cont.*)
standard deviation, 16
variance, 16
sampling distribution, 244, 250, 253, 263
for proportions, 263
sampling error, 5, 240
sampling frame, 236
scatterplot, 43, 45, 46
scatterplot matrix, 46
Scheffé's method, 794
scientific method, 228
second-order autoregressive process, 637
second-order exponential smoothing, 630
secondary data sources, 339
significance level, 352
simple exponential smoothing, 628
simple linear regression model, 496
simple random sampling, 236
Simpson's paradox, 485
simulating from distributions, 207
simulation methods for Bayesian statistics, 599
simultaneous confidence intervals, 789
simultaneous confidence region for simple linear regression model parameters, 519
single-factor analysis of variance (ANOVA), 409
slope estimator in simple linear regression, 499
small sample confidence intervals, 312
standard deviation, 91, 101
sample, 17
standard error of an estimator, 309, 537, 645
standard normal distribution, 181
standardizing, 183
stationary time series, 623
statistics, 2, 5, 226
stratified random sampling, 236, 333
studentized range, 791
subjective probability, 64
sufficient statistics, 295
sum of squares decomposition, 412, 413
survey error, 239
survival function, 193

t distribution, 198
Tchebysheff's theorem, 265
test statistic, 352
test statistics for the slope and intercept, simple linear regression, 515
testing multiple hypotheses, 775
time series plots, 47, 618, 623
tolerance interval, 303, 306
training data, 755
transformations, 205
trimmed mean, 14
Tukey's method, 586, 791, 792
two-factor analysis of variance, 425
two-level factorial design, 431
two-way table, 27
Type I error, 350, 351
Type II error, 350, 351

unbiased estimator, 275
uniform distribution, 93, 170
union and intersection, 55, 56
union of events, 57

validation data set, 755
variance, 91, 101
variogram, 625

weak law of large numbers, 266
Weibull distribution, 189, 192, 194
whiskers on box blots, 40
Winters' method for seasonal time series, 631

© in this web service Cambridge University Press & Assessment www.cambridge.org