Index

adjacency matrix, 250
alignment, sequence, 264
alternative hypothesis, 14, 145
analysis of variance (ANOVA), 204, 340
Anna Karenina principle, xviii
annotation data, **see also** metadata; supplementary information, 60, 198, 253
ANOVA (analysis of variance), 204

balanced design, 197, 348
banking, 69
batch effects, 126, 200, 211, 225, 340
Bayesian paradigm, 38
Bernoulli distribution, 3
beta distribution, 39
beta–binomial mixture, 102
between-groups sum of squares (BSS), 124
bias, 130, 195, 326, 339
bimodality, 49, 65, 86
binomial distribution, 4, 39
  and maximum likelihood, 25
Bioconductor, 17
biological replicates, 341
bipartite graph, 252
biplot, 179, 186, 229
bivariate distribution, 87
blocking (experimental design), 345
blocking factor, 209, 339
Bonferroni correction, 152
bootstrap, 93, 127, 263
  cluster validation using, 127
  nonparametric, 93
breakdown point, 205, 223
BSS (between-groups sum of squares), 124

Calinski–Harabasz index, 124
canonical correlation analysis (CCA), 239
canonical correspondence analysis, **see** constrained correspondence analysis
categorical variables, 2
  faceting by, 70
  multiple, 32, 229
CCA (canonical correlation analysis), 239
CCpnA (constrained correspondence analysis), 242
CDF (cumulative distribution function), 7, 65
CDs (clusters of differentiation), 115
centering, 166, 220
chain rule, 66
chi-squared distance, 229
chi-squared distribution, 15, 29
chi-squared statistic, 30, 232
ChIP-Seq, 90, 191
chunking, 361
classification, **see also** clustering, 140, 295, 310, 355
CLIP-Seq, 191, 216
clustering, 107
  baseline frequencies and, 130
  validation, 123
clusters of differentiation (CDs), 115
co-occurrence, 111, 229, 273
codon, 50
codon bias, 50
complementary events, 4, 8
complete linkage, 121
conditional probability, 7
confirmatory data analysis, xviii
confounding, 225, 227, 342
confusion table, 324
conservative testing approach, 21, 146, 214
constrained correspondence analysis (CCpnA), 242
contingency table, 32, 229
controlled experiment, 338
Cook's distance, 214
correlation, 164, 179, 238
correlation coefficient, 164, 166
  multitable version, 238
correspondence analysis (CA), 232
cost function, 325
count data, 27, 90, 96, 153, 193
count table, 193
covariance, 238
cross-validation, 321
cumulative distribution function (CDF), 7, 65
curse of dimensionality, 119, 323

Darwin, Charles R., 92, 249, 345
data augmentation, 87
data representation, 62, 357
data summarization, 164, 360
data transformation, 66, 78, 99, 117, 166, 212, 252, 351
  variance stabilizing, 99, 117, 351
dbscan, 118
deduction, 10, 20
delta method, 101
dendrogram, 76, 120
dense graph, 251
density-reachability, 119
dependencies, 37, 44, 149, 231, 289, 304, 353
  levels of replication and, 149
  sequential, 37
  spatial, 289, 304
design matrix, 203
diagnostic plots, 153, 157, 223, 329
  p-value histogram, 153
differential expression analysis, 198, 254
dimension reduction, 167, 217
directed graph, 250
discrete event, 2

discrete probability distribution, 2
discriminant analysis
  quadratic (QDA), 318
discriminant analysis, linear (LDA), 312
dispersion, 98, 194, 201, 211
dissimilarity, **see** distances
distances, 110, 120, 310
  correlation between, 238
  representing, 252
DNA-Seq, 191
dual scaling, 232
dynamic range, 100, 110, 166, 193, 286

ECDF (empirical cumulative distribution function), 30, 65, 92
EDA (exploratory data analysis), xviii, 199
effect size, 342, 346, 349
efficiency through design, 345
efficient computation, 360
eigen-decomposition, 180
elastic net, 327
EM algorithm, 86
empirical Bayes, 155, 211
empirical cumulative distribution function (ECDF), 30, 65, 92
epigenetics, 44
epitope detection, 6
ERGM (exponential random graph model), 275
error, 146, 151, 339
Escherichia coli, 43
evolutionary analysis, **see** phylogenetic tree
exchangeability, 3
expectation–maximization (EM) algorithm, 86
expected value, 12, 154
experimental design, 338
  multifactorial, 202
ExperimentHub, 327
exploratory data analysis (EDA), xviii, 199
exponential distribution, 44, 260
exponential random graph model (ERGM), 275
extreme value analysis, 8

faceting, 70
factor, **see also** categorical variables, 3
false discovery rate (FDR), 153
  local (fdr), 155, 159
  tail area (Fdr), 156
false negative rate (FNR), 146
false positive rate (FPR), 146
family-wise error rate (FWER), 151
feature extraction, image, 296
filtering operation, 137, 266
filtering, image, 286
finite mixture, 84
Fisher's exact test, 255
Fisher, Ronald A., xviii, 92, 337, 345
fitness for purpose, 352
flow cytometry, 115
fragment, 192
Friedman–Rafsky test, 271

Galton, Francis, 92
gamma distribution, 96
gamma–exponential distribution, 102
gamma–Poisson distribution, 97, 102, 195, 208
gap statistic, 125
Gene Ontology, 254
gene trees, 259
generalized linear model, 25, 208, 326
generative model, 10
generator matrix, 260
genetic code, 50
genetic screens, 191
goodness of fit, 21, 29, 35
  visual evaluation, 21, 29, 35, 97
grammar of graphics, 57
graph, 250
graph layout, 251, 269
graphics, 53
  faceting, 70
  grammar of graphics, 57
  heatmap, 75
  interactive, 72, 229, 267
GSEA (gene set enrichment analysis), 254

haplotype estimation, 38
Hardy–Weinberg equilibrium, 34
HARKing, 150
heatmap, 75, 109
heteroscedasticity, 193
heteroscedasticity, 100
hidden Markov models, 354
hierarchical clustering, 110, 120, 165, 259
hierarchical model, 38, 95
Hotelling's weighting method, 343
hotspot, 257
hypergeometric test, 255
hypothesis switching, 150
hypothesis testing, xviii, 11, 141, 255, 266, 271
  multiple, 150, 266
hypothesis weighting, 157, 354

identifiability, 87, 260
image data, 279
indel, 132, 264
independent filtering, 157
independent hypothesis weighting, 157
inertia, 170, 237
inference, 20
infinite mixture, 94, 102
interactive graphics, 72, 229, 267
intercept, 202

Jaccard distance, 111
Jaccard index, 111, 325
jitter, 269

k-means, 113
k-medoids, 113
kernel method, 310
kernels, 235

Laplace distribution, 95
large-*p* small-*n* problem, xviii
lasso, 95, 327
latent variable, 87, 233, 340, 352
layers (graphics), 60
LDA (linear discriminant analysis), 312
least absolute deviations, 205
least quantile of squares (LQS), 206
least sum of squares, 204
least trimmed sum of squares (LTS), 206
levels, 2
likelihood function, 22
linear combination, 170
linear discriminant analysis (LDA), 312
linear model, 204
linear regression, 168
linear technique, 167
loadings, 170, 177
local false discovery rate (fdr), 155, 159
log-likelihood ratio, 47
logistic regression, 207, 326
long format (data), 358
longitudinal data, 353

M-estimation, 205
Mantel coefficient, 238
MAP (maximum a posteriori) estimate, 41
marginal distribution, 40
marginal likelihood, 88
Markov chain, 37, 44, 253, 260
mass cytometry, 115
mass function, 2
mass spectroscopy, 163, 184, 360
matched design, 348
matrix, 12, 162
  rank of, 171
  singular value decomposition, 171
  trace of, 180
matrix association, 238
maximum a posteriori (MAP) estimate, 41
maximum jump, 121
maximum likelihood, 25, 207
maximum likelihood estimator (MLE), 21
MDS (multidimensional scaling), 218
mean squared error (MSE), 325
mean–variance relationship, 351
meta-analysis, 339
metadata, **see also** annotation data; supplementary information, 183, 197, 224, 265, 359
method hacking, 334
microbiome data, 137, 242, 274, 275, 327
minimal jump, 121
minimum spanning tree (MST), 268
misclassification rate, 320, 324
mixture model, 84
MLE (maximum likelihood estimator), 21
model averaging, 90
model complexity, 331
Monte Carlo (simulation), 9
Monte Carlo integration, 40
MSE (mean squared error), 325
MST (minimum spanning tree), 268
multidimensional scaling (MDS), 218
  non-metric, 223
multifactorial design, 202
multinomial distribution, 11, 27, 37
multiple testing, 150, 266
  hierarchical, 266
multitable correlation coefficient, 238

negative binomial distribution, 97
network, 251
noise, 129, 203, 287, 339
non-metric multidimensional scaling (NMDS), 223
nonparametric bootstrap, 93
nonparametric methods, 93
norm, 172
normalization, 193, 195
null distribution, 13, 144
null hypothesis, 12, 144, 201
null model, 20

objective, 310
objective function, 325
observational study, 338
Occam's razor, 141
one-sided test, 146
ordination, 233
orthogonal, 167
orthonormal, 175
OTU (operational taxonomic unit), 130, 133, 163, 258
out-of-memory data, 361
outlier, 205, 214, 223
overfitting, 309, 321

p-value, 15
p-value hacking, 150
p-value histogram, 153
paired design, 348
pairing (experimental design), 345
pairs plot, 165, 317
PAM (partitioning around medoids), 113
parallelization, 361
parameter, 4, 20
parameter tuning, 333
partitioning methods, 110, 113
PCA (principal component analysis), 167
penalization, 95, 205, 240, 326
performance, 331
permutation test, 149
  Mantel, 238
phylogenetic tree, 258
plotly, 73
point mass, 92
Poisson distribution, 5
Poisson process, 303
position weight matrix, 36
position-specific scoring matrix, 36
posterior distribution, 38
power, 11, 146, 346, 349
precision-recall curves, 324
predictor, 168, 309
preliminary data, 338
premature optimization, 360
preprocessing, 116, 136, 166, 184, 263
principal component, 169, 175, 180
principal component analysis (PCA), 167
  scatterplot, 69
principal coordinates analysis (PCoA), 221
principal plane, 177
prior distribution, 38
probabilistic model, 2
probability density function, 16, 85, 92
probability mass distribution, 5
probability model, 2, 10
projection, 167
prospective trial, 339
pseudocounts, 212

QDA (quadratic discriminant analysis), 318
QQ-plot, 29
quadratic discriminant analysis (QDA), 318
quality metric, 353
quantile, 29
quantile–quantile (QQ) plot, 29

randomization, 348
randomized controlled trial, 338
rank of a matrix, 171, 174
rank statistic, 8
rare event, 2, 8, 132, 194, 325
read, 192
recall, 324
receiver operating characteristic (ROC), 324
rectangular gating, 116
recursive partitioning, 120
regression, 25, 310
regression lines, 168
regularization, 326
regularized logarithm (rlog), 200, 212
rejection region, 142, 144
residuals, 25, 203
response, 168, 310
ridge regression, 95, 327
Ripley's K function, 305
risk function, 325
rlog (regularized logarithm), 200, 212
Rmarkdown, 355
RNA-Seq, 152, 163, 191, 194
robustness, 205, 223

ROC (receiver operating characteristic), 324
rootogram, 21, 97
RStudio, xxii, 355

sampling depth, 215
sampling distribution, 13, 92, 141, 263
sampling without replacement, 195
scaling, 166, 178
segmentation, image, 287, 291, 294
sensitivity, 146, 324
sequence logo, 36
sequencing library, 192
sequential design, 363
shrinkage estimation, 211
significance level, 9, 145
silhouette index, 135
single linkage, 121
single-cell data, 115, 228, 234
singular value, 172
singular value decomposition (SVD), 171, 175
slots, 281
sparsity, 163, 241, 251
spatial dependencies, 289, 304
spatial point process, 300
specificity, 146, 324
standardizing, 166
statistical learning, 309
statistical model, 10
  versus probabilistic model, 20
status of a variable, 108, 167
study, observational, 338
sufficiency, 3, 145, 360
supervised learning, 168, 309
supplementary information, **see also** annotation data; metadata; 182, 183, 227, 244, 265

technical replicates, 149, 341
temps perdu, 233
test statistic, 12, 144
tibble, 65, 218
tidy data, 357
trace of a matrix, 180
training data, 309
transformation, image, 284
true negative rate, 146, 324
true positive rate, 11, 146, 324
two-sided test, 146

undirected graph, 270
unsupervised learning, 108, 167

variability, 12, 92, 102, 204, 340, 349
variance, 326
variance-stabilizing transformation, 99, 117, 212
vectorization, 22, 361
vignette, xxii, 43
visualization, **see** graphics
Voronoi tessellation, 292

Ward's method, 122
wide format (data), 357
within-groups sum of squares (WSS), 123
workflow, 144, 335, 360
WSS (within-groups sum of squares), 123

zero-inflated counts, 90