General Index

SYMBOLS

R^2, 350
z-score, 113, 157, 251, 253, 269

A

absolute value, 73
addition rule, 200
adjacency matrix, 365
adjusted R^2, 176, 350
age-specific death rate, 31
age-specific fertility rate, 29
alternative hypothesis, 315
AND, 39
animation, 398
association, 50
asymptotic theorems, 264
average treatment effect, 50
axioms, 200

B

bag-of-words, 406
bar plot, 87
Bayes' rule, 227
Bayesian, 198
before-and-after design, 63
Bernoulli random variable, 241
betweenness, 372, 378
bias, 135, 278
bin, 89
binary random variable, 241
binary variable, 14, 37
binomial distribution, 246
binomial theorem, 248
birthday problem, 203
box plot, 92
butterfly ballot, 164

C

categorical variable, 42
causal effects, 47
causal inference, 47
ceiling effects, 104
census, 96
centering, 117
central limit theorem, 266, 290, 306, 330, 348, 351
centrality, 368
centroid, 117
ceteris paribus, 171
classification, 139
classification error, 417
closeness, 370, 378
clustering algorithms, 117
clusters, 117
coefficient of determination, 160, 176, 350
coefficients, 148
combinations, 208
complement, 201
complete randomization, 279, 308
computational revolution, 1
conditional cash transfer program, 193
conditional expectation, 337
conditional expectation function, 338
conditional independence, 222, 235
conditional probability, 213, 215
confidence bands, 291
confidence interval, 291
confidence level, 291
confounders, 60, 387
confounding bias, 60
confusion matrix, 139
consistent, 278
contingency table, 19
continuous random variable, 241, 242
control group, 50, 57
correlation, 112, 147
correlation coefficient, 112
cosine similarity, 425
counterfactual, 47
covariance, 351
coverage probability, 293
critical value, 291, 306
cross-section comparison design, 56
cross-section data, 63
cross-tabulation, 19
crude birth rate, 28
crude death rate, 30

cumulative distribution function (CDF), 242, 243
cumulative sum, 265

D
data revolution, 1
data-generating process, 148, 206, 281, 336
decile, 71
degree, 368, 376
degrees of freedom, 176, 288, 305
density, 89, 243
descriptive statistics, 67
dichotomous, 14
difference-in-differences, 64
difference-in-means estimator, 50, 169, 279
directed network, 366, 375
discrete random variable, 241
dissimilarity index, 423
disturbance, 148
document frequency, 409
document–term matrix, 411
dot product, 426
dummy variable, 14
DW-NOMINATE scores, 106

E
ecological inference, 361
edges, 367
Electoral College, 128
error, 148
error bands, 291
estimation error, 277
estimator, 276
event, 199
exogeneity, 337
expectation, 256, 278
experiment, 199
experimental data, 33
exploratory data analysis, 364
external validity, 51, 56, 189

F
factor, 42
factor variable, 42, 87
factorial, 202
factorial variable, 42, 87
false discoveries, 327
false discovery rate, 232
false negative, 139
false positive, 139, 232
farness, 370
file drawer bias, 359
first moment, 258
first quartile, 70
Fisher's exact test, 315
fitted value, 148
floor effects, 104
frequentist, 197
function, 10
fundamental problem of causal inference, 48, 308

G
Gaussian distribution, 249
get-out-the-vote, 51
Gini coefficient, 109
Gini index, 109
Google, 381
graph, 367
graph strength, 421

H
Hawthorne effect, 52, 55
heterogeneous treatment effects, 177
heteroskedastic, 348
heteroskedasticity-robust standard errors, 348
histogram, 89, 136
homoskedasticity, 346
hypothesis testing, 307

I
i.i.d., 246
ideology, 105
idf, 409
if qualifier, 38
immutable characteristics, 49
in-sample prediction, 167, 417
indegree, 377
independence, 218
independently and identically distributed, 246
indicator, 171
indicator function, 263
Institutional Review Board, 103
integration, 257
interaction effect, 178
intercept, 148
internal validity, 51, 56, 188, 189
interquartile range (IQR), 70
inverse document frequency, 409
inverse function, 187
item count technique, 103
item nonresponse, 101
item response theory, 106
iterations, 131
iterative algorithm, 117

J
joint independence, 222
joint probability, 214

K
Kish grid, 99

L
large sample theorems, 264
law of iterated expectation, 345

law of large numbers, 264, 278, 281
law of total probability, 201, 211, 214, 222
law of total variance, 346
least squares, 151
leave-one-out cross validation, 417
level of test, 314
limit, 197
linear model, 148
linear regression, 144
linear relationship, 147
list experiment, 103
logarithmic transformation, 99, 204
logical conjunction, 39
logical disjunction, 39
logical operators, 39
longitudinal data, 63
longitudinal study, 75
loop, 131, 418
Lorenz curve, 109
lower quartile, 70

M
macros, 129
maps, 386
margin of error, 296
marginal probability, 212
matrices, 219
mean-squared-error, 286
measurement models, 105
median, 67, 91
misclassification, 139
misreporting, 102
Monte Carlo error, 207, 284
Monte Carlo simulation, 205, 225, 245, 260, 265, 281
Monty Hall problem, 224
moving average, 191
multiple testing, 327
multistage cluster sampling, 98

N
natural experiment, 79, 386, 387
natural logarithm, 99
nearness, 370
network data, 364
network density, 421
no omitted variables, 339
nodes, 367
nonlinear relationship, 148
nonresponse, 278
normal distribution, 249, 266
null hypothesis, 313
numeric variable, 89

O
observational studies, 56, 338
one-sample *t*-test, 319
one-sample *z*-test, 319
one-sample tests, 316
one-sided *p*-values, 315
one-tailed *p*-values, 315
OR, 39
out-of-sample prediction, 167, 417
outcome variable, 33
outdegree, 377
outliers, 67, 122, 164
overfitting, 167, 417

P
packages, 23
PageRank, 381
panel data, 63, 141
parameter, 276
Pascal's triangle, 248
percentile, 71
permutations, 202
person-year, 28
placebo test, 188
political polarization, 109
polity score, 79
population average treatment effect, 280
population mean, 256
positive predictive value, 227
posterior probability, 227
potential outcomes, 48
power, 329
power analysis, 329
power function, 332
predicted value, 148
prediction error, 135, 148
pretreatment variables, 55, 60
prior probability, 227
probability, 197
probability density function (PDF), 243
probability distributions, 241
probability mass function (PMF), 242
probability model, 241
probability sampling, 96
Progresa, 193
proof by contradiction, 313
publication bias, 327, 359

Q
Q–Q plot, 115, 123, 253, 306
quadratic function, 182
quantile–quantile plot, 115, 123, 253, 306
quantile treatment effects, 76
quantiles, 67, 71, 115
quartiles, 70
quincunx, 267
quintile, 71
quota sampling, 97

R
random digit dialing, 98
random variables, 241

randomization inference, 313
randomized controlled trials, 49, 279, 338
randomized experiments, 49
randomized response technique, 104
rational number, 313
receiver, 366
reference distribution, 313
regression discontinuity design, 185
regression line, 148
regression toward the mean, 154, 253
relational operators, 38
representative, 97
residual, 148, 171
residual plot, 163
residuals, 253
root mean square (RMS), 72, 136, 152
root-mean-squared error, 136, 152, 286, 350
rule of thumb, 296

S
sample average treatment effect, 50, 279
sample average treatment effect for the treated, 65
sample correlation, 351
sample mean, 59, 256
sample selection bias, 51, 97
sample size calculation, 297
sample space, 199
sampling distribution, 278, 287, 313
sampling frame, 97, 98, 101
sampling variability, 260
sampling with replacement, 205
sampling without replacement, 97, 206
scalar, 59, 129
scaling, 117
scatterplot, 106, 145
scientific significance, 316, 320
scraping, 401
second moment, 258
second quartile, 70
selection bias, 61
selection on observables, 339
sender, 366
set, 199
sharp null hypothesis, 313
simple random sampling, 96, 206, 277
simple randomization, 279, 308
simulation, 205
slope, 148
social desirability bias, 27, 102
sparse, 411
spatial data, 386
spatial point data, 386
spatial polygon data, 386, 389
spatial voting, 105
spatial–temporal data, 386
standard deviation, 72, 73, 258
standard error, 287
standard normal distribution, 250, 253, 281
standardize, 117
standardized residuals, 253
statistical control, 61
statistical significance, 316, 320
step function, 247
Student's t-distribution, 304
Student's t-test, 333, 335
subclassification, 61
sum of squared residuals, 151, 171
supervised learning, 121, 407
support, 257
survey, 82
survey sampling, 96

T
tercile, 71
term frequency, 404, 406, 409
term frequency-inverse document frequency, 409
term–document matrix, 411
test statistic, 313
tf, 404
tf–idf, 409
third quartile, 70
time trend, 64
time-series, 141
time-series operators, 141
time-series plot, 108, 143
topics, 406
total fertility rate, 29
total sum of squares, 160
treatment, 48
treatment group, 50, 56
treatment variable, 33, 48
true positive rate, 227, 231
true positives, 231
two-sample t-test, 323
two-sample z-test, 322
two-sample tests, 316
two-sided p-value, 315, 317
two-tailed p-value, 315
type I error, 314
type II error, 314, 329

U
unbiased, 136, 278
unconfoundedness, 339
uncorrelated, 337
undirected network, 366, 375
uniform random variable, 242
unit nonresponse, 101, 301
unobserved confounders, 338
unsupervised learning, 121, 407
upper quartile, 70

V
value labels, 16

variable labels, 16
variables, 10
variance, 74, 258
Venn diagram, 200, 201
vertices, 367

W
weighted average, 236
with replacement, 97
word cloud, 406
working directory, 20