

# Index

correlation.py, 116  
age\_lm.py, 121  
agemodel.py, 121  
brfss.py, 48, 114  
brfss\_corr.py, 114, 116, 118  
brfss\_figs.py, 48  
brfss\_scatter.py, 113, 114  
Cdf.py, 30, 31, 34  
chi.py, 88  
class\_size.py, 26  
conditional.py, 22  
correlation.py, 108, 110, 114, 118  
decay.py, 101  
descriptive.py, 19  
erf.py, 43  
estimate.py, 100, 101  
first.py, 7, 8, 12  
gini.py, 69  
hypothesis.py, 81, 83, 85  
hypothesis\_analytic.py, 90  
irs.py, 49  
locomotive.py, 104  
myplot.py, 15, 31  
Pmf.py, 14, 16, 18, 77  
populations.py, 48  
populations\_cdf.py, 48  
pyplot, 15  
rankit.py, 46  
relay.py, 27, 46  
relay\_cdf.py, 31  
relay\_normal.py, 46  
relay\_soln.py, 27  
risk.py, 21  
score\_example.py, 29  
survey.py, 5, 7, 12  
thinkstats.py, 12, 60  
abstraction, 49  
Adams, Cecil, 23  
Adult Intelligence Scale, 44, 118  
adult weight, 47, 48, 114  
analysis, 49, 74, 88  
anecdotal evidence, 2, 9  
Anscombe's quartet, 113  
apparent effect, 8, 10, 20, 79  
artifact, 8, 10  
Australia, 37  
average, 11, 68  
bar plot, 15, 18  
baseball, 61  
basketball, 61  
Bayes factor, 86  
Bayes's theorem, 63  
Bayesian estimation, 77, 97, 104  
Bayesian probability, 84, 86  
Bayesianism, 54, 66, 98  
Behavioral Risk Factor Surveillance  
    System, 47, 48, 59, 110, 114,  
    116, 118  
belief, 54, 66, 86  
bias  
    confirmation, 2  
    oversampling, 26  
    selection, 2, 26  
biased estimator, 95, 96, 105  
bin, 21, 23, 77  
binning, 27  
binomial coefficient, 60

- binomial distribution, 60
- birth time, 38
- birth weight, 27, 32, 34, 35, 43, 46, 68, 85, 120
- birthday, 40
- bisection algorithm, 31
- Blue Man Group, 71
- booyah, 73
- bread police, 58
- BRFSS, 47, 48, 59, 110, 114, 116, 118
- Brisbane, 37
  
- cancer cluster, 61
- casino, 88
- causation, 118
- CCDF, 38, 42, 48
- CDF, 28, 29, 35, 40, 42, 46, 49, 69, 70, 76
  - complementary, 38, 42, 48
- Cdf object, 30
- cell, 87, 91
- censored data, 101, 105
- Central Limit Theorem, 75, 78
- central tendency, 11, 23, 34
- ch-square test, 91
- chance, 79, 83
- chapeau, 5
- chi-square distribution, 97
- chi-square statistic, 87
- Chi-square test, 87
- chi-square test, 88
- city size, 48
- class size, 25
- clinically significant, 22, 24
- cluster, 61
- clustering illusion, 61
- coefficient
  - binomial, 60
  - correlation, 107, 109, 110, 113, 114
  - determination, 117, 121
  - Gini, 68
  - skewness, 67
  - variation, 59, 66
- cohort, 10, 21, 33, 62, 85
- coin, 59, 60, 83
- Coleman, Norm, 101
- colormap, 113
- comic, 118
- complementary CDF, 38, 42, 48
- compression, 49
- computation, 1, 115
- conditional distribution, 32, 35
- conditional probability, 21, 24, 55, 63
- confidence interval, 97, 105
- confirmation bias, 2
- confused Monty problem, 58
- conjunction fallacy, 55
- continuous, 77
- continuous distribution, 37, 50
- contradiction, proof by, 80
- contributors, vii
- control group, 119, 121
- controlled trial, 119
- convergence, 76, 86
- convolution, 72, 75, 78
- cookie, 65
- corpus, 42, 51
- correlation, 107, 108, 118, 121
- correlation coefficient, 110
- cost function, 116
- covariance, 108, 110, 121
- credible interval, 99, 104, 105
- crooked die, 88
- cross-sectional study, 4, 9
- cross-validation, 85, 91
- cumulative distribution function, 28, 29, 69
- cycle, 4, 85
- dance, 59
- data collection, 3
- database, 5
- decay, 101

- density, 71  
 dependent variable, 117, 121  
 derivative, 70  
 descriptive statistics, 3, 11  
 deviation, 12, 67, 107, 108  
 diachronic, 63  
 dice, 53, 55, 59, 88  
 dictionary, 13  
 DiMaggio, Joe, 61  
 discrete, 77  
 distribution, 13, 23  
     binomial, 60  
     chi-square, 97  
     conditional, 32  
     continuous, 37, 50  
     empirical, 37, 40, 50  
     Erlang, 70, 73  
     exponential, 37, 40–42, 45, 49, 69–71, 73–77, 96–98, 100, 101  
     Gaussian, 42, 45, 46, 48, 58, 70, 74–77, 89, 93–95, 107, 113  
     Gumbel, 70, 74  
     lognormal, 46, 48, 59, 75, 76, 114  
     normal, 42, 45, 46, 48, 58, 70, 71, 74–77, 89, 93–95, 107, 113  
     Pareto, 40–42, 45, 48, 49, 75, 76  
     uniform, 34, 49, 98, 99, 101, 102, 104, 107, 112  
     Weibull, 42, 45, 50, 116  
 distribution framework, 76  
 distributions  
     operations, 67  
 drug testing, 64  
 election, 53  
 empirical distribution, 37, 40, 50  
 Erlang distribution, 70, 73  
 error, 94, 95  
 error function, 42, 50  
 error, Type I, 82  
 error, Type II, 82  
 estimation, 3, 93, 105  
     Bayesian, 77, 97  
 estimator, 93  
     biased, 95  
     unbiased, 95, 102  
 event, 53, 65  
     independent, 55  
 evidence, 63, 66  
 explanatory variable, 117  
 exploratory data analysis, 3, 20  
 exponential distribution, 41  
 exponential distribution, 37, 40, 42, 45, 49, 69–71, 73–77, 96–98, 100, 101  
 failure, 53, 66  
 fallacy  
     conjunction, 55  
     illusory superiority, 68, 78  
     sharpshooter, 62  
 false negative, 82, 91  
 false positive, 64, 82, 90  
 field, 10  
 field (database), 5  
 first babies, 1  
 fit, goodness, 117  
 Florida, girl named, 56  
 framework, distributions, 76  
 Franken, Al, 101  
 French, 5  
 frequency, 13, 23, 42, 53  
 frequentism, 54, 66, 98  
 Galton, Francis, 78  
 Gaussian distribution, 42, 45, 46, 48, 58, 70, 74–77, 89, 93–95, 107, 113  
 Gawande, Atul, 61  
 generative process, 49  
 German tank problem, 102  
 Gini coefficient, 68  
 girl named Florida, 56  
 goodness of fit, 117

- grayscale, 113
- Group, Blue Man, 71
- Gumbel distribution, 70, 74
  
- hapaxlegomenon, 42, 51
- height, 42, 59, 71, 112
- hexbin plot, 113
- Hist object, 14
- histogram, 13, 14, 23
- hitting streak, 61
- hot hand, 60
- hot spot, 60
- howzzat, 118
- hypothesis, 63, 87
- hypothesis testing, 3, 79, 90
  
- identical, 75
- identical trials, 53
- illusory superiority, 68, 78
- income, 48, 68
- independent, 55, 66, 73, 75
- independent variable, 117, 121
- inertia, 71
- init method, 5, 69
- intelligence, 44, 118
- interarrival time, 37, 50
- intercept, 115
- Internal Revenue Service, 48, 68
- interquartile range, 34, 35
- interval
  - confidence, 97, 105
  - credible, 99, 105
- intuition, 57
- inverse CDF algorithm, 34, 49
- IQ, 44, 118
- IRS, 48, 68
  
- JAMA, 64
- James Joyce Ramble, 33
- jitter, 112
- Journal of the American Medical Association, 64
  
- Lake Wobegon effect, 68
- Langan, Christopher, 44
- least squares fit, 114, 121
- length
  - pregnancy, 8, 15, 18, 19, 21, 44, 68, 79, 80, 82, 84, 87, 88
- likelihood, 63, 66, 86, 99
- likelihood ratio, 86, 91
- line plot, 18
- linear least squares, 114
- linear regression, 114
- linear relationship, 109
- linear transformation, 74
- locomotive problem, 102
- logarithm, 76
- logarithmic scale, 38, 41
- lognormal distribution, 46, 48, 59, 75, 76, 114
- longitudinal study, 4, 9
  
- MacKay, David, 65, 86, 101
- map, 13
- margin of error, 101
- margin of victory, 101
- Martin, Steve, 5
- mass, 71
- matplotlib, 15
- max, 74
- maximum likelihood estimator, 94, 96, 102, 105, 115
- mean, 11, 18, 37, 76, 87, 93, 94, 96, 107
  - trimmed, 20
  - truncated, 20
- mean squared error, 94, 102, 105
- mean, difference in, 80, 89
- median, 34, 37, 41, 94, 96
- medicine, 119
- method
  - init, 5, 69
- Minnesota Senate Race, 101
- MLE, 94, 96, 102, 105, 115
- Mlodinow, Leonard, 56

- mode, 15, 16, 23  
 model, 44–46, 48–50  
 moment of inertia, 71  
 Monte Carlo, 61, 66, 88  
 Monty Hall  
     confused, 58  
 Monty Hall problem, 56  
 Mosteller, Frederick, 102  
 MSE, 94, 102, 105  
 Munroe, Randall, 118  
 mutually exclusive, 59  
  
 National Survey of Family Growth,  
     3, 4, 19, 20, 27, 32, 34, 43, 79–  
     85, 88, 120  
 natural experiment, 120, 121  
 noise, 112  
 normal distribution, 42, 45, 46, 48, 58,  
     70, 71, 74–77, 89, 93–95, 107,  
     113  
 normal probability plot, 45, 50  
 normalization, 13, 23  
 normalize, 107, 121  
 normalizing constant, 63, 66  
 NSFG, 3, 4, 19, 20, 27, 32, 34, 43, 79–  
     85, 88, 120  
 null hypothesis, 80, 86, 90  
  
 oeuf, 5  
 one-sided test, 83, 91  
 operations on distributions, 67  
 outlier, 16, 19, 23, 67, 94, 96, 112, 114  
 oversampling, 4, 10, 26  
  
 p-value, 80, 89, 90  
 parameter, 37, 40, 41, 43, 47, 49, 50,  
     93, 96, 98  
 Pareto distribution, 40–42, 45, 48, 49,  
     75, 76  
 Pareto World, 42  
 Pareto, Vilfredo, 40  
 particles, 101  
  
 PDF, 70, 72, 77, 78, 100, 116  
 Pearson coefficient of correlation,  
     107, 109, 113  
 Pearson's median skewness coeffi-  
     cient, 67  
 Pearson, Karl, 109  
 percentile, 29, 34, 35, 44  
 percentile rank, 28, 33, 35, 107  
 plot  
     bar, 15, 18  
     hexbin, 113  
     line, 18  
     normal probability, 45, 50  
     scatter, 110  
 plotting, 15, 18  
 PMF, 13, 18, 23, 26, 27, 76  
 Pmf object, 16, 74, 99  
 point estimate, 97  
 point estimation, 105  
 pooled data, 120  
 population, 4, 9, 48  
 posterior, 63, 66, 99  
 posterior distribution, 102  
 posterior probability, 84  
 power, 90, 91  
 prediction, 94, 109  
 preferential attachment, 49  
 pregnancy length, 8, 15, 18, 19, 21, 44,  
     68, 79, 80, 82, 84, 87, 88  
 Presley, Elvis, 65  
 Prime Minister, 54  
 prior, 63, 66, 99  
 prior distribution, 104  
 prior probability, 84  
 probability, 1, 13, 23, 53  
     conditional, 21, 55  
     rules of, 54, 59  
 probability density function, 70, 78  
 probability mass function, 13  
 product, 76  
*Proofiness*, 101

- pumpkin, 12  
pyplot, 40, 110
- Q.E.D., 63  
quartile, 34
- race  
    relay, 46  
race time, 33  
random, 60  
random module, 34, 40–42, 50, 81  
random number, 33, 49  
random variable, 69, 72, 74, 78  
random variate, 70, 78  
randomized controlled trial, 119, 121  
rank, 107, 121  
    percentile, 33  
rankit, 45, 50  
raw data, 7, 10  
recode, 7, 10  
record, 10  
record (database), 5  
regression analysis, 120  
relative mean difference, 68  
relative risk, 21, 24, 87  
relay race, 26, 46  
replacement, 35  
replacement, sampling, 34  
representative, 4, 10  
resampling, 34, 35, 80, 81, 88  
residual, 115, 118, 121  
respondent, 4, 10, 48  
robust, 68, 77, 96, 107
- sample, 10  
sample mean, 89  
sample size, 2, 53, 82  
sample variance, 13, 95  
sampling, 34  
scatter plot, 110  
Seife, Charles, 101  
selection algorithm, 29  
selection bias, 2, 26  
shape, 16, 32  
sharpshooter fallacy, 62  
significance, 8, 22  
significance criterion, 82  
significant, 79, 90  
simulation  
    Monte Carlo, 61, 66, 88  
six-sigma event, 44  
skewness, 67, 77  
slope, 115  
slump, 60  
smoothing, 49  
spatial pattern, 61  
Spearman coefficient of correlation,  
    107, 113, 114  
sport, 60  
spread, 23, 34  
spurious relationship, 119  
standard deviation, 12, 23, 107  
standard score, 108, 121  
standard scores, 107  
statistically significant, 8, 10  
statistics, 1  
stick, 57  
Straight Dope, The, 23  
streak, 60  
study  
    cross-sectional, 4, 9  
    longitudinal, 4, 9  
subjective belief, 54, 66, 86  
success, 53, 66  
suite, 99  
sum, 72, 75  
summary statistic, 8, 10, 34  
summary statistics, 11  
survival analysis, 18  
swamp, 86  
switch, 57  
symmetric, 35, 58  
symmetry, 67

- table, 10
- table (database), 5
- tail, 116
- taxes, 48, 68
- test
  - one-sided, 83, 91
  - two-sided, 83, 91
- test set, 85
- test statistic, 87, 91
- testing set, 91
- Thailand, 54
- Thaksin Shinawatra, 54
- threshold, 82
- training set, 85, 91
- treatment group, 119, 121
- trial, 53, 66
- trim, 23
- trimmed mean, 20
- truncated mean, 20
- turbine, 116
- twin, 65
- two-sided test, 83, 91
  
- U.S. Census Bureau, 48
- unbiased estimator, 95, 102
- uniform distribution, 34, 49, 98, 99,  
101, 102, 104, 107, 112
- unique events, 53
- units, 107
- unreason, supreme law, 78
- unstable particles, 101
- update, 63, 66, 84, 99
  
- variability, 117
- variable
  - random, 69, 72, 74, 78
- variance, 12, 18, 23, 67, 76, 87, 94, 110
- variate
  - random, 70, 78
- visualization, 20
  
- WAIS, 44, 118
  
- Wechsler Adult Intelligence Scale, 44
- Weibull distribution, 42, 45, 50, 116
- weight, 112
  - adult, 47, 48, 114
  - birth, 27, 32, 34, 35, 43, 46, 68, 85,  
120
  - pumpkin, 11
  - sample, 7
- wind power density, 116
- wind speed, 116
- word frequency, 42
  
- xkcd, 118
  
- Zipf's law, 42