

# Index

- A-value, 160
- Accuracy, 57
- Adjusted  $R^2$ , 117, 123
- Agglomerative clustering, 150
- Alternative
  - hypothesis, 79
  - one-sided, 79
  - two-sided, 79
- Analysis of variance, 131
- ANOVA, 113
- ANOVA, 113, 131
  - assumptions, 140
  - F-test, 133
  - model, 60
  - one-way, 132
  - two-way, 136
- Association, 47
- Assumption, 7, 67
  - ANOVA, 140
  - constant variance, 125
  - normality, 125
- Average, 20
- Average linkage, 152
  
- Bar plot, 24
- Bayesian statistics, 145, 176
- Bell curve, 33
- Bias, 56
- Bin, 26
- Binomial
  - coefficient, 32
  - distribution, 31
- Biological
  - replication, 59
  - variation, 58
- BLAST, 110
- Bootstrap, 64
- Box plot, 28
  
- Categorical, 60
  - data, 19
  - variable, 113
- Causation, 48
- Cause and effect, 48
- Central Limit Theorem, 39, 66
  
- Checking model assumptions, 124
- Chi-square test, 85, 97
  - goodness-of-fit, 78
  - independence, 99
- Classification, 145–147
  - error, 147
  - rule, 147
- Clustering, 145, 148
  - $k$ -means, 153
  - agglomerative, 150
  - divisive, 150
  - hierarchical, 150
  - partitional, 150, 153
- Common response, 48, 49
- Complete linkage, 152
- Confidence, 72
- Confidence interval, 65, 113
  - computing, 72
  - interpretation, 67
  - large sample mean, 72
  - population proportion, 76
  - small sample mean, 73
- Confidence level, 67, 70, 72
- Confounding, 48
- Constant variance assumption, 125
- Contingency table, 19, 60, 98
- Continuous variable, 18
- Correlation, 46
- Correlation coefficient, 115, 116
- Critical value, 70
  - computing, 71
- Cumulative probability, 32
  
- Data
  - categorical, 19, 24, 25
  - quantitative, 19, 27, 28
  - transformation, 37
- Dependent variable, 51
- Descriptive statistics, 17
- Design of experiments, 51
- Deterministic model, 51
- Differential expression, 132
- Discrete variable, 17
- Discriminant function, 147
- Discrimination, 147

- Distance measure, 149
- Distribution, 18, 31, 52
  - binomial, 31
  - center, 23
  - Normal, 33
- Divisive clustering, 150
- Dot plot, 27
- Dummy variable, 142
- Dye-swap, 57, 162
  
- E*-value, 110
- Effect, 58
- Effect size, 61
- Error
  - bars, 44
  - standard, 21
- Estimate, 40
- Euclidean distance, 149
- Expected value, 20
- Experimental design, 6
- Explanatory variable, 51, 113
- Exploratory statistics, 5
  
- F-test, 85, 95, 122
  - ANOVA, 133
- Factor, 53
- Factor effect, 137
- False discovery rate, 169
- Fisher's exact test, 85, 91, 100, 105
- Fitting a model, 54
- Fold-change, 160
- Frequency, 26
  - relative, 25
- Frequentist statistics, 145, 176
  
- Global normalization, 161
- Goodness-of-fit test, 85, 97, 98
- Graphs, 24
  
- Heteroscedastic *t*-test, 88
- Hierarchical clustering, 150
- Histogram, 26
- Homoscedastic *t*-test, 87
- Hypothesis
  - alternative, 79
  - null, 79
- Hypothesis test, 79, 113
  - assumptions, 86
  - errors, 82
  - five step procedure, 81
  - power, 83
  
- Independence test, 97, 99, 105
- Independent variable, 51
- Inference, 65
- Inferential statistics, 5
- Interaction effect, 137
- Intercept, 55, 116
- Interquartile range, 21, 24
  
- K-means clustering, 153
  
- Least squares regression, 115
- Leverage, 127
- Likelihood, 101
- Likelihood ratio test, 101
- Linkage, 152
  - average, 152
  - complete, 152
  - single, 152
- Loading, 157
- Loess normalization, 161
- Log-odds score, 102
- Logistic regression, 60, 114, 120
  
- M*-value, 160
- MA-plot, 160
- Mahalanobis distance, 149
- Matched sample, 56
- Mathematical model, 51
- Maximum likelihood, 145, 174, 175
- Mean, 20, 23
- Means plot, 137
- Median, 20, 23
- Microarray
  - data analysis, 145, 158, 163
  - dye-swap, 164
  - experiment, 132
  - normalization, 160
- Model
  - deterministic, 51
  - mathematical, 51
  - statistical, 51, 53
- Model building, 123
- Model assumptions, 7
- Model parameter, 55, 113

- Model selection, 59
- Model utility test, 123
- Modified box plot, 29
- Monte Carlo, 109, 110
- Multiple linear regression, 121
- Multiple R-squared, 117
- Multiple regression, 114
- Multiple testing, 168
- Multivariate model, 113
  
- Next-generation sequencing (NGS), 169
  - barcoding, 172
  - base calling, 171
  - bridge amplification, 170
  - data analysis, 169
  - experimental overview, 170
  - flow cell, 170
  - library size, 170
  - sample preparation, 170
  - sequencing depth, 171
  - statistical issues, 172
- NGS, data analysis, 145
- Noise, 58
- Non-parametric test, 103
- Normal distribution, 33
  - mean, 34
  - parameters, 34
  - percentile, 36
  - shape, 33
  - standard, 34
  - standard deviation, 34
- Normality
  - assumption, 74, 125
  - checking for, 37
- Normalization, 160
  - dye-swap, 162
  - global, 161
  - loess, 161
  - print-tip, 162
- Null hypothesis, 79
  
- Observation, 18
- Oligonucleotide array, 159
- One-sample *t*-test, 85, 86
- One-sample *z*-test, 85, 91
- One-sided alternative, 79
  
- One-way ANOVA, 132
  - assumptions, 136
  - model, 114
- Ordinal variable, 17
- Outlier, 21, 125, 131
  
- p*-value, 81, 83, 110
- Paired *t*-test, 85, 88
- Parameter, 52, 55
  - population, 65
- Partitional clustering, 150, 153
- Pearson's product moment correlation measure, 46
- Percentile, 20, 36
- Permutation test, 85, 88, 108
- Pie chart, 25
- Plot
  - bar plot, 24, 30
  - box plot, 28, 30
  - box plot, modified, 29
  - dot plot, 27
  - histogram, 26, 30
  - pie chart, 25, 30
  - PP-plot, 36
  - QQ-plot, 36
  - rankit, 37
  - scatter plot, 27
  - strip chart, 27
- Population, 39, 53, 55, 65
  - parameter, 65
  - proportion, 91
- Post-hoc test, 96
- Posterior distribution, 177
- Power, 83
- Precision, 57, 71
- Predictor variable, 51, 113
- Principal component analysis, 145, 156
- Print-tip normalization, 162
- Prior distribution, 177
- Probability plot, 36
- Probability distribution, 31
- Probe set, 159
- Proportion, 91
  
- QQ-plot, 86
- Qualitative variable, 17
- Quantile plot, 37

- Quantitative, 60
  - data, 19
  - variable, 17, 113
- Quantitative trait locus, 177
- Quartile, 20
- Random sample, 56
- Random variable, 18, 30
- Range, 21, 24
- Rankit plot, 37
- Regression, 113
  - assumptions, 124
  - case study, 127
  - least squares, 115
  - logistic, 60, 114, 120
  - model building, 123
  - multiple, 114, 121
  - outliers, 125
  - simple linear, 115
- Regression model, 60
  - checking assumptions, 119
- Regression parameters
  - hypothesis testing, 118
  - interpretation, 117
- Reject, 80
- Relative frequency, 25
- Replication, 62
  - biological, 59
  - technical, 59
- Resampling, 62
- Residual, 115
- Residual standard error, 117
- Response variable, 51, 113
- Sample, 39, 53, 55, 65
  - biased, 56
  - matched, 56
  - random, 56
  - stratified, 56
- Sample mean, 40, 41
- Sample proportion, 39, 40, 91
- Sample size, 58, 60, 72
  - calculation, 77
- Sample statistic, 43, 55
- Sampling, 55
- Scatter plot, 27
- Scheffé test, 85, 96
- Side-by-side box plot, 29
- Significance, 83
- Significance level, 61, 70, 81–83
- Simple linear regression, 114, 115
- Single linkage, 152
- Slope, 55, 116
- Sources of variation, 6
- Spotted array, 159
- Spread, 19, 24
- Standard deviation, 20, 24, 43, 44
- Standard error, 21, 43, 45, 70
- Standard normal distribution, 34
- Statistic, 55, 65
- Statistical classification, 146
- Statistical inference, 65, 113
- Statistical model, 51, 53
- Statistical significance, 83, 84
- Statistical software, 8
- Stratified sample, 56
- Stratum, 56
- Strip chart, 27
- Summary statistics, 21
- Symmetry, 26
- t*-test, 86
- Table, contingency, 19
- Tail probability, 34, 35
- Technical replication, 59
- Technical variation, 59
- Test statistic, 6, 80, 81
- Training data, 147
- Transformation, 37
- Tree diagram, 150, 154
- Tukey's test, 85, 96
- Two-sample *t*-test, 85, 87
- Two-sample *z*-test, 85, 93
- Two-sided alternative, 79
- Two-way ANOVA, 136
  - model, 114
- Type I error, 82
- Type II error, 82
- Uncertainty, 65
- Univariate model, 113
- Variability, 24, 61, 72

- Variable, 17
  - categorical, 97, 113
  - continuous, 17
  - dependent, 51
  - discrete, 17
  - explanatory, 51, 113
  - factor, 17
  - independent, 51
  - ordinal, 17
  - predictor, 51, 113
  - qualitative, 17
  - quantitative, 113
  - random, 18, 30
  - response, 51, 113
  - selection, 124
  - transformation, 129
- Variance, 20, 24
- Variation
  - biological, 58
  - sources of, 6
  - technical, 59
- Wilcoxon-Mann-Whitney test, 85, 88, 104
- $z$ -test, 91

# Index of Worked Out Examples

- ANOVA in microarray experiment, 132
- ANOVA model for analysis of microarray data, 141
- Bar plot, 24
- Binomial distribution, 31
- Biological and technical replication, 59
- Categorical variables in regression, 142
- Causation, common response, and confounding, 48
- Central Limit Theorem, 42
- Chi-squared test for goodness-of-fit, 99
- Chi-squared test for independence, 99
- Classification, 146
- Computing normal percentiles, 36
- Computing normal probabilities, 35
- Confidence interval, 68
- Confidence interval for a population proportion, 76
- Confidence interval for Normal Data, 66
- Contingency table, 19, 98
- Correlation and causation, 47
- Determining outliers, 21
- Differential gene expression, 80
- Dissimilarity measure, 149
- Dye bias in microarray experiments, 57
- F*-test, 95
- Fisher's exact test, 107
- Frequentist and Bayesian QTL analysis, 177
- Hierarchical clustering, 150
- Histogram, 26
- Independent sample t-test, 88
- K-means clustering, 156
- Large sample confidence interval for mean, 72
- Likelihood, 101
- Maximum likelihood estimation of a population proportion, 175
- Maximum likelihood parameter estimation, 175
- Microarray ANOVA enumeration, 165
- Multiple linear regression, 127
- Multiple testing, 168
- One-sample *z*-test for a population proportion, 91
- One- and two-sided alternatives in hypothesis testing, 79
- One-way Anova, 132
- Parameter estimation in dye-swap microarray experiment, 166
- Permutation test, 108
- Predictor and response variables, 51
- Probability distribution, 30
- Publication bias, 57
- Random variable vs. observation, 18
- Resampling, 62
- Sample proportion, 39
- Sample size calculation, 77
- Sampling bias, 56
- Side-by-side box plot, 29
- Simple linear regression, 117
- Small sample confidence interval for a mean, 73
- Standard deviation vs. standard error, 43
- Two-sample *z*-test, 93
- Two-Way ANOVA, 136
- Variable types, 17
- Wilcoxon-Mann-Whitney test, 105

# Index of R Commander Commands

- Anova
  - one-way, 133
  - two-way, 137
- aov, 133
- Backward model selection, 124
- Bar plot, 26
- Binomial
  - coefficient, 32
  - probability, 32
  - tail probability, 32
- Box plot, 28
- cex, 14
- cex.lab, 14
- cex.main, 14
- Chi-squared
  - p*-value, 102
  - goodness-of-fit test, 98
  - test for independence, 100
- chisq.test, 98
- choose, 32
- col, 14
- Computing critical values for normal and *t*-distributions, 70
- Confidence interval for a population mean, 75
- Confidence interval for a population proportion, 77
- Contingency table, 98
- Converting numeric variables to factors, 13
- cor, 46
- Correlation coefficient, 46
- dbinom, 33
- Diagnostic plots, 125, 127
- Entering data, 12
- F*-test, 96
- Fisher's exact test, 107
- fisher.test, 107
- Forward model selection, 124
- Goodness-of-fit test, 98
- hclust, 152
- Help, 16
- Hierarchical clustering, 152
- Histogram, 26
- Importing data, 12
- Interquartile range, 21
- k-means clustering, 155
- KMeans, 155
- Likelihood ratio test, 102
- lm, 116
- Logistic regression, 121
- main, 14
- Maximum, 21
- Mean, 21
- Means plot, 137
- Median, 21
- Minimum, 21
- Missing data identifier, 13
- Model selection, 124
- Multiple linear regression, 122
- Normal
  - percentiles, 36
  - probabilities, 35
  - quantiles, 36
- One-sample *t*-test, 87
- One-way Anova, 133
- Output
  - save, 15
- plot, 14
- Paired sample *t*-test, 89
- pbinom, 33
- pch, 14

- pchisq, 102
- Percentile, 21
- Pie chart, 26
- Plot
  - bar plot, 26
  - box plot, 28
  - dot plot, 27
  - histogram, 26
  - pie chart, 26
  - scatter plot, 28
  - strip chart, 27
- pnorm, 92
- Principal component analysis, 157
- princomp, 157
- pt, 90
- qnorm, 70
- QQ-plot construction, 37
- qt, 70
- Range, 21
- Regression parameters, 117
- Residual check for linear model, 120
- Residual plot, 125, 127
- Save output, 15
- Save R script, 15
- Scatter plot, 28
- Script
  - save, 15
- Simple linear regression, 116
- Standard deviation, 21
- Strip chart, 27
- Summary statistics, 21
- t.test, 87–89
- Transformation, 129
- Tukey’s post hoc test, 97
- Two independent sample *t*-test, 88
- Two-way Anova, 137
- Variable
  - convert, 13
- Variable selection, 124
- Variable transformation, 129
- Variance, 21
- wilcox.test, 104
- Wilcoxon-Mann-Whitney test, 104
- xlab, 14
- ylab, 14