

---

# R for Starters

v0.55721



by Ole J. Forsberg



# R FOR STARTERS

ADVANCED METHODS OF RESEARCH INQUIRY FOR THE SOCIAL  
SCIENCES USING THE R STATISTICAL ENVIRONMENT

Version: 0.57721

OLE J. FORSBERG

*Oklahoma State University – Stillwater*

COPYRIGHT 2015 – OLE J. FORSBERG

---

---

ALL RIGHTS RESERVED. No part of this work covered by copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including, but not limited to photocopying, recording, scanning, digitizing, taping, Internet distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 US Copyright Act, without the prior written permission of the author.

The current draft version of this document is free (without cost). This document is distributed in the hope that it will be useful, but without any warranty; without even the implied warranty of merchantability or fitness for a particular use or for a particular purpose.

This document was typeset using the L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub> environment.

# Contents

<b>Preface</b>	<b>xvii</b>
<b>I Statistics and <i>The Search</i></b>	<b>1</b>
<b>1 An Introduction to R</b>	<b>3</b>
1.1 Installing R on your computer . . . . .	5
1.2 A quick, sample session . . . . .	6
1.3 A second example . . . . .	13
1.4 A third example . . . . .	16
1.5 Conclusion . . . . .	19
1.6 End of Chapter Materials . . . . .	20
<b>2 Statistics and the Scientific Method</b>	<b>29</b>
2.1 Start . . . . .	31
2.2 Conclusion . . . . .	33
2.3 End of Chapter Materials . . . . .	34
<b>3 Collecting Data</b>	<b>39</b>
3.1 Start . . . . .	41
3.2 Conclusion . . . . .	43
<b>4 Descriptive Statistics</b>	<b>49</b>
4.1 Variable Types . . . . .	51
4.2 Measures of Central Tendency . . . . .	55
4.3 Measures of Dispersion . . . . .	67
4.4 A Simple Univariate Analysis . . . . .	76
4.5 Conclusion . . . . .	81
4.6 End of Chapter Materials . . . . .	82

<b>II</b>	<b>Discrete Independent Variables</b>	<b>91</b>
<b>5</b>	<b>A Single Population</b>	<b>93</b>
5.1	Units of and Levels of Analysis . . . . .	95
5.2	The z-test . . . . .	96
5.3	The t-test . . . . .	106
5.4	Testing the Assumption . . . . .	113
5.5	Non-Parametric Means Tests I . . . . .	118
5.6	Non-Parametric Means Tests II* . . . . .	122
5.7	Further Examples . . . . .	123
5.8	Conclusion . . . . .	128
5.9	End of Chapter Materials . . . . .	129
<b>6</b>	<b>Comparing Two Groups</b>	<b>137</b>
6.1	Two independent samples; equal variance . . . . .	139
6.2	Two independent samples; unequal variance . . . . .	144
6.3	Testing the Assumption . . . . .	146
6.4	Non-Parametric Means Tests I . . . . .	147
6.5	Non-Parametric Means Tests II* . . . . .	154
6.6	Further Examples . . . . .	158
6.7	Conclusion . . . . .	163
6.8	End of Chapter Materials . . . . .	164
<b>7</b>	<b>Comparing three or more means</b>	<b>173</b>
7.1	The Multiple Comparisons Issue . . . . .	175
7.2	Analysis of Variance . . . . .	177
7.3	Non-Parametric Means Tests I . . . . .	190
7.4	Non-Parametric Means Tests II* . . . . .	195
7.5	Post-Hoc Testing . . . . .	196
7.6	Further Examples . . . . .	201
7.7	Conclusion . . . . .	208
7.8	End of Chapter Materials . . . . .	209
<b>8</b>	<b>Proportions Tests</b>	<b>219</b>
8.1	The One-Population Proportion Test . . . . .	221
8.2	The Two-Population Proportion Test . . . . .	229
8.3	Conclusion . . . . .	234
8.4	End of Chapter Materials . . . . .	235
<b>9</b>	<b>Two-by-Two Tables</b>	<b>241</b>

9.1	Conclusion . . . . .	242
9.2	End of Chapter Materials . . . . .	243
<b>10</b>	<b>Categorical Independence</b>	<b>247</b>
10.1	Conclusion . . . . .	248
10.2	End of Chapter Materials . . . . .	249
<b>11</b>	<b>Estimators and Intervals*</b>	<b>255</b>
11.1	Selecting Estimators . . . . .	257
11.2	Coverage and Intervals . . . . .	269
11.3	A Final Example . . . . .	273
11.4	Conclusion . . . . .	275
11.5	End of Chapter Materials . . . . .	276
<b>III</b>	<b>The Classical Linear Model</b>	<b>281</b>
<b>12</b>	<b>Linear Regression</b>	<b>283</b>
12.1	Scatterplots . . . . .	285
12.2	The Method of Ordinary Least Squares . . . . .	293
12.3	Goodness of Fit . . . . .	301
12.4	Maine and the Ballot Measure. . . . .	304
12.5	Conclusion . . . . .	316
12.6	End of Chapter Materials . . . . .	317
<b>13</b>	<b>Assumptions of Linear Regression</b>	<b>323</b>
13.1	Multicollinearity . . . . .	326
13.2	Normality . . . . .	331
13.3	Identically Distributed . . . . .	334
13.4	Independence . . . . .	339
13.5	Single Population . . . . .	342
13.6	A Full Example . . . . .	345
13.7	Conclusion . . . . .	350
13.8	End of Chapter Materials . . . . .	351
<b>14</b>	<b>Linear Regression and Transformations</b>	<b>359</b>
14.1	The Issue of Boundedness . . . . .	361
14.2	Data Bounded by 0 and 1 . . . . .	363
14.3	Data Bounded Below by 0 . . . . .	367
14.4	Additional Bounds . . . . .	371
14.5	Conclusion . . . . .	378

14.6	R Functions . . . . .	380
14.7	Extensions . . . . .	383
14.8	Applications . . . . .	384
14.9	References and Additional Readings . . . . .	385
<b>IV The Generalized Linear Model</b>		<b>387</b>
<b>15 Generalized Linear Models</b>		<b>389</b>
15.1	The CLM and the GLM . . . . .	391
15.2	The Requirements for GLMs . . . . .	392
15.3	Assumptions of GLMs. . . . .	396
15.4	The Gaussian Distribution . . . . .	397
15.5	Generalized Linear Models in R . . . . .	399
15.6	Conclusion . . . . .	404
15.7	End of Chapter Materials . . . . .	406
<b>16 Binary Dependent Variables</b>		<b>411</b>
16.1	Binary Dependent Variables . . . . .	413
16.2	Latent Variable Modeling . . . . .	416
16.3	The Mathematics. . . . .	418
16.4	Modeling with the Logit . . . . .	425
16.5	Prediction Accuracy . . . . .	428
16.6	Modeling with Other Links. . . . .	434
16.7	Model Selection . . . . .	436
16.8	Conclusion . . . . .	441
16.9	End of Chapter Materials . . . . .	443
<b>17 Nominal and Ordinal Dependent Variables</b>		<b>449</b>
17.1	Nominal Dependent Variable . . . . .	452
17.2	Ordinal Dependent Variable . . . . .	460
17.3	Conclusion . . . . .	464
17.4	R Functions . . . . .	465
17.5	Extensions . . . . .	466
17.6	Applications . . . . .	467
17.7	References and Further Readings . . . . .	468
<b>18 Count Dependent Variables</b>		<b>469</b>
18.1	Linear or Poisson Regression? . . . . .	471
18.2	The Mathematics. . . . .	472



18.3	Overdispersion . . . . .	477
18.4	Body counts . . . . .	482
18.5	The Bias-Variance trade-off . . . . .	488
18.6	Conclusion . . . . .	489
18.7	R Functions . . . . .	491
18.8	Extensions . . . . .	493
18.9	Applications . . . . .	494
18.10	References and Additional Readings . . . . .	495
<b>19</b>	<b>Lifetime Models</b>	<b>497</b>
19.1	The Mathematics . . . . .	499
19.2	Gamma or Logarithm? . . . . .	502
19.3	Conclusion . . . . .	504
19.4	R Functions . . . . .	506
19.5	Extensions . . . . .	507
19.6	Applications . . . . .	508
19.7	References and Additional Readings . . . . .	509
<b>V</b>	<b>The Appendices</b>	<b>511</b>
<b>Appendix A</b>	<b>Important Discrete Distributions</b>	<b>513</b>
A.1	Discrete Distributions . . . . .	515
A.2	Bernoulli . . . . .	521
A.3	Binomial . . . . .	524
A.4	Hypergeometric . . . . .	532
A.5	Geometric . . . . .	536
A.6	Negative Binomial (Pascal; Pólya) . . . . .	540
A.7	Poisson . . . . .	544
A.8	End of Appendix Materials . . . . .	548
<b>Appendix B</b>	<b>Important Continuous Distributions</b>	<b>557</b>
B.1	Continuous Distributions . . . . .	559
B.2	Uniform . . . . .	567
B.3	Normal (Gaussian) . . . . .	573
B.4	Chi-Squared (Helmert) . . . . .	577
B.5	Exponential . . . . .	580
B.6	Gamma (Erlang) . . . . .	584
B.7	End of Appendix Materials . . . . .	587

<b>Appendix C</b>	<b>The Standard Normal Distribution</b>	<b>595</b>
C.1	The Standard Normal Distribution . . . . .	597
C.2	The Z-Transform . . . . .	598
C.3	The Central Limit Theorem . . . . .	601
C.4	End of Chapter Materials . . . . .	603

# List of Figures

1	Bloom's Taxonomy . . . . .	xxi
1.1	The R Window after starting R. . . . .	7
1.2	The R Window after tiling the two sub-windows. . . . .	8
1.3	Results of preliminary analysis of the $\mathcal{U}(0,1)$ dataset. . . . .	12
1.4	Results of preliminary analysis of <code>positioningtube</code> dataset. . . . .	15
1.5	Histogram of p-values from the Monte Carlo experiment. . . . .	18
4.1	Histogram of GDP per capita . . . . .	63
4.2	Bar chart of the world region . . . . .	68
4.3	Pie chart of the world region . . . . .	69
4.4	Boxplot of GDP per capita . . . . .	72
5.1	The hypothesized distribution for the Niepołomice Forest example. . . . .	102
5.2	Monte Carlo results for the z-test. . . . .	106
5.3	Monte Carlo results for t-tests with Exponential data. . . . .	117
6.1	The rejection region for the Gender Height example. . . . .	142
6.2	A box-and-whiskers plot of the two groups of heights. Note that the male heights appear to be significantly higher than female heights, which is what the t-test indicated. . . . .	146
6.3	A box-and-whiskers plot of the mean fire return interval for two biomes, xeric shrubland and temperate broadleaf forest. Note the significant difference in the measures of center. . . . .	153
6.4	Back-to-back histogram of populations in two census regions . . . . .	160
6.5	Back-to-back histogram of populations in two census regions . . . . .	162
7.1	Distribution of the test statistic for the Rice Example . . . . .	182
7.2	NCAA Conference football score distribution. . . . .	183
7.3	Q-Q plots of non-Normal data. . . . .	186
7.4	Q-Q plots of two Normally distributed sets of data. . . . .	187

7.5	Q-Q plots of the team scores in the six conferences. . . . .	188
7.6	Boxplot of GDP per Capita . . . . .	192
7.7	Sir Ronald A. Fisher, FRS . . . . .	197
7.8	John W. Tukey, ForMemRS . . . . .	198
7.9	Bill Kruskal . . . . .	200
7.10	Boxplot of colony counts across the samples . . . . .	202
7.11	Boxplot of mfri across the biomes . . . . .	204
7.12	Boxplot of US perception across the alliances . . . . .	207
11.1	Mean square error for two estimators of $\pi$ . . . . .	262
11.2	Mean square error for two estimators of $\pi$ . . . . .	263
11.3	Mean square error for two estimators of $b$ . . . . .	265
11.4	Comparison of the mean estimator and the median estimator . .	267
11.5	Illustration of coverage rates . . . . .	272
11.6	Graphic of mean square error curves . . . . .	274
12.1	Scatterplots illustrating linear correlation . . . . .	286
12.2	Default scatterplots illustrating linear correlation . . . . .	287
12.3	Annotated scatterplot illustrating linear correlation . . . . .	290
12.4	Scatterplot of Nobel Prize rate against chocolate consumption . .	292
12.5	Height and weights of five males . . . . .	293
12.6	Heights and weights with regression line . . . . .	294
12.7	Height and weights with effect shown . . . . .	297
12.8	Pairwise plots between the three independent variables. . . . .	307
12.9	Prediction graphs of our <code>ssm</code> model. . . . .	312
12.10	Plot of the predicted vote outcomes from the MC experiment. . .	315
13.1	Normal Q-Q plot and histogram of the residuals . . . . .	332
13.2	Two residual plots . . . . .	335
13.3	An index plot of Cook's distance for each unit . . . . .	344
13.4	Residual plot and histogram . . . . .	346
13.5	Residual plot and histogram . . . . .	348
13.6	Prediction plot for model <code>mod2a</code> . . . . .	349
14.1	Schematic of the variable transformation procedure. . . . .	363
14.2	Schematic of the transformation procedure for Example 14.1. . .	366
14.3	Histogram of the MC experiment for Maine. . . . .	367
14.4	Histogram of the MC experiment for Papua New Guinea. . . . .	370
14.5	A scatterplot of the results of the South Sudan referendum. . . .	374
14.6	The results of the South Sudan referendum with predictions. . .	378

15.1	Graphs comparing CLM and GLM transformed results. . . . .	402
15.2	Comparison of CLM and GLM predictions. . . . .	404
16.1	Residual scatterplot for Example 16.1. . . . .	415
16.2	Schematic of logistic regression. . . . .	417
16.3	Three symmetric link functions. . . . .	423
16.4	Two asymmetric link functions. . . . .	424
16.5	The logistic curve. . . . .	425
16.6	Plot of predicted coin weightings. . . . .	428
16.7	The model accuracy against various thresholds. . . . .	431
16.8	The ROC curve for the coin flipping model. . . . .	432
16.9	The ROC curve using the <code>ROC</code> command. . . . .	433
16.10	Plot of logit and complementary loglog functions. . . . .	435
16.11	Plot of logit and loglog functions. . . . .	437
17.1	Explanatory schematic for thresholding. . . . .	462
18.1	Data plot with regression lines. . . . .	472
18.2	Plot of initiative use against state population. . . . .	478
18.3	Plot of the number of deaths due to terrorism. . . . .	489
19.1	GDP per capita predictions . . . . .	504
A.1	A plot of the pmf for Example A.2 . . . . .	519
A.2	The Bernoulli $Bern(\pi = 0.15)$ pmf . . . . .	521
A.3	The Binomial $Bin(n, \pi)$ pmf . . . . .	524
A.4	Epmf for the difference of two Binomials . . . . .	531
A.5	The Hypergeometric $\mathcal{H}(N, K, n)$ pmf. . . . .	532
A.6	The Geometric $\mathcal{G}eom(\pi)$ pmf. . . . .	536
A.7	The Geometric $\mathcal{G}eom(\pi = 0.60)$ pmf. . . . .	539
A.8	The Negative Binomial $NegBin(r, \pi)$ pmf. . . . .	540
A.9	The Poisson $\mathcal{P}(\lambda)$ pmf. . . . .	544
B.1	The pdf for the example . . . . .	560
B.2	The CDF for the example . . . . .	561
B.3	The Uniform $\mathcal{U}(0, 1)$ pdf . . . . .	567
B.4	The Uniform $\mathcal{U}(0, 1)$ CDF . . . . .	569
B.5	Empirical pdf for Example B.1 . . . . .	572
B.6	The Normal pdf . . . . .	573
B.7	The Normal $\mathcal{N}(0, 1)$ CDF . . . . .	575
B.8	The Chi-Squared $\chi^2(\nu)$ pdf . . . . .	577

B.9	The Chi-Squared $\chi^2(\nu)$ CDF . . . . .	579
B.10	The Exponential $\text{Exp}(\lambda)$ pdf . . . . .	580
B.11	The Exponential $\text{Exp}(\lambda)$ CDF . . . . .	582
B.12	The Gamma $\text{GAM}(a, s)$ pdf. . . . .	584
B.13	The Gamma $\text{GAM}(a, s)$ CDF. . . . .	586
C.1	The Standard Normal pdf, $\phi(z)$ . . . . .	597

# List of Tables

1	The Greek alphabet . . . . .	xxii
1.1	R download links . . . . .	5
4.1	Table of some measures of central tendency . . . . .	66
4.2	Table of the various measures of dispersion . . . . .	75
5.1	Schematic of the levels of analysis . . . . .	96
5.2	The two types of errors . . . . .	98
5.3	Sample of students and their pre- and post-test averages. . . . .	112
5.4	External debt for six selected States. . . . .	119
6.1	External debt for eleven selected States. . . . .	149
6.2	Mean fire return intervals for 12 areas in the US . . . . .	152
7.1	Yields from four varieties of rice. . . . .	178
7.2	NCAA Football ANOVA table results . . . . .	185
7.3	ANOVA table for the <code>gdpcap</code> data. . . . .	191
11.1	Table of estimated coverage rates . . . . .	273
12.1	Statistics on the <code>ssm</code> data. . . . .	305
12.2	Correlations between the variables in the <code>ssm</code> data . . . . .	306
12.3	The symbols and their meanings in the grammar of formulas. . . . .	308
12.4	Results table for the <code>ssm</code> example. . . . .	310
13.1	Several Normality tests . . . . .	333
14.1	Results table for the <code>ssm</code> vote model. . . . .	362
14.2	Results table for the <code>ssm</code> vote model (logit). . . . .	365
14.3	Results table for the GDP per capita model. . . . .	369
14.4	Results table for the South Sudan referendum model. . . . .	375

15.1	GLM distributions and links. . . . .	393
15.2	Results table for the <code>ssm</code> example. . . . .	400
15.3	Results table for the <code>ssm</code> vote model (logit). . . . .	401
15.4	Results table from fitting the GDP data with GLMs. . . . .	403
16.1	Insurance data to accompany Example 16.1. . . . .	414
16.2	Binary dependent variable link functions. . . . .	422
16.3	Results table for logit regression on the coin flip data. . . . .	427
16.4	Results table for cloglog regression on the coin flip data. . . . .	436
17.1	Correlation matrix. . . . .	456
17.2	Results table for the logit model. . . . .	456
17.3	Results table for the multinomial regression model. . . . .	458
17.4	Result of ordinal regression in R. . . . .	462
18.1	Initiative model, with adjusted standard errors. . . . .	479
18.2	Initiative model, fitted using QLE. . . . .	480
18.3	Results table for initiatives using the Negative Binomial. . . . .	481
18.4	Three terrorism models, days as independent variable. . . . .	484
18.5	Three terrorism models, using days as offset. . . . .	485
18.6	Two terrorism models with higher degrees. . . . .	488
C.1	Standard Normal Table . . . . .	606