

# Contents

<b>Preface</b>	<b>v</b>
<b>1 Statistical thinking for programmers</b>	<b>1</b>
1.1 Do first babies arrive late? . . . . .	2
1.2 A statistical approach . . . . .	3
1.3 The National Survey of Family Growth . . . . .	3
1.4 Tables and records . . . . .	5
1.5 Significance . . . . .	8
1.6 Glossary . . . . .	9
<b>2 Descriptive statistics</b>	<b>11</b>
2.1 Means and averages . . . . .	11
2.2 Variance . . . . .	12
2.3 Distributions . . . . .	13
2.4 Representing histograms . . . . .	14
2.5 Plotting histograms . . . . .	15
2.6 Representing PMFs . . . . .	16
2.7 Plotting PMFs . . . . .	18
2.8 Outliers . . . . .	19
2.9 Other visualizations . . . . .	20

---

2.10	Relative risk . . . . .	21
2.11	Conditional probability . . . . .	21
2.12	Reporting results . . . . .	22
2.13	Glossary . . . . .	23
<b>3</b>	<b>Cumulative distribution functions</b>	<b>25</b>
3.1	The class size paradox . . . . .	25
3.2	The limits of PMFs . . . . .	27
3.3	Percentiles . . . . .	28
3.4	Cumulative distribution functions . . . . .	29
3.5	Representing CDFs . . . . .	30
3.6	Back to the survey data . . . . .	32
3.7	Conditional distributions . . . . .	32
3.8	Random numbers . . . . .	33
3.9	Summary statistics revisited . . . . .	34
3.10	Glossary . . . . .	35
<b>4</b>	<b>Continuous distributions</b>	<b>37</b>
4.1	The exponential distribution . . . . .	37
4.2	The Pareto distribution . . . . .	40
4.3	The normal distribution . . . . .	42
4.4	Normal probability plot . . . . .	45
4.5	The lognormal distribution . . . . .	46
4.6	Why model? . . . . .	49
4.7	Generating random numbers . . . . .	49
4.8	Glossary . . . . .	50

<b>5</b>	<b>Probability</b>	<b>53</b>
5.1	Rules of probability . . . . .	54
5.2	Monty Hall . . . . .	56
5.3	Poincaré . . . . .	58
5.4	Another rule of probability . . . . .	59
5.5	Binomial distribution . . . . .	60
5.6	Streaks and hot spots . . . . .	60
5.7	Bayes’s theorem . . . . .	63
5.8	Glossary . . . . .	65
<b>6</b>	<b>Operations on distributions</b>	<b>67</b>
6.1	Skewness . . . . .	67
6.2	Random Variables . . . . .	69
6.3	PDFs . . . . .	70
6.4	Convolution . . . . .	72
6.5	Why normal? . . . . .	74
6.6	Central limit theorem . . . . .	75
6.7	The distribution framework . . . . .	76
6.8	Glossary . . . . .	77
<b>7</b>	<b>Hypothesis testing</b>	<b>79</b>
7.1	Testing a difference in means . . . . .	80
7.2	Choosing a threshold . . . . .	82
7.3	Defining the effect . . . . .	83
7.4	Interpreting the result . . . . .	83
7.5	Cross-validation . . . . .	85
7.6	Reporting Bayesian probabilities . . . . .	86

---

7.7	Chi-square test . . . . .	87
7.8	Efficient resampling . . . . .	88
7.9	Power . . . . .	90
7.10	Glossary . . . . .	90
<b>8</b>	<b>Estimation</b>	<b>93</b>
8.1	The estimation game . . . . .	93
8.2	Guess the variance . . . . .	94
8.3	Understanding errors . . . . .	95
8.4	Exponential distributions . . . . .	96
8.5	Confidence intervals . . . . .	97
8.6	Bayesian estimation . . . . .	97
8.7	Implementing Bayesian estimation . . . . .	99
8.8	Censored data . . . . .	101
8.9	The locomotive problem . . . . .	102
8.10	Glossary . . . . .	105
<b>9</b>	<b>Correlation</b>	<b>107</b>
9.1	Standard scores . . . . .	107
9.2	Covariance . . . . .	108
9.3	Correlation . . . . .	108
9.4	Making scatterplots in pyplot . . . . .	110
9.5	Spearman's rank correlation . . . . .	113
9.6	Least squares fit . . . . .	114
9.7	Goodness of fit . . . . .	117
9.8	Correlation and Causation . . . . .	118
9.9	Glossary . . . . .	121