



Vital Statistics

Probability and Statistics
for Economics and Business

William H. Sandholm

University of Wisconsin-Madison

Brett A. Saraniti

Northwestern University

New York Oxford
OXFORD UNIVERSITY PRESS

Contents

Asterisks indicate sections tangential to the main line of argument.

Preface for Students	xxvii
Preface for Instructors	xxix
Supplements	xxxv
Acknowledgments	xxxvii

1. Random Reasoning 1

1.1	Introduction	1
1.2	Probability	2
1.3	Statistics	7
1.4	Conclusion	10

2. Probability Models 11

2.1	Ex Ante vs. Ex Post	12
2.2	Sample Spaces	14
2.2.1	Sample spaces, outcomes, and events	14
2.2.2	New events from old	16
2.3	Probability Measures	18
2.3.1	The axioms of probability	18
2.3.2	Further properties of probability measures	20
2.3.3	Interpreting and assigning probabilities	23
2.4	Conditional Probability	24
2.4.1	What is conditional probability?	25
2.4.2	Joint, marginal, and conditional probabilities	29
2.4.3	The total probability rule	30
2.4.4	Bayes' rule	31

2.5	Independence	36
2.5.1	Independence of pairs of events	36
2.5.2	Independence of many events	38
2.5.3	Independence of many events: A formal treatment*	41
2.6	Constructing Probability Models*	44
2.6.1	Two probability problems	44
2.6.2	Discussion of the Linda problem	45
2.6.3	Discussion of the Monty Hall problem	46
2.A	Appendix: Finite and Countable Additivity	50
2.E	Exercises	51

3. Random Variables 65

3.1	Random Variables	65
3.1.1	What exactly is a random variable?	66
3.1.2	Ex ante vs. ex post revisited	68
3.1.3	The distribution of a random variable	68
3.2	Traits of Random Variables	70
3.2.1	Expected value	70
3.2.2	Variance and standard deviation	73
3.2.3	An alternate formula for expected values*	77
3.3	Functions of Random Variables	79
3.4	Independent Random Variables	86
3.4.1	Independence of two random variables	86
3.4.2	Independence of many random variables	88
3.4.3	Sums of independent random variables	89
3.4.4	New independent random variables from old	93
3.E	Exercises	95

4. Multiple Random Variables 106

4.1	Multiple Random Variables	106
4.1.1	Joint distributions and marginal distributions	108
4.1.2	Conditional distributions	110
4.1.3	Conditional traits and the law of iterated expectations	112

4.2	Traits of Random Variable Pairs	114
4.2.1	Covariance	114
4.2.2	Correlation	116
4.2.3	Some useful facts	120
4.2.4	Independence and zero correlation	121
4.3	Functions of Multiple Random Variables	123
4.4	Portfolio Selection*	127
4.4.1	A simple model of a financial market	127
4.4.2	Portfolio selection and diversification	128
4.4.3	Efficient portfolios	131
4.4.4	The benefits of diversification	133
4.A	Appendix	138
4.A.1	Definitions, formulas, and facts about random variables	138
4.A.2	Derivations of formulas and facts	141
4.B	The Capital Asset Pricing Model	online
4.E	Exercises	145

5.

Bernoulli Trials Processes and Discrete Distributions 158

5.1	Families of Distributions	159
5.1.1	Indicator random variables	159
5.1.2	Bernoulli distributions	160
5.1.3	Traits of Bernoulli random variables	161
5.2	Bernoulli Trials Processes	163
5.3	How to Count	165
5.3.1	Choice sequences	165
5.3.2	Orderings	166
5.3.3	Permutations	167
5.3.4	Combinations	169
5.4	Binomial Distributions	170
5.4.1	Definition	171
5.4.2	Another way to represent binomial distributions	174
5.4.3	Traits of binomial random variables	175

5.5	Simulation and Mathematical Analysis of Probability Models*	177
5.5.1	The birthday problem	177
5.5.2	Simulations	177
5.5.3	Mathematical analysis	178
5.5.4	Simulation versus mathematical analysis	180
5.E	Exercises	181

6.

Continuous Random Variables and Distributions 191

6.1	Continuous Probability Models	192
6.1.1	Why bother with continuous probability models?	192
6.1.2	“Probability zero” and “impossible”	192
6.2	Continuous Random Variables and Distributions	194
6.2.1	Cumulative probabilities	194
6.2.2	Density functions	197
6.2.3	Density functions: Intuition	204
6.2.4	Percentiles of continuous distributions	205
6.2.5	Traits of continuous random variables	206
6.3	Uniform Distributions	206
6.3.1	Definition	207
6.3.2	Traits	209
6.3.3	Shifting and scaling	209
6.4	Normal Distributions	212
6.4.1	Shifting, scaling, and the standard normal distribution	212
6.4.2	Standard normal probabilities	214
6.4.3	Normal probabilities	217
6.5	Calculating Normal Probabilities Using the Table	220
6.5.1	The standard normal distribution table	221
6.5.2	Calculating standard normal probabilities	223
6.5.3	Calculating normal probabilities	225
6.6	Sums of Independent Normal Random Variables	228
6.6.1	Distributions of sums of independent random variables	228
6.6.2	Brownian motion*	231

6.A	Continuous Distributions (using calculus)	online
6.B	Continuous Joint Distributions (using calculus)	online
6.E	Exercises	235

7.**The Central Limit Theorem 248**

7.1	I.I.D. Random Variables	249
7.2	Sums and Sample Means of I.I.D. Random Variables	252
7.2.1	Definition	252
7.2.2	Traits of sums and sample means of i.i.d. random variables	254
7.3	The Law of Large Numbers	258
7.3.1	Statement of the law of large numbers	258
7.3.2	The law of large numbers and the “law of averages”	260
7.3.3	Proving the law of large numbers*	261
7.4	The Central Limit Theorem	262
7.4.1	Convergence in distribution	263
7.4.2	Statement of the central limit theorem	264
7.4.3	Simulations with continuous trials	266
7.4.4	The continuity correction	269
7.4.5	Simulations with discrete trials	275
7.5	The Central Limit Theorem: Applications	276
7.5.1	Normal approximation of binomial distributions	276
7.5.2	Gambling	279
7.5.3	Queues	279
7.5.4	Statistical inference	282
7.A	Proof of the Central Limit Theorem	online
7.E	Exercises	284

8.**Poisson and Exponential Distributions 295**

8.1	Poisson Distributions and the Poisson Limit Theorem	296
8.1.1	e	297
8.1.2	Poisson distributions	300
8.1.3	The Poisson limit theorem	303

8.2	Exponential Distributions	309
8.2.1	Definition	309
8.2.2	Probabilities and traits	311
8.2.3	Peculiar properties	313
8.3	The Exponential Interarrival Model and the Poisson Process*	318
8.A	Appendix	321
8.E	Exercises	322

9. The Psychology of Probability 332

9.1	Thought Experiments	334
9.2	Framing Effects	335
9.3	Overconfidence	339
9.4	Misestimating the Impact of Evidence	342
9.5	The “Law of Small Numbers”	345
9.6	Gambling Systems and Technical Trading Strategies	351
9.E	Exercises	356

10. How to Lie with Statistics 365

10.1	Introduction	366
10.2	Variation	367
10.2.1	Variation within a population	367
10.2.2	Variation within subgroups: Simpson’s paradox	369
10.2.3	Variation in the results of random samples	372
10.3	Polls and Sampling	373
10.3.1	Sampling from the wrong population	373
10.3.2	Designing polls: Wording of questions	374
10.3.3	Designing polls: Selection of response alternatives	376
10.3.4	Designing polls: Arrangement of questions	377
10.3.5	Administering polls: Ensuring honest reporting	378
10.3.6	When can I trust a poll?	379
10.4	Endogenous Sampling Biases	380

10.5 Causal Inference and Extrapolation	382
10.5.1 Confounding variables	383
10.5.2 Spurious correlation and data mining	384
10.5.3 Linear extrapolation of nonlinear data	385
10.E Exercises	387

11. Data Graphics 393

11.1 Data	394
11.1.1 Types of variables	395
11.1.2 Types of data sets	397
11.1.3 Sources of economic and business data	398
11.2 Graphics for Univariate Data	399
11.2.1 Graphics that display every observation	399
11.2.2 Graphics for absolute and relative frequencies	402
11.2.3 Graphics for cumulative frequencies	408
11.3 Graphics for Multivariate Data	410
11.3.1 Graphics for frequencies	410
11.3.2 Graphics that display every observation	411
11.4 Principles for Data Graphics Design	418
11.4.1 First, do no harm	418
11.4.2 Infographics	419
11.4.3 One step beyond	421
11.A Appendix: Creating Data Graphics in Excel	online
11.E Exercises	427

12. Descriptive Statistics 435

12.1 Descriptive Statistics for Univariate Data	436
12.1.1 Measures of relative standing: Percentiles and ranges	436
12.1.2 Measures of centrality: Mean and median	440
12.1.3 Measures of dispersion: Variance and standard deviation	441

12.2	Descriptive Statistics for Bivariate Data	446
12.2.1	Measures of linear association: Covariance and correlation	446
12.2.2	Visualizing correlations	448
12.2.3	Computing correlations: Arithmetic, pictures, or computer	451
12.2.4	The road ahead: Regression analysis	456
12.E	Exercises	457

13. Probability Models for Statistical Inference 464

13.1	Introduction	465
13.2	The I.I.D. Trials Model for Statistical Inference	467
13.3	Inference about Inherently Random Processes	468
13.3.1	Bernoulli trials	469
13.3.2	Trials with an unknown distribution	470
13.4	Random Sampling and Inference about Populations	470
13.4.1	Random sampling	470
13.4.2	The trials' traits equal the data set's descriptive statistics	472
13.4.3	Bernoulli trials	474
13.4.4	Trials with an unknown distribution	475
13.5	Random Sampling in Practice	476
13.E	Exercises	482

14. Point Estimation 487

14.1	Parameters, Estimators, and Estimates	488
14.2	Desirable Properties of Point Estimators	490
14.3	The Sample Mean	492
14.3.1	Unbiasedness and consistency	493
14.3.2	Efficiency	495
14.3.3	The distribution of the sample mean	498
14.4	The Sample Variance	499
14.4.1	Defining the sample variance	500
14.4.2	Unbiasedness and consistency of the sample variance	502
14.5	Classical Statistics and Bayesian Statistics*	505

14.A	Appendix: A Short Introduction to Bayesian Statistics	507
14.B	Appendix: Derivations of Properties of the Sample Variance	515
14.E	Exercises	517

15. Interval Estimation and Confidence Intervals 527

15.1	What Is Interval Estimation?	528
15.2	Constructing Interval Estimators	529
15.2.1	The 95% interval estimator for μ when σ^2 is known	530
15.2.2	The 95% interval estimator for μ when σ^2 is unknown	534
15.2.3	The $(1 - \alpha)$ interval estimator for μ when σ^2 is unknown	535
15.2.4	Looking ahead: Standard errors and t distributions	538
15.3	Interval Estimators for Bernoulli Trials	539
15.4	Interpreting Confidence	541
15.5	Choosing Sample Sizes	548
15.5.1	Sample sizes for general i.i.d. trials	548
15.5.2	Sample sizes for Bernoulli trials processes	550
15.6	A Better Interval Estimator for Bernoulli Trials*	552
15.E	Exercises	557

16. Hypothesis Testing 567

16.1	What Is Hypothesis Testing?	568
16.2	Hypothesis Testing: Basic Concepts	569
16.2.1	The probability model	570
16.2.2	Null and alternative hypotheses	571
16.2.3	One-tailed and two-tailed tests	573
16.2.4	Hypothesis tests and their significance levels	574
16.3	Designing Hypothesis Tests	575
16.3.1	Hypothesis tests for μ when σ^2 is known	575
16.3.2	Hypothesis tests for μ when σ^2 is unknown	581
16.3.3	Hypothesis tests for Bernoulli trials	582
16.4	Two-Tailed Hypothesis Tests	585
16.4.1	Two-tailed tests vs. one-tailed tests	587
16.4.2	Comparing two-tailed hypothesis tests and confidence intervals	588

16.5	Alternate Ways of Expressing Hypothesis Tests	591
16.5.1	z -statistics	591
16.5.2	P-values	593
16.6	Interpreting Hypothesis Tests	597
16.6.1	The meaning of significance	597
16.6.2	“Do not reject” vs. “accept”	598
16.6.3	Statistical significance versus practical significance	599
16.6.4	P-value .049 vs. P-value .051	602
16.6.5	Hypothesis testing in a vacuum	602
16.7	Significance and Power	604
16.7.1	Type I and Type II errors	605
16.7.2	Evaluating error probabilities	606
16.7.3	The power curve	612
16.7.4	Underpowered studies	616
16.8	Choosing Sample Sizes	619
16.8.1	Sample sizes for general i.i.d. trials	619
16.8.2	Sample sizes for Bernoulli trials processes	621
16.9	Summary and Preview	623
16.E	Exercises	625

17.**Inference from Small Samples 641**

17.1	The t-Statistic	642
17.2	t Distributions	644
17.3	Small-Sample Inference about the Mean of Normal Trials	648
17.3.1	The t -statistic and the t distribution	648
17.3.2	Interval estimation	648
17.3.3	Hypothesis testing	650
17.4	Sort-of-Normal Trials: The Robustness of the t-Statistic	652
17.5	Evaluating Normality of Trials*	657
17.A	Appendix: Descendants of the Standard Normal Distribution	online
17.E	Exercises	663

18.**Inference about Differences in Means 671**

18.1	Inference from Two Separate Samples	672
18.1.1	The basic two-sample model	672
18.1.2	Bernoulli trials	676
18.1.3	Small samples, normal trials, equal variances*	679
18.2	Inference from Paired Samples	683
18.2.1	Constructing paired samples	683
18.2.2	The basic paired-sample model	684
18.2.3	Small samples, normal trials*	686
18.3	Choosing between Separate and Paired Samples	687
18.3.1	A general rule	687
18.3.2	Paired sampling using two observations per individual	689
18.3.3	Pairing samples using observable characteristics*	691
18.4	Causal Inference: Treatment Effects*	697
18.4.1	Randomized controlled experiments and observational studies	697
18.4.2	Interventions and causal assumptions	699
18.4.3	Potential outcomes and average treatment effects	700
18.4.4	A probability model of an observational study	701
18.4.5	Selection bias in observational studies	702
18.4.6	Random assignment eliminates selection bias	704
18.4.7	Controlling for observable confounding variables	705
18.A	Appendix: Decomposition of Variance in the Separate Sample Model	706
18.B	Appendix: The Distribution of the Pooled Sample Variance	online
18.E	Exercises	708

19.**Simple Regression: Descriptive Statistics 722**

19.1	The Regression Line	724
19.1.1	A brief review of descriptive statistics	725
19.1.2	The regression line	726
19.1.3	Examples, computations, and simulations	726

19.2	Prediction and Residuals	731
19.2.1	Predictors, predictions, and residuals	731
19.2.2	Best-in-class predictors	734
19.2.3	Further characterizations of the regression line	737
19.2.4	Deriving the best constant and best linear predictors*	739
19.3	The Conditional Mean Function	740
19.3.1	Best unrestricted prediction	740
19.3.2	Best linear prediction of conditional means	746
19.4	Analysis of Residuals	747
19.4.1	Sums of squares and variances of residuals for best-in-class predictors	747
19.4.2	Relative quality for best-in-class predictors	749
19.4.3	Decomposition of variance for regression	753
19.4.4	Sums of squares revisited	754
19.5	Pitfalls in Interpreting Regressions	755
19.5.1	Nonlinear relationships	755
19.5.2	Regression to the mean	756
19.5.3	Correlation and causation	761
19.6	Three Lines of Best Fit*	765
19.6.1	The reverse regression line	765
19.6.2	The neutral line	767
19.6.3	The three lines compared	771
19.A	Appendix	774
19.A.1	Equivalence of the characterizations of the regression line	774
19.A.2	Best linear prediction of conditional means	775
19.A.3	Relative quality for best-in-class predictors: Derivation	776
19.A.4	Decomposition of variance for regression: Derivation	777
19.B	Appendix: Characterization of the Neutral Line	online
19.E	Exercises	778