

Texts in Statistical Science

Statistics for Epidemiology

Nicholas P. Jewell



CHAPMAN & HALL/CRC

Contents

1	Introduction	1
1.1	Disease processes	1
1.2	Statistical approaches to epidemiological data	2
1.2.1	Study design	3
1.2.2	Binary outcome data	4
1.3	Causality	5
1.4	Overview	5
1.4.1	Caution: what is not covered	7
1.5	Comments and further reading	7
2	Measures of Disease Occurrence	9
2.1	Prevalence and incidence	9
2.2	Disease rates	12
2.2.1	The hazard function	13
2.3	Comments and further reading	15
2.4	Problems	16
3	The Role of Probability in Observational Studies	19
3.1	Simple random samples	20
3.2	Probability and the incidence proportion	21
3.3	Inference based on an estimated probability	22
3.4	Conditional probabilities	24
3.4.1	Independence of two events	26
3.5	Example of conditional probabilities—Berkson's bias	26
3.6	Comments and further reading	28
3.7	Problems	29
4	Measures of Disease–Exposure Association	31
4.1	Relative risk	31
4.2	Odds ratio	32
4.3	The odds ratio as an approximation to the relative risk	33
4.4	Symmetry of roles of disease and exposure in the odds ratio	34
4.5	Relative hazard	35
4.6	Excess risk	37
4.7	Attributable risk	38

4.8	Comments and further reading	40
4.9	Problems	41
5	Study Designs	43
5.1	Population-based studies	45
5.1.1	Example—mother's marital status and infant birthweight	46
5.2	Exposure-based sampling—cohort studies	47
5.3	Disease-based sampling—case-control studies	48
5.4	Key variants of the case-control design	50
5.4.1	<i>Risk-set sampling of controls</i>	51
5.4.2	Case-cohort studies	53
5.5	Comments and further reading	55
5.6	Problems	56
6	Assessing Significance in a 2×2 Table	59
6.1	Population-based designs	59
6.1.1	Role of hypothesis tests and interpretation of p-values	61
6.2	Cohort designs	62
6.3	Case-control designs	64
6.3.1	Comparison of the study designs	65
6.4	Comments and further reading	68
6.4.1	Alternative formulations of the χ^2 test statistic	69
6.4.2	When is the sample size too small to do a χ^2 test?	70
6.5	Problems	71
7	Estimation and Inference for Measures of Association	73
7.1	The odds ratio	73
7.1.1	Sampling distribution of the odds ratio	74
7.1.2	Confidence interval for the odds ratio	77
7.1.3	Example—coffee drinking and pancreatic cancer	78
7.1.4	Small sample adjustments for estimators of the odds ratio	79
7.2	The relative risk	81
7.2.1	Example—coronary heart disease in the Western Collaborative Group Study	82
7.3	The excess risk	83
7.4	The attributable risk	84
7.5	Comments and further reading	85
7.5.1	Measurement error or misclassification	86
7.6	Problems	90
8	Causal Inference and Extraneous Factors: Confounding and Interaction	93
8.1	Causal inference	94
8.1.1	Counterfactuals	94
8.1.2	Confounding variables	99

8.1.3	Control of confounding by stratification	100
8.2	Causal graphs	102
8.2.1	Assumptions in causal graphs	105
8.2.2	Causal graph associating childhood vaccination to subsequent health condition	106
8.2.3	Using causal graphs to infer the presence of confounding	107
8.3	Controlling confounding in causal graphs	109
8.3.1	Danger: controlling for colliders	109
8.3.2	Simple rules for using a causal graph to choose the crucial confounders	111
8.4	Collapsibility over strata	112
8.5	Comments and further reading	116
8.6	Problems	119
9	Control of Extraneous Factors	123
9.1	Summary test of association in a series of 2×2 tables	123
9.1.1	The Cochran–Mantel–Haenszel test	125
9.1.2	Sample size issues and a historical note	128
9.2	Summary estimates and confidence intervals for the odds ratio, adjusting for confounding factors	128
9.2.1	Woolf’s method on the logarithm scale	129
9.2.2	The Mantel–Haenszel method	130
9.2.3	Example—the Western Collaborative Group Study: part 2	131
9.2.4	Example—coffee drinking and pancreatic cancer: part 2	133
9.3	Summary estimates and confidence intervals for the relative risk, adjusting for confounding factors	134
9.3.1	Example—the Western Collaborative Group Study: part 3	135
9.4	Summary estimates and confidence intervals for the excess risk, adjusting for confounding factors	136
9.4.1	Example—the Western Collaborative Group Study: part 4	137
9.5	Further discussion of confounding	138
9.5.1	How do adjustments for confounding affect precision?	138
9.5.2	An empirical approach to confounding	142
9.6	Comments and further reading	143
9.7	Problems	144
10	Interaction	147
10.1	Multiplicative and additive interaction	148
10.1.1	Multiplicative interaction	148
10.1.2	Additive interaction	149
10.2	Interaction and counterfactuals	150
10.3	Test of consistency of association across strata	152
10.3.1	The Woolf method	153
10.3.2	Alternative tests of homogeneity	155

10.3.3 Example—the Western Collaborative Group Study: part 5	156
10.3.4 The power of the test for homogeneity	158
10.4 Example of extreme interaction	160
10.5 Comments and further reading	161
10.6 Problems	162
11 Exposures at Several Discrete Levels	165
11.1 Overall test of association	165
11.2 Example—coffee drinking and pancreatic cancer: part 3	167
11.3 A test for trend in risk	167
11.3.1 Qualitatively ordered exposure variables	169
11.3.2 Goodness of fit and nonlinear trends in risk	170
11.4 Example—the Western Collaborative Group Study: part 6	171
11.5 Example—coffee drinking and pancreatic cancer: part 4	173
11.6 Adjustment for confounding, exact tests, and interaction	175
11.7 Comments and further reading	176
11.8 Problems	176
12 Regression Models Relating Exposure to Disease	179
12.1 Some introductory regression models	181
12.1.1 The linear model	181
12.1.2 Pros and cons of the linear model	183
12.2 The log linear model	183
12.3 The probit model	184
12.4 The simple logistic regression model	186
12.4.1 Interpretation of logistic regression parameters	187
12.5 Simple examples of the models with a binary exposure	188
12.6 Multiple logistic regression model	190
12.6.1 The use of indicator variables for discrete exposures	191
12.7 Comments and further reading	196
12.8 Problems	196
13 Estimation of Logistic Regression Model Parameters	199
13.1 The likelihood function	199
13.1.1 The likelihood function based on a logistic regression model	201
13.1.2 Properties of the log likelihood function and the maximum likelihood estimate	204
13.1.3 Null hypotheses that specify more than one regression coefficient	206
13.2 Example—the Western Collaborative Group Study: part 7	207
13.3 Logistic regression with case-control data	212
13.4 Example—coffee drinking and pancreatic cancer: part 5	215
13.5 Comments and further reading	218
13.6 Problems	219

14	Confounding and Interaction within Logistic Regression Models	221
14.1	Assessment of confounding using logistic regression models . . .	221
14.1.1	Example—the Western Collaborative Group Study: part 8 . . .	223
14.2	Introducing interaction into the multiple logistic regression model	225
14.3	Example—coffee drinking and pancreatic cancer: part 6	227
14.4	Example—the Western Collaborative Group Study: part 9	230
14.5	Collinearity and centering variables	230
14.5.1	Centering independent variables	233
14.5.2	Fitting quadratic models	233
14.6	Restrictions on effective use of maximum likelihood techniques . .	235
14.7	Comments and further reading	236
14.7.1	Measurement error	237
14.7.2	Missing data	237
14.8	Problems	240
15	Goodness of Fit Tests for Logistic Regression Models and Model Building	243
15.1	Choosing the scale of an exposure variable	243
15.1.1	Using ordered categories to select exposure scale	244
15.1.2	Alternative strategies	245
15.2	Model building	246
15.3	Goodness of fit	250
15.3.1	The Hosmer–Lemeshow test	252
15.4	Comments and further reading	254
15.5	Problems	255
16	Matched Studies	257
16.1	Frequency matching	257
16.2	Pair matching	258
16.2.1	Mantel–Haenszel techniques applied to pair-matched data	262
16.2.2	Small sample adjustment for odds ratio estimator	264
16.3	Example—pregnancy and spontaneous abortion in relation to coronary heart disease in women	264
16.4	Confounding and interaction effects	265
16.4.1	Assessing interaction effects of matching variables	265
16.4.2	Possible confounding and interactive effects due to nonmatching variables	266
16.5	The logistic regression model for matched data	269
16.5.1	Example—pregnancy and spontaneous abortion in relation to coronary heart disease in women: part 2	271
16.6	Example—the effect of birth order on respiratory distress syndrome in twins	274
16.7	Comments and further reading	276

16.7.1	When can we break the match?	277
16.7.2	Final thoughts on matching	278
16.8	Problems	279
17	Alternatives and Extensions to the Logistic Regression Model	285
17.1	Flexible regression model	285
17.2	Beyond binary outcomes and independent observations	289
17.3	Introducing general risk factors into formulation of the relative hazard—the Cox model	290
17.4	Fitting the Cox regression model	293
17.5	When does time at risk confound an exposure–disease relationship?	295
17.5.1	Time-dependent exposures	296
17.5.2	Differential loss to follow-up	296
17.6	Comments and further reading	297
17.7	Problems	298
18	Epilogue: The Examples	301
	References	303
	Glossary of Common Terms and Abbreviations	311
	Index	319