STATISTICS IN MEDICINE, VOL. 8, 795–802 (1989)

SAMPLE SIZE TABLES FOR LOGISTIC REGRESSION

F. Y. HSIEH*

Department of Epidemiology and Social Medicine, Albert Einstein College of Medicine, Bronx, NY 10461, U.S.A.

SUMMARY

Sample size tables are presented for epidemiologic studies which extend the use of Whittemore's formula. The tables are easy to use for both simple and multiple logistic regressions. Monte Carlo simulations are performed which show three important results. Firstly, the sample size tables are suitable for studies with either high or low event proportions. Secondly, although the tables can be inaccurate for risk factors having double exponential distributions, they are reasonably adequate for normal distributions and exponential distributions. Finally, the power of a study varies both with the number of events and the number of individuals at risk.

KEY WORDS Logistic regression Sample size

INTRODUCTION AND ASSUMPTIONS

Logistic regression is commonly used in the analysis of epidemiologic data to examine the relationship between possible risk factors and a disease. In follow-up studies the proportion of individuals with disease (event) is usually low, but it is higher in case-control studies. In this paper I present tables of the required number of subjects in such studies for event proportions ranging from 0·01 to 0·50, covering most follow-up and case-control studies.

In the logistic regression model, the dependent variable (the disease status) is a dichotomous variable taking the values 0 for non-occurrence and 1 for occurrence. If the independent variable (such as the risk factor) is also dichotomous, the approximate required sample size can be found from published sample size tables for the comparison of two proportions.[1] For matched case-control studies, sample size calculations can be obtained from Dupont.[2] The sample size tables which I present in this paper are derived from Whittemore's formula.[3] The tables assume that the risk factors are continuous and have a joint multivariate normal distribution. The following section describes the sample size tables and their use.

THE SAMPLE SIZE TABLES

Tables I to V display the required sample size for a study using logistic regression with only one covariate (that is, risk factor). To use the tables one must specify (1) the probability P of events at the mean value of the covariate, and (2) the odds ratio r of disease corresponding to an increase of one standard deviation from the mean value of the covariate.

The tables give five choices of percentage values for the one-tailed significance level α per cent and the power $1-\beta$ per cent: (I) $\alpha=5$, $1-\beta=70$ (II) $\alpha=5$, $1-\beta=80$ (III) $\alpha=5$, $1-\beta=90$

* Current address: Anaquest, BOC Health Care, 100 Mountain Avenue, Murray Hill, NJ 07974, U.S.A.

0277–6715/89/070795–08$05.00
© 1989 by John Wiley & Sons, Ltd.

Received September 1988
Revised January 1989

(IV) $\alpha=5$, $1-\beta=95$ (V) $\alpha=1$, $1-\beta=95$. As explained in Appendix I, the sample size for an odds ratio r is the same as that required for an odds ratio $1/r$. For example, the sample sizes for odds ratios of 2 and 2·5 are the same as those required for odds ratios 0·5 and 0·4, respectively.

When there is more than one covariate in the model, multiple logistic regression may be used to estimate the relationship of a covariate to disease, adjusting for the other covariates. The sample size required to detect such a relationship is greater than that listed in Tables I to V. For the calculation of sample size, let ρ denote the multiple correlation coefficient relating the specific covariate of interest to the remaining covariates. One must specify (1) the probability P of an event at the mean value of all the covariates, and (2) the odds ratio r of disease corresponding to an increase of one standard deviation from the mean value of the specific covariate, given the mean values of the remaining covariates. The sample size read from Tables I to V should then be divided by the factor $1-\rho^2$ to obtain the required sample size for the multiple logistic regression model. This method yields an approximate upper bound rather than an exact value for the sample size needed to detect a specified association. Unlike Whittemore's formula,[3] this method does not require the user to specify the coefficients of the remaining covariates.

EXAMPLES

Whittemore[3] used Hulley's data[4] to calculate the sample size for a follow-up study designed to test whether the incidence of coronary heart disease (CHD) among white males aged 39–59 is related to their serum cholesterol level. For this study, the probability of a CHD event during an 18-month follow-up for a man with a mean serum cholesterol level is 0·07. To detect an odds ratio of 1·5 for an individual with a cholesterol level of one standard deviation above the mean using a one-tailed test with a significance level of 5 per cent and a power of 80 per cent, we need 614 individuals (from Table II).

To detect the same effect while controlling for the effects of triglyceride, and assuming that the correlation coefficient of cholesterol level with log triglyceride level is 0·4, we would need $614/(1-0{\cdot}16)=731$ individuals for the study.

DISCUSSION

These sample size tables do not explicitly require knowledge of the number of covariates in the regression model. The results in Appendix I indicate that the number of covariates is not important, and that the inclusion of new covariates which do not increase the multiple correlation coefficient with the covariate of interest does not affect sample size. An adjustment of the overall P-value for multiple significance testing may be needed when several covariates are of interest as potential risk factors for the disease.

Where one is interested in the effect of one specific covariate, Appendix I also shows that the sample sizes given in the tables may be used whatever the values of the coefficients of the remaining covariates.

The results of Monte Carlo simulations in Appendix II indicate that, when there is only one covariate in the model, the given sample sizes are reasonably accurate for both normal and exponential distributions of the covariate, although the tables can be inaccurate for some distributions, such as the double exponential. When there are two covariates having a bivariate normal distribution, the values in the tables overestimate the required sample size, but to an acceptable degree. The simulations also show that the power of the test varies both with the number of events and with the number of individuals at risk.

Whittemore[3] has found the required sample size to be very sensitive to the distribution of the covariates. I recommend that when a covariate is not normally distributed, leaving the adequacy

Table I. Sample size required for univariate logistic regression having an overall event proportion P and an odds ratio r at one standard deviation above the mean of the covariate when $\alpha = 5$ per cent (one-tailed) and $1\text{-}\beta = 70$ per cent

	Odds ratio r															
P	0·6	0·7	0·8	0·9	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·5	3·0
0·01	1799	3732	9601	43222	52828	14403	6933	4198	2878	2132	1665	1351	1128	964	546	386
0·02	925	1909	4900	22040	26938	7349	3540	2147	1474	1094	856	696	583	500	291	214
0·03	633	1301	3334	14980	18308	4997	2410	1463	1006	748	587	478	402	345	205	157
0·04	487	997	2550	11450	13993	3822	1844	1121	772	575	452	369	311	268	163	129
0·05	400	815	2080	9332	11404	3116	1505	916	631	471	371	304	256	222	137	112
0·06	341	694	1767	7920	9678	2646	1279	779	538	402	317	260	220	191	120	100
0·07	300	607	1543	6911	8445	2310	1117	681	471	352	278	229	194	169	108	92
0·08	268	542	1375	6154	7520	2058	996	608	421	315	249	206	175	152	99	86
0·09	244	491	1245	5566	6801	1862	902	551	382	286	227	187	160	139	92	81
0·10	225	451	1140	5095	6225	1705	827	505	351	263	209	173	147	129	86	77
0·12	195	390	984	4389	5362	1470	714	437	304	229	182	151	129	114	78	72
0·14	175	346	872	3885	4746	1302	633	388	270	204	163	135	116	103	72	68
0·16	159	314	788	3507	4284	1176	572	351	245	185	148	124	107	94	67	65
0·18	147	289	723	3213	3924	1078	525	323	226	171	137	115	99	88	64	62
0·20	137	268	670	2977	3636	1000	487	300	210	160	128	107	93	83	61	60
0·25	120	232	576	2554	3119	859	420	259	182	139	112	94	82	73	56	57
0·30	108	207	514	2271	2773	765	374	232	163	125	101	86	75	67	52	55
0·35	100	190	469	2069	2527	697	342	212	150	115	93	79	70	63	50	53
0·40	93	177	435	1918	2342	647	318	198	140	108	88	75	66	60	48	52
0·45	89	167	409	1801	2198	608	299	186	132	102	83	71	63	57	47	51
0·50	85	159	388	1706	2083	576	284	177	126	97	80	68	60	55	45	50

Note: To obtain sample sizes for multiple logistic regression, divide the number from the table by a factor of $1-\rho^2$, where ρ is the multiple correlation coefficient relating the specific covariate to the remaining covariates.

Table II. Sample size required for univariate logistic regression having an overall event proportion P and an odds ratio r at one standard deviation above the mean of the covariate when $\alpha = 5$ per cent (one-tailed) and $1-\beta = 80$ per cent

	Odds ratio r															
P	0·6	0·7	0·8	0·9	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·5	3·0
0·01	2334	4872	12580	56741	69359	18889	9076	5485	3751	2771	2158	1746	1453	1237	690	480
0·02	1199	2492	6421	28935	35367	9637	4635	2804	1921	1422	1110	900	751	642	367	267
0·03	821	1699	4368	19666	24037	6554	3155	1911	1311	972	760	618	517	444	260	196
0·04	632	1302	3342	15031	18371	5012	2414	1464	1006	747	585	477	401	344	206	160
0·05	518	1064	2726	12251	14972	4086	1970	1196	823	612	481	392	330	285	174	139
0·06	443	905	2315	10397	12706	3470	1674	1018	701	522	411	336	284	245	152	125
0·07	389	792	2022	9073	11087	3029	1463	890	614	458	361	296	250	217	137	115
0·08	348	707	1802	8080	9873	2699	1304	794	548	410	323	266	225	196	125	107
0·09	317	641	1631	7307	8929	2442	1181	720	497	372	294	242	206	179	116	101
0·10	291	588	1494	6689	8174	2236	1082	660	457	342	271	223	190	166	109	96
0·12	254	509	1289	5762	7041	1928	934	571	396	297	236	195	167	146	98	89
0·14	227	452	1142	5100	6231	1708	828	507	352	265	211	175	150	132	91	84
0·16	206	410	1032	4604	5624	1542	749	459	320	241	192	160	137	121	85	80
0·18	191	377	947	4218	5152	1414	687	422	294	222	178	148	128	113	80	77
0·20	178	350	878	3909	4774	1311	638	392	274	207	166	139	120	106	77	75
0·25	155	303	755	3352	4095	1126	549	339	237	180	145	122	106	94	70	71
0·30	140	271	673	2982	3641	1003	490	303	213	162	131	111	96	86	66	68
0·35	129	248	614	2717	3318	915	448	277	195	149	121	103	90	81	63	66
0·40	121	231	570	2518	3075	848	416	258	182	140	114	96	85	76	61	64
0·45	115	218	536	2364	2886	797	391	243	172	132	108	92	81	73	59	63
0·50	110	207	509	2240	2735	756	372	231	164	126	103	88	78	70	57	62

Note: To obtain sample sizes for multiple logistic regression, divide the number from the table by a factor of $1-\rho^2$, where ρ is the multiple correlation coefficient relating the specific covariate to the remaining covariates.

of the calculated sample size in doubt, one should perform a transformation[5] of the covariate to achieve normality before using Tables I to V.

In conclusion, the methods in this paper provide slightly conservative estimates of the required sample size for normally distributed covariates. The tables are simple to use and are suitable for a variety of epidemiologic studies.

Table III. Sample size required for univariate logistic regression having an overall event proportion P and an odds ratio r at one standard deviation above the mean of the covariate when $\alpha = 5$ per cent (one-tailed) and $1-\beta=90$ per cent

	Odds ratio r															
P	0·6	0·7	0·8	0·9	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·5	3·0
0·01	3192	6706	17383	78551	96029	26120	12529	7554	5154	3797	2948	2377	1972	1674	917	627
0·02	1640	3430	8873	40056	48966	13327	6398	3863	2639	1948	1516	1225	1020	869	488	349
0·03	1123	2338	6036	27225	33279	9063	4355	2632	1801	1332	1038	842	702	600	345	256
0·04	864	1792	4618	20809	25435	6930	3333	2017	1382	1024	800	650	544	466	274	210
0·05	709	1465	3767	16959	20729	5651	2720	1648	1131	839	657	534	448	385	231	182
0·06	605	1246	3199	14393	17591	4798	2311	1402	963	715	561	458	385	332	202	163
0·07	532	1090	2794	12560	15350	4189	2019	1226	843	627	493	403	340	293	182	150
0·08	476	973	2490	11185	13670	3732	1800	1094	753	561	442	362	306	265	167	140
0·09	433	882	2254	10116	12362	3377	1630	991	683	510	402	330	279	242	155	132
0·10	398	810	2065	9260	11317	3092	1494	909	628	469	370	304	258	224	145	126
0·12	347	700	1781	7977	9748	2666	1289	786	544	407	322	266	226	197	131	117
0·14	310	622	1578	7061	8627	2361	1143	698	484	363	288	238	203	178	121	110
0·16	282	564	1426	6373	7787	2133	1034	632	439	330	263	218	186	164	113	105
0·18	261	518	1308	5839	7133	1955	949	581	404	305	243	202	173	153	107	101
0·20	243	482	1214	5411	6610	1813	881	540	376	284	227	189	163	144	102	98
0·25	212	417	1043	4641	5669	1557	758	466	326	247	198	166	144	128	94	93
0·30	192	373	930	4128	5042	1387	676	417	292	222	179	151	131	117	88	89
0·35	177	342	849	3761	4593	1265	618	382	268	205	166	140	122	109	84	86
0·40	166	318	788	3486	4257	1173	574	355	250	192	155	131	115	103	81	84
0·45	157	300	741	3272	3996	1102	540	335	236	181	147	125	110	99	78	83
0·50	150	286	703	3101	3787	1045	513	319	225	173	141	120	105	95	76	81

Note: To obtain sample sizes for multiple logistic regression, divide the number from the table by a factor of $1-\rho^2$, where ρ is the multiple correlation coefficient relating the specific covariate to the remaining covariates.

Table IV. Sample size required for univariate logistic regression having an overall event proportion P and an odds ratio r at one standard deviation above the mean of the covariate when $\alpha = 5$ per cent (one-tailed) and $1-\beta=95$ per cent

	Odds ratio r															
P	0·6	0·7	0·8	0·9	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·5	3·0
0·01	4001	8439	21927	99209	121290	32967	15795	9511	6478	4765	3692	2971	2461	2084	1130	764
0·02	2055	4316	11192	50591	61847	16820	8066	4863	3318	2445	1898	1532	1272	1081	601	425
0·03	1407	2942	7614	34384	42033	11438	5490	3314	2264	1671	1301	1052	876	747	425	312
0·04	1083	2255	5825	26281	32126	8747	4202	2539	1737	1284	1002	812	678	580	337	255
0·05	888	1843	4751	21419	26182	7132	3429	2074	1421	1052	822	668	559	480	284	221
0·06	759	1568	4036	18178	22219	6056	2914	1764	1210	898	703	572	480	413	249	199
0·07	666	1372	3525	15863	19389	5287	2546	1543	1060	787	617	504	424	365	224	183
0·08	597	1225	3141	14127	17266	4710	2270	1377	947	704	553	452	381	329	205	170
0·09	543	1110	2843	12776	15614	4262	2055	1248	859	640	503	412	348	301	190	161
0·10	499	1019	2604	11696	14293	3903	1883	1145	789	588	464	380	322	279	179	153
0·12	435	881	2247	10075	12312	3365	1626	990	684	511	404	332	282	246	161	142
0·14	388	783	1991	8918	10897	2980	1442	879	608	456	361	298	254	222	148	134
0·16	353	710	1799	8049	9835	2692	1304	796	552	414	329	272	233	204	139	128
0·18	326	652	1650	7374	9010	2468	1196	732	508	382	304	252	216	190	132	123
0·20	305	607	1531	6834	8349	2288	1110	680	473	356	284	236	203	179	126	120
0·25	266	524	1316	5861	7160	1965	956	587	410	310	248	207	179	159	115	113
0·30	240	469	1173	5213	6368	1750	853	525	367	279	224	188	163	145	108	108
0·35	222	430	1071	4750	5802	1596	779	481	337	257	207	175	152	136	103	105
0·40	208	401	994	4403	5377	1481	724	448	315	240	194	164	143	129	99	103
0·45	197	378	935	4133	5047	1391	681	422	297	228	185	156	137	123	96	101
0·50	188	359	887	3917	4783	1319	647	401	283	217	177	150	132	119	94	99

Note: To obtain sample sizes for multiple logistic regression, divide the number from the table by a factor of $1-\rho^2$, where ρ is the multiple correlation coefficient relating the specific covariate to the remaining covariates.

APPENDIX I: SAMPLE SIZE FORMULAE

Let Y denote the disease status, and let $Y=1$ if the disease occurs and $Y=0$ otherwise. Let X_1, $X_2, \ldots, X_k$ denote the covariates, which are assumed to have a joint multivariate normal distribution. The logistic regression model specifies that the conditional probability of disease

Table V. Sample size required for univariate logistic regression having an overall event proportion P and an odds ratio r at one standard deviation above the mean of the covariate when $\alpha = 1$ per cent (one-tailed) and $1-\beta = 95$ per cent

	Odds ratio r															
P	0·6	0·7	0·8	0·9	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·5	3·0
0·01	5897	12367	32029	144672	176857	48120	23090	13930	9509	7011	5447	4395	3650	3100	1706	1172
0·02	3030	6326	16349	73774	90182	24552	11792	7123	4870	3597	2801	2266	1887	1609	908	651
0·03	2074	4312	11122	50141	61290	16695	8026	4854	3323	2459	1919	1556	1300	1111	642	478
0·04	1596	3305	8508	38325	46844	12767	6143	3719	2550	1890	1478	1201	1006	863	509	391
0·05	1309	2701	6940	31235	38177	10411	5013	3038	2086	1549	1213	988	830	714	429	339
0·06	1118	2299	5895	26508	32398	8839	4260	2584	1777	1321	1037	846	712	614	376	305
0·07	982	2011	5148	23132	28271	7717	3722	2260	1556	1158	911	745	628	543	338	280
0·08	879	1795	4588	20600	25175	6875	3318	2017	1390	1037	816	669	565	490	310	261
0·09	800	1627	4153	18631	22768	6221	3004	1828	1261	942	743	610	516	448	288	247
0·10	736	1493	3804	17055	20842	5697	2753	1677	1158	866	684	562	477	415	270	235
0·12	640	1292	3282	14692	17953	4911	2376	1450	1003	752	596	491	418	366	243	218
0·14	572	1148	2908	13004	15889	4350	2107	1288	893	671	533	441	376	330	224	206
0·16	521	1040	2628	11738	14341	3929	1906	1166	810	610	485	403	345	303	210	196
0·18	481	956	2410	10753	13137	3602	1749	1072	746	562	449	373	321	283	199	189
0·20	449	889	2236	9966	12174	3340	1623	996	694	524	419	349	301	266	190	183
0·25	392	768	1923	8548	10441	2869	1397	860	601	456	366	307	266	236	174	173
0·30	354	688	1713	7602	9285	2554	1247	769	539	411	331	278	242	216	163	166
0·35	326	630	1564	6927	8460	2330	1139	704	495	378	306	258	225	202	156	161
0·40	306	587	1452	6421	7840	2162	1058	656	462	354	287	243	213	192	150	157
0·45	290	553	1365	6027	7359	2031	996	618	436	335	272	231	203	183	146	154
0·50	277	527	1295	5712	6974	1926	945	587	416	320	260	222	195	177	142	152

Note: To obtain sample sizes for multiple logistic regression, divide the number from the table by a factor of $1-\rho^2$, where ρ is the multiple correlation coefficient relating the specific covariate to the remaining covariates.

occurrence $P = P(X) = P(Y = 1 | X_1, \ldots, X_k)$ is related to $X_1, \ldots, X_k$ by

$$\log[P/(1-P)] = \theta_0 + \theta_1 X_1 + \ldots + \theta_k X_k.$$

Assume, without loss of generality, that among the k covariates X_1 is the covariate of primary interest. We wish to test the null hypothesis of H_0: $\boldsymbol{\theta} = [0, \theta_2, \ldots, \theta_k]$ against the alternative hypothesis H_1: $\boldsymbol{\theta} = [\theta^*, \theta_2, \ldots, \theta_k]$.

Let ρ denote the multiple correlation coefficient of X_1 with $X_2, \ldots, X_k$. If each of the covariates X_i has been normalized to have mean zero and variance one, the sample size N needed to test at level α and power $1-\beta$ can be approximated, according to Whittemore,[3] by

$$N \exp(\theta_0) = [V^{1/2}(\boldsymbol{\theta}^0) Z_\alpha + V^{1/2}(\boldsymbol{\theta}^*) Z_\beta]^2 / \theta^{*2}, \qquad (1)$$

where $V(\boldsymbol{\theta}) = [\exp(\boldsymbol{\theta}'\Sigma\boldsymbol{\theta}/2)(1-\rho^2)]^{-1}$, $\boldsymbol{\theta}^0 = (0, \theta_2, \ldots, \theta_k)$, $\boldsymbol{\theta}^* = (\theta^*, \theta_2, \ldots, \theta_k)$, Σ is the correlation matrix of $X_1, \ldots, X_k$, and Z_α and Z_β are standard normal deviates with probabilities α and β, respectively, in the upper tail. When there is only one covariate in the model, (1) reduces to

$$NP = [Z_\alpha + \exp(-\theta^{*2}/4) Z_\beta]^2 / \theta^{*2}. \qquad (2)$$

Note that (2) relates the power of the test directly to the expected number of events NP, implying that power will be independent of sample size N given a fixed number of events. However, because of deviations from the approximate formula (1), the above statement is not accurate. Monte Carlo simulations in Appendix II show that the power of the test is an increasing function of sample size even when the number of events remains constant.

Whittemore[3] suggested that the approximation could be improved by a multiplying factor of $1 + 2P\delta$, where

$$\delta = [1 + (1 + \theta^{*2})\exp(5\theta^{*2}/4)][1 + \exp(-\theta^{*2}/4)]^{-1}. \qquad (3)$$

The required sample size may then be written as a function of P and θ^* as follows:

$$N_1 = [Z_\alpha + \exp(-\theta^{*2}/4)Z_\beta]^2 (1 + 2P\delta)/(P\theta^{*2}). \quad (4)$$

In this formula the value of θ^* represents the log odds ratio of disease corresponding to an increase in X of one standard deviation from the mean. In practice, as in Tables I to V, one would specify the value of the odds ratio r instead of the value of θ^*.

Because the standard normal distribution is symmetric about the mean 0, the sample size to detect a log odds ratio θ^*, or an odds ratio $r = \exp(\theta^*)$, is the same as a log odds ratio $-\theta^*$, or an odds ratio $1/r = \exp(-\theta^*)$.

According to Whittemore's formula,[3] the sample size calculation for the multivariate case requires specification of the coefficients for each covariate and their correlation matrix. This is impractical for routine use. Whittemore has already pointed out that the inclusion of covariates which are correlated with X_1 but independent of the event (that is, $\theta_2 = \ldots = \theta_k = 0$) leads to a loss of power when testing the relationship of X_1 to the event. Since $V(\boldsymbol{\theta}) = V(-\boldsymbol{\theta}) \geqslant 0$, it can be shown from (1) that the sample size required for $\theta_2 = \ldots = \theta_k = 0$ is an approximate upper bound after applying the correction of (3). Therefore, I suggest using approximate sample size N_M, derived from (1) by substituting $\theta_2 = \ldots = \theta_k = 0$, for multiple logistic regression:

$$N_M = N_1/(1 - \rho^2). \quad (5)$$

This formula provides maximum sample sizes and does not require the specification of coefficients for each covariate, or the full correlation matrix.

APPENDIX II: MONTE CARLO POWER SIMULATIONS

To check the accuracy of the calculated sample size tables, Monte Carlo simulations were carried out for various proportions of events and for covariates with different distributions. A set of 1000 trials was generated for five event proportions, four odds ratios and three distributions. Let m and s be the mean and standard deviation, respectively, of the covariates; let U and G be standard uniform and standard normal variables, respectively. The distributions of covariates were generated as follows:

1. Normal distribution: $m + sG$.
2. Double exponential distribution: $m + Is\log U$, where $I = -1$ if $U^* < 0{\cdot}5$ and $I = 1$ otherwise. U^* is a standard uniform variable independent of U.
3. Exponential distribution: $m - s - s\log U$.

Under the alternative hypothesis that the specified association between covariate and disease occurrence exists, the above mean and standard deviation are specified by the odds ratios through the following equations:

Diseased group: $m = \log(\text{odds ratio})$ and $s = \exp(-m^2/4)$,

Non-diseased group: $m = 0$ and $s = 1$.

When two covariates (X_1, X_2) have a joint bivariate normal distribution with a correlation coefficient ρ, the first covariate (X_1) was generated by the above procedure 1 and the second covariate (X_2) was obtained from $X_2 = \rho X_1 + G(1 - \rho^2)^{1/2}$.

The results are shown in Table VI. Formula (4) seems to slightly underestimate the power for a covariate with a normal distribution and severely overestimate the power for a double exponential covariate. For an exponential covariate, formula (4) overestimates the power when both odds ratio

Table VI. Estimated power from Monte Carlo simulations (1000 repetitions) for normal, double exponential and exponential covariates, compared with formula (4)

Event proportion P	No. of events/ sample size	Source	Odds ratio r 1·30 Sig. level 0·05	1·30 0·01	1·50 Sig. level 0·05	1·50 0·01	1·70 Sig. level 0·05	1·70 0·01	2·00 Sig. level 0·05	2·00 0·01
0·02	20/1000	Formula (4)	0·296	0·110	0·533	0·265	0·741	0·466	0·923	0·744
		Normal	0·326	0.111	0.532	0.281	0.742	0.475	0.923	0.757
		Double exp.	0·217	0·078	0·338	0·125	0·522	0·266	0·729	0·434
		Exponential	0·265	0·112	0·485	0·226	0·718	0·420	0·947	0·755
0·05	30/600	Formula (4)	0·388	0·164	0·680	0·404	0·874	0·661	0·980	0·902
		Normal	0·418	0·163	0·723	0·463	0·907	0·702	0·989	0·945
		Double exp.	0·263	0·107	0·439	0·193	0·672	0·384	0·851	0·648
		Exponential	0·364	0·151	0·675	0·376	0·903	0·667	0·994	0·936
0·10	40/400	Formula (4)	0·443	0·201	0·751	0·487	0·920	0·749	0·990	0·941
		Normal	0·456	0·206	0·813	0·546	0·949	0·834	0·999	0·979
		Double exp.	0·294	0·107	0·547	0·291	0·733	0·463	0·905	0·732
		Exponential	0·431	0·208	0·787	0·516	0·955	0·785	1·000	0·978
0·20	60/300	Formula (4)	0·520	0·260	0·832	0·599	0·959	0·843	0·996	0·972
		Normal	0·562	0·283	0·894	0·711	0·990	0·917	1·000	0·994
		Double exp.	0·350	0·139	0·630	0·331	0·823	0·588	0·970	0·868
		Exponential	0·603	0·298	0·869	0·652	0·977	0·898	1·000	0·996
0·50	100/200	Formula (4)	0·568	0·301	0·866	0·655	0·969	0·871	0·996	0·972
		Normal	0·599	0·327	0·888	0·697	0·986	0·929	1·000	0·995
		Double exp.	0·368	0·152	0·674	0·397	0·855	0·643	0·975	0·895
		Exponential	0·627	0·336	0·891	0·687	0·991	0·921	1·000	1·000

Table VII. Estimated power from Monte Carlo simulations (1000 repetitions) for bivariate normal covariates compared with formula (4)

Event proportion P	No. of events/ sample size	Source	Correlation coefficient	Odds ratio r 1·50 Sig. level 0·05	1·50 0·01	1·70 Sig. level 0·05	1·70 0·01	2·00 Sig. level 0·05	2·00 0·01
0·05	30/600	Formula (4)	0·0	0·680	0·404	0·874	0·661	0·980	0·902
			0·3	0·644	0·366	0·845	0·611	0·970	0·867
			0·7	0·438	0·193	0·624	0·339	0·827	0·568
		Normal	0·0	0·723	0·463	0·907	0·702	0·989	0·945
			0·3	0·659	0·392	0·897	0·661	0·987	0·923
			0·7	0·457	0·217	0·647	0·377	0·870	0·627
0·20	60/300	Formula (4)	0·0	0·832	0·599	0·959	0·843	0·996	0·972
			0·3	0·799	0·551	0·943	0·801	0·993	0·955
			0·7	0·578	0·304	0·769	0·502	0·917	0·703
		Normal	0·0	0·894	0·711	0·990	0·917	1·000	0·994
			0·3	0·849	0·635	0·975	0·890	0·999	0·992
			0·7	0·657	0·378	0·842	0·607	0·976	0·878
0·50	100/200	Formula (4)	0·0	0·866	0·655	0·969	0·871	0·996	0·972
			0·3	0·836	0·606	0·955	0·833	0·993	0·956
			0·7	0·619	0·341	0·796	0·538	0·918	0·732
		Normal	0·0	0·888	0·697	0·986	0·929	1·000	0·995
			0·3	0·894	0·633	0·978	0·906	1·000	0·995
			0·7	0·640	0·351	0·871	0·658	0·979	0·897

Table VIII. Estimated power from Monte Carlo simulations (1000 repetitions) for different sample sizes and number of events

Number of events	Sample size	Odds ratio r: 1·30, Sig. level 0·05	1·30, 0·01	1·50, Sig. level 0·05	1·50, 0·01	1·70, Sig. level 0·05	1·70, 0·01	2·00, Sig. level 0·05	2·00, 0·01
20	40	0·170	0·042	0·343	0·088	0·508	0·187	0·756	0·372
	100	0·272	0·084	0·482	0·200	0·698	0·420	0·887	0·668
	400	0·325	0·131	0·551	0·256	0·736	0·465	0·931	0·754
50	100	0·360	0·140	0·671	0·347	0·847	0·626	0·977	0·881
	250	0·507	0·243	0·831	0·571	0·959	0·826	0·999	0·990
	1000	0·566	0·316	0·869	0·662	0·984	0·927	1·000	1·000
100	200	0·583	0·301	0·883	0·702	0·986	0·932	1·000	0·998
	500	0·749	0·509	0·987	0·915	1·000	0·994	1·000	1·000
	2000	0·840	0·610	1·000	1·000	1·000	1·000	1·000	1·000

and event rate are low and underestimates the power otherwise. In general, if the covariate is normally distributed, we are assured that the sample size obtained from the tables will be slightly conservative. Table VII shows that formula (4) underestimates the power for bivariate normal covariates, but to an acceptable degree. Table VIII shows the results of simulations using normal covariates relating the number of events and the sample size to the power of the test. They show that when the number of events remains constant, the power of the test varies with sample size.

ACKNOWLEDGEMENTS

I thank the reviewers for very helpful comments.

REFERENCES

1. Fleiss, J. *Statistical Methods for Rates and Proportions*, Wiley, New York, 1981.
2. Dupont, W. D. 'Power calculations for matched case-control studies', *Biometrics*, **44**, 1157–1168 (1988).
3. Whittemore, A. 'Sample size for logistic regression with small response probability', *Journal of the American Statistical Association*, **76**, 27–32 (1981).
4. Hully, S. B., Rosenman, R. A., Bowol, R. and Brand, R. 'Epidemiology as a guide to clinical decisions: the association between triglyceride and coronary heart disease', *New England Journal of Medicine*, **302**, 1383–1389 (1980).
5. Sokal, R. R. and Rohlf, F. J. *Biometry*, Freeman, San Francisco, 1969.