The Philippine Statistician
Vol. xxvi, Nos. 3 & 4 1977

SAMPLING CONSIDERATION FOR DISEASES WITH LOW PREVALENCE

By ILDEFONSO T. CRUZ[1] and ELIZABETH TAN[2]

An important consideration in the planning of a survey is sample size determination. And in this process the usual question which evolves following thoughtful statistical deliberations bears upon how large a sample should be studied in order for the results to meet certain requirements, such as specified precision of estimates for parameters of interest. To the practitioners of the sampling art and to most administrators of survey projects, this is a very crucial question since it is evidently wasteful to have too large a sample, and useless to have one which is too small. A rational answer, as if to underscore its importance, is not always easy to find for in the majority of cases, we do not possess enough information to guide us in the choice of a sample size which could be considered "best" in some sense, on account of our lack of knowledge about certain properties of the underlying population. Nevertheless the usual and most immediate objective of an investigation for setting sample size is the determination of a minimum number of units to constitute the sample so as to fulfill certain specifications, such as the desired precision or non-exceedence of error we are willing to tolerate in the estimates.

Now in a situation where the condition to be studied is relatively rare in the population, the main interest may not be in the estimation of the minuscule prevalence per se, but in ascertaining how extensive the sampling should be so that there will be a good chance of discovering at least one or a specified number of cases. The important considerations relative to this type of sampling outlook appears to be the following:

(i) the rarer the prevalence of the disease or condition, the more difficult it is to encounter a case, and

---

[1] Professor and Chairman, Department of Epidemiology and Biostatistics, Institute of Public Health, University of the Philippines System.

2 Master of Statistics, 1977.

(ii) even with a very large sample, there is always a non-zero probability that not even a single case will be seen, with this probability increasing markedly as the prevalence goes down.

In this context therefore, one can only talk of probabilities of including at least one or a specified number of cases in any given sample.

A more precise formulation of the problem then is: What must be the size of a study group from a large population in order to achieve a high probability, say, not less than $1-\alpha$ (e.g. 95% or α=.05), so that

(i) at least one case is included in the sample, or more generally, so that
(ii) at least r cases ($r > 1$) are present in the sample?

METHODOLOGY

Some Results From Binomial Sampling

In sampling problems involving a disease of a given prevalence, say p, the number X of cases found in a sample of size n follows a binomial distribution. This is true regardless of whether p is small or not, provided n is small relative to the size of the population so that sampling is in effect, with replacement. On this basis formulas for minimum n which will yield at least m cases, with probability $(1-\alpha)$ or more, are derived as follows:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n. \tag{1}$$

Then

$$P(X < m) = \sum_{x=0}^{m-1} \binom{n}{x} p^x (1-p)^{n-x}, \tag{2}$$

and

$$P(X \geqslant m) = 1 - \sum_{x=0}^{m-1} \binom{n}{x} p^x (n-x)^{n-x} \geqslant 1 - \alpha, \tag{3}$$

as specified.

For $m = 1$, the solution for n turns out to be a closed expression obtained as follows:

$$P(X \geq 1) \geq 1 - \alpha, \tag{4}$$

$$\text{i.e. } 1 - (1 - p)^n \geq 1 - \alpha. \tag{5}$$

$$\text{Hence} \qquad n = \ln \alpha \,/\, \ln(1 - p), \tag{6}$$

where the sign " $>$" has been omitted with the understanding that n here is the smallest sample size to achieve the problem specifications.

For $m=2$, the value of n will be given by the solution to

$$(1-p)^{n-1}\,(1-p+np) = \alpha, \tag{7}$$

and for the general case where $m=r$, n is the solution to the equation

$$(1-p)^n + \binom{n}{1}(1-p)^{n-1}p + \ldots + \binom{n}{r-1}(1-p)^{n-r+1}p^{r-1} = \alpha. \tag{8}$$

It will be noted that when $m=1$, the solution is straightforward. When m exceeds 1, the equations have to be solved by trial and error or by some iterative procedure, such as the method of false position or the more popular Newton-Raphson technique. Since the labor involved in the process of iteration increases tremendously with every rise in m, the only practicable way is through a computerized approach. Several such trials were made in solving (3); the Newton-Raphson method in particular turned out to be feasible at the lower levels of m. However, a serious underflow problem cropped up at the higher values, since n will be large correspondingly and hence, a number of infinitesimal magnitude results when p (which is itself assumed to be rather small) is raised to a large exponent. An alternative approach is to use the Poisson approximation, which in this case turns out to be extremely accurate in view of the low levels stipulated for p. In addition, there is considerable simplication of the equations used in the iteration process, together with the disappearance to a large enough degree, of the underflow problem encountered earlier.

The methodology for this alternative is developed more fully in the next section.

The Use of the Poisson Approximation

The binominal distribution with parameters n and p, under the circumstances where n approaches infinity and p approaches zero but such that np remains constant, say equal to λ, approximately obeys the Poisson probability law with parameter $\lambda=np$; that is

$$P(X=x) = \binom{n}{x} p^x(1-p)^{n-x} \approx e^{-\lambda}\lambda^x/x! \qquad (9)$$

Then $$Pr(X<m) = \sum_{x=0}^{m=1} \binom{n}{x} p^x(1-p)^{n-x} \approx \sum_{x=0}^{m-1} e^{-\lambda}\lambda^x/x! \qquad (10)$$

for any fixed integers $x = 0, 1, 2, \ldots$.

This leads to results parallel to those of equations (6), (7) and (8). Thus for

i) $m = 1$,

$$e^{-\lambda} = \alpha$$

or

$$n = -\ln\alpha/p. \qquad (11)$$

Note that this reduces to equation (6) with the use of the well-known approximation $\ln(1-p) \approx -p$ for small p.

ii) $m = 2$, n is calculated from

$$e^{-\lambda}(1 + \lambda) = \alpha \qquad (12)$$

or $$\ln(1 + \lambda) - \lambda - \ln\alpha = 0\ ; \qquad (13)$$

iii) $m = r$, n is the solution to the equation

$$e^{-\lambda}\{1+\lambda+\lambda^2/2!+\lambda^3/3+\ldots+\lambda^{m-1}/(m-1)!\} = \alpha \qquad (14)$$

or

$$\ln\{1+\lambda+\lambda^2/2!+\ldots+\lambda^{r-1}/(r-1)!\} - \lambda - \ln\alpha = 0. \qquad (15)$$

Looking at the general case (case iii) it may be noted that (14)

is easily expressed in terms of an incomplete Γ-function ratio[3], since

$$e^{-\lambda}[1+\lambda+\lambda^2/2! + \ldots +\lambda^{m-1}/(m-1)!]$$
$$= 1 - \int_0^{\lambda} u^{m-1} e^{-u}\, du \Big/ \int_o^{\infty} u^{m-1} e^{-u}\, du$$
$$= 1 - I(\lambda/\sqrt{m}, m-1). \qquad (16)$$

where, using Pearson's notation for the ratio,

$$I(\lambda/\sqrt{m}, m-1) = \int_0^{\lambda} u^{m-1} e^{-u}\, du \Big/ \int_0^{\infty} u^{m-1} e^{-u}\, du. \qquad (17)$$

Thus we may restate (14) as

$$I(\lambda/\sqrt{m}, m-1) = 1-\alpha. \qquad (18)$$

Unfortunately, this result, while seemingly elegant leads to a laborious process which does not circumvent the repetitive nature of the calculations even with the use of tables. Thus computerization of this procedure does not appear to be promising nor practicable. Another approach is the development of a Newton-Raphson routine to the iterative solution of equation (15). An outline for this is as follows:

$$\text{Let } F(\lambda) = \ln\{1+\lambda+\lambda^2/2!+\ldots+\lambda^{r-1}/(r-1)!\} - \lambda - \ln\alpha. \qquad (19)$$

$$\text{Then } F'(\lambda) = \partial F/\partial\lambda = -\lambda^{r-1}/[(r-1)!\{1+\lambda+\lambda^2/2!+\ldots+\lambda^{r-1}/(r-1)!\}] \qquad (20)$$

If λ_i is a provisional root of $F(\lambda)$ then a better approximation is given by

$$\lambda_{i+1} = \lambda_i + \delta(\lambda_i) \qquad (21)$$

where $\delta(\lambda_i) = -F(\lambda_i)/F'(\lambda_i)$, an additive correction applied to the provisional root λ_i to arrive at the next iterate. $F(\lambda_i)$ and $F'(\lambda_i)$ are understood to be values of the functions $F(\lambda)$ and $F'(\lambda)$ evaluated at the point λ_i. As the process continues, we obtain a succession of approximations which should converge to the real root. Under convergence conditions, the difference between λ_i and λ_{i+1}, i.e. $|\lambda_{i+1} - \lambda_i|$, diminishes rapidly as i increases and a practical operating rule is to terminate the iteration when $|\lambda_{i+1} - \lambda_i|$ becomes less than some small number, here taken to be 10^{-6}. The choice of the starting value λ_o is oftentimes critical in keeping the number of iterations down to a reasonable level. In this case it was found that taking $\lambda_o \approx 1.8$ m will hold that number to a value less than 10. A detailed investigation revealed, at the least for the first few cases, that there will be no problems in attaining convergence.

A FORTRAN — IV computer program based on the Newton-Raphson solution was developed and values of n were generated at the IBM 360 facility at the U.P. Computer Center at Diliman. A compilation of the results is shown in Tables 1 and 2, for stated values of the prevalence p and number of cases m, at $(1-\alpha)$ levels of 90% and 95%.

To see how close the approximation is to the exact results from the binomial, the example below is worked out, using analogous equations (6) and (11).

For $p = 10/100{,}000 = .0001$ and $\alpha = .05$,

$n = 29{,}955.8$ by equation (6) while equation (11) yields

$n = 29{,}957.3$.

These values do not differ by any appreciable degree.

DISCUSSION

The sample size n may be read directly from the tables for listed levels of p and m. However not all intervening values between the limits chosen for these parameters are given and therefore in the applications, we need to note if

(i) the number of cases m and prevalence p are both listed in the table or

(ii) the desired number of cases is given while the specified p is not,

(iii) both m and p are unlisted in the tables.

If it is (i) then the sample size may be read directly from the table, while if it is (ii) we need to use the relation $n = \hat{\lambda}/p$ in solving for n, where $\hat{\lambda}$ is the solution obtained for λ at that particular m. Case (iii) can be handled by interpolation but a better method is fashioned on the basis of the observation that the plot of λ on m is nearly linear on double logarithmic paper, notably in the range $m \geqslant 10$, where the estimation for non-tabulated n will be necessary. The charts shown in figures 1 and 2 show that extent of this linearity for α levels of 5% and 10%respectively. Least squares fitting applied to log λ on log m yielded the equations

$$\hat{\lambda}_m = 2.014013\ m^{0.878967} \tag{22}$$

for $\alpha = .05$, and

$$\hat{\lambda}_m = 1.723966\ m^{0.9059051} \tag{23}$$

for $\alpha = .10$.

These may be used for estimating λ for $m \geqslant 10$, from which n is easily obtained. As an example, consider the situation where p is thought to be around 5/100,000 and it is desired to draw a sample which will yield at least 16 cases at the .95 probability level. Since m=16 is not tabulated, we use (22) to estimate λ.

Thus $\hat{\lambda}_{16} = 2.014013\ \ (16)^{0.878967} = 23.03796$,

and $n = (23.03796/5)\ \ 100{,}000 = 46076$,

a result which appears quite reasonable when compared to the nearest tabular entries.

A general idea of how far the results of this procedure compare with the computer-generated values may be obtained by taking an m for which the sample size can be read directly from tables 1 (or 2) and then applying the above procedure to get a parallel estimate for n. Thus from Table 1 for m = 20 and p = 5/100,000, n = 557,585. On the other hand equation (22) yields

$$\hat{\lambda}_{20} = 2.014013 \quad (20)^{.878967} \quad = 28.0301,$$

and

$$n = (28.0301/5)\ 100{,}000 = 560{,}602.$$

The percentage error is

560,602 — 557,585)/557,585 = .54%, which appears to be tolerable considering the levels of sample size requirements involved.

AN APPLICATION

One of the important developments in the health scenario in recent years is the increasing attention devoted to cancer research and control, resulting in improved survival rates of patients. There has been a noticeable rise in the rates since the 1960's and this is continuing into the present decade. In fact, the prognosis of patients with certain forms of cancer is considerably brighter now than ten years ago. These improvements are due to developments in surgical and supportive techniques, in radio theraphy and in diagnostic procedures. Indeed one of the recognized measures for the effective control of cancer is by prevention and prophylactic treatment of invasive forms. In order to achieve this, accurate and practicable diagnostic tests were and are still being developed. Now the clinical usefulness of such a test rests on the attainment of a happy balance between its so-called sensitivity and its specificity, for, an insensitive test gives too many negative results for the disease it is supposed to pick up while a non-specific test gives many positive results among individuals free of the disease it is supposed to diagnose.

Many diagnostic tests suffer from at least one of the above shortcomings. Hence the evaluation of the usefulness of a particular diagnostic test requires careful study — a study which

by its very nature has to deal with a broad base of subjects and thus transcend the clinic level out into the realm of statistics.

Lingao et. al (1975)[4] proposed a modified Alpha-Fetoprotein (AFP) test for the diagnosis of primary hepatoma (liver cancer). In their report, an attempt to assess the sensitivity and specificity of the test was made. Some 753 patients were subjected to the test and the results were reported prior to the diagnoses of the attending physicians. Of these patients only 119 proceeded to a state where conclusive diagnoses for various illnesses were arrived at, either by autopsy, exploratory surgery or needle biopsy. The study centered on this latter group of patients so that there can be no question as to the correctness of diagnosis.

In discussing the results of this and similar studies it is convenient to introduce the following notation:

Let D be the event that a person has the disease in question, say hepatoma,
$\bar{D}$ the event that he does not have the disease,
T the event that he gives a positive AFP test results, and
$\bar{T}$ the event that he gives a negative test response.

If the test is applied to samples of individuals known to have the disease (D's) and not to have the disease ($\bar{D}$'s), the results may be displayed in the following manner:

GROUP	AFP TEST RESULTS	
	Positive T	Negative $\bar{T}$
Sick (D)	P(T/D)	P($\bar{T}$/D)
Not Sick ($\bar{D}$)	P(T/$\bar{D}$)	P($\bar{T}$/$\bar{D}$)

where P(T/D) = probability of a positive test result given that the individual has the disease,

[4] Lingao, Augusto et al. "A Modified Alpha-Fetoprotein Test for the Diagnosis of Primary Hepatoma," Phil. Jour. Internal Medicine, Volume 13, (July-September 1975) pp. 109-123.

$P(\bar{T}/\bar{D})$ = probability of a negative test given that the individual does not have the disease.

The other conditional probabilities are interpreted in a similar manner.

Let P(D) be the unconditional probability or proportion of the population who are sick (prevalence of the disease),

P(T) be the overall proportion responding positive to the test.

With these formulation we can now lay down more formal definitions of the concepts of sensitivity and specificity of a diagnostic test: (P(T/D) is sensitivity and expresses the ability of the test to pick up those who are really sick. Specificity on the other hand is $P(\bar{T}/\bar{D})$, which measures the ability of the test to detect an individual who is in reality free of the disease. In practice, greater concern is placed on the error rates associated with the diagnostic test if it were to be used in a survey or a screening program. This in turn leads to a lot of misconceptions among many researchers, particularly in the health field, since misclassification is of serious dimensions usually when the overall prevalence of the disease is low. The problem is compounded when one attempts to use the findings of the test to estimate this prevalence in a survey.

The initial difficulty is on sample size. In the case of hepatoma, no reliable figures on prevalence for the Philippines are available and one has to rely on data from other Asiatic populations published elsewhere and spotty reports of local investigations. It appears from these sources that a resonable fix on the overall prevalence of liver cancer is anywhere from 10/100,000 to 45/100,000 population. Suppose it is 30/100,000 and the investigator wants to see at least 15 cases. From Table 2, it is seen that he will need about 73,000 (72,955 exactly) to attain this minimum yield with 95% assurance. Many health researchers will be amazed (if not shocked) by this seemingly voluminous requirement and the reason is not too difficult to see. Most of them have been trained in if not actually working within the confines of a hospital or medical laboratory and hence are accustomed to applying a diagnostic test to individuals who are at least suspected, if not clinically identified, as having the disease. They are thus conditioned to

seeing the test pick out a lot of cases among individuals which are in many respects highly selected. They need therefore some orientation on what the performance would be if the test is tried out under field conditions and the findings of this study, notably, Tables 1 and 2, provide useful information which will now allow most to appreciate the situation from that perspective.

Having gotten around this problem, the next one is concerned about the nature of the yield of the test. And here, in the case of low prevalence disease, it appears that the specificity becomes very crucial.

Going back to the 119 patients with confirmed diagnoses, 67 turned out to have primary hepatoma while the rest (52) were found to have other diseases. The findings are summarized below:

RESULTS OF MODIFIED AFP TEST ON 119 PATIENTS WITH CONFINED DIAGNOSES

DISEASE STATUS	AFP TEST		Total
	Positive	Negative	
Hepatoma	57	10	67
Non-hepatoma	9	43	52
Total	66	53	119

Thus,

$$\text{sensitivity} = (57/67)\ 100 = 85.1\%, \text{ and}$$

$$\text{specificity} = (43/52)\ 100 = 82.7\%.$$

Some caution should be exercised in projecting this specificity estimate to field conditions, since the non-hepatoma group appeared to be overloaded with other liver conditions which, though non-primary liver cancer, nevertheless yield weak but positive AFP test. There is therefore some grounds to suspect underestimation of specificity in this case. This is further supported by a run of negative test results on a series of 10 healthy subjects reported in the same study. Thus perhaps a more realistic estimate, though possibly stil on the low side, is

$$((43+10)/(52+10)) \times 100 = 85.5\%,$$

with the inclusion of the healthy group of indivduals tested. It is interesting to note in this regard that the standard AFP

test proved to be very specific in hands of other workers.[5]

To see what sort of difficulty arises with the use of the test with the assumption of specificity even up to the level of 85.5= as recomputed, consider the problem above in its original context where a requisite sample of 73,000 individuals is to be tested.[6] The expectation here is at least 15 primary hepatoma cases. The total number of positive results expected is

$$\begin{aligned} &15 \times \text{sensitivity level} + (73{,}000 - 15) \\ &\times (1 = \text{specificity level}) \\ &= 15\ (0.851) + 72{,}985\ (1 - 0.855) \\ &= 15 + 10{,}583 \\ &= 10{,}596, \end{aligned}$$

of which the larger component (10,583) constitute the false positives. Hence the proportion of false positives, or false positivity rate is

$$\frac{10{,}583}{10{,}596} \times 100 = 99.9\%$$

Therefore, nearly all positives are false positives, in this situation where a moderately specific test is applied in a mass survey for low prevalence disease. There is serious misclassification error in this direction. The false negatives, on the other hand, will not be much of a problem since the total negative results expected is

$$\begin{aligned} &15\ (1\text{-sensitivity}) + (73{,}00 - 15)\ (\text{specificity}) \\ &= 2 + 62{,}402 \\ &= 62{,}404 \end{aligned}$$

of which only 2 (the smaller component) are false.

---

[5] See for instance, Application of Serum Alpha Feto-Protein in Mass Survey of Primary Carcinoma of the Liver. The co-ordinating Group for the Research of Liver Cancer, People's Republic of China, Am. J. Chinese Med. 2,: No. 3, pp. 241-245, 1974.

[6] An appeal to Bayes' theorem at this point would have led to a more rigid presentation and the same findings, but the simplified approach adopted here appears to be more understandable in an intuitive sense.

TABLE 1. MINIMUM SAMPLE SIZE WHICH WILL YIELD WITH 90% PROBABILITY THE STATED NUMBER OF CASES OR MORE, FOR VARIOUS LEVELS OF EXPECTED PREVALENCE

	No. of Cases	EXPECTED PREVALENCE, CASES/100,000								
λ	M	5	10	15	20	25	30	35	40	45
2.3026	1	46052	23026	15351	11513	9210	7675	6579	5756	5117
3.8897	2	77794	38897	25931	19449	15559	12966	11113	9724	8644
5.3223	3	106446	53223	35482	26612	21289	17741	15207	13306	11827
6.6808	4	133616	66808	44539	33404	26723	22269	19088	16702	14846
7.9936	5	159872	79936	53291	39968	31974	26645	22839	19984	17764
9.2747	6	185494	92747	61831	46373	37099	30916	26499	23187	20610
10.5321	7	210642	105321	70214	52660	42128	35107	30092	26330	23405
11.7709	8	235418	117709	78473	58855	47084	39236	33631	29427	26158
12.9947	9	259894	129947	86631	64974	51979	43316	37128	32487	28877
14.2060	10	284120	142060	94707	71030	56824	47353	40589	35515	31569
20.1280	15	402560	201280	134187	100640	80512	67093	57509	50320	44729
25.9025	20	518051	259025	172684	129513	103610	86342	74007	64756	57561
31.5836	25	631671	315836	210557	157918	126334	105279	90239	78959	70186
37.1985	30	743970	371985	247990	185993	148794	123995	106281	92996	82663
42.7685	35	855271	427635	285090	213818	171054	142545	122182	106909	95030
48.2891	40	965782	482891	321927	241446	193157	160964	137969	120723	107309
53.7825	45	1075650	537825	358550	268913	215130	179275	153664	134456	119517
59.2490	50	1184980	592490	394993	296245	236996	197497	169283	148123	131664
64.6926	55	1293852	646926	431284	323463	258770	215642	184836	161732	143761
70.1163	60	1402326	701163	467442	350581	280465	233721	200332	175291	155814
75.5226	65	1510452	755226	503484	377613	302091	251742	215779	188807	167828
80.9135	70	1618270	809135	539423	404567	323654	269712	231181	202284	179808
86.2906	75	1725812	862906	575271	431453	345163	287635	246545	215727	191757
91.6553	80	1833106	916553	611035	458276	366621	305518	261872	229138	203678
97.0087	85	1940174	970087	646725	485044	388035	323362	277168	242522	215575
102.3518	90	2047037	1023518	682346	511759	409408	341173	292434	255880	227449
107.6855	95	2153711	1076855	717904	538428	430742	358952	307673	269214	239301
113.0105	100	2260211	1130105	753404	565053	452042	376702	322887	282526	251135

TABLE 1. MINIMUM SAMPLE SIZE WHICH WILL YIELD WITH 90% PROBABILITY THE STATED NUMBER OF CASES OR MORE, FOR VARIOUS LEVELS OF EXPECTED PREVALENCE

	No. of Cases	EXPECTED PREVALENCE, CASES/100,000								
λ	M	50	55	60	65	70	75	80	85	90
2.3026	1	4605	4187	3838	3542	3289	3070	2878	2709	2558
3.8897	2	7779	7072	6483	5984	5557	5186	4862	4576	4322
5.3223	3	10645	9677	8871	8188	7603	7096	6653	6262	5914
6.6808	4	13362	12147	11135	10278	9544	8908	8351	7860	7423
7.9936	5	15987	14534	13323	12298	11419	10658	9992	9404	8882
9.2747	6	18549	16863	15458	14269	13250	12366	11593	10911	10305
10.5321	7	21064	19149	17553	16203	15046	14043	13165	12391	11702
11.7709	8	23542	21402	19618	18109	16816	15695	14714	13848	13079
12.9947	9	25989	23627	21658	19992	18564	17326	16243	15288	14439
14.2060	10	28412	25829	23677	21855	20294	18941	17757	16713	15784
20.1280	15	40256	36596	33547	30966	28754	26837	25160	23680	22364
25.9025	20	51805	47096	43171	39850	37004	34537	32378	30474	28781
31.5836	25	63167	57425	52639	48590	45119	42111	39479	37157	35093
37.1985	30	74397	67634	61998	57228	53141	49598	46498	43763	41332
42.7685	35	85527	77752	71273	65790	61091	57018	53454	50310	47515
48.2891	40	96578	87798	80482	74291	68984	64385	60361	56811	53655
53.7825	45	107565	97786	89638	82742	76832	71710	67228	63274	59758
59.2490	50	118498	107725	98748	91152	84641	78999	74061	69705	65832
64.6926	55	129385	117623	107821	99527	92418	86257	80866	76109	71881
70.1163	60	140233	127484	116860	107871	100166	93488	87645	82490	77907
75.5226	65	151045	137314	125871	116189	107889	100697	94403	88850	83914
80.9135	70	161827	147115	134856	124482	115591	107885	101142	95192	89904
86.2906	75	172581	156892	143818	132755	123272	115054	107863	101518	95878
91.6553	80	183311	166646	152759	141008	130936	122207	114569	107830	101839
97.0087	85	194017	176379	161681	149244	138584	129345	121261	114128	107787
102..3518	90	204704	186094	170586	157464	146217	136469	127940	120414	113724
107.6855	95	215371	195792	179476	165670	153836	143581	134607	126689	119651
113.0105	100	226021	205474	188351	173862	161444	150681	141263	132954	125567

TABLE 1. MINIMUM SAMPLE SIZE WHICH WILL YIELD WITH 90% PROBABILITY THE STATED NUMBER OF CASES OR MORE, FOR VARIOUS LEVELS OF EXPECTED PREVALENCE

	No. of Cases	EXPECTED PREVALENCE, CASES/100,000								
λ	M	95	100	150	200	250	300	350	400	500
2.3026	1	2424	2303	1535	1151	921	768	658	576	461
3.8897	2	4094	3890	2593	1945	1556	1297	1111	972	778
5.3223	3	5602	5322	3548	2661	2129	1774	1521	1331	1064
6.6808	4	7032	6681	4454	3340	2672	2227	1909	1670	1336
7.9936	5	8414	7994	5329	3997	3197	2665	2284	1998	1599
9.2747	6	9763	9275	6183	4637	3710	3092	2650	2319	1855
10.5321	7	11086	10532	7021	5266	4213	3511	3009	2633	2106
11.7709	8	12390	11771	7847	5885	4708	3924	3363	2943	2354
12.9947	9	13679	12995	8663	6497	5198	4332	3713	3249	2599
14.2060	10	14954	14206	9471	7103	5682	4735	4059	3552	2841
20.1280	15	21187	20128	13419	10064	8051	6709	5751	5032	4026
25.9025	20	27266	25903	17268	12951	10361	8634	7401	6476	5181
31.5836	25	33246	31584	21056	15792	12633	10528	9024	7896	6317
37.1985	30	39156	37199	24799	18599	14879	12400	10628	9300	7440
42.7685	35	45014	42764	28509	21382	17105	14255	12218	10691	8553
48.2891	40	50831	48289	32193	24145	19316	16096	13797	12072	9658
53.7825	45	56613	53783	35855	26891	21513	17928	15366	13446	10757
59.2490	50	62367	59249	39499	29625	23700	19750	16928	14812	11850
64.6926	55	68097	64693	43128	32346	25877	21564	18484	16173	12939
70.1163	60	73807	70116	46744	35058	28047	23372	20033	17529	14023
75.5226	65	79497	75523	50348	37761	30209	25174	21578	18881	15105
80.9135	70	85172	80913	53942	40457	32365	26971	23118	20228	16183
86.2906	75	90832	86291	57527	43145	34516	28764	24654	21573	17258
91.6553	80	96479	91655	61104	45826	36662	30552	26187	22914	18331
97.0087	85	102114	97009	64672	48504	38803	32336	27717	24252	19402
102.3518	90	107739	102352	68235	51176	40941	34117	29243	25588	20470
107.6855	95	113353	107686	71790	53843	43074	35895	30767	26921	21537
113.0105	100	118958	113011	75340	56505	45204	37670	32289	28253	22602

TABLE 2. MINIMUM SAMPLE SIZE WHICH WILL YIELD WITH 95% PROBABILITY THE STATED NUMBER OF CASES OR MORE, FOR VARIOUS LEVELS OF EXPECTED PREVALENCE

	No. of Cases	EXPECTED PREVALENCE, CASES/100,000								
λ	M	5	10	15	20	25	30	35	40	45
2.9957	1	59915	29957	19972	14979	11983	9986	8559	7489	6657
4.7439	2	94877	47439	31626	23719	18975	15813	13554	11860	10542
6.2958	3	125916	62958	41972	31479	25183	20986	17988	15739	13991
7.7537	4	155073	77537	51691	38768	31015	25846	22153	19384	17230
9.1535	5	183070	91535	61023	45768	36614	30512	26153	22884	20341
10.5130	6	210261	105130	70087	52565	42052	35043	30037	26283	23362
11.8424	7	236848	118424	78949	59212	47370	39475	33835	29606	26316
13.1481	8	262962	131481	87654	65741	52592	43827	37566	32870	29218
14.4347	9	288693	144347	96231	72173	57739	48116	41242	36087	32077
15.7052	10	314104	157052	104701	78526	62821	52351	44872	39263	34900
21.8865	15	437730	218865	145910	109432	87546	72955	62533	54716	48637
27.8792	20	557585	278792	185862	139396	111517	92931	79655	69698	61954
33.7524	25	675048	337524	225016	168762	135010	112508	96435	84381	75005
39.5410	30	790820	395410	263606	197705	158164	131803	112974	98852	87869
45.2656	35	905312	452656	301771	226328	181063	150885	129330	113164	100590
50.9397	40	1018795	509397	839598	254699	203759	169799	145542	127349	113199
56.5726	45	1131453	565726	377151	282863	226291	188575	161636	141432	125717
62.1711	50	1243421	621711	414474	310855	248684	207237	177632	155428	138158
67.7401	55	1354802	677401	451601	338700	270961	225800	193543	169350	150534
73.2837	60	1465674	732837	488558	366418	293135	244279	209382	183209	162853
78.8050	65	1576100	788050	525366	394025	315220	262683	225157	197012	175122
84.3065	70	1686130	843065	562043	421532	337226	281022	240876	210766	187348
89.7903	75	1795807	897903	598602	448952	359162	299301	256544	224476	199534
95.2582	80	1905165	952582	635055	476291	381033	317527	272166	238146	211685
100.7117	85	2014234	1007117	671411	503558	402847	335706	287748	251779	223804
106.1520	90	2123040	1061520	707680	530760	424608	353840	303291	265380	235893
111.5801	95	2231603	1115801	743868	557901	446321	371934	318800	278950	247956
116.9971	100	2339943	1169971	779981	584986	467989	389990	334278	292493	259994

TABLE 2. MINIMUM SAMPLE SIZE WHICH WILL YIELD WITH 95% PROBABILITY THE STATED NUMBER OF CASES OR MORE, FOR VARIOUS LEVELS OF EXPECTED PREVALENCE

No. of Cases		EXPECTED PREVALENCE, CASES/100,000								
λ	M	50	55	60	65	70	75	80	85	90
2.9957	1	5991	5447	4993	4609	4280	3994	3745	3524	3329
4.7439	2	9488	8625	7906	7298	6777	6325	5930	5581	5271
6.2958	3	12592	11447	10493	9686	8994	8394	7870	7407	6995
7.7537	4	15507	14098	12923	11929	11077	10338	9692	9122	8615
9.1535	5	18307	16643	15256	14082	13076	12205	11442	10769	10171
10.5130	6	21026	19115	17522	16174	15019	14017	13141	12368	11681
11.8424	7	23685	21532	19737	18219	16918	15790	14803	13932	13158
13.1481	8	26296	23906	21914	20228	18783	17531	16435	15468	14609
14.4347	9	28869	26245	24058	22207	20621	19246	18043	16982	16039
15.7052	10	31410	28555	26175	24162	22436	20940	19632	18477	17450
21.8865	15	43773	39794	36477	33672	31266	29182	27358	25749	24318
27.8792	20	55758	50690	46465	42891	39827	37172	34849	32799	30977
33.7524	25	67505	61368	56254	51927	48218	45003	42191	39709	37503
39.5410	30	79082	71893	65902	60832	56487	52721	49426	36519	43934
45.2656	35	90531	82301	75443	69639	64665	60354	56582	53254	50295
50.9397	40	101879	92618	84900	78369	72771	67920	63675	69929	56600
56.5726	45	113145	102859	94288	87035	80818	75430	70716	66556	62858
62.1711	50	124342	113038	103618	95648	88816	82895	77714	73142	69079
67.7401	55	135480	123164	112900	104216	96772	90320	84675	79694	75267
73.2837	60	146567	133243	122139	112744	104691	97712	91605	86216	81426
78.8050	65	157610	143282	131342	121238	112579	105073	98506	92712	87561
84.3065	70	168613	153285	140511	129702	120438	112409	105383	99184	93674
89.7903	75	179581	163255	149651	138139	128272	119720	112238	105636	99767
95.2582	80	190517	173197	158764	146551	136083	127011	119073	112069	105842
100.7117	85	201423	183112	167853	154941	143874	134282	125890	118484	111902
106.1520	90	212304	193004	176920	163311	151646	141536	132690	124885	117947
111.5801	95	223160	202873	185967	171662	159400	148774	139475	131271	123978
116.9971	100	233994	212722	194995	179996	167139	155996	146246	137644	129997

TABLE 2. MINIMUM SAMPLE SIZE WHICH WILL YIELD WITH 95% PROBABILITY THE STATED NUMBER OF CASES OR MORE, FOR VARIOUS LEVELS OF EXPECTED PREVALENCE

	No. of Cases	EXPECTED PREVALENCE, CASES/100,000								
λ	M	95	100	150	200	250	300	350	400	500
2.9957	1	3153	2996	1997	1498	1198	999	856	749	599
4.7439	2	4994	4744	3163	2372	1898	1581	1355	1186	949
6.2958	3	6627	6296	4197	3148	2518	2099	1799	1574	1259
7.7537	4	8162	7754	5169	3877	3101	2585	2215	1938	1551
9.1535	5	9635	9154	6102	4577	3661	3051	2615	2288	1831
10.5130	6	11066	10513	7009	5257	4205	3504	3004	2628	2103
11.8424	7	12466	11842	7895	5921	4737	3947	3384	2961	2368
13.1481	8	13840	13148	8765	6574	5259	4383	3757	3287	2630
14.4347	9	15194	14435	9623	7217	5774	4812	4124	3609	2887
15.7052	10	16532	15705	10470	7853	6282	5235	4487	3926	3141
21.8865	15	23038	21886	14591	10943	6755	7295	6253	5472	4377
27.8792	20	29347	27879	18586	13940	11152	9293	7965	6970	5576
33.7524	25	35529	33752	22502	16876	11501	11251	9644	8438	6750
39.5410	30	41622	39541	26361	19770	15816	13180	11297	9885	7908
45.2656	35	47648	45266	30177	22633	18106	15089	12933	11316	9053
50.9397	40	53621	50940	33960	25470	20376	16980	14554	12735	10188
56.5726	45	59550	56573	37715	28286	22629	18858	16164	14143	11315
62.1711	50	65443	62171	41447	31086	24868	20724	17763	15543	12434
67.7401	55	71305	67740	45160	33870	27096	22580	19354	16935	13548
73.2837	60	77141	73284	48856	36642	29313	24428	20938	18321	14657
78.8050	65	82953	78805	52537	39402	31522	26268	22516	19701	15761
84.3065	70	88744	84306	56204	42153	33723	28102	24088	21077	16861
89.7903	75	94516	89790	59860	44895	35916	29930	25654	22448	17958
95.2582	80	100272	95258	63505	47629	38103	31753	27217	23815	19052
100.7117	85	106012	100712	67141	50356	40285	33571	28775	25178	20142
106.1520	90	111739	106152	70768	53076	42461	35384	30329	26538	21230
111.5801	95	117453	111580	74387	55790	44632	37193	31880	27895	22316
116.9971	100	123155	116997	77998	58499	46799	38999	33428	29249	23399

FIGURE 1

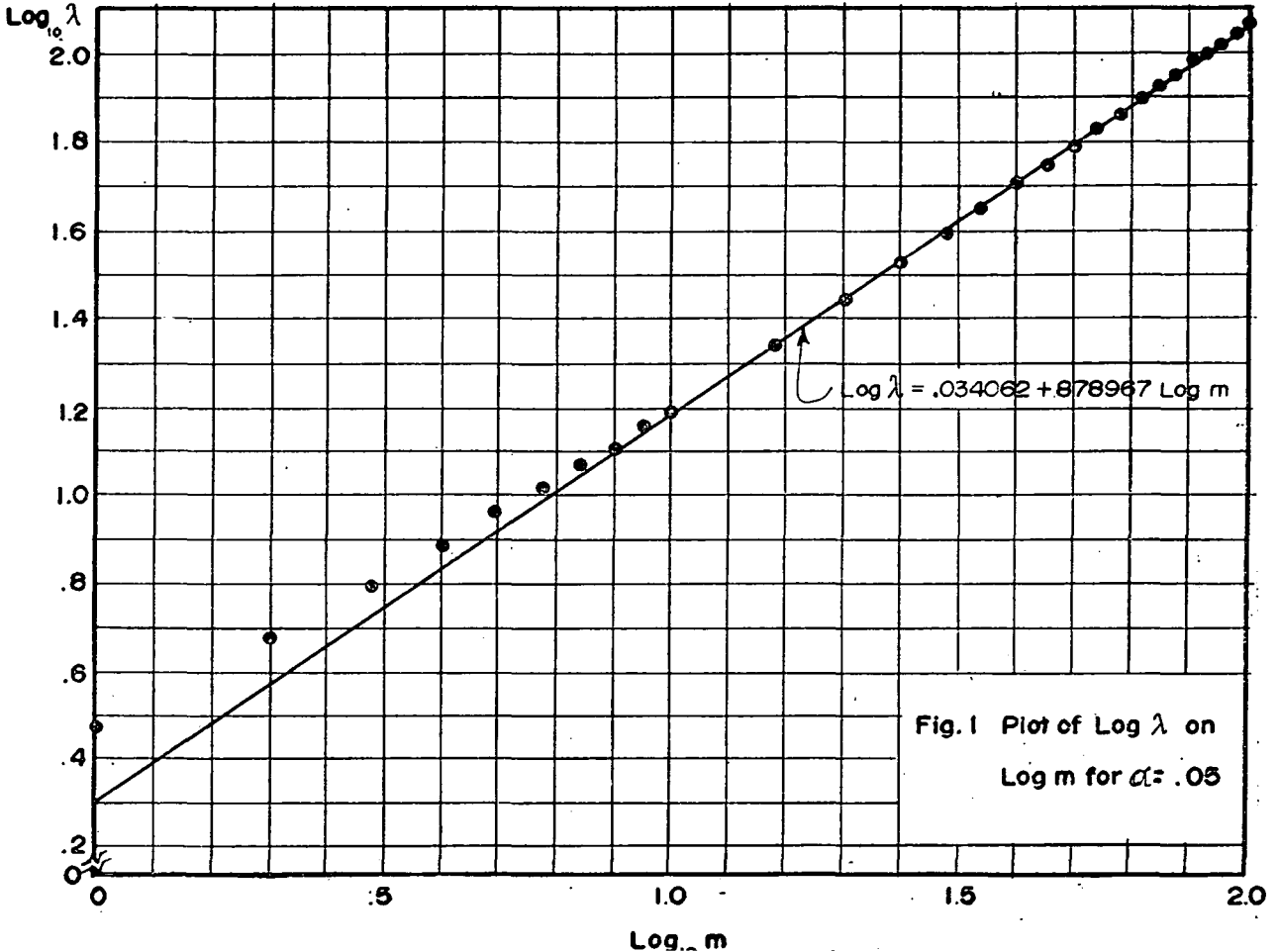

Fig. 1 Plot of Log λ on Log m for α = .05

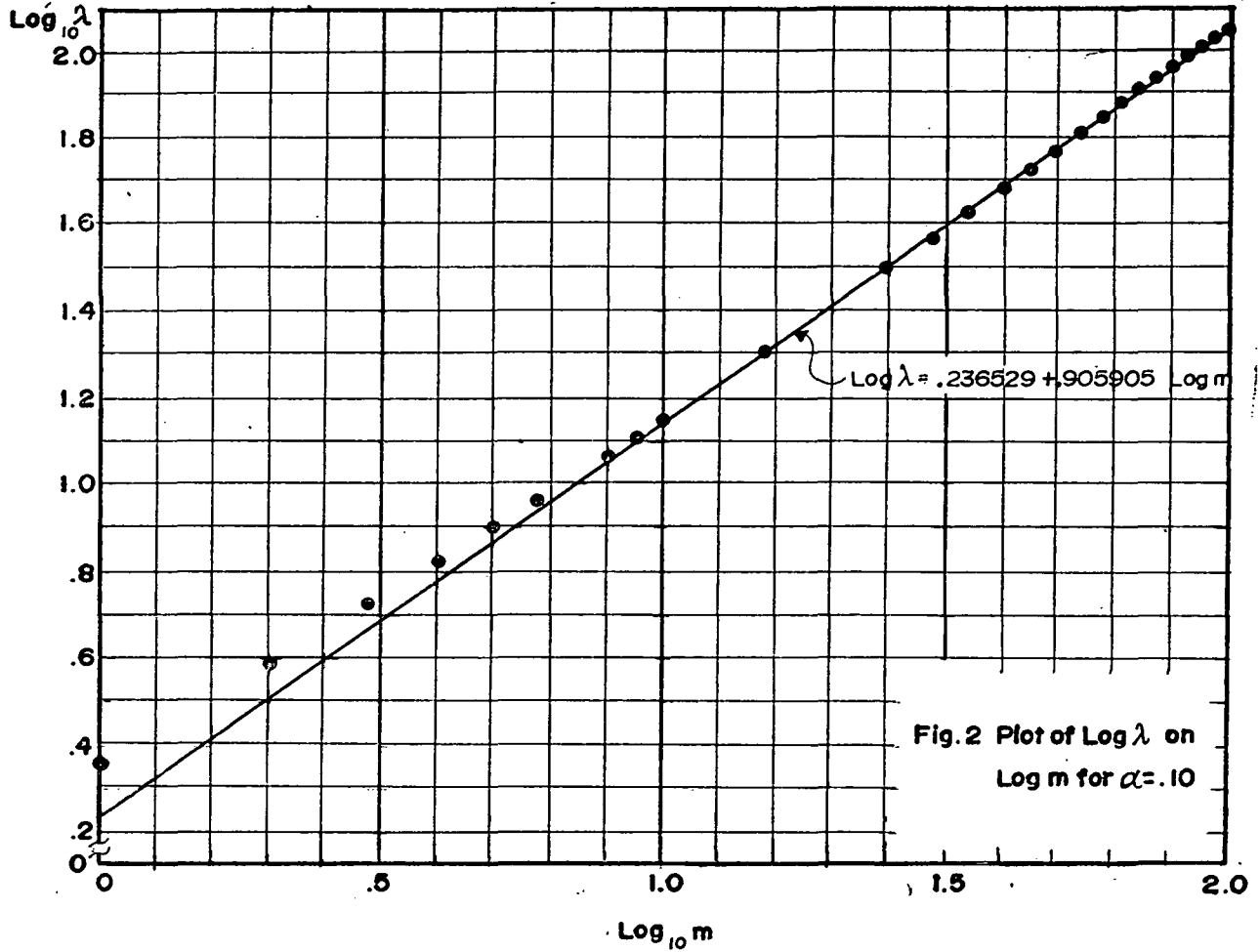

Fig. 2 Plot of Log λ on Log m for α = .10