I. Wong

891

Analysis of Completeness of the Earthquake Sample in the Puget Sound Area and Its Effect on Statistical Estimates of Earthquake Hazard

by

J. C. Stepp
National Oceanic and Atmospheric Administration
Environmental Research Laboratories
Boulder, Colorado 80302

In this paper the 100-year sample of earthquakes known to have occurred in the Puget Sound area between 1870 and 1969 is evaluated for completeness and the question of fitting the frequency formula $\log N = a + bI_0$ to biased samples that are short with respect to the recurrence interval of the largest earthquakes contained in them is studied. The usual bias in earthquake catalogs against small shocks is found to be particularly severe for the Puget Sound sample. A sample that is homogeneous in maximum intensity V and larger quakes would be confined to the most recent 15-year interval. This sample is too short to obtain satisfactory estimates of long-term earthquake occurrence. A method is described for determining the interval in an intensity class over which that class is homogeneous. Using this interval to determine the mean rate of occurrence for that class, one is able to fit the frequency formula, $\log N = A+bI_0$, for entirely homogeneous samples. The procedure permits the complete 100-year sample to be used to fit the frequency formula.

Introduction:

The validity of the empirical recurrence relation for earthquakes (1, 2)

$$\log N = a + b M, \qquad (1)$$

where N is the number of earthquakes, M is magnitude, and a and b are constants established from an observed data sample, has been confirmed in many seismicity studies. Recently the formula has been shown to hold for microearthquakes (3) and for microfractures (4, 5), suggesting that it must reflect some fundamental physical property of the fracture process. Presumably, then, the relation also holds for tectonic regions of interest in detailed earthquake hazard mapping. Indeed, some earthquake hazard zoning in the USSR has been based on this assumption (6).

In practice the incompleteness of available samples of earthquake data make it difficult to obtain fits of equation (1) that are thought to represent true long-term recurrence rates. All earthquake catalogs are biased against small shocks, because of seismograph station density or, in the early records, population density. Accordingly, the bias is more severe in successively earlier reporting periods. For example, it has been shown (7) that the

southern California catalog is homogeneous in magnitude 4.0 and larger events since 1933 and in magnitude 3.0 and larger quakes since 1953. To fit equation (1) using an earthquake catalog, the choice must be between using a short sample that is complete in small events or a longer sample that is complete in only large quakes. It has been suggested (9) that a 29-year sample drawn from small regions of the dimension of interest in earthquake hazard mapping may not give earthquake recurrence estimates that represent long term seismicity. Therefore, it is necessary to use longer samples that give more accurate statistical averages of the large earthquakes of primary engineering interest (10).

In this paper the historical sample of earthquakes known to have occurred in the Puget Sound region (Figure 1) between 1870 and 1969 is analyzed for completeness and a method is proposed for fitting the frequency formula that makes use of the complete 100-year sample. Because the sample of events for which magnitudes are known is extremely small, maximum modified Mercalli intensity is used as the measure of earthquake size.

Statement of Problem:

The nature of the problem can be illustrated by attempting to fit the formula, $\log N = a + bI_0$, to the recurrence rates shown in Figure 2. In this illustration the convention of grouping events in intensity increments is used; $N(I_0)$ is the number of quakes per year having maximum intensity I_0 and a and b are constants to be determined by fitting the points $\log N(I_0)$ on I_0 by the method of least-squares. For a given sample the problem is to select the interval of intensities to be used in the least squares fit. The proper lower bound is determined by the completeness of the data sample, while the upper bound is governed by the length of the sample. Although a statistical test has recently been suggested for determining the lower bound of completeness of a sample (7), the usual method in practice has been to choose the interval of data to be fitted by inspection of a plot of $\log N(I_0)$ on I_0. The smallest intensity is usually selected as the value where $\log N(I_0)$ clearly departs from a straight line plot, while the largest earthquake in the sample is usually included. If this procedure is applied to the 100-year sample from the Puget Sound area shown in Figure 2, one cannot determine whether the lower bound above which the data may be considered homogeneous should be V or VI. Thus, while maximum intensity IV quakes can be excluded on inspection as being incompletely reported in the 100-year sample interval, no clear choice can be made between intensity V, represented by case II, and intensity VI, represented by case III, as the lowest intensity above which the sample is completely reported. Indeed, based on the confidence intervals on the coefficient b, each of the three cases can be said to represent the sample equally well.

General Considerations of Data Reporting:

A number of definitive catalogs of historical earthquakes have been compiled for the Puget Sound and adjoining area (11, 12, 13, 14). The data used in this study were drawn from these definitive listings and updated through 1969 with data from the NOAA serial "United States Earthquakes."

Figure 3 shows the number of earthquakes per decade grouped in three intensity ranges, $I_0 < IV$, $IV \leq I_0 < VI$, and $I_0 > VI$. In addition, the total number of events per decade are plotted. The numerical data corresponding to Figure 3 are listed in Table I for the complete record from 1840 through 1969.

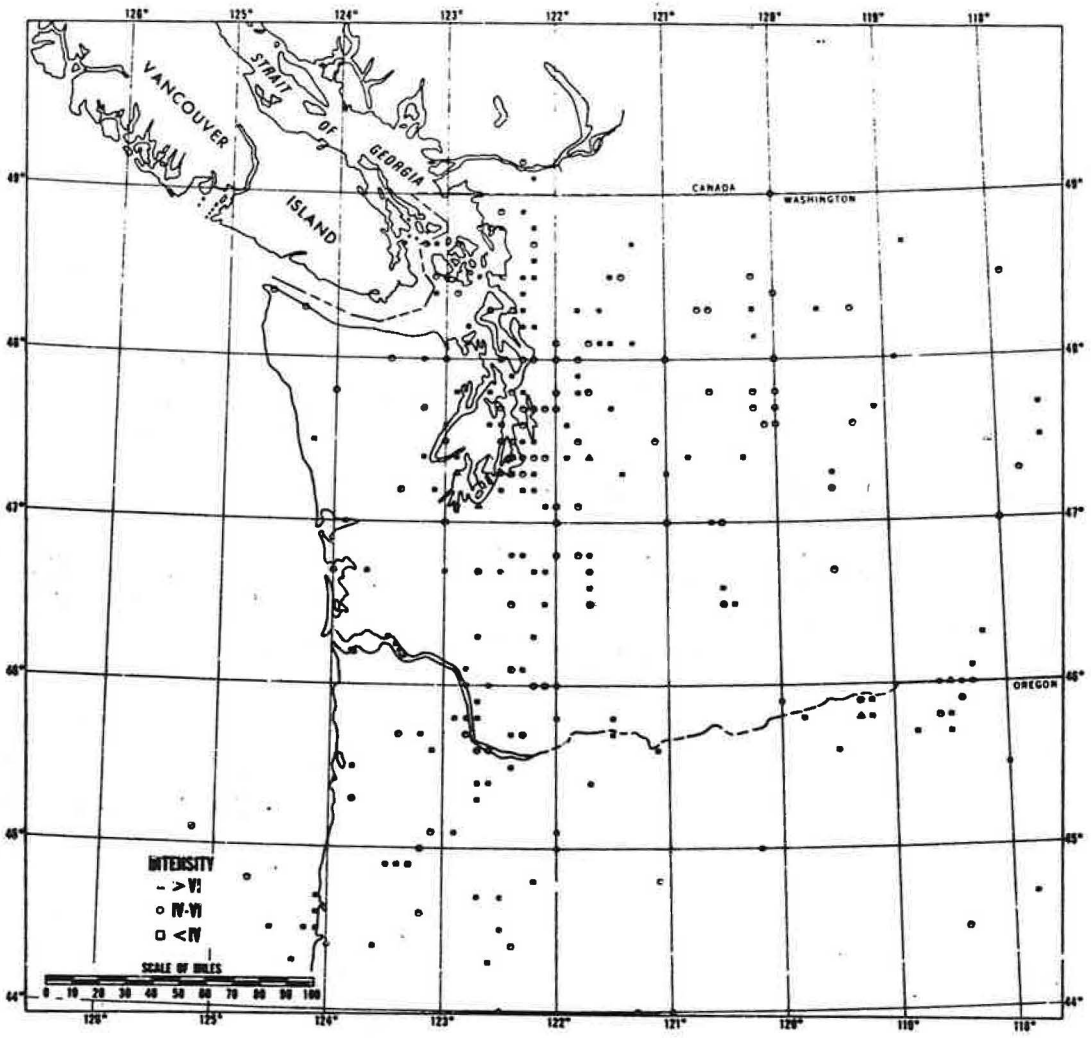

FIGURE 1. Geographic Distribution of Earthquake Epicenters in the Vicinity of the Puget Sound Through 1969

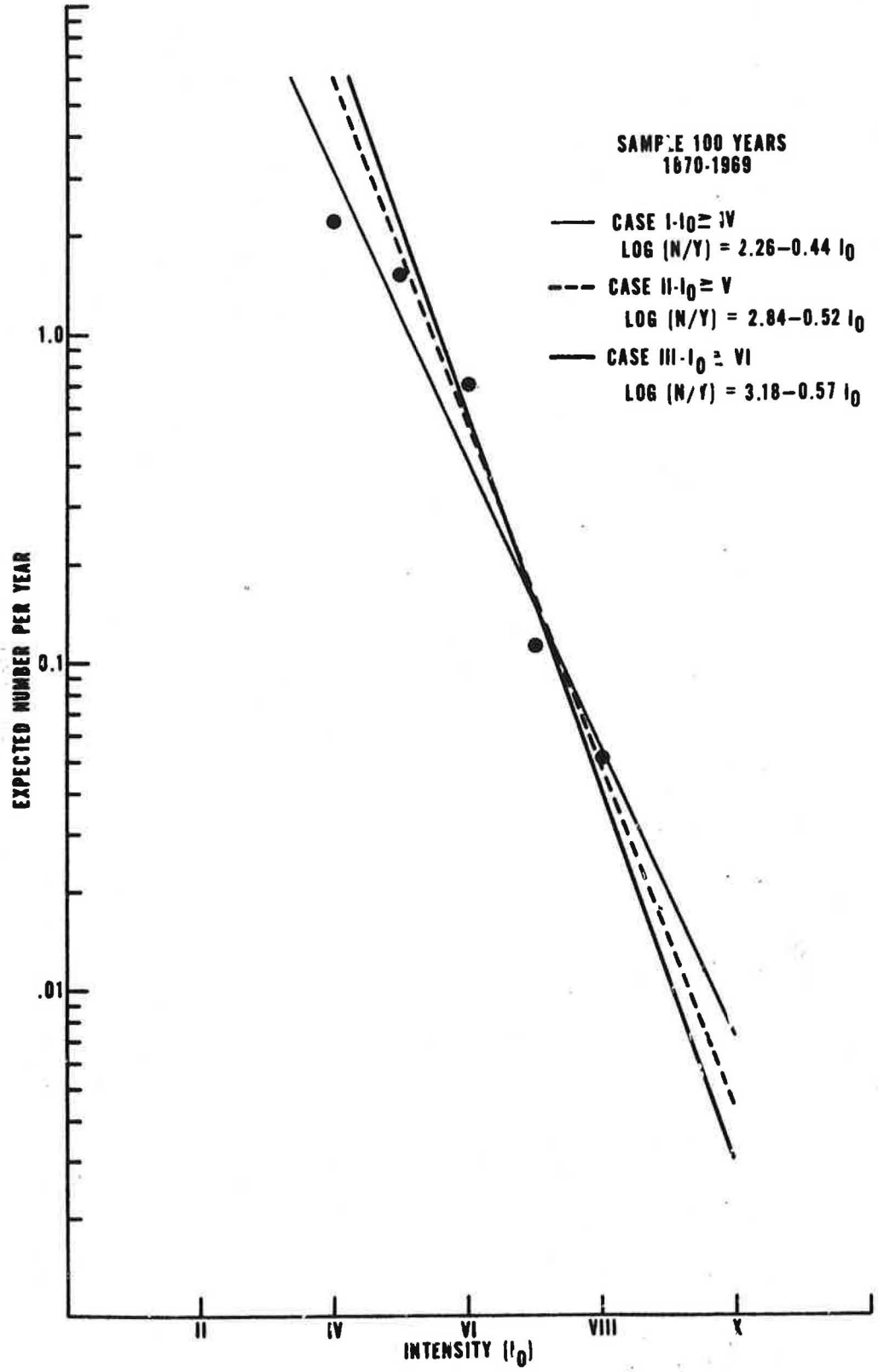

FIGURE 2. Conventional Least-Squares Fit of Log (N/yr) = $a + bI_o$ for 100-Year Sample, Assuming all Events of $I_o \geq$ IV to be Equally Completely Reported

Table I.

Number of Earthquakes Reported in Each Decade Since the Beginning of the Available Historical Record

Decade	$I_0 < IV$	$IV \leq I_0 \leq VI$	$I_0 > VI$	Total
1840-1849	1	1	0	2
1850-1859	1	1	0	2
1860-1869	4	1	0	5
1870-1879	8	2	3	13
1880-1889	18	7	0	25
1890-1899	20	19	1	40
1900-1909	9	16	1	26
1910-1919	28	34	0	62
1920-1929	19	22	1	42
1930-1939	112	61	2	175
1940-1949	53	70	5	128
1950-1959	39	68	0	107
1960-1969	43	82	1	126
	355	384	14	753

The three earliest decades of the record are so obviously lacking in data that it did not seem worthwhile plotting them.

A number of observations that relate to the completeness of the sample can be made. The first feature to note is that there is no reason, based on the results presented in the figure, to question the completeness of the large earthquakes ($I_0 > VI$) since 1870. The fluctuation in the number of maximum intensity VII and larger earthquakes reported per decade shows no trend in the 100-year sample period from 1870 through 1969. Because these events have an average felt area of about 200,000 square kilometers, they are widely experienced. They, accordingly, would have been widely reported even in the early period. It is likely, therefore, that these large earthquakes have been completely reported during the past 100 years, even though their true maximum intensities may not always have been observed.

Secondly, the most significant jump in the total number of reported events occurs in the decade 1930-39. This can be explained by the added interest in earthquake reporting in consequence of adding all postmasters to the questionnaire canvass in 1929. Approximately 71 percent of the earthquakes in the total historical record were reported in the 40-year interval from 1930 through 1969. This is in contrast to 23 percent in the next older 40-year interval from 1890 through 1929, and six percent in the first 50 years of the record.

A third important feature of Figure 3 is the near constant number of total reported events since the decade beginning 1930. Indeed, if one excludes the 36 aftershocks associated with magnitude 5.8 earthquake of July 15, 1936, centered near the Washington-Oregon border at the 46°N, 119°W, the near constancy of the total number of reported events is even more apparent. This is not interpreted as meaning that there is a complete record of earthquakes in the Puget Sound region since 1930. Rather, it means that the intensity reports are

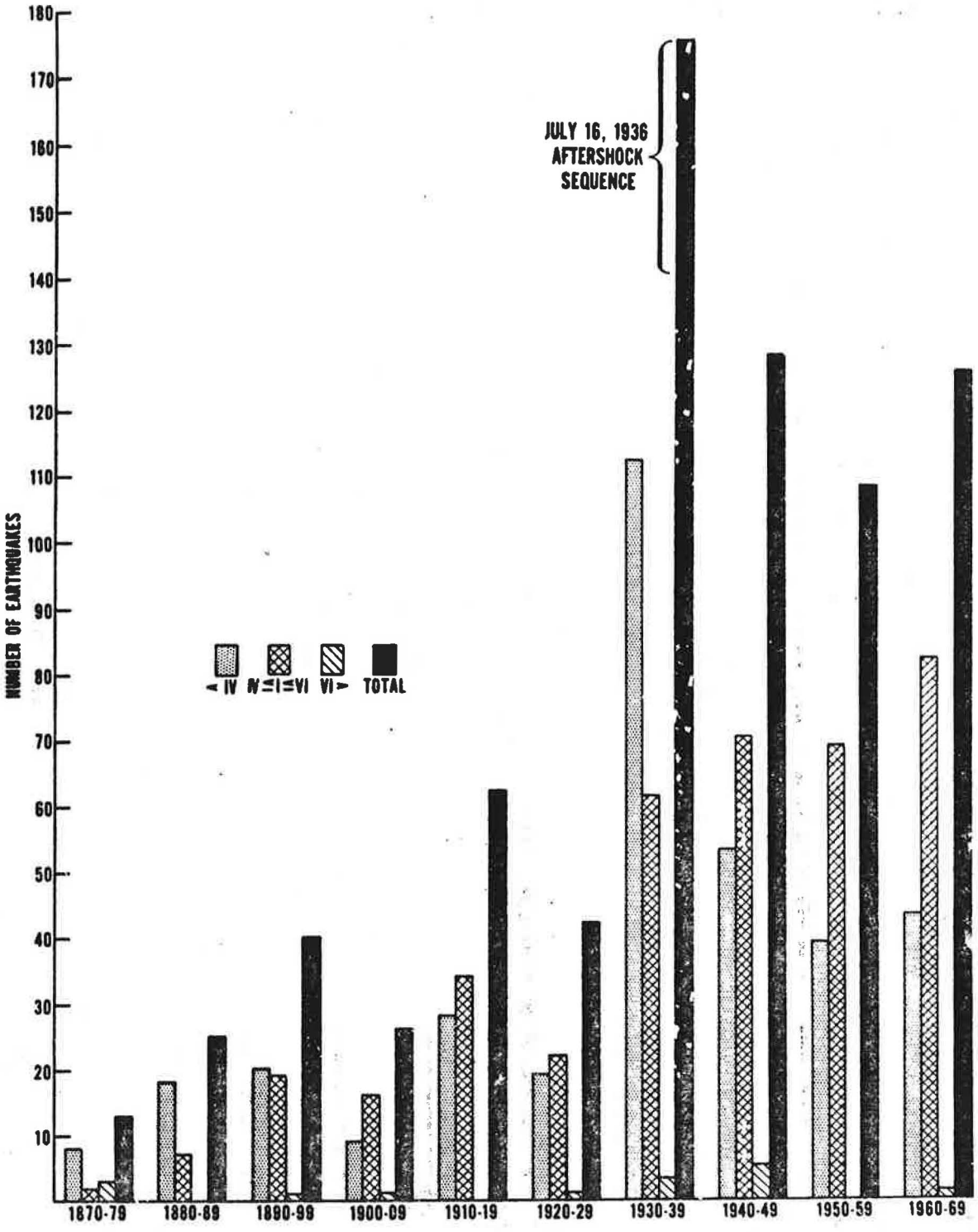

FIGURE 3. Reported Earthquakes in Each Decade Grouped in Three Intensity Ranges

the most complete available index of seismicity in the area through 1969, and that reporting, in terms of total number of events, has not been significantly improved since about 1930.

The fourth feature of Figure 3 of significance to the present study is the slow, but constant, increase in the number of intermediate earthquakes beginning with the decade 1930-39. This is coupled with a decline in the reported number of small quakes. A reasonable interpretation is that with increasing population density, the maximum intensity of these intermediate earthquakes has been reported more frequently. An alternative interpretation is that the observed behavior is due to statistical fluctuations in activity. The possibility of a temporal trend in activity cannot be rejected on the basis of Figure 3 alone. However, the near constancy of overall activity supports the first interpretation.

Analysis of Sample Completeness:

The analysis of the previous section suggests that the Puget Sound area earthquake sample is severely incomplete below intensity VI before 1930. Thus, to determine the mean rates of occurrence, λ = N/year, from the complete 100-year sample leads to serious underestimates of λ for the middle and low intensity levels. On the other hand, if the sample is shortened to the time interval in which the lowest intensity class included in the computation is completely reported, mean rates of occurrence cannot be established for the largest observed earthquakes because of lack of data. To overcome this problem we seek to determine the subinterval of the 100-year sample in which λ is stable for each intensity class, and assume that this represents the interval of complete reporting. A separate mean rate of occurrence can then be determined from the interval of complete data for each intensity class.

To analyze the nature of the incompleteness of the data sample in this detail earthquakes are grouped in intensity classes and each intensity class is modeled as a point process in time. Use is made of the property of statistical estimation that the variance of the estimate of a sample mean is inversely proportional to the number of observations in the sample. Thus the variance can be made as small as desired by making the number of observations in the sample large enough, provided that reporting is complete in time and the process is stationary, i.e., the mean, variance and other moments of each observation stay the same. To obtain an efficient estimate of the variance of the sample mean, it is assumed that the earthquake sequence can be modeled by the Poisson distribution. If $k_1, k_2, k_3, \ldots k_n$, are the number of quakes per unit time interval, then an unbiased estimate of the mean rate per unit time interval of this sample is (15)

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} k_i, \qquad (2)$$

and its variance is

$$\sigma_\lambda^2 = \lambda/n, \qquad (3)$$

where n is the number of unit time intervals. Taking the unit time interval to be one year gives

$$\sigma_\lambda = \sqrt{\lambda}/\sqrt{T} \qquad (4)$$

as the standard deviation of the estimate of the mean, where T is the sample length.

Thus, assuming stationarity, we expect that σ_λ behaves as $1/\sqrt{T}$ in the subinterval of the 100-year sample in which the mean rate of occurrence in an intensity class is constant. If the mean rate of occurrence is constant we expect stability to occur only in the subinterval that is long enough to give a good estimate of the mean but short enough that it does not include intervals in which reports are incomplete.

The rates of earthquake occurrence as a function of time interval are listed in Table II for maximum intensities IV through VIII. The rate is given as N/T where N is the cumulative number of earthquakes in the time interval T, for subintervals of the 100-year sample shown in the first column. These data are used to compute the standard deviation of the estimate of the mean through equation (4). The results are shown in Figure 4.

Figure 4 and Table II reveal several features significant to statistical treatment of earthquake data, regardless of whether one uses the empirical formula $\log N = a + bI_0$, the extreme-value distribution, or other statistical approaches. First, the postulated behavior of σ_λ is observed, at least over a subinterval of the total 100-year interval, for all intensity classes V and larger. The fact that maximum intensity IV events fail to follow the postulated behavior can be sufficiently explained by these events being incompletely reported even during the most recent interval, 1955-69. Secondly, a minimum time interval is required to reach a stable estimate of the mean recurrence rate; this interval is a function of intensity class, being successively longer with each higher maximum intensity class. This minimum interval ranges from at least five mean return periods for the largest earthquakes in the sample to about 15 mean return periods for the maximum intensity VI events. Thus, for maximum intensity V quakes, 5 to 10 years of homogeneous observations are sufficient to establish a stable mean rate; for maximum intensity VI, the minimum

Table II.

Rate of Earthquake Occurrence for Given Maximum Intensity I_0 and Time Interval T Based Upon the Historical Record of Reported Earthquakes

	T	IV		V		VI		VII		VIII	
	Years	N	N/T	N	N/T	N	N/T	N	N/T	N	N/T
1965-1969	5	21	4.20	20	4.0	3	0.60	0	0	1	0.20
1960-1969	10	46	4.60	45	4.5	18	1.80	1	0.10	1	0.10
1955-1969	15	64	4.27	59	3.93	27	1.80	1	0.07	1	0.07
1950-1969	20	78	3.90	69	3.45	29	1.45	1	0.05	1	0.05
1940-1969	30	117	3.90	89	1.97	44	1.47	4	0.13	3	0.10
1930-1969	40	157	3.92	110	2.75	48	1.20	6	0.15	3	0.08
1920-1969	50	178	3.56	120	2.40	50	1.00	7	0.14	3	0.06
1910-1969	60	191	3.18	136	2.27	56	0.93	7	0.12	3	0.05
1900-1969	70	199	2.84	143	2.04	59	0.84	8	0.11	3	0.04
1890-1969	80	209	2.61	148	1.85	64	0.80	10	0.13	3	0.04
1880-1969	90	212	2.36	150	1.67	66	0.73	10	0.11	3	0.03
1870-1969	100	214	2.14	150	1.50	69	0.69	11	0.11	5	0.05

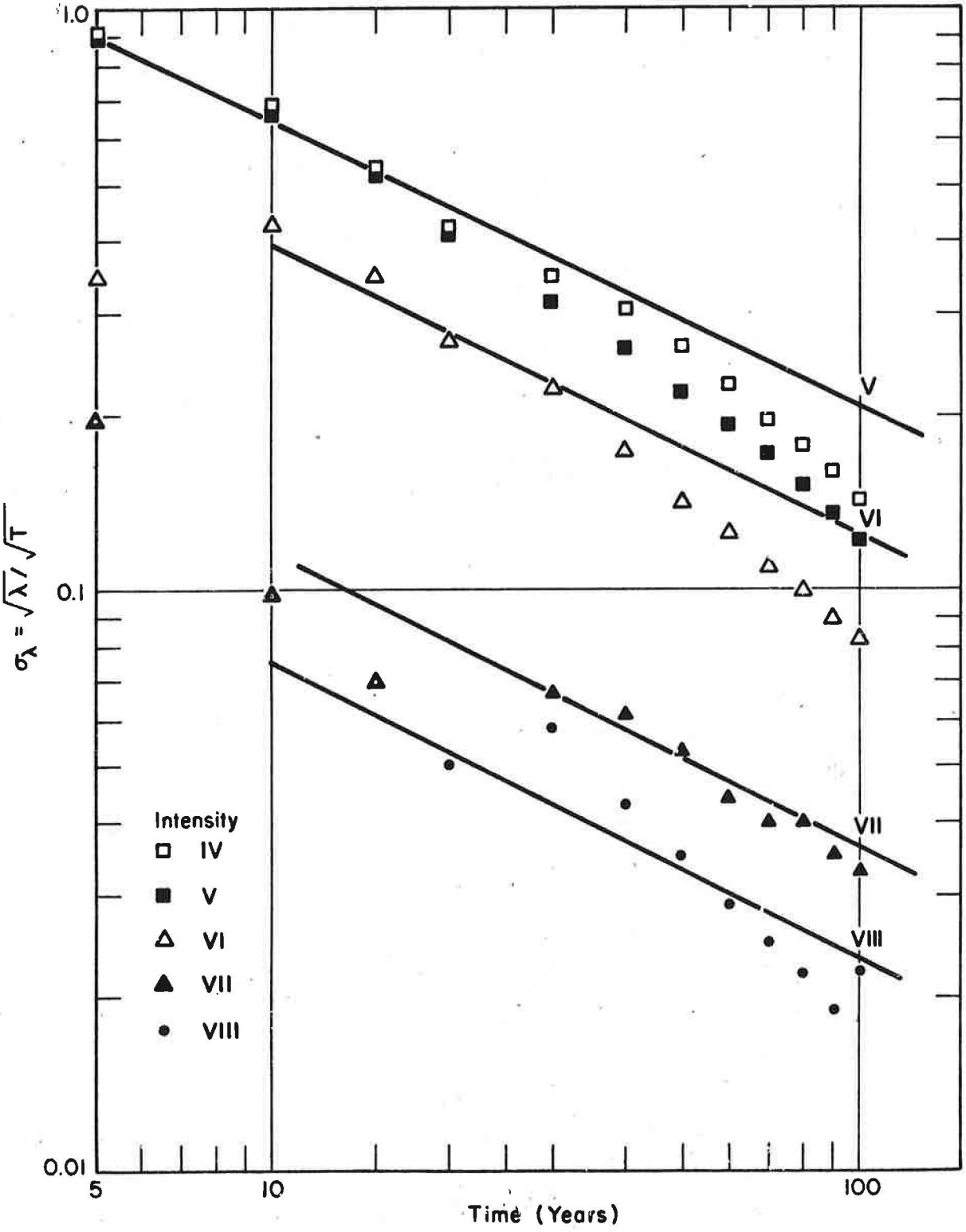

FIGURE 4. Standard Deviation of the Estimate of the Mean of the Annual Number of Events as a Function of Sample Length and Intensity Class.

observation period is between 10 and 20 years. A stable estimate of the mean recurrence rate of maximum intensity VII events is obtained in about 30 to 40 years of observations, while a stable estimate of the mean recurrence rate of maximum intensity VIII quakes is not completely definitive even for the complete 100-year sample. Third, departure of observed values of σ_λ from $1/\sqrt{T}$ behavior with increasing sample length (decreasing real time) occurs for all except intensity class VIII. The point in time at which departure from $1/\sqrt{T}$ behavior occurs is a function of maximum intensity.

The observed departure of σ_λ from $1/\sqrt{T}$ behavior can be sufficiently explained by incomplete reporting of earthquakes as the early data are incorporated into the sample. Alternatively, it can be explained by a trend toward increasing frequency in the data. However, if the latter explanation were the true one, departure of σ_λ from $1/\sqrt{T}$ behavior would be expected to occur at the same time in all intensity classes. The fact that this behavior is not observed forms the explanation that departure is due to incomplete reporting. It is concluded that maximum intensity V events are completely reported only during the most recent 15- to 20-year interval, intensity VI quakes during the most recent 30- to 40-year interval, intensity VII during at least the past 80 years, and intensity VIII over the complete 100-year sample interval.

It is seen from the above analysis that we may create an artificially homogeneous data sample by carefully determining the intervals over which earthquakes in different intensity classes are completely reported. For each intensity class the interval must be long enough to establish a stable mean rate of occurrence and short enough that it does not include intervals in which the data are incompletely reported. This amounts, in practice, to minimizing the error of estimate in the mean rate of occurrence of each earthquake class.

Discussion:

Two questions are frequently raised about the use of the formula $\log N = a + bI_0$ that relate to assessing earthquake hazard. The first concerns the minimum geographic region for which the formula holds. The second concerns the minimum sample length needed to be representative of long-term earthquake recurrence. For example, estimates of earthquake recurrence in southern California based on a 29-year sample are found to be consistent with historical experience; but when the area is reduced to about 8500 square kilometers, recurrence estimates apparently break down (9). Thus, there is an apparent trade-off between the size of the sample area and the sample length in time. An increase in the size of the geographic region from which a sample is drawn increases the rate at which data are acquired. The time required to obtain an adequate sample of observations needed to determine mean rate of occurrence of earthquakes in a given intensity class is reduced accordingly. In practice we are constrained in both regards. The statistical estimate of the earthquake hazard at a site, is made from a sample that is drawn from a small geographic region and limited in time to the available historical record. The method discussed here, while not using the entire historical sample, allows us to use the largest appropriate portion in each intensity class.

Figure 5 illustrates regression on what is considered to be the minimum error mean expected numbers of events per year based on the results presented in Table II and Figure 4. It is obvious on inspection that maximum intensity IV quakes are incompletely reported ever for the most recent five years, from 1965 through 1969. They are, accordingly, excluded in the determination of the

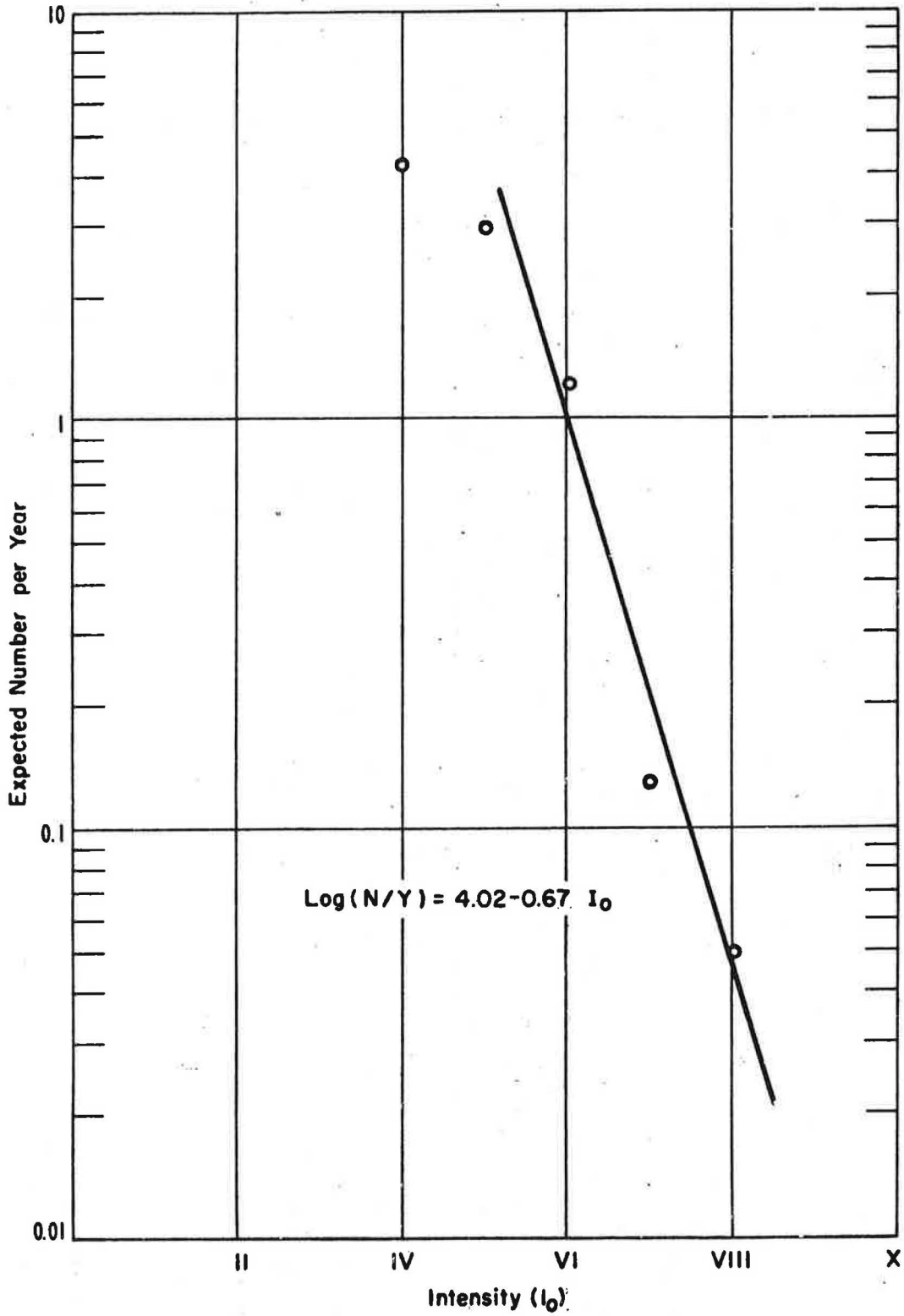

FIGURE 5. Conventional Least-Squares Fit of log (N/yr) = a + bI_o for 100-year Sample Corrected for Incompleteness

frequency equation coefficients. The slope b of the frequency formula of Figure 5 is approximately 30 percent larger than that of Case II, Figure 1, computed from the same data set uncorrected for incomplete reporting. Because the slope describes the distribution of earthquakes in size, it is a crucial parameter in earthquake risk calculations. Failure to correct for incomplete reporting in the data sample causes the recurrence rates of large earthquakes to be overestimated, while recurrence rates of small quakes are underestimated.

Acknowledgements:

I am grateful to David Perkins for suggesting the use of the moments of the Poisson distribution to estimate sample completeness and for helpful discussions. This work was taken from the writer's PhD thesis, written at the Pennsylvania State University. I am grateful to Dr. B. F. Howell, Jr., for his helpful direction of the thesis work.

Bibliographic References:

1. Ishimoto, M., and K. Iida, 1939, "Observations sur les seisms euregistré par le microseismograph construite dernierment (I), *Bull. Earthq. Res. Inst.*, Vol. 17, pp. 443-478.
2. Gutenberg, B., and C. F. Richter (1944), "Frequency of earthquakes in California," *Bull. Seism. Soc. Am.*, Vol. 34, pp. 185-188.
3. Sanford, A. R., and C. R. Holmes (1962), "Microearthquakes near Socorro, New Mexico," *Jour. Geophys. Res.*, Vol. 67, pp. 4449-4459.
4. Mogi, K. (1962), "Study of the elastic shocks caused by the fracture of heterogeneous materials and its relation to earthquakes phenomena," *Bull. Earthq. Res. Inst.*, Vol. 40, pp. 125-173.
5. Scholz, C. H. (1968), "The frequency-magnitude relation of micro-fracturing in rock and its relation to earthquakes," *Bull. Seis. Soc. Am.*, Vol. 56, pp. 185-200.
6. Riznichenko, J. V. (1959), "On quantitative determination and mapping of seismic activity," *Ann. Geophys.*, Vol. 12, pp. 227-237.
7. Knopoff, L., and J. K. Gardner (1969), "Homogeneous catalogs of earthquakes," *Proc. Nat. Acad. Sci.*, Vol. 63, pp. 1051-1054.
8. Aki, K. (1965), "Maximum likelihood estimate of b in the formula log N = a + bm and its confidence limits," *Bull Earthq. Res. Inst.*, Vol. 43, pp. 237-239.
9. Allen, C. R., P. St. Amand, C. F. Richter, and J. M. Nordquist (1965), "Relationship between seismicity and geologic structure in the southern California region," *Bull. Seis. Soc. Am.*, Vol. 50, pp. 447-471.
10. Benjamin, J. R. (1968), "Probabilistic models for seismic force design," *Jour. of the Structural Div., ASCE*, Vol. 94, pp. 1175-1196.
11. Berg, J. W., and C. D. Baker (1963), "Oregon earthquakes, 1841 through 1958," *Bull. Seis. Soc. Am.*, Vol. 53, pp. 95-108.
12. Rasmussen, N. H. (1967), "Washington state earthquakes, 1840 through 1965," *Bull. Seis. Soc. Am.*, Vol. 57, pp. 463-476.
13. Milne, W. G. (1963), "Seismicity of western Canada, west of the 113th meridian, 1841-1951," *Pub. Dominion Obs.*, Vol. XVIII, No. 7, pp. 119-146.
14. Milne, W. G., and K. A. Lucas (1961), "Seismic activity in western Canada, 1955-1959 inclusive," *Pub. Dominion Obs.*, Vol. XXVI, No. 1, pp. 1-23.
15. Hamilton, W. C. (1964), *Statistics in Physical Science*, The Ronald Press Co., New York, 230 pp.

16. Stepp, J. C. (1971), "An investigation of earthquake risk in the Puget Sound area by use of the Type I Distribution of largest extremes," PhD thesis, Pennsylvania State Univ., Dept. of Geosciences.