

Coastal Engineering Technical Note

CONFIDENCE LIMITS ON PREDICTED VALUES

PURPOSE: To determine the probable accuracy of a predicted value.

GENERAL: Values used in engineering design are often predicted using data from laboratory experiments or field measurements. A common means used is to statistically fit a regression line to plotted data. This regression line will give a predicted average value, i.e., the real value has an equal probability of being above or below the predicted value. It is desirable to have a confidence limit on the real value. This confidence limit around the predicted value will provide a range of values within which there is a given probability of finding the real value.

Taking the independent variable as x and the dependent variable (to be predicted) as y , a data point is given as (x_i, y_i) . For a given value of the independent variable, x_k , the predicted value of the dependent variable is \hat{y}_k and the real value is y_k . Regression lines giving a formula relating y to x , such as:

$$y = a + bx \quad (1)$$

for linear regression, can be obtained by established means (see Draper and Smith, 1966; Mendenhall, 1966; Daniel and Wood, 1971; or Walpole and Myers, 1978). Various manufacturers produce programable calculators which have standard programs for linear regression.

Care must be exercised in applying regression analysis and confidence limits to particular sets of data. The data sets should be plotted to determine if the relationships appear to be linear or curvilinear. Also, there must be a clear physical relationship between the dependent and the independent variables. Some of the references discuss the underlying assumptions and the precautions which should be observed in applying regression analysis.

Where linear regression is used for predictions, the confidence limits can be attained as follows:

For a given probability, P_1 that $y \leq (\hat{y} + e_s t)$, or probability, P_2 , that $(\hat{y} - e_s t) \leq y \leq (\hat{y} + e_s t)$, t is obtained from the table of t values. Table 1 has been developed from the t -distribution (see Draper and Smith, 1966) and may be used directly to obtain probabilities for the example problem to be shown. (Note that Table 1 is not a t -distribution table since one enters the table directly with the number of data points rather than with the number of data points less two. This assumes that a t -distribution with two degrees of freedom is appropriate for the solution). The value e_s is the standard error and for a given predicted value y is given as:

$$e_s = \left[\frac{\sum (y_i - \hat{y}_i)^2}{n - 2} \right]^{1/2} \left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]^{1/2} \quad (2)$$

where \hat{y}_i is the predicted value of y corresponding to the data point (x_i, y_i) , n is the number of data points, \bar{x} the mean value of x_i , and x_k is the value of x for which we are predicting y_k . For a given probability, the confidence limits for the various values of y_k will form curves above and below the regression line defining the values of \hat{y}_k .

***** EXAMPLE *****

GIVEN: At a hypothetical coastal location, there exists a 150-year record of tsunami flood levels. A total of seven tsunamis occurred having measured flood levels above mean sea level of 4 feet, 5 feet, 5.5 feet, 7 feet, 8 feet, 12 feet, and 16 feet. All flood levels are for a location 200 feet shoreward of the coastline.

FIND: A flood level that has a 99 percent probability of not being exceeded by the 100-year tsunami.

Table 1. t Values

number of data points, n	Probability P_1 that $y \leq \hat{y} + e_s t$									
	0.55	0.65	0.75	0.85	0.9	0.95	0.975	0.99	0.995	0.9995
	Probability P_2 that $\hat{y} - e_s t \leq y \leq \hat{y} + e_s t$									
	0.1	0.3	0.5	0.7	0.8	0.9	0.95	0.98	0.99	0.999
3	0.158	0.510	1.000	1.963	3.078	6.314	12.706	31.821	63.657	636.619
4	0.142	0.445	0.816	1.386	1.886	2.920	4.303	6.965	9.925	31.598
5	0.137	0.424	0.765	1.250	1.638	2.353	3.182	4.541	5.841	12.924
6	0.134	0.414	0.741	1.190	1.533	2.132	2.776	3.747	4.604	8.610
7	0.132	0.408	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
8	0.131	0.404	0.718	1.134	1.440	1.943	2.447	3.143	3.707	5.959
9	0.130	0.402	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
10	0.130	0.399	0.706	1.108	1.397	1.860	2.306	2.896	3.355	5.041
11	0.129	0.398	0.703	1.100	1.383	1.833	2.262	2.821	3.250	4.781
12	0.129	0.397	0.700	1.093	1.372	1.812	2.228	2.764	3.169	4.587
13	0.129	0.396	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437
14	0.128	0.395	0.695	1.083	1.356	1.782	2.179	2.681	3.055	4.318
15	0.128	0.394	0.694	1.079	1.350	1.771	2.160	2.650	3.012	4.221
16	0.128	0.393	0.692	1.076	1.345	1.761	2.145	2.624	2.977	3.140
17	0.128	0.393	0.691	1.074	1.341	1.753	2.131	2.602	2.947	4.073
18	0.128	0.392	0.690	1.071	1.337	1.746	2.120	2.583	2.921	4.015
19	0.128	0.392	0.689	1.069	1.333	1.740	2.110	2.567	2.898	3.965
20	0.127	0.392	0.688	1.067	1.330	1.734	2.101	2.552	2.878	3.922
21	0.127	0.391	0.688	1.066	1.328	1.729	2.093	2.539	2.861	3.883
22	0.127	0.391	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.850
23	0.127	0.391	0.686	1.063	1.323	1.721	2.080	2.518	2.831	3.819
24	0.127	0.390	0.686	1.061	1.321	1.717	2.074	2.508	2.819	3.792
25	0.127	0.390	0.685	1.060	1.319	1.714	2.069	2.500	2.807	3.767
26	0.127	0.390	0.685	1.059	1.318	1.711	2.064	2.492	2.797	3.745
27	0.127	0.390	0.684	1.058	1.316	1.708	2.060	2.485	2.787	3.725
28	0.127	0.390	0.684	1.058	1.315	1.706	2.056	2.479	2.779	3.707
29	0.127	0.389	0.684	1.057	1.314	1.703	2.052	2.473	2.771	3.690
30	0.127	0.389	0.683	1.056	1.313	1.701	2.048	2.467	2.763	3.674
40	0.126	0.388	0.681	1.051	1.304	1.686	2.025	2.429	2.710	3.564
50	0.126	0.388	0.680	1.048	1.299	1.678	2.012	2.407	2.684	3.510
100	0.126	0.386	0.677	1.041	1.289	1.659	1.982	2.365	2.627	3.393
∞	0.126	0.385	0.674	1.036	1.282	1.645	1.960	2.326	2.576	3.291

SOLUTION: The probability of occurrence of each flood level, h, is given by $P(h) = m/(n + 1)$ where m is the rank and n is the period of record in years. This gives:

$$P(16) = 1/(150 + 1) = 0.0066$$

$$P(12) = 2/(150 + 1) = 0.0132$$

$$P(8) = 3/(150 + 1) = 0.0199$$

$$P(7) = 4/(150 + 1) = 0.0265$$

$$P(5.5) = 5/(150 + 1) = 0.0331$$

$$P(5) = 6/(150 + 1) = 0.0397$$

$$P(4) = 7/(150 + 1) = 0.0464$$

Flood levels are plotted against probability in the Figure.

Using the equation for tsunami flooding given by Houston, et al.,(1977) and Camfield (1980) that

$$h = -B - A \log_{10} P(h) , \quad (3)$$

linear regression from standard methods (references 2,3,5) gives

$$h = -15.58 - 14.42 \log_{10} P(h) . \quad (4)$$

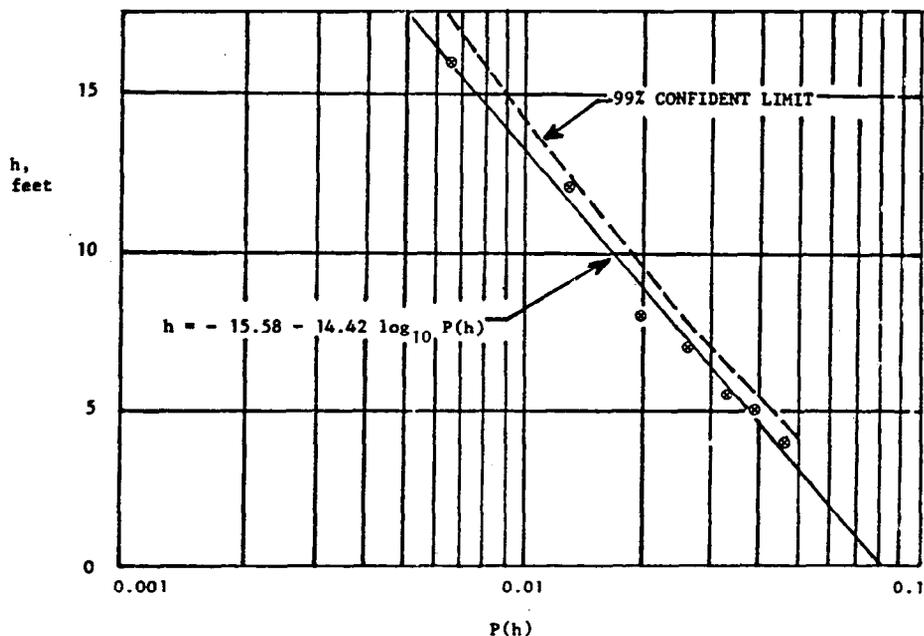


Figure 1. Probability of occurrence of tsunami flood levels.

Note: This figure is only applicable to the given set of data and should not be used for design purposes.

For $P(h) = 0.01$ (the 100-year tsunami), from equation (4)

$$\begin{aligned}\hat{h} &= -15.58 - 14.42 \log_{10} (0.01) = -15.58 - 14.42 (-2) \\ \hat{h} &= 28.84 - 15.58 = 13.26 \text{ feet}\end{aligned}$$

which is the predicted flood level from linear regression. To calculate e_s , note that the independent variable "x" is $\log_{10} P(h)$ where $P(h)$ has been previously calculated, and the dependent variable "y" is h .

$$e_s = \left[\frac{\sum (h_i - \hat{h}_i)^2}{n - 2} \right]^{1/2} \left[\frac{1}{n} + \frac{(\log_{10} P(h)_k - \overline{\log_{10} P(h)})^2}{\sum (\log_{10} P(h)_i - \overline{\log_{10} P(h)})^2} \right]^{1/2} \quad (5)$$

$n = 7$ (the number of data points)

The term $\left[\frac{\sum (h_i - \hat{h}_i)^2}{n - 2} \right]^{1/2}$ is obtained directly from existing programs for linear regression available on programable calculators (individual users should refer to manufacturers handbooks for the calculators that they have available). In this case, $h_1 = 16$, $h_2 = 12$, $h_3 = 8$, $h_4 = 7$, $h_5 = 5.5$, $h_6 = 5$, $h_7 = 4$ then

$$\left[\frac{\sum (h_i - \hat{h}_i)^2}{n - 2} \right]^{1/2} = 0.5579$$

(Note: working this problem without a programable calculator is a long, tedious process which is beyond the scope of this technical note).

Equation (5) for the standard error now becomes

$$e_s = 0.5579 \left[\frac{1}{7} + \frac{(\log_{10} 0.01 - \overline{\log_{10} P(h)})^2}{\sum (\log_{10} P(h)_i - \overline{\log_{10} P(h)})^2} \right]^{1/2}$$

Using a programable calculator, and inserting values of h and $\log_{10} P(h)$, the mean value of $\log_{10} P(h)$ is

$$\overline{\log_{10} P(h)} = -1.6501$$

and the summation is given as

$$\Sigma \left(\log_{10} P(h)_i - \overline{\log_{10} P(h)} \right)^2 = 0.5307$$

The standard error is now

$$e_s = 0.5579 \left[\frac{1}{7} + \frac{(0.3499)^2}{0.5307} \right]^{1/2}$$

$$e_s = 0.341$$

To solve for h, for the 100-year tsunami

$$h \leq \hat{h} + e_s t$$

from the table, for n = 7, and P₁ = 0.99, t = 3.365

$$h \leq 13.26 + 0.341(3.365) = 13.26 + 1.14$$

$$h \leq 14.4 \text{ ft.}$$

There is a 99 percent probability that the 100-year tsunami will not exceed a flood level of 14.4 feet. To obtain a confidence limit curve (or the 99 percent confidence limit line in Figure 1), one repeats the computation for a number of frequency intervals.

REFERENCES:

1. CAMFIELD, Frederick E., "Tsunami Engineering," Special Report No. 6, U.S. Army Coastal Engineering Research Center, Fort Belvoir, VA., February 1980.
2. DANIEL, Cuthbert and WOOD, Fred, *Fitting Equations to Data*, John Wiley and Sons, Inc., New York, New York, 1971.
3. DRAPER, N.R., and SMITH, H., *Applied Regression Analysis*, John Wiley and Sons, Inc., New York, New York, 1966.
4. HOUSTON, J.R., CARVER, R.D., and MARKLE, D.G., "Tsunami-Wave Elevation Frequency of Occurrence for the Hawaiian Islands," Technical Report H-77-16, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS, August 1977.
5. MENDENHALL, William, *Introduction to Statistics*, Wadsworth Publishing Co., Inc., 1966.
6. WALPOLE, R.E., and MYERS, R.H., *Probability and Statistics for Engineers and Scientists*, 2nd Ed., MacMillan Publishing Co., Inc., New York, New York, 1978.