Sample Size Determination for Engineering

Burcharth, Hans F.; Liu, Zhen

Publication date:
1994

Document Version
Accepted author manuscript, peer reviewed version

Sample size determination for engineering

Burcharth,H.F. and Liu,Z.

(-) Introduction

The slender concrete armour units, such as dolosse and tetrapods, have been widely used as the packing layer of breakwaters. However the recent breakwater failures revealed that their design should incorporate both the hydraulic stability and structural integrity of an individial armour unit. With the purpose of calculating maximum static tensile stress of dolosse at different positions and different slopes (simply abbreviated as stress hereafter), a large vibration ramp with changable slopes is built. About 70 dolosse weighing 200 kg are randomly put into two layers on the ramp. How to correlate the results in the presence of waves with the corresponding structural integrity assessment is beyond the scope of this artical. At the moment we are only interested in the appropriate sample sizes.

At each position and each slope the dolos stresses will be randomly distributed due to the complicated shape of dolosse and their random placement and orientation. In order to get the stress distribution at each case the corresponding experiment needs to be repeated many times with the fully rebuilt ramp. Since every experiment is expensive and time consuming, it is necessary to decide how many times one experiment should be repeated to get a reliable corresponding stress distribution.

From the physical point of view we know that all stress distributions will follow the same type, with different parameters. Therefore the feasible sample sizes at each case may be determined by the detailed study on one case.

(=) Getting the stress distribution of one case

The 80 stress measurements of the dolos at the bottom position of the bottom layer in the slope 1 : 1.5 are made. The results show an agreement with log--normal distribution, cf Fig.(1). The stress density function and related parameters fitted according to the most likelihood method are:

$$f(\ln\tau) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{\ln\tau-\mu}{\sigma}\right)^2} \qquad (1)$$

with

$$\mu = \frac{1}{80}\sum_{i=1}^{80} \ln\tau_i = 2.45$$

$$\sigma^2 = \frac{1}{80-1}\sum_{i=1}^{80} (\ln\tau_i - \mu) = 1$$

For such large sample it would be reasonable to reckon them as the the corresponding population.

(三) Theoretical consideration of appropriate sample size

Supposing a random variable X is normally distributed with the mean μ the variance σ^2, simply expressed as:

$$X \sim N(\mu , \sigma^2) \qquad (2)$$

Supposing the sample size is n. It is natural to regard the mean $\overline{X}$ and variance S^2 of a sample as an approximation of the mean μ and variance σ^2 of the corresponding population respectively. That is:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (3)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}) \qquad (4)$$

If we pick up n-size sample repeatedly, the stochastical theory tells us that the mean $\overline{X}$ of the sample will be normally distributed with the mean and variance being μ and σ^2/n respectively, while $\frac{\sigma^2}{n-1}S^2$ will be Chi-squre distributed with (n-1) number of freedom, simply written as:

$$\overline{X} \sim N(\mu , \frac{\sigma^2}{n}) \qquad (5)$$

$$\frac{\sigma^2}{n-1} S^2 \sim \chi^2(n-1) \qquad (6)$$

Now looking back to expression (3), defining x_α by:

$$F(x_\alpha) = P(x < x_\alpha) = \alpha \qquad (7)$$

where F(x) is the distribution function of X.

By variable substitution:

$$Y = \frac{X - \mu}{\sigma} \qquad (8)$$

getting standard normal distribution:

$$Y \sim N(0 , 1) \qquad (9)$$

y_α can be obtained from the standard normal distribution table. And x_α is decided by the relation:

$$x_{\alpha} = y_{\alpha} * \sigma + \mu \tag{10}$$

Since $\bar{X}$ is still noamally distributed, as expressed in Eq.(6). Similarly we get:

$$\bar{x}_{\alpha} = y_{\alpha} * \frac{\sigma}{\sqrt{n}} + \mu \tag{11}$$

$\bar{x}_{\alpha}$ means that $\bar{x}$'s value of a sample will be smaller than $\bar{x}_{\alpha}$ with α probability.

Normally engineers would rather use x_{β} , say $\tau_{0.9}$, instead of $\bar{x}$ ($x_{0.5}$) in the design process. Hence it is valuable to consider x_{β}'s distribution of a random sample.

The random variable X_{β} is:

$$X_{\beta} = y_{\beta} * S + \bar{X} \tag{12}$$

Theoretically the distribution of X_{β} is obtainable, but in practice even with the help of numerical calculations it is still difficult to obtain it due to the complex joint distribution of $\bar{X}$ and S. On the other hand, now that we assure that the value of $\bar{X}$ and S will be less than x_{α_1}, and s_{α_2} , with α_1 and α_2 probability respectively, by their combination we can get the value $(x_{\beta})_r$ of X_{β}, which means that the value of X_{β} will be less than $(x_{\beta})_r$, with r probability. Since $\bar{X}$ and S are independent variables, it is sure that $r > \alpha_1 * \alpha_2$. In other words, the results of this estimation will be a little too conservative.

Since s > 0, Eq.(14) can be rewritten as:

$$X_{\beta} = \frac{y_{\beta}\,\sigma}{\sqrt{n-1}} \sqrt{\frac{\sigma^{2}}{n-1} S^{2}} + \bar{X} = y_{\beta}\,\sigma \sqrt{\frac{V}{n-1}} + \bar{X} \tag{13}$$

where V is a new random variable, defined as:

$$V = \frac{\sigma^{2}}{n-1} S^{2} \sim \chi^{2}(n-1) \tag{14}$$

Therefore:

$$(x_{\beta})_r = y_{\beta}\,\sigma \sqrt{\frac{V_{\alpha_1}}{n-1}} + \bar{x}_{\alpha_2} \tag{15}$$

$$r > \alpha_1 * \alpha_2$$

v_{α_2} is obtained from the Chi--square distribution table.

Example: Find $(\tau_{0.9})_{0.9}$ of the above stress distribution if the sample size is 50.

since: $X = \ln\tau \sim N(-2.45, 1)$

defining: $Y = \dfrac{X+2.45}{1}$ $\quad Y \sim N(0, 1)$

we have:

$$\beta = 0.9$$

$$\gamma \doteq \alpha_1 * \alpha_2 = 0.9$$

selecting: $\alpha_1 = 0.95$, $\alpha_2 = 0.95$

from the standard normal distribution table getting:

$$y_{0.9} = 1.28 , \quad y_{0.95} = 1.64$$

the mean of samples $\bar{X}$:

$$\bar{X} \sim N(-2.45, \frac{1}{\sqrt{50}})$$

$$\bar{x}_{0.95} = y_{0.95} * \frac{1}{\sqrt{50}} + (-2.45) = -2.22$$

from the Chi--square distribution table obtaining:

$$v_{0.95} = 67.2$$

therefore

$$(x_{0.9})_{0.9} = y_{0.9}\,\sigma\sqrt{\frac{V_{0.95}}{n-1}} + \bar{x}_{0.95} = -0.75$$

$$(\tau_{0.9})_{0.9} = \exp\,(X_{0.9})_{0.9} = 0.47 \qquad \text{(mpa)}$$

$(\tau_{0.9})_{0.9}$ means that $\tau_{0.9}$ of a 50--size sample wiil be smaller than $(\tau_{0.9})_{0.9}$ with at least 90% probability.

(IV) Comparison of the approximation with practice

In order to get a real distribution of $\tau_{0.9}$ of samples, a random sampling computer procedure is designed in such a way that all the data of the population have the same chance to be picked up.

Every time the computer randomly pick up 50 data, which is then fitted to log--normal distribution according to the most likelihood method and $\tau_{0.9}$ is calculated from the fitted function. This is repeated for 5000 times. The results, shown in Fig.2, shows that $\tau_{0.9}$ is less than 0.47 with 92% probability and with 90% probability $\tau_{0.9}$ is less than 0.46, which are very close to the approximate results.

Peahaps the most intuitive way of expressing the suitable sample size is to give confidence zone of the random variable p($\ln\tau$), which will give one general impression about whether the sample size is big enough. Fig.3 obtained from randomly picked up data analysis, shows one of the 90%

confidence zone of $p(\ln\tau)$ of 50--size samples.

(五) Representative case

If enfineers say the 50--size sample of the stress is big enough to represent the corresponding population in this case (the dolosse of bottom position of bottom layer in the slope 1:1.38), how about other cases ? To answer the question let's consider the defference between x_β of a sample (with r probability) and x_β of the population.

$$\frac{(x_\beta)_r - x_\beta}{x_\beta} = \frac{y_\beta \sigma \sqrt{\frac{V_{\alpha_2}}{n-1}} + \bar{x}_{\alpha_1} - (y_\beta \sigma + \mu)}{y_\beta \sigma + \mu} = \frac{y_\beta \sigma \left(\sqrt{\frac{V_{\alpha_2}}{n-1}} - 1\right) + y_{\alpha_1} \frac{\sigma}{\sqrt{n}}}{y_\beta \sigma + \mu} \tag{16}$$

$$= \frac{y_\beta \left(\sqrt{\frac{V_{\alpha_2}}{n-1}} - 1\right) + y_{\alpha_1} \frac{1}{\sqrt{n}}}{y_\beta + \frac{\mu}{\sigma}}$$

From the above equation we know that the bigger the ratio of the mean against the deviation of the population, the greater the defference, other conditions being the same. In other words, the satisfied sample size determined from the case with small μ/σ will also be suitable for other cases. From the practical experience we judge that the stress distribution of the chosen case may have the smallest μ/σ in the experiment series, therefore 50--size sample will be the suitable one for other cases.

(六) conclusion

A method is presented for determination of the appropriate sample size in practical engineering and research. Its main idea is to find the feasible sample size by a big enough sample analysis. It is suggested that the large sample be taken in the case with small μ/σ. The principle holds for other kinds of distribution although the current analysis is only based on the log--normal distribution.

Acknowledgemant

The dolos static stress study is finacially supported by the Danish scientifical research council. The valuable discussions with Mr.P.Frigaard, PH.D.student in Department of Civil Engineering of Aalborg University is highly appreciated.

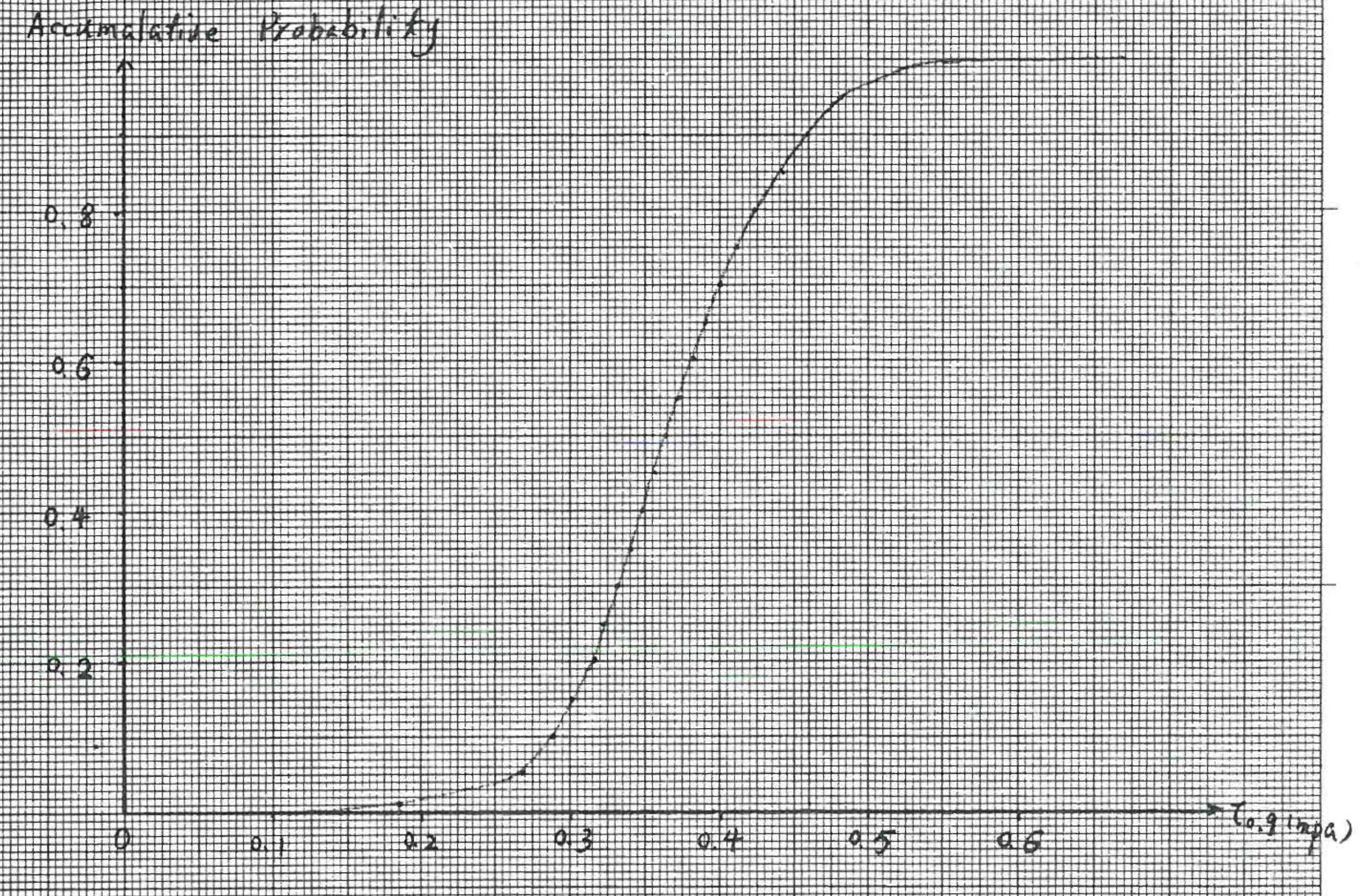

Fig.2 Distribution function curve of $\tau_{0.9}$

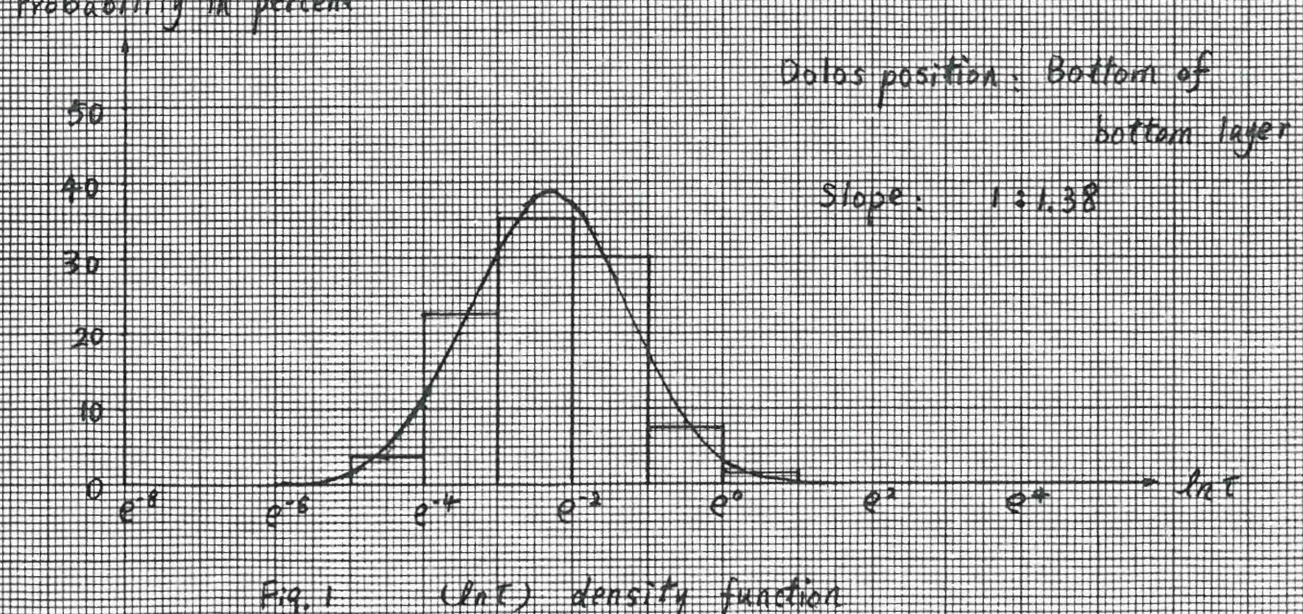

Fig. 1 ($ln\tau$) density function

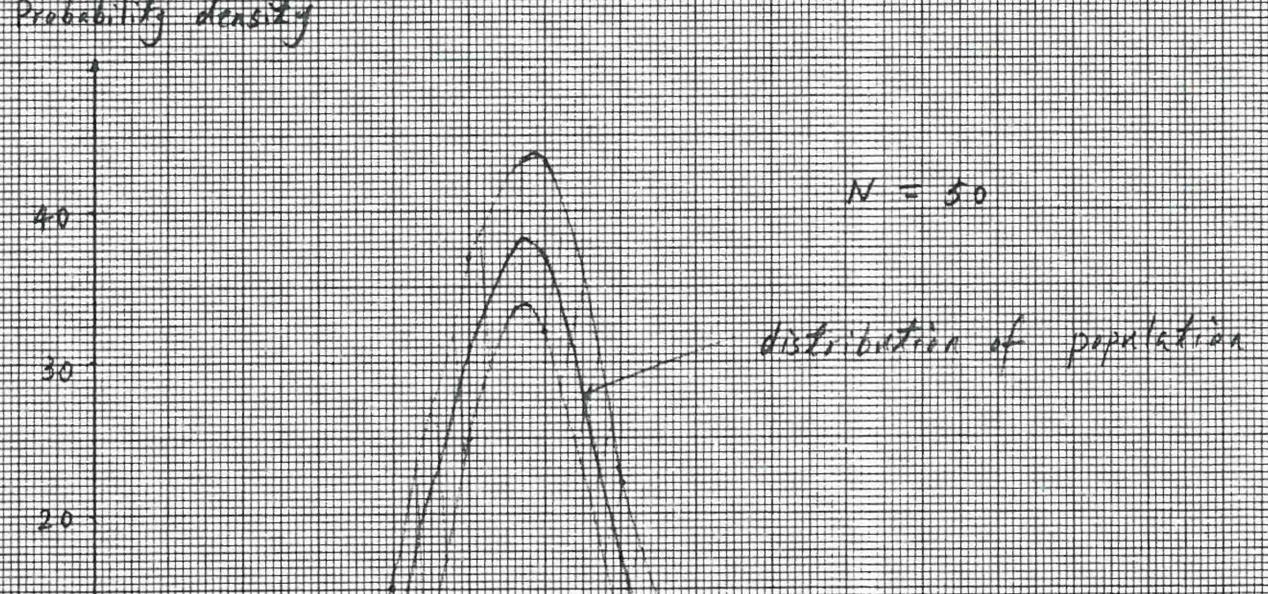

Fig 3 90% confidence zone of P(ln τ)