The 14th World Conference on Earthquake Engineering
October 12-17, 2008, Beijing, China

STATISTICAL ANALYSIS TOWARDS THE IDENTIFICATION OF ACCURATE PROBABILITY DISTRIBUTION MODELS FOR THE COMPRESSIVE STRENGTH OF CONCRETE

S. Silvestri[1], G. Gasparini[1], T. Trombetti[2] and C. Ceccoli[3]

[1] *Ph.D Researcher, DISTART Dept. of Structural Engineering , University of Bologna, Italy*
[2] *Associate Professor, DISTART Dept. of Structural Engineering , University of Bologna, Italy*
[3] *Full Professor, DISTART Dept. of Structural Engineering , University of Bologna, Italy*
Email: tomaso.trombetti@unibo.it; stefano.silvestri@unibo.it; giada.gasparini@mail.ing.unibo.it; claudio.ceccoli@mail.ing.unibo.it

ABSTRACT :

Structural design deeply involves material strengths, which are obviously affected by uncertainty. As far as the material "concrete" is concerned, building codes (Italian codes, Eurocodes, ACI regulations) typically state that it should be identified and classified by means of its conventional characteristic uniaxial compressive cubic strength at 28 days.

Although the fundamental importance of the problem of the evaluation of the characteristic compressive strength of concrete in structural design, the statistical analysis geared to identify an accurate probabilistic model of the concrete strength has not been a central issue of the research works in the field for many years.

This paper describes the results of an investigation performed to obtain the compressive strength statistic characteristics of a production of about half a million cubic meters of concrete. This amount of concrete was produced over a five-year period. The results obtained indicate that the statistical distribution that best captures the characteristics of the available experimental data is the Shifted Lognormal. On the other hand, the Italian code and, to some extent, the Eurocode substantially base the evaluation of concrete properties upon a Normal distribution. It is therefore advisable that design codes will encompass the possibility for the engineer to evaluate the concrete properties based upon these more refined statistical models.

KEYWORDS: concrete compressive strength, experimental data, probability distribution models, statistical analysis, design codes

1. INTRODUCTION

Real world problems involve uncertainties (Melchers 1999, Conte 2001, Ang and Tang 2007). In order to face engineering problems under uncertainty, it is necessary to prearrange proper probabilistic models capable of providing the quantification of such uncertainty, so that it can be taken into account in the process of decision making for engineering planning and design (Melchers 1999, Conte 2001).
The construction of an accurate probabilistic model requires different steps which are well-established (Ang and Tang 2007) and summarised hereafter:

- First, it is necessary to clearly define the event of interest of the engineering problem at hand which is affected by uncertainty.
- A random variable is introduced as a mathematical vehicle for representing the event in analytical form.
- The probability of the event expressed in terms of the selected random variable depends on the values of the parameters as well as on the form of the distribution model (Probability Density Function or Cumulative Distribution Function).
- In general, the probability distribution model appropriate to describe a random phenomenon is not known.
- However, the required probability distribution model may be determined empirically based on the available observational data (probability papers).
- Then, the assumed probability distribution model may be verified (or disapproved) in the light of available data using certain statistical tests (goodness-of-fit tests for distribution).
- If tests are positive, the probabilistic model may be consciously and used in the engineering problem.

Structural design deeply involves material strengths (Neville 2006), which are obviously affected by uncertainty. As far as the material "concrete" is concerned, building codes (Italian codes, Eurocodes, ACI regulations) typically state that it should be identified and classified by means of its conventional characteristic uniaxial compressive cubic strength at 28 days.
In this paper, the common procedure described above is applied to the problem of the evaluation of the characteristic compressive cubic strength of concrete at 28 days.
Although the fundamental importance of this matter in structural design, the statistical analysis geared to identify an accurate probabilistic model of the concrete strength has not been a central issue of the research works in the field for many years. Away back in 1974, Kameswara Rao and Swamy formulated a first statistical theory for the strength of concrete. Only Dayaratnam and Ranganathan, away back in 1976, faced directly this same issue collecting data on tests of compressive strength of 150 mm concrete cubes and concluding that the strength of most of the examined groups of the concrete follow Normal distribution with 1% significance level except for a small group of 150 kg/cm^2 concrete. On the same research topic but with reference to specific kind of concrete, Corradini *et al.* (1984) developed a statistical evaluation of the mechanical properties of superplasticised concrete, while, more recently, Bhanja and Sengupta (2002) used statistical methods to investigate the compressive strength of silica fume concrete.

2. DETAILS OF THE ANALYSED CONCRETE

Compressive strength values of roughly 10,000 concrete cubic specimens have been examined; each specimen corresponding to each 50 cubic meters sample of produced concrete. The production of about half a million cubic meters of concrete relates to an important engineering work. The number of specimens is therefore consistent with (even larger than) the minimum requested by the more recent Italian code, D.M. 14/09/2005, (at least 15 samples for 1500 m^3) for large-size building sites.
The production relates to two facilities which are installed in the same construction site and operate with the same machinery, the same design mixture of concrete, the same cement and aggregates from the same gravel pit. This allows to count the concrete as "homogenous", according to the common definition of "concrete homogeneity" stated by most of the design codes (e.g. the Italian code and the Eurocode).
The concrete production has been obtained over a five-year period. Data have been collected separately for every year. This allows to check also the time invariability of the concrete mixture properties.
In the following sections, we will carry out a detailed statistical inference analysis upon the concrete

compressive strength in order to identify the most accurate probability distribution which best describes such quantity. The aim is to evaluate the probability of the event "compressive strength of concrete less than a prescribed value" or, the other way round, to obtain the value of compressive strength which is characterized by a given probability of exceedance. In more detail, we want to verify if the examined concrete, which was designed in order to obtain a characteristic strength, R_{ck}, at least equal to $25\ \mathrm{N/mm^2}$, is actually characterized by such value. The characteristic value being defined by the code as the value such that only 5% of the observations are less than this value.

3. RANDOM VARIABLES: BASIC CONCEPTS

The concrete compressive strength, *R*, is taken as continuous random variable. Throughout the paper we will adopt the usual random variable notation (Ang and Tang 2007) indicating a random variable with capital letter (e.g. *R*) and its possible values with lowercase letters (e.g. *r* or r_i). With reference to the basic probability theory (Ang and Tang 2007), the Cumulative Distribution Function (CDF), $F_R(r)$, of the random variable *R* provides the probability that *R* assumes values lower than a specific value *r* $F_R(r) = P[R \le r]$, while the Probability Density Function (PDF), $f_R(r)$, of the random variable *R* provides the density of probability of *R*, i.e. $f_R(r)dr$ provides the probability that *R* assumes values in the interval $(r, r+dr]$: $f_R(r)dr = P[r < R \le r+dr]$

The 5% percentile of *R*, which is the characteristic compressive strength recalled by the codes, is the value R_{ck} such that $F_R(R_{ck}) = P[R \le R_{ck}] = 0.05$.

Given that the exact form of the probability distribution may not be known, a distribution may be approximately described by a number of derived properties, commonly called "moments". Other important descriptors of a probability distribution are the mode and the median. The mode, $\tilde{r}$, is the most probable value of a random variable; the median, r_m, is the value at which $F_R(r_m) = 0.50$ and thus larger and smaller values are equally probable. Similarly, also the statistical population of observational data from real world (composed of *n* measures r_i of the random variable *R*) are usually characterised by means of the experimental moments.

4. SELECTION OF ADEQUATE PROBABILITY DENSITY FUNCTIONS

For the purposes of this research work, it is worth distinguishing between (1) distributions which present symmetric PDF with respect to the mean value (i.e. characterised by skewness coefficient equal to zero) and (2) distributions which present asymmetric PDF with respect to the mean value (i.e. characterised by non-zero skewness coefficient). Among symmetric PDF distributions we recall the Uniform and the Normal ones, whilst among asymmetric PDF distributions we recall the Lognormal, the Exponential, the Gamma, and the Gumbel ones. The Beta distribution, which is particularly appropriate for a random variable whose range of possible values is bounded, is quite versatile, given that it can be symmetric or asymmetric depending on the values assumed by its parameters.

With the aim of obtaining the more adequate representation of the actual probability distribution of the concrete compressive cubic strength, the three following distributions are selected and considered in the statistical analysis reported in this research work: (a) the Normal distribution, (b) the Shifted Lognormal distribution, (c) the Gumbel distribution.

The Normal distribution is characterised by the following PDF (Ang and Tang 2007):

$$f_R(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{r-\mu}{\sigma}\right)^2\right] \qquad -\infty < r < \infty \qquad (4.1)$$

where μ and σ are the parameters of the distribution and coincide with the mean value and the standard

deviation of the random variable, respectively (i.e. $\mu \equiv \mu_R$ and $\sigma \equiv \sigma_R$). By reason of the Central Limit Theorem, the Normal distribution is well suited for the representation of a random variable which derives from the sum of a large number of different contributions, none of which is dominant and which are all characterized by the same distribution and by not so different mean values and standard deviations.
The Normal distribution has been selected because it is the most widely used for the representation of random variables, many building codes are based upon this distribution, it is one which arises frequently in practice as a limiting case of other probability distributions, it well represents random variables which are affected by different factors, which is the case of the concrete strength which is affected, for example, by the strengths and the dimensions of the aggregates, by the quality of the cement, by the mixing technologies, by the quantity and the chemical properties of the water in the mixture.
The Lognormal distribution, in which the natural logarithm of the random variable *R*, rather than *R* itself, has a normal distribution, is characterized by the following PDF (Ang and Tang 2007):

$$f_R(r) = \frac{1}{\zeta \cdot r \cdot \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln r - \lambda}{\zeta}\right)^2\right] \qquad r \geq 0 \tag{4.2}$$

where λ and ζ are the parameters of the distribution and coincide with the mean value and the standard deviation of $\ln R$, respectively (i.e. $\lambda \equiv \mu_{\ln R}$ and $\zeta \equiv \sigma_{\ln R}$). The parameters λ and ζ are related to the mean value and the standard deviation of the random variable *R* as $\lambda = \ln \mu_R - \frac{1}{2}\zeta^2$, where:

$$\zeta^2 = \ln\left(1 + \left(\frac{\sigma_R}{\mu_R}\right)^2\right) = \ln\left(1 + \delta_R^{\,2}\right) \cong \delta_R^{\,2} \;\text{ if }\; \delta_R \leq 0.3 \tag{4.3}$$

The Shifted Lognormal distribution differs from the Lognormal one in that it is translated along the abscissas axis, i.e. it starts from a given value *a* and not from zero and it is defined for $x \geq a$.
The Shifted Lognormal distribution has been selected because of the positiveness of its values, it is especially useful in those applications where the values of the random variable are known, from physical consideration, to be strictly positive (on the contrary of the Normal distribution, as recalled previously), as it is the case of the strength of materials, it is widely adopted for the statistical representation of the strength of materials.
The Gumbel distribution, which is also known as the type I asymptotic distribution of the largest value (i.e. the random variable constituted by the largest value, R, from samples of size *n* of a set of initial random variables X_1, X_2, …, X_n), is characterized by the following CDF (Ang and Tang 2007):

$$F_R(r) = \exp\left[e^{-\alpha(r-u)}\right] \tag{4.4}$$

where *u* and α are the parameters of the distribution and represent the most probable value of *R* and an inverse measure of the dispersion of the values of *R*, respectively. The parameters u and α are related to the mean value and the standard deviation of the random variable *R* as $\mu_R = u + 0.577216/\alpha$ and $\sigma_R = \pi/\sqrt{6} \cdot \alpha$.
Parameter *u* is commonly referred to as the "characteristic largest value". It is worth noticing that, for the Gumbel distribution, the skewness coefficient is always equal to 1.1396 and therefore independent from *u* and α. The Gumbel distribution has been selected due to the fact that it has been expressly developed, from a theoretical point of view, for the representation of the extreme values of random variables (as it is the case of the compressive strength values of concrete).

5. STATISTICAL INFERENCE FROM OBSERVATIONAL DATA

The techniques (1) for estimating the parameters values from available observational data (measurements) of a random variable and (2) for deriving the appropriate probabilistic model for a random variable are embodied in the methods of statistical inference, in which information obtained from sampled data is used to represent the corresponding information regarding the population from which the samples are derived. Inferential methods of statistics, therefore, provide a link between the real world and the idealized probability models assumed in a

probabilistic analysis.
In this research work, the classical method called Point Estimation (Ang and Tang 2007) has been adopted.
In this research work, the characteristic value R_{ck} of the compressive strength is searched. Given that R_{ck} can be computed only once the CDF of *R* is known, the attention is focused on the correct representativeness of the CDF, rather than that of the PDF, of the random variable *R*. Consequently, the Kolmogorov-Smirnov (K-S) test (which compares the experimental cumulative frequency with the CDF of an assumed theoretical distribution) has been here adopted.
In the K-S test, for a ordered set of sample data (concrete compressive strengths) of size *n*, a stepwise experimental cumulative frequency function is developed as follows:

$$S_R(r) = \begin{cases} 0 & r < r_1 \\ k/n & r_k \le r < r_{k+1} \\ 1 & r \ge r_n \end{cases} \tag{5.1}$$

where $r_1, r_2, \ldots, r_n$ are the observed values of the ordered set of data (i.e. the experimental compressive strengths of concrete sorted in increasing order). In the K-S test, the maximum difference between the experimental $S_R(r)$ and the assumed theoretical $F_R(r)$, over the entire range of *R*, is the measure of discrepancy between the observed data and the assumed theoretical model. This maximum difference is usually denoted by $D = \max\left|F_R(r) - S_R(r)\right|$.
Then, at various significance levels which are identified by the scalar α, *D* is compared with the critical value D^{α}, as defined by $P\left[D \le D^{\alpha}\right] = 1 - \alpha$. If the observed *D* is less than the critical value D^{α}, the assumed theoretical model is acceptable at the specified significance level α. It is clear that the smaller *D* is, the higher the significance level α is.

6. EXPERIMENTAL DATA ANALYSIS: DISCUSSION ON DIFFERENT CODE FORMULAS FOR THE CONCRETE CHARACTERISTIC STRENGTH

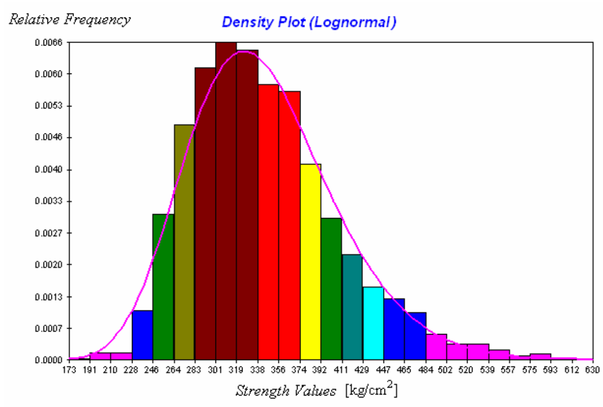
Table 1 reports the parameters values estimated from the available measurements of the compressive strength of the produced concrete. It can be seen that the examined statistical population is characterized by a positive (and considerably larger than zero) value of the skewness coefficient.

Table 1: estimated parameters from the available data

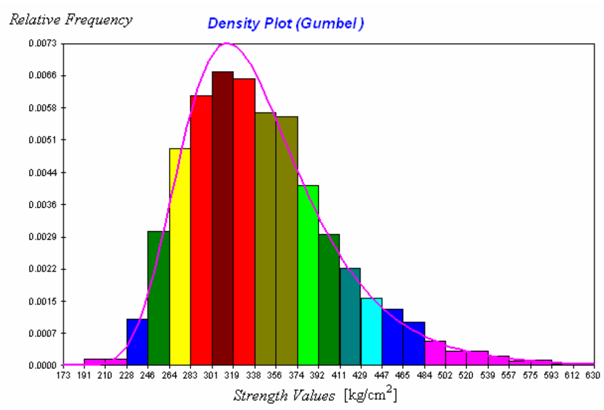
minimum value	19.60 N/mm^2
maximum value	63.00 N/mm^2
experimental mean value	$\mu_{R,\exp} = 34.04$ N/mm^2
experimental variance	$\sigma_{R,\exp}{}^2 = 42.16$ (N/mm^2)^2
experimental standard deviation	$\sigma_{R,\exp} = 6.50$ N/mm^2
experimental coefficient of variation	$\delta_{R,\exp} = 0.191$
experimental second order central moment	42.16 (N/mm^2)^2
experimental third order central moment	219 (N/mm^2)^3
experimental skewness coefficient	$\theta_{R,\exp} = 0.80367$

This property indicates immediately that distribution models which are characterized by symmetric PDF cannot be suitable for this statistical population.
Then, in accordance with the prescriptions of the Comité Euro International du Beton (CEB) regarding the evaluation and the check of the structural concrete, a comparison study of different distribution models has been carried out. Several distribution models have been taken into account: the Normal, the Lognormal, the Gumbel,

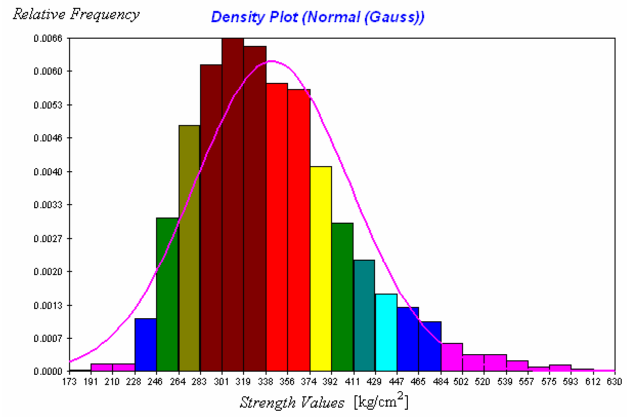
the Gamma, the Beta, the Shifted Lognormal and the Shifted Gamma distributions.
The comparison between the significance levels obtained in the K-S test conducted upon all theoretical distribution models considered showed that the probability distribution which provides the highest significance level is the Shifted Lognormal one, whilst the one which provides the worst is the Normal one.
The results have shown that the Shifted Lognormal distribution is capable of capturing the characteristics of the experimental data much better than all the other distributions examined. The Gumbel distribution has also shown good results. On the other hand, the Normal distribution, which is so often used for statistical representations, has displayed worse results in the capacity of capturing the characteristics of the experimental data.
In more detail Figs. 1 (a), (b) and (c) show the comparison between the experimental frequency histogram and the assumed theoretical PDF as obtained with reference to the Shifted Lognormal, the Gumbel and the Normal distributions, respectively.



(a)



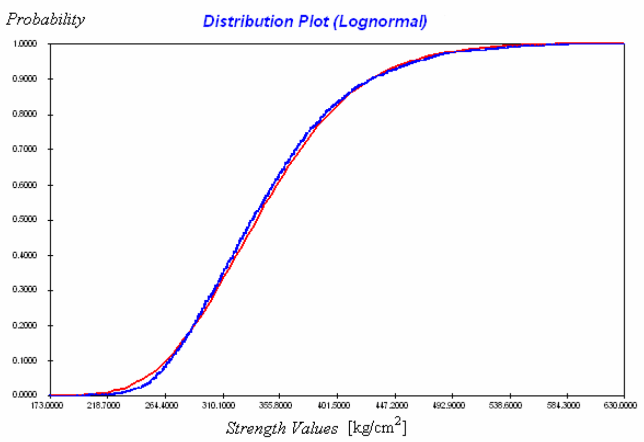
(b)



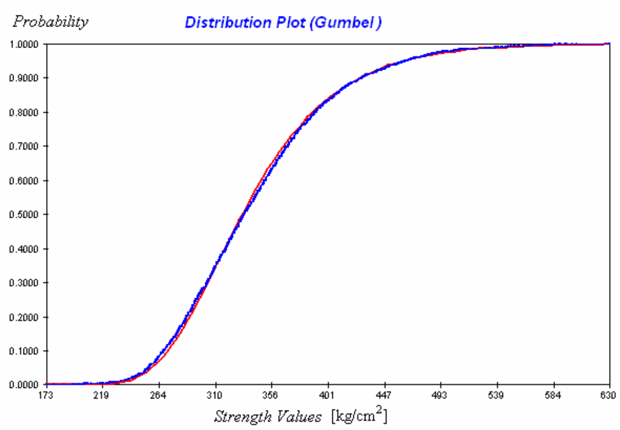
(c)

Figure 1: Theoretical PDF and experimental frequency histogram (a) Shifted Lognormal distribution, (b) Gumbel distribution, (c) Normal distribution

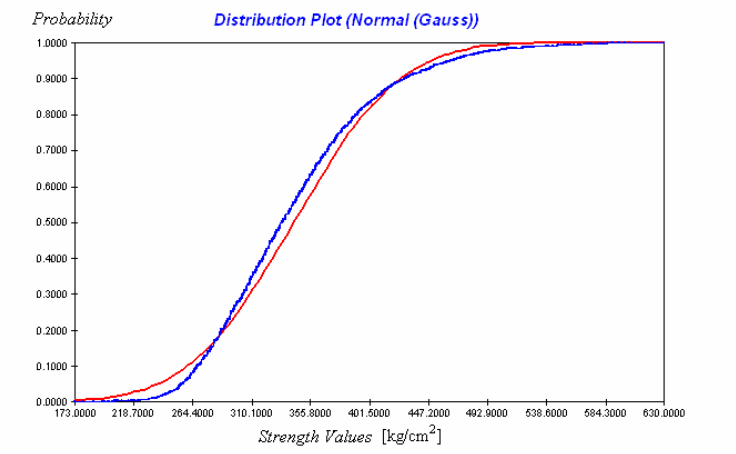
Figs. 2 (a), (b) and (c) show the comparison between the experimental cumulative diagram and the assumed theoretical CDF as obtained with reference to the Shifted Lognormal, the Gumbel and the Normal distributions, respectively.



(a)



(b)



(c)

Figure 2: Theoretical CDF and experimental cumulative diagram (a) Shifted Lognormal distribution, (b) Gumbel distribution, (c) Normal distribution

It is clear that the theoretical PDFs and CDFs of the Shifted Lognormal and the Gumbel distributions fit almost exactly the experimental cumulative diagram. On the other hand, the theoretical PDF and CDF of the Normal distribution differ significantly from the experimental frequency histogram and cumulative diagram respectively, especially in correspondence of the smallest values of concrete strength, which indeed are of fundamental importance for the evaluation of the characteristic value of the concrete strength.
The issue that the Normal distribution model may not be capable of capturing the essence of the sample population at hand should be carefully considered in design codes.

In fact, Eurocode 1, in the operational rules for the evaluation of the characteristic value of material strengths, explicitly states that the formulas to be used are based upon the assumption of Normal distributions for the material strengths. This is also confirmed by the fact that the characteristic value, $x_{k(n)}$, (which includes the statistical uncertainty) is to be evaluated, in the case of a sample of large size *n*, according to $x_{k(n)} = \mu_{X,\exp} - 1.64 \cdot \sigma_{X,\exp}$ where $\mu_{X,\exp}$ and $\sigma_{X,\exp}$ are the experimental mean value and the experimental standard deviation.
What proposed by Eurocode 1, is more penalizing than the corresponding formula proposed by the Italian code, which gives $x_k = \mu_{X,\exp} - 1.40 \cdot \sigma_{X,\exp}$
For sake of comparison, it is convenient to recall what proposed by the American Concrete Institute, which gives $x_k = \mu_{X,\exp} - 1.34 \cdot \sigma_{X,\exp}$ and $x_k = \mu_{X,\exp} - 2.32 \cdot \sigma_{X,\exp} + 500$ in which the strengths are measured in psi.
The ACI regulations pay specific attention upon a fundamental issue for the evaluation of the strength

characteristics of concrete: "the homogeneity in time of the strength characteristics in a given concrete production". In fact, it is reasonable and desirable not only to define the 5% percentile value, but also to impose the further condition that the strengths less than the 5% percentile value are uniformly distributed over all concrete production. This issue is still not explicitly faced by the Italian code.

6.1 Homogeneity in time of the strength characteristics of the examined sample

In order to evaluate the homogeneity properties of the concrete production, the available data have been elaborated as data related to a stochastic process. The objective is to assess whether the process is time-invariant (distribution characteristics constant in time) or time-variant (distribution characteristics which vary in time). The data have been collected in four different groups belonging to four corresponding consecutive temporal intervals (of length slightly larger than 1 year) and a statistical analysis (as the general one described in previous sections) has been carried out separately upon each single group. The results obtained indicate that each single analysis has provided statistical information similar to that obtained in the general case: the distribution which best captures the experimental data is always the Shifted Lognormal one. Therefore, the examined concrete production seems to be time-invariant and facilities which work over the years following constant methodologies (same machinery, same design mixture of concrete, same cement and aggregates from the same gravel pit) are capable of providing homogeneous productions.

7. CONCLUSIONS

In this paper, a statistical inference analysis has been carried out upon the compressive strength values of an extensive population set of concrete cubic specimens, which have been obtained with reference to an homogenous production of about half a million cubic meters of concrete. Such production has been obtained over a five-year period.

The results of the statistical analysis clearly show that the probability density function which best interprets the experimental measurements is the Lognormal one and not the Normal (Gaussian) one. The Italian code and, in certain measure, the Eurocode make explicit reference to Normal distributions, thus leading to a usually penalising evaluation of the characteristic strength (5% percentile). It is therefore advisable that deign codes will encompass the possibility for the engineer to evaluate the concrete characteristics based upon these more refined statistical models.

REFERENCES

Ang, A.H.S. and Tang, W.H. (2007), Probability Concepts in Engineering. Emphasis on Applications to Civil and Environmental Engineering, 2nd Edition, John Wiley and Sons, Inc.

Bhanja, S., and Sengupta, B. (2002), "*Investigations on the compressive strength of silica fume concrete using statistical methods*", Cement and Concrete Research, **32**(9), 1391-1394.

Conte, J.P., (2001). *Statistics, Probability and Reliability*. Course at the University of California at San Diego, fall 2001, 4.68-4.74.

Corradini, G., Scoccia G., Volpe, R. and Tavano S. (1984), "*Statistical evaluation of mechanical properties of superplasticized concrete*", Cement and Concrete Research, **14**(3), 375-385.

Dayaratnam, P., and Ranganathan, R. (1976), "*Statistical analysis of strength of concrete*", *Building and Environment*, **11**(2), 145-152.

Eurocode 1, Basis of design and actions on structures, UNI-EN 1991-1 and UNI-EN 1991-2.

Kameswara Rao, C.V.S. and Swamy, R.N. (1974), "*A statistical theory for the strength of concrete*", Cement and Concrete Research, **4**(4), 669-681.

Melchers, R.E. (1999), Structural Reliability Analysis and Prediction, 2nd Edition, John Wiley and Sons, Inc.

Neville, A. (2006), Neville on Concrete: An Examination of Issues in Practice, BookSurge LLC, USA.