Environmental Modeling

Tutorial 1

Prabhas K. Yadav, PhD

2016

Statistical Distribution

Water-quality data typically consist of measurements of stochastic variables that can only be characterized by probability distributions.

A probability function defines the relationship between the outcome of a random process and the probability of occurrence of that outcome.

For environmental modelling most important probability distributions are:

1. Normal Distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ – Check here
2. Log-Normal Distribution: $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$ for $x > 0$ – Check here
3. Uniform Distribution: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$ – Check here
4. Chi-Square Probability Distribution: $f(x) = \frac{x^{-(1-v/2)} \exp(-x/2)}{2^{v/2}\Gamma(v/2)}$ for $x, v > 0$ – here

Normal Distribution and Standard Normal Distribution

It is usually more convenient to work with the standard normal deviate, z, which is defined by

$$z = \frac{x - \mu_x}{\sigma_x}$$

where x is normally distributed. The probability density function of z, where $\mu = 0$ and $\sigma = 1$ is therefore given by

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$$

The value of $f(z)$ are tabulated and can be found here, but it can also be approximated using:

$$f(z) = \begin{cases} B & z \leq 0 \\ 1 - B & z \geq 0 \end{cases}$$

where,

$$B = \frac{1}{2}\left[1 + 0.19685|z|^2 + 0.115194|z|^2 + 0.000344|z|^3 + 0.019527|z|^4\right]^{-4}$$

For plots check here

Normal Distribution and Standard Normal Distribution

Problem 1

Water-quality samples in a lake show that the concentrations of a pollutant fluctuate randomly and can be approximated by a normal distribution with a mean of 10 mg/L and a standard deviation of 3 mg/L. (a) Estimate the probability that the concentration of the pollutant will exceed 12 mg/L; and (b) estimate the concentration of the pollutant that is likely to be exceeded only 5% of the time.

Solution, part (a)

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{12 - 10}{3} = 0.6667$$

i.e., $z > 0$. To find B, we use:

$$B = \frac{1}{2}\left[1+0.196854|z|+0.115194|z|^2+0.000344|z|^3++0.019527|z|^4\right]^{-4}$$

Substituting $z = 0.6667$, we get:

$$B = 0.2524$$

Normal Distribution and Standard Normal Distribution

Problem 1–continue

We find $f(z)$ from:

$$f(z) = f(0.6667) = 1 - B = 1 - 0.2524 = 0.7476$$

The probability that the concentration exceeds 12 mg/L is therefore equal to $10.7476 = 0.2524$, or approximately 25%.

Solution, part (b)

Let x_{95} be the concentration that is exceeded 5% of the time and z_{95} be the corresponding standard normal deviate, then:

$$f(z_{95}) = 0.95$$

$$B_{95} = 1 - f(z_{95}) = 0.05$$

B_{95} can be found using:

$$B_{95} = \frac{1}{2}\Big[1 + 0.196854|z_{95}| + 0.115194|z_{95}|^2 + \\ + 0.000344|z_{95}|^3 + 0.019527|z_{95}|^4\Big]^{-4}$$

Normal Distribution and Standard Normal Distribution

Problem 1–continue

$$0.05 = \frac{1}{2}\Big[1 + 0.196854|z_{95}| + 0.115194|z_{95}|^2 + \\ + 0.000344|z_{95}|^3 + 0.019527|z_{95}|^4\Big]^{-4}$$

Numerical methods are required to obtain z_{95}. This can be done using Goal Seek in Spreadsheet (check here) of using Matlab or Python (check here). We get $z_{95} = 1.643$, and hence

$$z_{95} = \frac{x_{95} - \mu_x}{\sigma_x}$$

$$1.643 = \frac{x_{95} - 10}{3}$$

which yields $x_{95} = 14.9$ mg/L. Hence, a concentration of 14.9 mg/L will be exceeded only 5% of the time.

Log-Normal Distribution

In cases where the random variable, X, is equal to the product of n random variables $X_1, X_2, \ldots X_n$, such that

$$X = X_1, X_2, \ldots X_n$$

then logarithm of X is equal to the sum of n random variables, where

$$\ln X = \ln X_1 + \ln X_2 + \ldots + X_n$$

then if Y is normally distributed, the theory of random functions can be used to show that the probability density function of X, the log-normal distribution (for $x > 0$), is given by

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu_y)^2}{2\sigma_y^2}\right)$$

For the plot check here

Log-Normal Distribution

The mean, variance and skewness of a log-normally distributed variable, X, in terms of the parameters of the log-transformed variable, μ_y and σ_y, are given by:

$$\mu_x = \exp\left(\mu_y + \frac{\sigma_y^2}{2}\right)$$

$$\sigma_x^2 = \mu_x^2[\exp(\sigma_y^2) - 1]$$

$$g_x = 3C_v + C_v^3 \quad \text{where} \quad C_v = \frac{\sigma_x}{\mu_x}$$

Problem 2

The natural logarithms of concentration data collected in a coastal water follow a normal distribution with a mean of 2.97 and a standard deviation of 0.301, where the concentrations are measured in mg/L. (a) Estimate the mean, standard deviation, and skewness of the measured concentration data; (b) What is the probability of the concentration exceeding 30 mg/L?

Log-Normal Distribution

Solution 2a

From the given data: $\mu_y = 2.97$ and $\sigma_y = 0.301$.
Using Equations from the last slide yields:

$$\mu_x = \exp\left(\mu_y + \frac{\sigma_y^2}{2}\right) = \exp\left(2.97 + \frac{0.301^2}{2}\right) = 20.4 \text{ mg/L}$$

$$\sigma_x = \sqrt{\mu_x^2[\exp(\sigma_y^2) - 1]} = \sqrt{20.4^2[\exp(0.301^2) - 1]} = 6.28 \text{ mg/L}$$

$$C_v = \frac{\sigma_x}{\mu_x} = \frac{6.28}{20.4} = 0.308$$

$$g_x = 3C_v + C_v^3 = 3(0.308) + (0.308)^3 = 0.953$$

Solution 2b

(b) For a concentration of $C = 30$ mg/L, the exceedance probability is determined from the following calculations,

$$\ln C = \ln(30) = 3.40$$

Log-Normal Distribution

Solution 2b continue

$$z = \frac{\ln C - \mu_y}{\sigma_y} = \frac{3.40 - 2.97}{0.301} = 1.43$$

$$B = \frac{1}{2}\Big[1 + 0.196854|z| + 0.115194|z|^2 + 0.000344|z|^3 + \\ + 0.019527|z|^4\Big]^{-4} \\ = \frac{1}{2}[1 + 0.196854(1.43) + 0.115194(1.43)^2 + 0.000344(1.43)^3 + \\ + 0.019527(1.43)^4]^{-4} = 0.075$$

Hence, the probability of the sample concentration exceeding 30 mg/L is 0.075 or 7.5%.

Uniform Distribution

The uniform distribution describes the behaviour of a random variable in which all possible outcomes are equally likely within the range $[a, b]$. For a continuous random variable, x, the uniform probability density function, $f(x)$, is given by:

$$f(x) = \frac{1}{b-a} \quad \text{for } a \leq x \leq b$$

where the parameters a and b define the range of the random variable.

The mean, μ_x, and variance, σ_x^2 , of a uniformly distributed random variable are given by:

$$\mu_x = \frac{1}{2}(a+b), \; \sigma_x^2 = \frac{1}{12}(b-a)^2$$

For the plot check here

Uniform Distribution

Problem 3

Anecdotal evidence based on historical sampling in a river indicate that Escherichia coli concentrations are in the range of 1100 mg/L. Based on this anecdotal report, estimate the E.coli concentration that is likely to have an exceedance rate of 10%.

Solution 3

Since other values are not indicated, it is appropriate to assume a uniform probability distribution between 1 and 100 mg/L. Hence, a = 1 mg/L, b = 100 mg/L, and the probability distribution of the concentration in given by the equation:

$$f(c) = \frac{1}{b-a} = \frac{1}{100-1} = \frac{1}{99} \quad (\text{ mg/L})^{-1}$$

Therefore, if c_{90} is the concentration with a 10% exceedance probability:

$$(100c_{90})\frac{1}{99} = 0.10$$

which yields $c_{90} = 90.1$ mg/L.

Chi-Square Distribution

In probability theory and statistics, the chi-squared distribution (also chi-square or χ^2-distribution) with v degrees of freedom is the distribution of a sum of the squares of v independent *standard* normal random variables.

The chi-squared distribution is used in the common chi-squared tests for *goodness of fit* of an observed distribution to a theoretical one. then the probability density function of χ^2 is defined as the chi-square distribution and is given by

$$f(x) = \frac{x^{-(1-v/2)} \exp(-x/2)}{2^{v/2}\Gamma(v/2)} \quad \text{for } x, v > 0$$

where $x = \chi^2$, and v is called the number of degrees of freedom. The mean and variance of the chi-square distribution are given by

$$\mu_x^2 = v \quad \text{and } \sigma_{x^2}^2 = 2v$$

The shape of the χ^2 can be found here and the tabulated value of $f(x)$ can be found here

Chi-Square Distribution

Chi-Square Goodness-of-Fit Criteria

It is defined as (more here):

$$\chi^2 = \sum_{i=1}^{n} = \frac{(\text{observed value}_i - \text{simulated value}_i)^2}{\text{simulated value}_i}$$

In order to accept the model results as a good fit we need:

$$P(\chi^2 \leq \chi_0^2) = 1 - \alpha$$

Where α is confidence level, $1 - \alpha$ is significance level. The probability values are obtained from the distribution table (check last slide).

Problem 4

A waste discharge with biochemical oxygen demand (BOD) at Km 0.0 causes a depletion in dissolved oxygen in a stream. Model calibration results are tabulated below (DO model) together with field measurements (DO field).

Chi-Square Distribution

Distance (Km)	Concentration (mg/L)	
	Measured	Simulated
0	8	8
5	6.6	6.3
10	5.5	5.4
20	4.4	4.58
30	4.6	4.64
40	4.5	5.1
50	5.2	5.5
60	6	6
80	7	6.7
100	7.3	7.2

Determine if the model calibration is acc the following statistical criteria: Chi-square 0.10 significance level (a 90% confidence level.

Chi-Square Distribution

Distance (Km)	DO Measured (mg/L)	Do Simulated (mg/L)	Observed-simulated	(Observed-simulated)^2	(Observed-simulated)^2 /Simulated
0	8	8	0.00	0.00	0.0000
5	6.6	6.3	0.30	0.09	0.0143
10	5.5	5.4	0.100	0.010	0.0019
20	4.4	4.58	-0.180	0.032	0.0071
30	4.6	4.64	-0.040	0.002	0.0003
40	4.5	5.1	-0.600	0.360	0.0706
50	5.2	5.5	-0.300	0.090	0.0164
60	6	6	0.000	0.000	0.0000
80	7	6.7	0.300	0.090	0.0134
100	7.3	7.2	0.100	0.010	0.0014
Total					$\chi^2 = 0.1253$

Degree of freedom = Total observation -1 = 10 -1 = 9

Confidence level α = 1 - significance level $=1 - 0.10 = 0.90$

From the tabulation data of χ^2*distribution*, it can be found that $\chi_0^2 = 4.168$

Since $\chi^2 < \chi_0^2 = 4.168$, the model passes the goodness of fit test at a 0.10 significance level. Check here for details.

Linear Regression

Linear regression is an approach for modelling the relationship between a scalar dependent variable y and one or more independent variables denoted X. Check here for details.

Linear regression of paired data for model predictions and field observations at the same time requires:

Sample covariance given by:

$$s_{x,y} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Sample standard deviation given by:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}} \quad s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})}{n-1}}$$

And **Sample correlation coefficient** given by

$$R = \frac{s_{x,y}}{s_x \cdot s_y}$$

Linear Regression

Problem 5

A waste discharge with biochemical oxygen demand (BOD) at Km 0.0 causes a depletion in dissolved oxygen in a stream. Model calibration results are tabulated below (DO model) together with field measurements (DO field).

Distance (Km)	Concentration (mg/L)	
	Measured	Simulated
0	8	8
5	6.6	6.3
10	5.5	5.4
20	4.4	4.58
30	4.6	4.64
40	4.5	5.1
50	5.2	5.5
60	6	6
80	7	6.7
100	7.3	7.2

Perform linear least- squares regression of model results (DO simulation on x-axis) versus observed data (DO measurement on y-axis) with $R^2 > 0.8$.

Linear Regression

Solution

Distance (Km)	Measured (mg/L) (y)	Simulated (mg/L) (x)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
0	8	8	2.058	2.090	4.235	4.368	4.301
5	6.6	6.3	0.358	0.690	0.128	0.476	0.247
10	5.5	5.4	-0.542	-0.410	0.294	0.168	0.222
20	4.4	4.58	-1.362	-1.510	1.855	2.280	2.057
30	4.6	4.64	-1.302	-1.310	1.695	1.716	1.706
40	4.5	5.1	-0.842	-1.410	0.709	1.988	1.187
50	5.2	5.5	-0.442	-0.710	0.195	0.504	0.314
60	6	6	0.058	0.090	0.003	0.008	0.005
80	7	6.7	0.758	1.090	0.575	1.188	0.826
100	7.3	7.2	1.258	1.390	1.583	1.932	1.749
Total					**11.272**	**14.629**	**12.614**

$$s_x = 1.119 \quad s_y = 1.275 \quad s_{x,y} = 1.402$$

$$R = \frac{s_{x,y}}{s_x \cdot s_y} = \frac{1.402}{1.119 \times 1.275} = 0.987$$

Since $R^2 = 0.966 > 0.80$, the calibration is good.

Complete calculation is here

Assignment Problems

1. Water-quality samples in a lake show that the concentrations of a pollutant fluctuate randomly and can be approximated by a normal distribution with a mean of 15 mg/L and a standard deviation of 2 mg/L. (a) Estimate the probability that the concentration of the pollutant will exceed 10 mg/L; and (b) estimate the concentration of the pollutant that is likely to be exceeded only 15% of the time.
2. The natural logarithms of concentration data collected in a coastal water follow a normal distribution with a mean of 2 and a standard deviation of 0.5, where the concentrations are measured in mg/L. (a) Estimate the mean, standard deviation, and skewness of the measured concentration data; (b) What is the probability of the concentration exceeding 10 mg/L?
3. Anecdotal evidence based on historical sampling in a river indicate that Escherichia coli concentrations are in the range of 220 mg/L. Based on this anecdotal report, estimate the E.coli concentration that is likely to have an exceedance rate of 15%.

Assignment

4. A waste discharge with biochemical oxygen demand (BOD) at km 0.0 causes a depletion in dissolved oxygen in a stream. Model calibration results are tabulated below (DO model) together with field measurements (DO field).

Distance (Km)	Concentration (mg/L)	
	Measured	Simulated
0	8	8
5	6.6	6.3
10	5.5	5.4
20	4.4	4.58
30	4.6	4.64
40	4.5	5.1
50	5.2	5.5
60	6	6
80	7	6.7
100	7.3	7.2

Multiply column two with 1.5 and column three with 1.8 and then perform χ^2 goodness test and the regression test to check whether the calibration can be considered good.