Statistics and Probability Theory

in

Civil, Surveying and Environmental Engineering

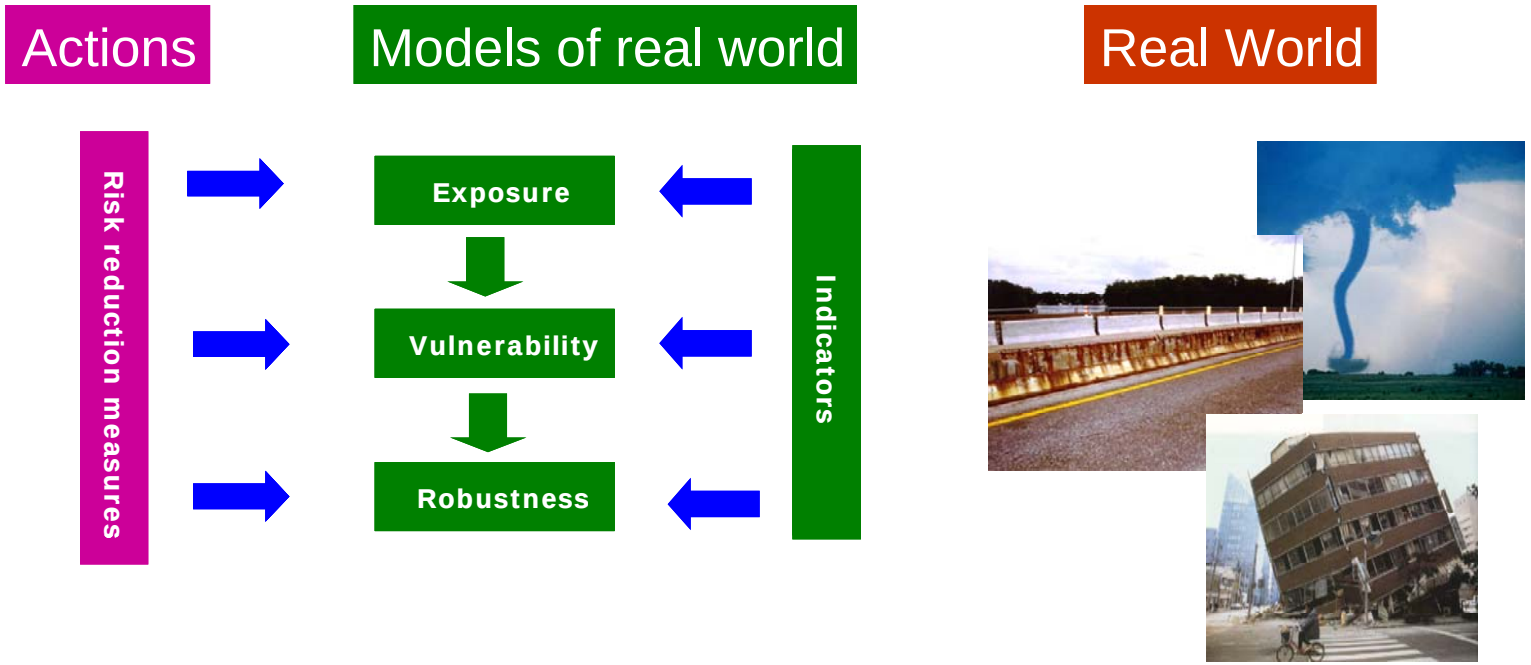
Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Overview of Uncertainty Modeling

- Random variables and their characteristics



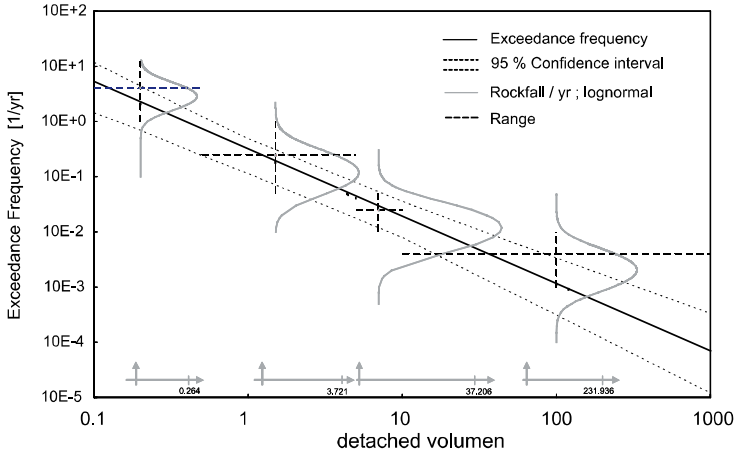
Overview of Uncertainty Modeling

- Random variables and their characteristics

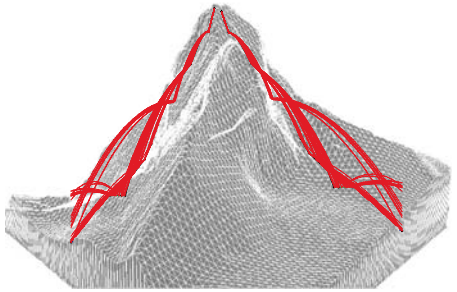
  Design of rock-fall galleries

Overview of Uncertainty Modeling

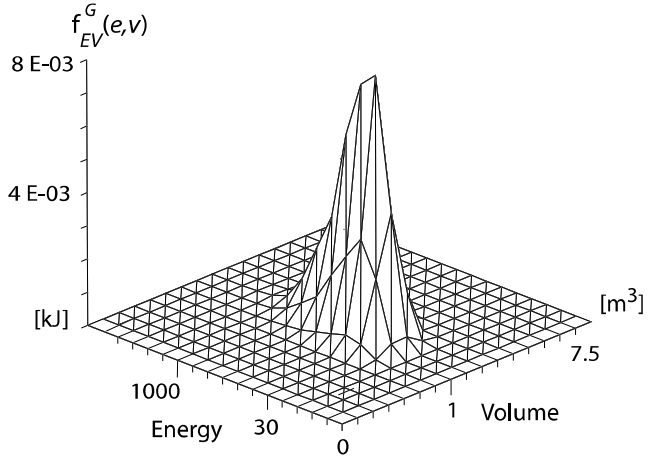
- Random variables and their characteristics



Detachment modeling



Fall modeling



Tools in Uncertainty Modeling

- Engineering problems – also those involving uncertainty are very often specific – unique !

  Being able to solve such problems requires

  - basic tools (physical, mathematical, natural sciences, human sciences, engineering,…)

  - innovation (being able to identify ways of solving problems)

  - training !

  Training is important because it provides experience.

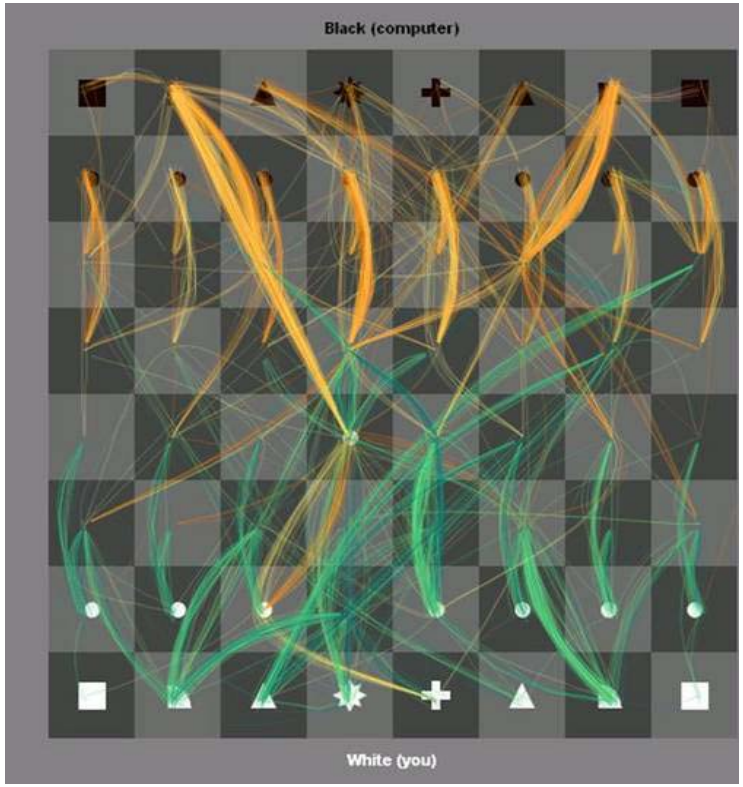
  By training we start to recognize patterns !

Tools in Uncertainty Modeling

- Pattern recognition helps to identify:

the usefulness of solution strategies from previous problems

the potential of the available tools in a given context



Tools in Uncertainty Modeling

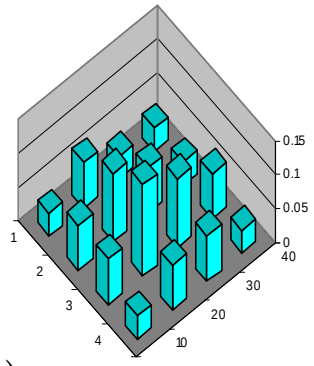
- **Random variables and their characteristics**

The expectation operator

$$E[c] = c$$
$$E[cX] = cE[X]$$
$$E[a + bX] = a + bE[X]$$
$$E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$$

The variance operator

$$Var[c] = 0$$
$$Var[cX] = c^2 Var[X]$$
$$Var[a + bX] = b^2 Var[X]$$



Jointly distributed random variables

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \le x_1 \cap X_2 \le x_2 \cap \ldots \cap X_n \le x_n)$$

Tools in Uncertainty Modeling

- Random variables and their characteristics

  Functions of random variables

  - sum of two random variables

$$Y = X_1 + X_2$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1$$

  - non-linear function of random variables

$$Y = g(X)$$

$$f_Y(y) = \left| \frac{\partial x}{\partial y} \right| f_X(x)$$

Tools in Uncertainty Modeling

- Random variables and their characteristics

  Functions of random variables

$$\mathbf{Y} = (Y_1, Y_2, .. Y_n)^T$$

$$Y_i = g_i(\mathbf{X}), \qquad X_i = f_i(\mathbf{Y})$$

$$f_{\mathbf{Y}}(\mathbf{y}) = \left|\mathbf{J}\right| f_{\mathbf{X}}(\mathbf{x})$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

Contents of Todays Lecture

- Random variables
  - The Central Limit Theorem
  - The Normal distribution
  - The Log-Normal distribution

- Stochastic Processes and Extremes
  - Random sequences (Bernoulli trials)
  - Binomial distribution
  - Geometric distribution

Random Variables

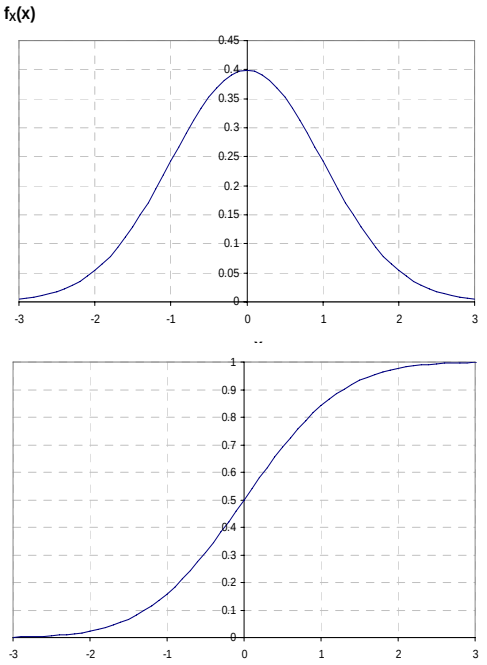
- The Central Limit Theorem states:

  The probability distribution function of a sum of a number of random variables approaches the Normal (Gaussian) distribution as the number becomes large

$$Y = X_1 + X_2 + ... + X_n$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$



Random Variables

- The Central Limit Theorem

  Conditions for the validity of the theorem:

$$Y = X_1 + X_2 + \ldots + X_n$$

  The sum should not be dominated by one or a few components

  The statistical dependency between components should not be strong

  No requirements to the type of distribution of the components

  If the components have skew distributions the number increases

Random Variables

- Illustration:

  A structural member is measured using a ruler.
  - The ruler has limited length (2 m).
  - The smallest unit on the ruler is 1 mm.

  All measurements are rounded to the closest unit on the ruler.

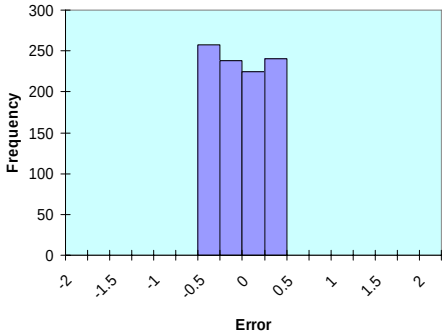
  Each measurement is subject to a measurement uncertainty uniformly distributed in the range of +/- 0.5 mm.

  We now consider the accumulated error associated with measurements over lengths
  - up to 2 m (one measurement)
  - between 2 and 4 m (two measurements)
  - between 6 and 8 m (four measurements)
  - between 14 and 16 m (eight measurements)

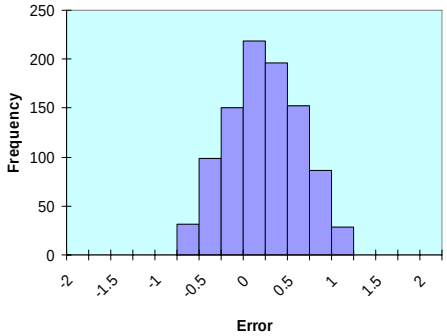
Random Variables

- Illustration:

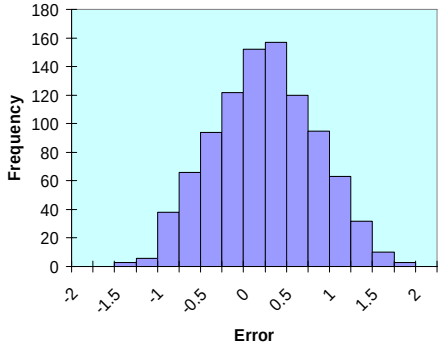
N=1



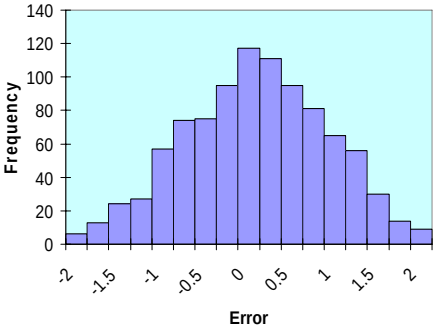
N=2



N=4



N=8



Random Variables

- The Normal distribution

  The analytical form of the Normal distribution may be derived by repeated use of the result regarding the probability density function for the sum of two random variables

  The Normal distribution is very frequently applied in engineering modeling when a random quantity can be assumed to be composed as a sum of a number of individual contributions: $X_i, i = 1,2,..,n$

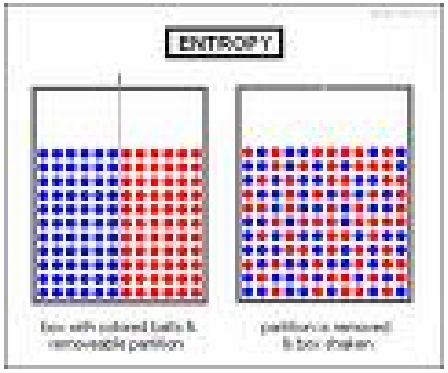
  A linear combination *S* of *n* Normal distributed random variables $S = a_0 + \sum_{i=1}^{n} a_i X_i$ is thus also a Normal distributed random variable

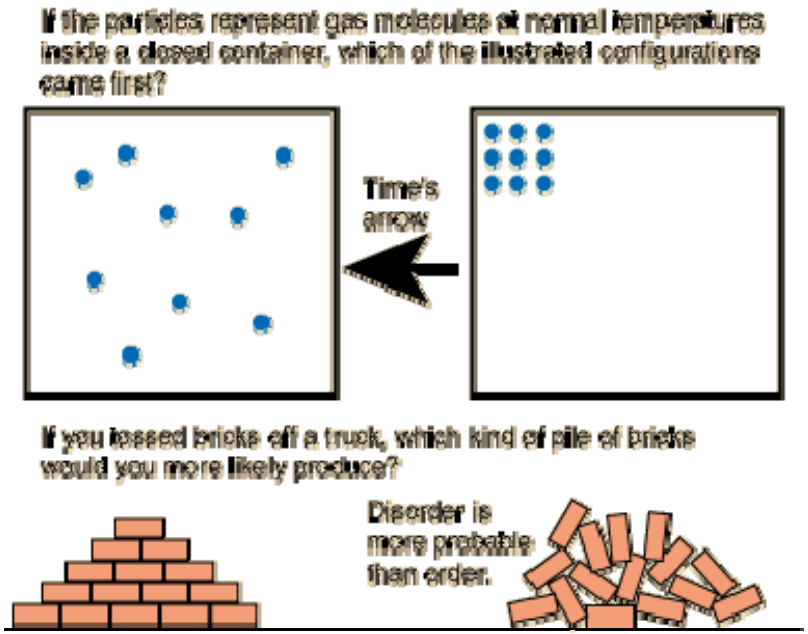
Random Variables

- The Normal distribution

  The Normal distribution also results from other considerations

  The distribution of energy in an isolated system

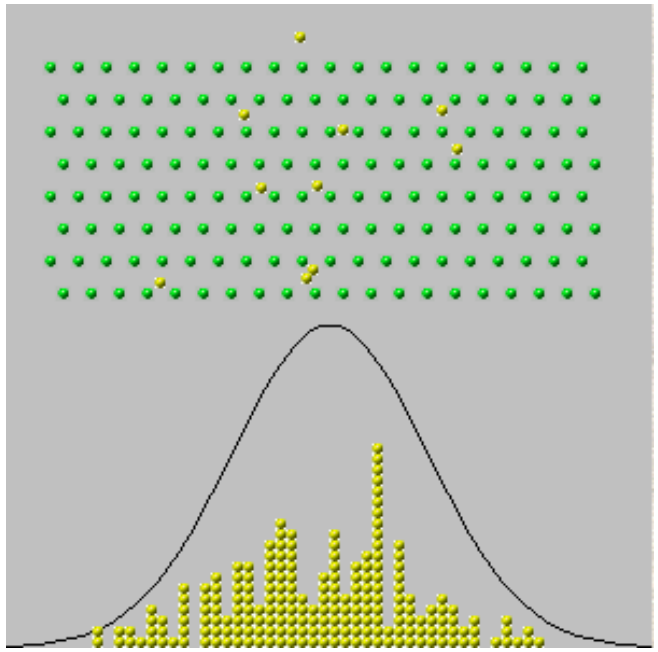




Random Variables

- The Normal distribution

  The accumulation of random movements

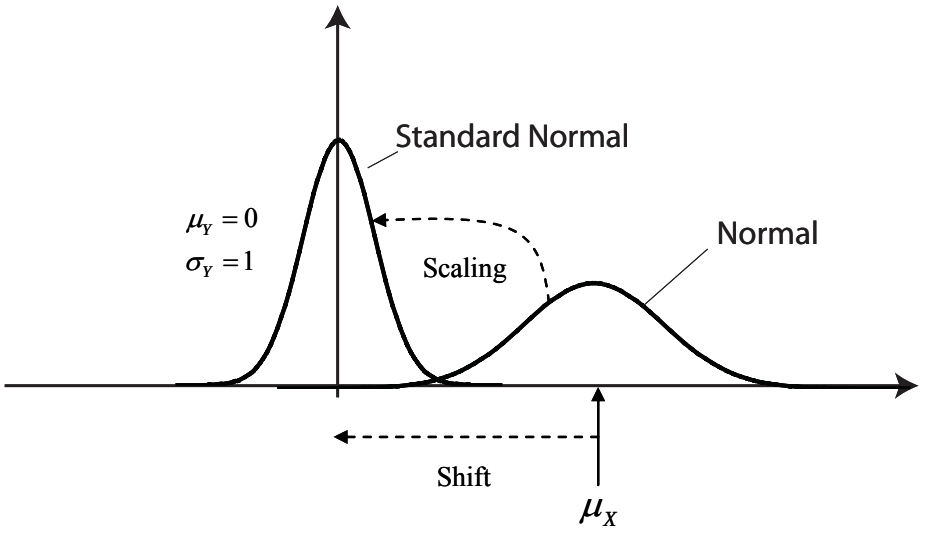


Random Variables

- The Normal distribution:

  In the case where the mean value is equal to zero and the standard deviation is equal to 1 the random variable is said to be *standardized*.

$$Y = \frac{X - \mu_X}{\sigma_X}$$ **Standardized random variable**



Random Variables

- The Normal distribution:

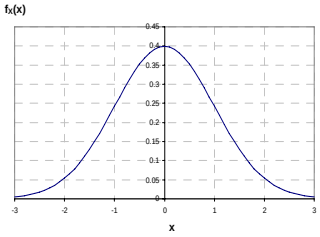
  In the case where the mean value is equal to zero and the standard deviation is equal to 1 the random variable is said to be *standardized*.

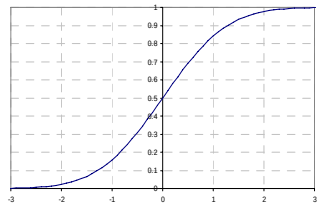
$$Y = \frac{X - \mu_X}{\sigma_X}$$ **Standardized random variable**

$$f_Y(y) = \varphi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)$$

$$F_Y(y) = \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{1}{2}x^2\right) dx$$

Standard normal





Random Variables

When the logarithm of a random variable X i.e.

$Y = \ln(X), \qquad Y : N(\mu_Y, \sigma_Y)$

is Normal distributed the random variable X is said to be Log-Normal distributed

$$X : LN(\lambda, \zeta)$$

$$f_X(x) = \frac{1}{x\zeta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(x)-\lambda}{\zeta}\right)^2\right) \qquad \mu_X = \exp\left(\lambda + \frac{\zeta^2}{2}\right)$$

$$F_X(x) = \Phi\left(\frac{\ln(x)-\lambda}{\zeta}\right) \qquad \sigma_X = \exp\left(\lambda + \frac{\zeta^2}{2}\right)\sqrt{\exp(\zeta^2)-1}$$

Random Variables

Where the Normal distribution follows from the sum of random variables – **Central Limit Theorem**

the Log-Normal distribution follows from the product of random variables

$$\ln(X_1 \cdot X_2 \cdots X_n) = \ln(\prod_{i=1}^{n} X_i) = \sum_{i=1}^{n} \ln(X_i)$$

Random Variables

The Log-Normal distribution has the useful property that if

$$P = \prod_{i=1}^{n} Y_i^{a_i}$$

and all Y_i are independent Log-Normal distributed random variables with parameters ζ_i , λ_i and $\varepsilon_i = 0$ then P is also Log-Normal with parameters

$$\lambda_P = \sum_{i=1}^{n} a_i \lambda_i \qquad \zeta_P^2 = \sum_{i=1}^{n} a_i^2 \zeta_i^2 \qquad f_P(p) = \frac{1}{p\zeta_P\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(p)-\lambda_P}{\zeta_P}\right)^2\right)$$

Random Variables

The Log-Normal distribution is often used to model

- uncertain parameters which cannot have negative realizations
- fatigue lifes
- steel and concrete resistance
- daily river flows
- whenever a random variable results as a product of several random variables

Random Variables

Concrete compression strength

$$\mu_X = \exp\left(\lambda + \frac{\zeta^2}{2}\right)$$

$$\sigma_X = \exp\left(\lambda + \frac{\zeta^2}{2}\right)\sqrt{\exp(\zeta^2) - 1}$$

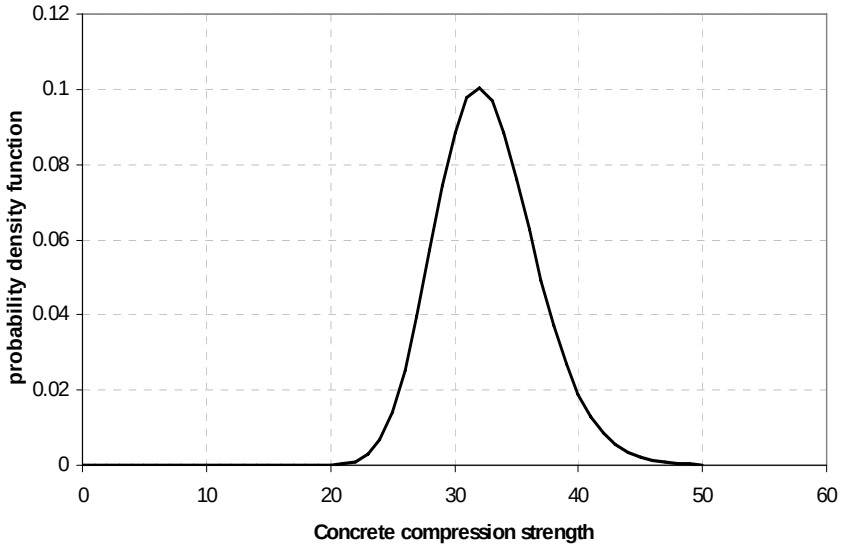
$$\Downarrow$$

$$V_X = \frac{\sigma_X}{\mu_X} = \sqrt{\exp(\zeta^2) - 1} = \frac{4.05}{32.67} = 0.12 \Rightarrow \zeta = 0.12, \lambda = 3.48$$

Probability of value lower than 25 MPa

$$F_X(25) = \Phi\left(\frac{\ln(25) - 3.48}{0.12}\right) = 0.018$$

i	x_i
1	24.4
2	27.6
3	27.8
4	27.9
5	28.5
6	30.1
7	30.3
8	31.7
9	32.2
10	32.8
11	33.3
12	33.5
13	34.1
14	34.6
15	35.8
16	35.9
17	36.8
18	37.1
19	39.2
20	39.7



Random Variables

There exist a large number of different probability density and cumulative distribution functions:

Uniform
Normal
Log-normal
Exponential
Beta
Gamma
...

...

Distribution type	Parameters	Moments
Uniform, $a \le x \le b$ $f_X(x) = \frac{1}{b-a}$ $F_X(x) = \frac{x-a}{b-a}$	a b	$\mu = \frac{a+b}{2}$ $\sigma = \frac{b-a}{\sqrt{12}}$
Normal $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ $F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt$	μ $\sigma > 0$	μ σ
Shifted Lognormal, $x > \varepsilon$ $f_X(x) = \frac{1}{(x-\varepsilon)\zeta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right)^2\right)$ $F_X(x) = \Phi\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right)$	λ $\zeta > 0$ ε	$\mu = \varepsilon + \exp\left(\lambda + \frac{\zeta^2}{2}\right)$ $\sigma = \exp\left(\lambda + \frac{\zeta^2}{2}\right)\sqrt{\exp(\zeta^2)-1}$
Shifted Exponential, $x \ge \varepsilon$ $f_X(x) = \lambda \exp(-\lambda(x-\varepsilon))$ $F_X(x) = 1 - \exp\left(-\lambda\left(x-\varepsilon\right)\right)$	ε $\lambda > 0$	$\mu = \varepsilon + \frac{1}{\lambda}$ $\sigma = \frac{1}{\lambda}$
Gamma, $x \ge 0$ $f_X(x) = \frac{b^p}{\Gamma(p)} \exp(-bx) x^{p-1}$ $F_X(x) = \frac{\Gamma(bx, p)}{\Gamma(p)}$	$p > 0$ $b > 0$	$\mu = \frac{p}{b}$ $\sigma = \frac{\sqrt{p}}{b}$

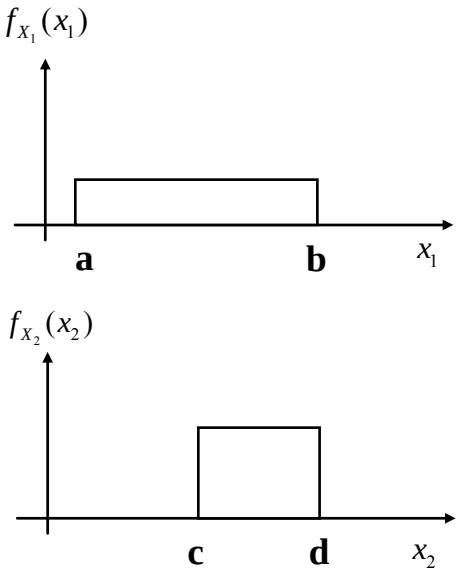
Small Example 1

We remember the convolution integral which we used for establishing the probability density function for the sum of two random variables:

$$Y = X_1 + X_2$$

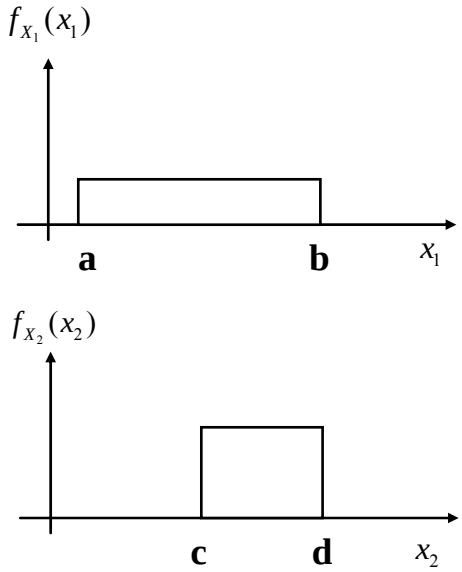
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1$$

Let us see how easily this works for two uniformly distributed random variables:



Small Example 1

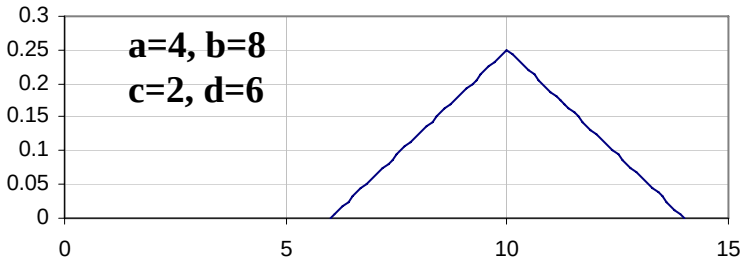
Assuming that the two random variables are independent we can write the convolution integral as:

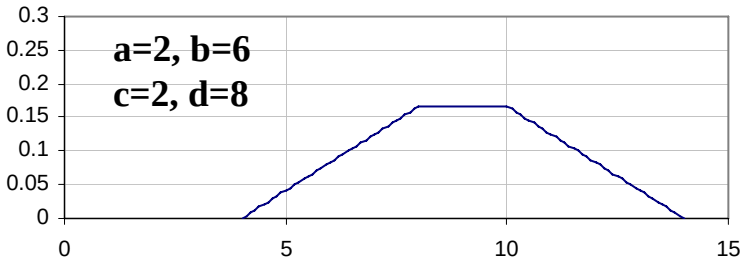


$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1)\, f_{X_1}(x_1)\, dx_1$$

$$= \frac{1}{(b-a)(d-c)} \int_a^b \Uparrow (y - x_1 \in [c;d])\, dx_1$$

$$= \frac{1}{(b-a)(d-c)} \left[x_1 \right]_{\max(c,y-a)}^{\min(d,y-b)}, \quad a + c \le y \le b + d$$





Stochastic Processes and Extremes

- Random quantities may be "time variant" in the sense that they take new values at different times or at new trials.

- If the new realizations occur at discrete times and have discrete values the random quantity is called a **random sequence**

  failure events, traffic congestions,…

- If the new realizations occur continuously in time and take continues values the random quantity is called a **random process** or **stochastic process**

  wind velocity, wave heights,…

Stochastic Processes and Extremes

- Random sequences

  - A sequence of experiments with only two possible and mutually exclusive outcomes is called a **Bernoulli trial**
  - Typically the outcomes of Bernoulli trials are denoted **successes** or **failures**

  If the probability of success in one trial is constant and equal to *p* the probability density of *Y* successes in *n* trials, i.e. $p_Y(y)$ is given by:

$$p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1,2...n$$

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

Binomial probability density function

Binomial operator

Stochastic Processes and Extremes

- **Random sequences**

  - A sequence of experiments with only two possible and mutually exclusive outcomes is called a Bernoulli trial

  The Binomial cumulative distribution function then follows as:

$$P_Y(y) = \sum_{i=0}^{y} \binom{y}{i} p^i (1-p)^{n-i}, \qquad y = 0,1,2,\ldots n$$

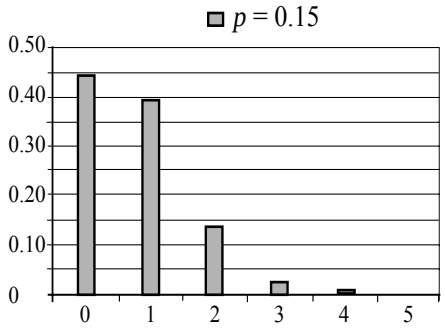
Stochastic Processes and Extremes

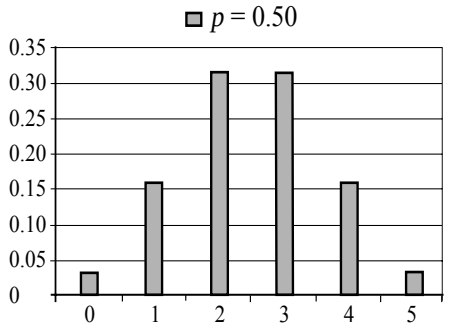
- Random sequences

  - A sequence of experiments with only two possible and mutually exclusive outcomes is called a **Bernoulli trial**

  Illustration:

  Binomial probability density function for n=5 and p=0.15 and p=0.5





Small Example 2

We remember that we can establish the probability density function of a function of a random variable through:

$$Y = g(X)$$

$$f_Y(y) = \left| \frac{\partial x}{\partial y} \right| f_X(x)$$

Small Example 2

Let us see how easily this works:

$$Y = X^2$$

$$\Downarrow$$

$$X = \sqrt{Y}$$

$$f_Y(y) = \left|\frac{\partial x}{\partial y}\right| f_X(x)$$

$$\frac{\partial x}{\partial y} = \frac{\partial \sqrt{y}}{\partial y} = \frac{1}{2} y^{-\frac{1}{2}}$$

$$f_Y(y) = \left|\frac{1}{2} y^{-\frac{1}{2}}\right| f_X(\sqrt{y})$$

Stochastic Processes and Extremes

- **Random sequences**

  The expected value and the variance of a **binomially distributed** random variable *Y* is given by:

$$E[Y] = np$$

$$Var[Y] = np(1-p)$$

Stochastic Processes and Extremes

- **Random sequences**

  The probability density function for the number of (independent) trials before the first success can be given as:

$$p_N(n) = p(1-p)^{n-1}$$

Geometric probability density

  and the corresponding cumulative distribution function is thus

$$P_N(n) = \sum_{i=1}^{n} p(1-p)^{i-1} = 1-(1-p)^n$$

Geometric cumulative distribution

Small Example 3

We remember that we could establish the probability density function of a vector of random variables *Y* which were given as functions of a vector of random variables *X*

$$\mathbf{Y} = (Y_1, Y_2, ..Y_n)^T$$

$$\mathbf{X} = (X_1, X_2, ..X_n)^T \qquad Y_i = g_i(\mathbf{X}) \qquad X_i = f_i(\mathbf{Y})$$

$$f_{\mathbf{Y}}(\mathbf{y}) = \left|\mathbf{J}\right| f_{\mathbf{X}}(\mathbf{x})$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

Small Example 3

Let us see how easily this approach can be applied for the following problem:

$$Y_1 = X_1 + X_2 \qquad X_1 = Y_1 - Y_2$$

$$Y_2 = X_2 \qquad X_2 = Y_2$$

$$\mathbf{J} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \qquad \det(\mathbf{J}) = 1 \times 1 - 0 \times 1 = 1 \Rightarrow \left|\mathbf{J}\right| = 1$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots\cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots\cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

$$f_{\mathbf{Y}}(y_1, y_2) = f_{\mathbf{X}}(y_1 - y_2, y_2)$$

Stochastic Processes and Extremes

The median of the geometric distribution provides information in regard to how "long" we need to play a game with probability *p* of winning per time unit.

Time units might be

- tosses (dices)
- years (earthquakes)

The median is defined through

$$P_N(n) = 0.5 = 1 - (1-p)^n$$

All we need to determine is *n* as a function of *p*

Stochastic Processes and Extremes

The median of the geometric distribution provides information in regard to how "long" we need to play a game with probability *p* of winning per time unit.

$$P_N(n) = 0.5 = 1 - (1-p)^n$$

We take the natural logarithm on both sides and get:

$$\ln(0.5) = n \ln(1-p)$$

$$\Downarrow$$

$$0.7 \approx -n \ln(1-p)$$

Now we use that the natural logarithm of

$$\ln(1-p) = -p + \frac{1}{2}p^2 - \frac{1}{3}p^3 + \ldots = \sum_{k=1}^{\infty} (-1)^k \frac{p^k}{k}$$

$$\Downarrow$$

$$\ln(1-p) \approx -p \quad \text{for small } p$$

$$0.7 \approx np \Rightarrow n = \frac{0.7}{p}$$

Stochastic Processes and Extremes

We can now apply this result:

50% chance of getting a 6 requires (n tosses):
$n = 0.7 \times 6 = 4$ tosses

50% chance of getting two 6 (with 2 dices) requires:
$n = 0.7 \times 36 = 25$ tosses

50% chance experiencing an earthquake with an annual probability of 0.001 requires (n years):
$n = 0.7 \times 1000 = 700$ years

Stochastic Processes and Extremes

- **Random sequences**

  The expected value and the variance of a random variable with a *Geometrically* distributed random variable are given by:

$$E[N] = \frac{1}{p}$$

$$Var[N] = \frac{1-p}{p^2}$$

If *p* is the annual probability of e.g. an extreme earthquake *E[N]* is the return period of such earthquakes