



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



LECTURE NOTES

Fundamentals of Probability and Statistics

Chapter 1: Lecture 1

Spring 2018

Written by marco broccardo
IBK, ETHZ

1 Element of set theory

Set theory is a branch of mathematics and probability theory which enables to tackle class of problems dealing with random phenomena.

1.1 Definitions

Random phenomenon a phenomenon that has more than one possible outcome. The true outcome is unknown until it is observed.

Sample space collection of all possible outcomes of a random phenomenon. Notation: S .

Sample point each element of the sample space. Notation: x .

Event a collection of sample points which represents a subset of the sample space. Notation: E , $E \subseteq S$.

These definitions are typically visualized via Venn's diagram. Figure 1 shows a typical Venn diagram. Observe that the rectangular shape is used for S , the oval shape for E , and the dot for x .

1.2 Operation on Events

Union: given the events E_1 and E_2 , the union event, denoted with $E_1 \cup E_2$, is the event that contains all sample points in either E_1 or E_2 . Figure 2(i).

Intersection: given the events E_1 and E_2 , the intersection event, denoted with $E_1 \cap E_2$ or simply $E_1 E_2$, is the event that contains the sample points both in E_1 and E_2 Figure 2(ii).

These operations obey certain properties:

Commutative property of union: $E_1 \cup E_2 = E_2 \cup E_1$.

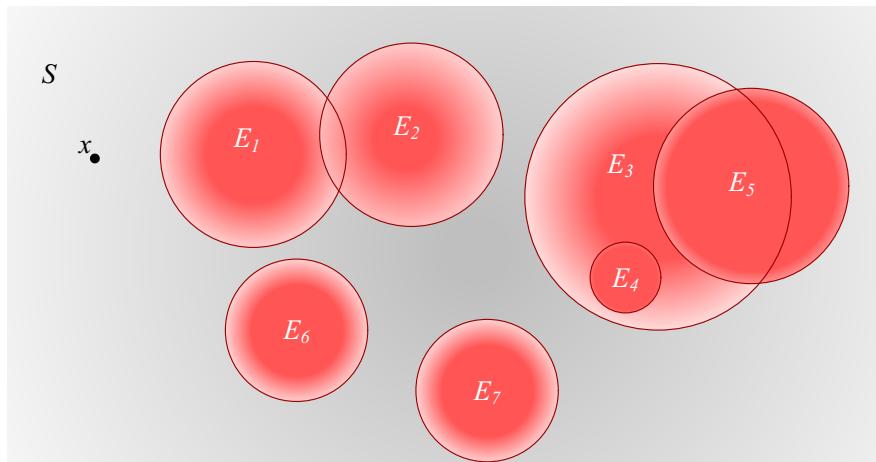


Figure 1: Venn diagram

Commutative property of intersection: $E_1E_2 = E_2E_1$.

Associative property of union: $E_1 \cup (E_2 \cup E_3) = (E_1 \cup E_2) \cup E_3 = E_1 \cup E_2 \cup E_3$.

Associative property of intersection: $E_1(E_2E_3) = (E_1E_2)E_3 = E_1E_2E_3$.

Distributive property: $E_1(E_2 \cup E_3) = (E_1E_2) \cup (E_1E_3) = E_1E_2 \cup E_1E_3$.

Observe that intersection takes precedence over union. It follows, that in the application of the distributive property the intersection operations must be performed prior to the union operations.

1.3 Special Events

Certain event: the event that contains all possible sample points of the sample space. Then, the sample space S is the certain event.

Null Event: the event that contains no sample points. Notation \emptyset .

Mutually exclusive events: the events E_1 and E_2 are mutually exclusive when they have no common sample points. Then, $E_1E_2 = \emptyset$.

Collectively exhaustive events: the events E_1, E_2, \dots, E_N are collectively exhaustive when their union spans the entire sample space. Then, $E_1 \cup E_2 \cup \dots \cup E_N = S$.

Complementary events: the complement of the event E is the event \bar{E} contains all the sample space that are not in the event E . Then, $E\bar{E} = \emptyset$ and $E \cup \bar{E} = S$, i.e. E and \bar{E} are mutually exclusive and collectively exhaustive.

1.4 De Morgan's rules

Given a series of events E_1, E_2, \dots, E_N , it is easy to prove the following rules:

$$\overline{E_1 \cup E_2 \cup \dots \cup E_N} = \bar{E}_1 \bar{E}_2 \dots \bar{E}_N, \text{ i.e. } \overline{\bigcup_{n=1}^N E_n} = \bigcap_{n=1}^N \bar{E}_n, \quad (1)$$

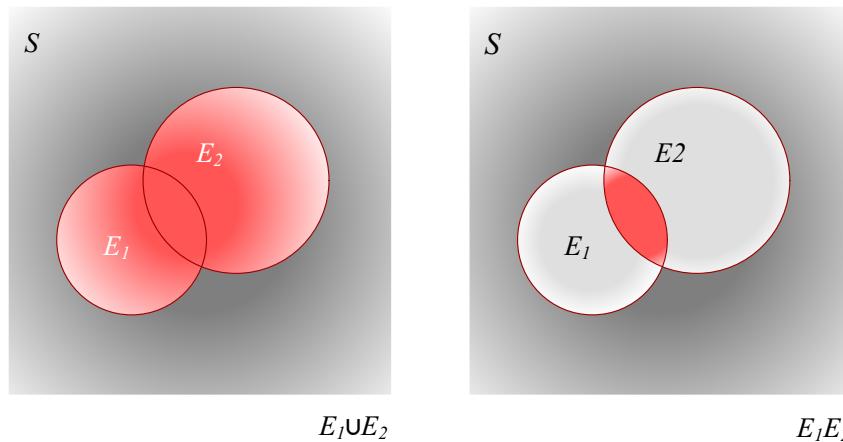


Figure 2: (i) Union and (ii) Intersection Events

and

$$\overline{E_1 E_2 \dots E_N} = \bar{E}_1 \cup \bar{E}_2 \cup \dots \cup \bar{E}_N, \text{ i.e. } \overline{\bigcap_{n=1}^N E_n} = \bigcup_{n=1}^N \bar{E}_n. \quad (2)$$

2 Element of Probability theory

The probability of an event E in a sample space S , is a measure, a weight, of the likelihood of occurrence of the event relative to other events in S . Notation: $P(E|S)$ or $P(E)$. Probability theory is based on the following Kolmogorov ¹ three axioms:

- I. $0 \leq P(E) \leq 1$.
- II. $P(S) = 1$.
- III. $P(\bigcup_{n=1}^N E_n) = \sum_{n=1}^N P(E_n)$, for mutually exclusive E_1, E_2, \dots, E_N .

The following results are derived based on the above axioms:

- i. $P(\bar{E}) = 1 - P(E)$.
- ii. $P(\emptyset) = 0$.
- iii. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2)$.

Common pitfalls: when approaching for the first time probability theory, a common mistake among students is to confuse the null event, \emptyset , with the real number 0. Observe that 0 can be a sample point, and an event, it follow that $0 \neq \emptyset$, and $P(\emptyset) \neq P(0)$! Moreover, observe that $P(E) = 0 \not\Rightarrow E = \emptyset$.

The last rule can be easily generalized for N events, i.e.

$$P(E_1 \cup E_2 \cup \dots \cup E_N) = \sum_{n=1}^N P(E_n) - \sum_{m=1}^N \sum_{n>m}^N P(E_n E_m) + \sum_{l=1}^N \sum_{m>l}^N \sum_{n>m}^N P(E_n E_m E_l) - \dots + (-1)^{N-1} P(E_1 E_2 \dots E_N). \quad (3)$$

Moreover based on (1) we can write

$$P(E_1 \cup E_2 \cup \dots \cup E_N) = 1 - P(\overline{E_1 \cup E_2 \cup \dots \cup E_N}) = 1 - P(\bar{E}_1 \bar{E}_2 \dots \bar{E}_N). \quad (4)$$

2.1 Conditional Probability

Conditional Probability is a simple, yet very profound, concept in Probability theory. We actually argue that this is the most important concept for this class. In fact, it sets the foundation for a good understanding of all the lecture notes, and the basis for a gently introduction to Bayesian statistics. Given two events E_1 and E_2 , the conditional probability of E_1 given E_2 , denoted with $P(E_1|E_2)$, defines the the probability of observing sample points of E_1 , given that

¹Andrey Nikolaevich Kolmogorov (*1903 †1987) was a Russian and Soviet mathematician who is the father of modern Probability Theory. His contributions spans from abstract mathematics to physics. It is considered one of the greatest minds of the 20th century

the event E_2 has occurred. This is a way of redefining the sample space, since by assuming E_2 has occurred the possible sample points are now confined within E_2 . By definition, the conditional probability is defined as

$$\begin{aligned} P(E_2|E_1) &= \frac{P(E_1 E_2)}{P(E_1)}, \text{ if } P(E_1 > 0) \\ &= 0, \text{ if } P(E_1 = 0), \end{aligned} \quad (5)$$

or rearranging the last equation as:

$$P(E_1 E_2) = P(E_2|E_1)P(E_1), \quad (6)$$

or by using the commutative property as

$$P(E_1 E_2) = P(E_1|E_2)P(E_2), \quad (7)$$

For a set of events E_1, E_2, \dots, E_N the probability of intersection can be written in different way, depending on the order of the conditioning, e.g.

$$P(E_1 E_2 \dots E_N) = P(E_N|E_1 \dots E_{N-1})P(E_{N-1}|E_1 \dots E_{N-2}) \dots P(E_2|E_1)P(E_1), \quad (8)$$

$$= P(E_1|E_2 \dots E_N)P(E_2|E_3 \dots E_N) \dots P(E_{N-1}|E_N)P(E_N), \quad (9)$$

In contrast to the conditional probability $P(E_2|E_1)$, the unconditional probability $P(E_1)$ is named marginal probability.

2.2 Statistical Independence

Two events E_1 and E_2 are statistically independent, $E_1 \perp\!\!\!\perp E_2$, if

$$P(E_2|E_1) = P(E_2), \quad (10)$$

i.e., the occurrence or the knowledge of E_2 does not affect the probability of occurrence of E_1 . It follows that for two statistically independent events

$$P(E_1 E_2) = P(E_1)P(E_2). \quad (11)$$

Common pitfalls: often students confuse independence with mutually exclusive. The two notions are mathematically and conceptually very different. Mutually exclusive relates to share sample points between two events, while statically independence relates to the probability properties expressed by (10). It is easy to show that $P(E_1 E_2) = P(E_1)P(E_2) \iff E_1 \perp\!\!\!\perp E_2$. However, for a set of events E_1, E_2, \dots, E_N , statistically independence requires that

$$P\left(E_n \mid \bigcap_{n \neq m} E_m\right) = P(E_n), \quad (12)$$

i.e., the conditional probability of E_n given any set of remaining events must be equal to the unconditional probability of E_n . Equivalently, the events are statistically independent if for any selections of indices the joint probability of the events is equal to the product of their marginal probabilities. For example for three events E_1, E_2, E_3 , statistically independence, $E_1 \perp\!\!\!\perp E_2 \perp\!\!\!\perp E_3 \Rightarrow$

$$\begin{aligned} P(E_1 E_2 E_3) &= P(E_1)P(E_2)P(E_3), \\ P(E_1 E_2) &= P(E_1)P(E_2), \\ P(E_2 E_3) &= P(E_2)P(E_3), \\ P(E_3 E_1) &= P(E_3)P(E_1), \end{aligned} \quad (13)$$

however, $P(E_1 E_2 E_3) = P(E_1)P(E_2)P(E_3)$ alone $\not\Rightarrow E_1 \perp\!\!\!\perp E_2 \perp\!\!\!\perp E_3$.

2.3 Theorem of Total Probability

Consider a set of mutually exclusive and collectively exhaustive events E_1, E_2, \dots, E_N , i.e. $E_n E_m = \emptyset$ and $\bigcup_{n=1}^N E_n = S$. Then, consider an event A , according to the theorem of total probability,

$$P(A) = \sum_{n=1}^N P(A|E_n)P(E_n). \quad (14)$$

The proof is given by taking in follow

$$\begin{aligned} P(A) &= P(AE_1 \cup AE_2 \cup \dots \cup AE_N) \\ &= \sum_{n=1}^N P(AE_n) = \sum_{n=1}^N P(A|E_n)P(E_n) \end{aligned} \quad (15)$$

Figure 4 shows a Venn diagram representation of the events. The theorem of total probability is pivotal in risk assessment, since often it is easier to compute conditional probabilities than marginal probabilities.

2.4 Bayes' Theorem

Consider two events A and B . Then, considering the commutative property $P(AB) = P(A|B)P(B) = P(B|A)P(A)$ and re-arranging terms we can write

$$P(B|A) = \frac{P(A|B)}{P(A)} P(B). \quad (16)$$

The last equation is known as Bayes's² theorem. This theorem is at the base of Bayesian statistics. Its significance lies in the fact that the probability of event A appears in the unconditional form $P(A)$ on the right, and in its conditional form $P(B|A)$ on the left side. It follows,

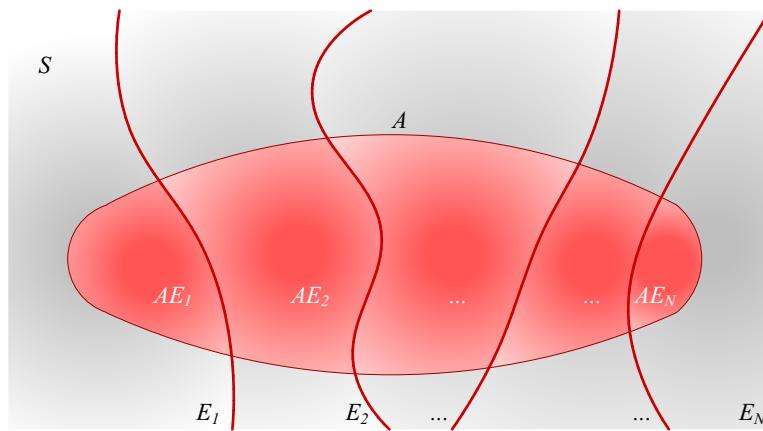


Figure 3: Representation of total probability

²Thomas Bayes (*1701 †1761) was a British Presbyterian minister, statistician, and philosopher. It is curious to know that Mr Bayes never published his work on Bayes' theorem. His notes were revised and published by the Welsh philosopher and mathematician Richard Price after Bayes death. However, it must be said that the great French mathematician Pierre-Simon Laplace (*1749 †1827) pioneered and popularised the modern Bayesian probability.

that Bayes' Theorem sets the foundation for updating the probability of an event A when the observation of event B is made.

Given a set of mutually exclusive and collectively exhaustive events E_1, E_2, \dots, E_N , we can write

$$P(E_n|A) = \frac{P(A|E_n)}{P(A)} P(E_n), \quad (17)$$

and by using the total probability theorem (23)

$$P(E_n|A) = \frac{P(A|E_n)}{\sum_{n=1}^N P(A|E_n)P(E_n)} P(E_n). \quad (18)$$

All rules of probability apply to conditional probabilities, provided a proper resize of the sample space is made. For example:

$$P(E_1 \cup E_2|A) = P(E_1|A) + P(E_2|A) - P(E_1E_2|A) \quad (19)$$

$$P(E_1E_2|A) = P(E_1|E_2A)P(E_2|A) = P(E_2|E_1A)P(E_1|A), \quad (20)$$

and given a set of mutually exclusive and collectively exhaustive events E_1, E_2, \dots, E_N , we can rewrite (23)

$$P(B|A) = \sum_{n=1}^N P(B|E_nA)P(E_n|A). \quad (21)$$

Example I

Suppose that we would like to know the probability of failing of the Bay Bridge, Figure 4, which connects the city of Oakland with the city of San Francisco. The Bay bridge is a series system composed of two spans. The east span connects the city of Oakland with Yerba Buena island, and the west span connects Yerba Buena island with the city of San Francisco. The sample space of the intensity measures of the earthquake are defined as $S_{im} = \{IV, VI, VIII, X\}$, where IV is a low intensity earthquake, VI a medium intensity earthquake, $VIII$ is a strong earthquake, and X is a devastating earthquake. Seismologists and earthquake engineers compute the probability of occurrence of such events. For this case, the values³ are reported in Table 1.

Let's denote with F_S the system failure event, F_W is the west span failure event, and F_E is the east span failure event. Structural engineers compute the probability of failure of structures conditional to a level of hazard, i.e. $F_W|IM$, and $F_E|IM$. These values³ are reported in Table 2.

The Bay Bridge system fails (i.e. F_S) if the east span fails (i.e. F_E) OR the west span fails (i.e. F_W). Then, we can write this event as

$$P(F_S) = P(F_W \cup F_E) = P(F_W) + P(F_E) - P(F_WF_E). \quad (22)$$

We can compute $P(F_W)$, $P(F_E)$, and $P(F_EF_W)$, by the total probability theorem. In fact, the earthquake intensities represent a collection of mutually exclusive and collectively exhaustive events. Then, for example, we can write $P(F_W)$ as

$$P(F_W) = \sum_{n=1}^N P(F_W|IM_n)P(IM_n). \quad (23)$$

In the same way, we can use the total probability theory to compute F_E and F_EF_W .

³Numbers reported here are purely academic.

Problems

- i Compute $P(F_W)$, $P(F_E)$, and $P(F_E F_W)$, and $P(F_S)$.
- ii Are the events $F_W | IM$ and $F_E | IM$ statistically independent?
- iii Are the events F_W and F_E statistically independent? Which conclusions can you draw?

Suppose that we know that the Bay Bridge system has failed; however, we do not know which of the two spans has failed. Given this information we would like to update the probability that the west span has failed. Then, we can use the Bayes' theorem as

$$P(F_W | F_W \cup F_E) = \frac{P(F_W \cup F_E | F_W)}{P(F_W \cup F_E)} P(F_W), \quad (24)$$

it should be clear that $P(F_W \cup F_E | F_W) = 1$ since if we know that the west span has failed the system is surely failed. Then we can write

$$P(F_W | F_W \cup F_E) = \frac{1}{P(F_W \cup F_E)} P(F_W), \quad (25)$$

and finally solve the problem with the values from the previous problems.

Problems

- i Compute $P(F_W | F_W \cup F_E)$, $P(F_E | F_W \cup F_E)$. Which conclusions can you draw?
- ii What is the probability that the west component ALONE was the cause of failure?
- iii Given that the Bay Bridge system failed, what is the probability that $IM = VIII$ (i.e. $P(IM = VIII | F_W \cup F_E)$)?
- iv Given that the Bay Bridge system failed, what is the most likely earthquake intensity? Is it an “intuitive” result? Which conclusions can you draw?



Figure 4: Bay Bridge. Source: Wikipedia.

$P(IM = IV)$	$P(IM = VI)$	$P(IM = VIII)$	$P(IM = X)$
0.8	0.15	0.045	0.005

Table 1: IM probabilities

IM	$P(F_W IM = im)$	$P(F_E IM = im)$	$P(F_W F_E IM = im)$
IV	$1.00E-4$	$1.00E-4$	$1.00E-8$
VI	$5.00E-3$	$1.00E-3$	$5.00E-6$
$VIII$	$8.00E-2$	$5.00E-3$	$4.00E-4$
X	$2.00E-1$	$5.00E-2$	$1.00E-2$

Table 2: Conditional probabilities of failure



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



LECTURE NOTES

Fundamentals of Probability and Statistics

Chapter 1: Lecture 2

Spring 2018

Written by marco broccardo
IBK, ETHZ

3 Random Variables

A random variable is a *numerical representation* of a random phenomenon. In particular, the random variable is defined as a one-to-one mapping of its sample space on the real line, Figure 5. Random variables are convenient in many engineering applications since outcomes of engineering problems are often given in numerical values (e.g. magnitude of an earthquake, peak-ground acceleration, yielding point, etc.). Random variables can also have a pure conventional mapping. For example, the sample space of the state of a building after a seismic event can be defined as follow $S \in \{D_0, D_1, D_2, D_3, C\}$, where D_0 is the no damage state, D_1 is light damage state, D_2 the moderate damage state, D_3 the heavy damage state, and C the collapse state. A possible description of this random phenomenon in terms of a random variable X is: $X = 0$ for D_0 , $X = 1$ for D_1 , $X = 2$ for D_2 , $X = 3$ for D_3 , and $X = 4$ for C . Observe that this mapping is totally arbitrary and it can be defined differently by a different analyst. **Notation:** we use UPPERCASE letter to denote a random variable, e.g. X . We use italic lowercase to denote the sample points (outcomes), e.g. $x_1, x_2, \dots, x_n, \dots$ are the possible outcomes of X . In this section, there are two exceptions to this rule, that are the letters N and M . These are used to indicate the upper bound of a sequence of natural numbers characterized by the indexes m and n .

3.1 Probability Distributions

3.1.1 Discrete Random Variables

Given a discrete random variable X with possible outcomes $x_1, x_2, \dots, x_n, \dots$, the probability mass function, PMF, is defined as

$$p_X(x_n) = P(X = x_n). \quad (26)$$

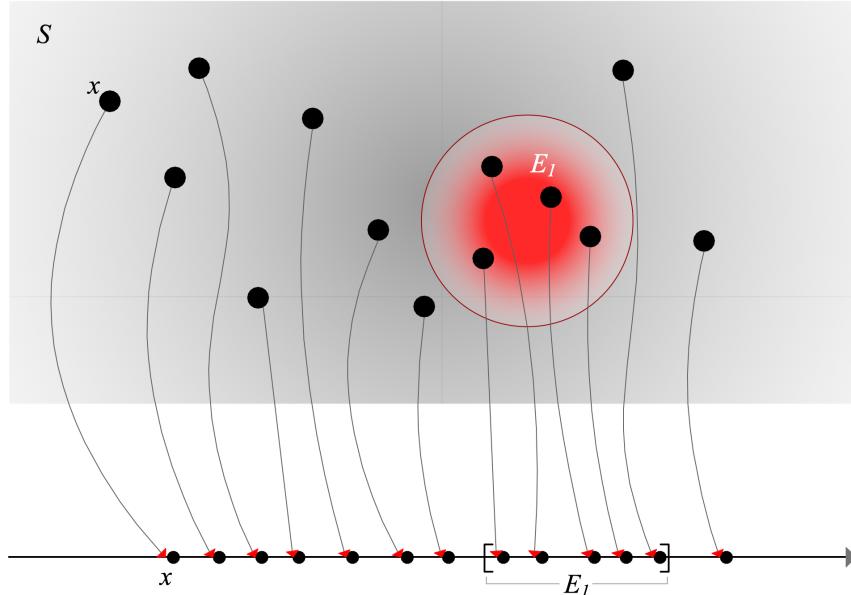


Figure 5: Random variable, mapping.

Observe that $p_X(x) = 0$ for $x \neq x_n$ and $p_X(x_n) = P(X = x_n)$ for any n . Moreover, $x_1, x_2, \dots, x_n, \dots$ is a set of mutually exclusive and collectively exhaustive events; then,

$$\sum_n p_X(X = x_n) = 1, \quad (27)$$

and $0 \leq p_X(x) \leq 1$. Given the PMF, the probability of any event can be determined by rules of probability theory. For example, the probability that X belongs to the interval $(a, b]$ is given by

$$P(a < X \leq b) = \sum_{a < x_n \leq b} p_X(x_n). \quad (28)$$

The probability distribution of a random variable can be expressed also in terms of the cumulative distribution function, CDF, defined as

$$F_X(x) = \sum_{x'_n \leq x} p_X(x'_n). \quad (29)$$

Example II

Suppose that you design a block of six identical buildings in a seismic area. Given an earthquake ground motion the probability of damage of one building is denoted with p . Since the buildings are close, it can be assumed that they will experience the same ground motion. It is also assumed that given the ground motion, the damaging events are statistically independent, i.e. if we know the ground motion, knowledge of the state of one building does not affect the probability of damage of another one. Observe that this is equivalent of assuming that the structural uncertainties are statistically independent.

We are interested in determining the probability distribution of the number of damage buildings. Let us define a random variable, X , representing this random phenomenon with possible outcomes $0, 1, 2, \dots, 6$. A possible way to have x damage buildings out of 6 is to consider the first x buildings to be damaged and the remaining $(6 - x)$ not damaged. Given the independence assumption, this is equivalent to write $p^x \times (1 - p)^{6-x}$. However, this combination is not the only possible one, and there are other ways of selecting x item out of n . In particular, the total number of combinations is given by the binomial factor

$$\binom{6}{x} = \frac{6!}{x!(6-x)!}; \quad (30)$$

then, we can write

$$p(x) = \binom{6}{x} p^x (1-p)^{(6-x)}, \text{ for } x = 0, 1, 2, \dots, 6. \quad (31)$$

The (31) is known as the binomial distribution. Later in this chapter, we will examine this distribution in greater details. Verify that $\sum_{x=0}^6 p(x) = 1$.

Problems-III

- i Compute the PMF of X given $p = 0.15$, what is the most likely number of buildings that are damaged?
- ii What is the probability that two or more buildings are damaged?

- iii Suppose you are the owner of the first building and you are in vacation. Suppose that you receive a call saying that there was an earthquake and at least 2 buildings are damaged. Given this information, what is the probability that your building is damaged?

3.1.2 Continuous Random Variables

A continuous random variable X has an infinite number of possible outcomes. It follows that we can not list out the probability associated with a specific sample point x , since x is one of the infinite number of sample points. To overcome this issue, we characterize probability within an interval. Then, we define the probability density function (PDF), $f_X(x)$, as a non-negative function such that

$$f_X(x) = \frac{P(x < X \leq x + dx)}{dx}, \quad (32)$$

where dx is a differential element of infinitesimal length.

Common pitfalls: given a sample point x , often students confuse the probability with the probability DENSITY of x . Notice the difference between the two. The PDF gives the probability per unit length! Moreover, observe that $f_X(x) \geq 0$ does not have an upper bound, so it is possible that $f(x) > 1$ for some x .

Given the definition (32), the probability that X is within the interval $(l, u]$ is written as

$$P(l < X \leq u) = \int_l^u f_X(x) dx. \quad (33)$$

Moreover, $f_X(x)$ must satisfy

$$\int_{-\infty}^{\infty} f_X(x) dx = 1. \quad (34)$$

An alternative description for the probability distribution is the cumulative distribution function (CDF), $F_X(x)$, defined as

$$F_X(x) = \int_{-\infty}^x f_X(x') dx'. \quad (35)$$

The CDF is related to the PDF by the following relationship

$$f_X(x) = \frac{dF(x)}{dx}; \quad (36)$$

moreover, the CDF must satisfy the following relationships: $F_X(-\infty) = 0$, $F_X(\infty) = 1$ and $F_X(l) \leq F_X(u)$ for $l \leq u$.

Example III

Suppose that you design an important structure which is located 10 km from the Hayward fault⁴, Figure 6. We would like to derive the PDF and the CDF of the epicentral distance r , from the design site to the epicenter of an earthquake. r can be defined as random variable, R , because

⁴click these links to see simulations of earthquakes occurring in the Hayward fault for different epicenters: [San Pablo Bay epicenter](#), [Oakland epicenter](#), [Fremont epicenter](#).

the epicenter occurs randomly along the fault. In this example, it is easier to define first the CDF, $F(r)$, and then derive the PDF with (36). Given that all the locations along the fault are equally likely to be an epicenter we can write

$$F_R(r) = P(R \leq r) = \frac{\text{length of segment with distance} \leq r}{\text{length of the fault}}. \quad (37)$$

Given the geometry of the site (Figure 6), the numerator can be written as $2\sqrt{r^2 - 10^2}$, and $10 \leq r \leq \sqrt{40^2 + 10^2} = 41.23$. Thus,

$$\begin{aligned} F_R(r) &= 0, \text{ for } r < 10 \\ &= \frac{2\sqrt{r^2 - 10^2}}{80}, \text{ for } 10 \leq r \leq 41.23 \\ &= 1 \text{ for } r > 41.23. \end{aligned} \quad (38)$$

The PDF is obtained by differentiation of the (38) as

$$\begin{aligned} f_R(r) &= 0, \text{ for } r < 10 \\ &= \frac{2r}{80\sqrt{r^2 - 10^2}}, \text{ for } 10 \leq r \leq 41.23 \\ &= 0 \text{ for } r > 41.23. \end{aligned} \quad (39)$$

Figure 7 shows the CDF and PDF of R .

Problems-IV

- i Which conclusions can you draw from this PDF? Where the earthquake is more likely to occur ? (Give just a rough description, e.g. far from the site, near the site, etc.)
- ii Now suppose that you do not know the exact location of the building, can you write a parametric code that, given a location, automatically give you the PDF and the CDF of R ? (Assume the site is within the extension of the fault).

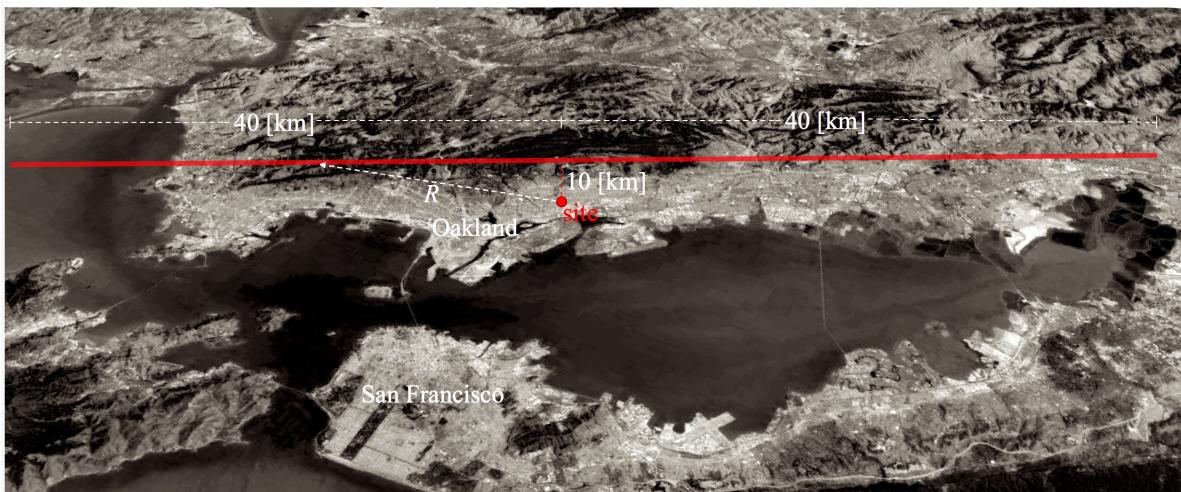


Figure 6: Hayward fault. Source: Google earth.

- iii Suppose that your site is located in Sion, Switzerland. In this case there is not a clear fault geometry, so the epicenter can occur anywhere randomly within 80 [km] from the site. Assume that all outcome points have equal likelihood. What is the CDF and PDF of R ? Make a plot of the functions.
- iv Compare this last PDF with (39), which conclusions can you draw?

3.1.3 Mixed Random variables

In some engineering application we deal with random variables that are of mixed type. In particular, these can take both finite probabilities on specific sample points and probability densities function in one or more intervals. Such variables can be encompassed within the continuous random variable probability density functions by using the Dirac⁵ delta function, $\delta(x)$, here defined as

$$\delta(x - a) = \lim_{\sigma \rightarrow 0} \frac{1}{\sigma \sqrt{\pi}} e^{-(x-a)^2/\sigma^2}. \quad (40)$$

$$(41)$$

Given the above definition it is easy to verify that

$$\int_{-\infty}^{\infty} \delta(x - a) dx = 1. \quad (42)$$

A little more thoughts (which are however beyond the purpose of this class) are required to understand the following equation

$$f(a) = \int_{-\infty}^{\infty} f(x) \delta(x - a) dx. \quad (43)$$

Roughy speaking the above equation essentially define a spike of unit area at a specific point a .

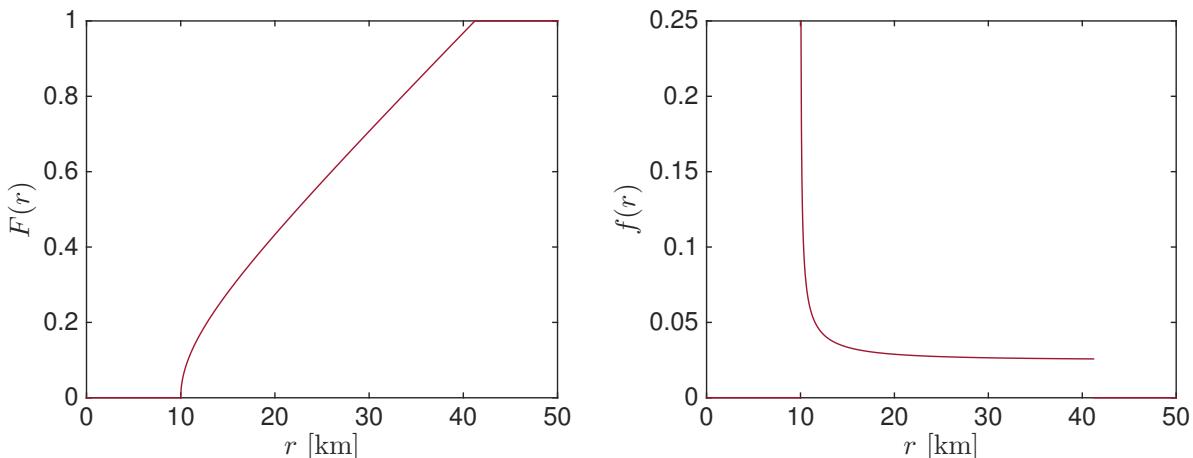


Figure 7: CDF and PDF of the epicentral distance for the given site

⁵Paul Adrien Maurice Dirac (*1902 †1984) was an English physicist who made fundamental contributions in quantum mechanics and electrodynamics. He is regarded as one of the most important physicist of the 20th century. Observe that the history of delta function is precedent than Dirac, and its origin can be found in the work of Jean-Baptiste Joseph Fourier (*1749 †1827). In particular in the famous treatise *Théorie analytique de la chaleur*, where for the first time it was presented what we know now as Fourier integral theorem.

The generalized PDF of a mixed random variable X , with probability mass, $p(x_n)$ at discrete points x_1, x_2, \dots, x_N and a probability density function $\bar{f}(x)$ for $x \in (-\infty, \infty)$, can be written as

$$f_X(x) = \bar{f}_X(x) + \sum_{n=1}^N p_{X_n}(x_n) \delta(x - x_n) \quad (44)$$

Example IV

Different researchers argue that the epicentral distance is not the “most appropriate” site-source distance to consider in seismic risk analysis. In fact, it would be more realistic to represent the earthquake as a finite rupture rather than a single point. In this case, a “more appropriate” site-source distance is given by the distance from the site to the nearest point on the finite rupture. In earthquake engineering, this distance is known as Joyner-Boore distance, r_{JB} . Let s denote the length of the rupture; moreover, assume that $s \leq l/2 = 40[\text{km}]$, where l is the total length of the fault. It follows that in case the rupture extends between San Pablo Bay and Oakland the r_{JB} is the distance from the site to the right end of the rupture. Conversely, if the rupture extends between Oakland and Fremont the r_{JB} is the distance from the site to the left end of the rupture. Moreover, if the rupture crosses the foot of the normal from the site, then $r_{JB} = d = 10[\text{km}]$ (where d is the perpendicular distance between the site to the fault). Since the location of the rupture is unknown, this is a random phenomenon. Therefore, it is convenient to define the random variable R_{JB} . Given the geometry of the problem (Figure 8, first subplot), the realizations r_{jb} of R_{JB} are belonging to the following interval $[d, \sqrt{d^2 + (l/2 - s)^2}]$ (Figure 8, third subplot).

A convenient way to solve this problem is to fix a reference point along the rupture s . Let us define this reference point as the right end of the rupture. It follows that the sample space of the reference point is the interval $[s, l]$ of length $l - s$ (Figure 8 second subplot). Within this interval the realizations of the event $\{R_{JB} \leq r_{JB}\}$ occurs when the reference point belong to the interval $[l/2 - \sqrt{r^2 - d^2}, l/2 + \sqrt{r^2 - d^2} + s]$ of length $2\sqrt{r^2 - d^2} + s$ (Figure 8 last subplot). Given the equal likelihood of the points within the two intervals $P(R_{JB} \leq r_{JB})$ is equal to the ratio of the length of the two intervals, i.e.

$$\begin{aligned} F_{R_{JB}}(r) &= 0, \text{ for } r < 10 \\ &= \frac{2\sqrt{r^2 - 10^2} + s}{l - s}, \text{ for } 10 \leq r \leq \sqrt{100 + (40 - s)^2} \\ &= 1 \text{ for } r > \sqrt{100 + (40 - s)^2}. \end{aligned} \quad (45)$$

Observe that while $F_{R_{JB}}(R = 10) = 0$ in Example III,

$$F_{R_{JB}}(R_{JB} = 10) = \frac{s}{l - s}, \quad (46)$$

which is the finite probability that the rupture passes through the foot of the normal from the site. This events occurs when the reference point fell in the interval $[l/2, l/2 + s]$. $F_{R_{JB}}(r_{JB})$ has a discontinuity at $R_{JB} = d$, which makes R_{JB} a mixed random variable with a probability mass $s/(l - s)$ at $r = d$ and probability density $f_{R_{JB}}(r_{JB}) = dF(r_{JB})/dr_{JB}$ in the interval $[d, \sqrt{d^2 + (l/2 - s)^2}]$. Owning Eq. (44)

$$f_{R_{JB}}(r_{JB}) = \frac{2r}{(80 - s)\sqrt{r^2 - 100}} + \delta(r_{JB} - d)\frac{s}{l - s}, \text{ for } 10 \leq r \leq \sqrt{100 + (40 - s)^2} \quad (47)$$

One can easy verify that for $s = 0$, Eq. (45) and Eq. (47) are equal to Eq. (38) and (39). Figure 9 show the CDF and PDF for different s lengths.

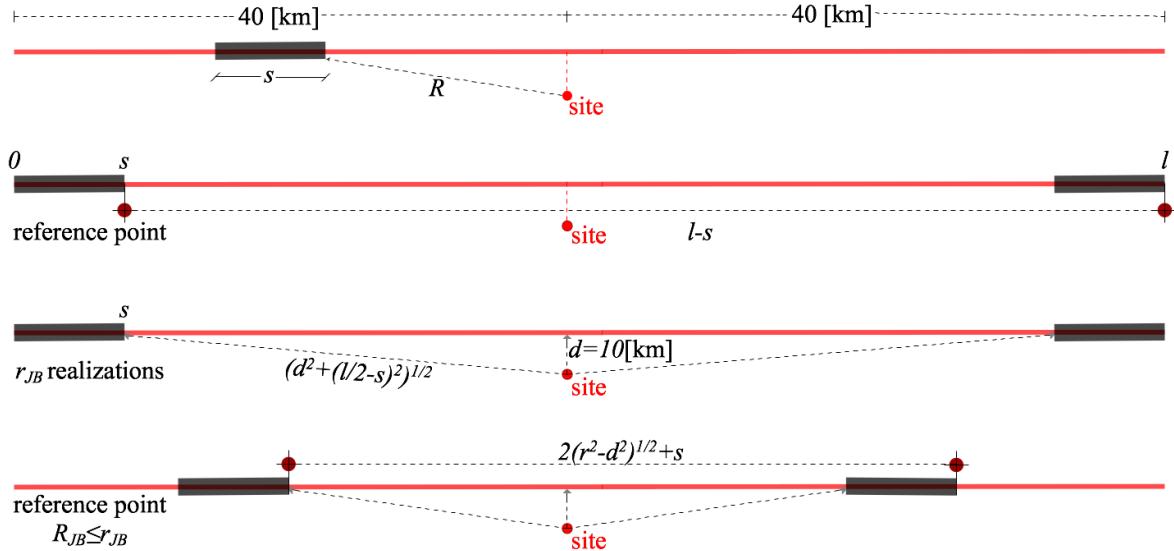


Figure 8: Hayward fault. Example IV

4 Multiple random variables

A lot of engineering problems deal with more than one random variable at time. For example, safety of structural components involve analysis of both load and a structural capacity, which are usually random variables. The fundamental question when we deal with problems involving multiple random variables is: if we know individually the PMF or PDF of single random variables, do we have complete information about the problem? The answer is in general NO! What we are missing is the statistical relationship between the random variables. This relationship is completely defined in terms of a joint probability distribution.

Given two random variables, the joint PDF is defined as

$$f_{XY}(x, y) = \frac{P(x < X \leq x + dx \cap y < Y \leq y + dy)}{dxdy}. \quad (48)$$

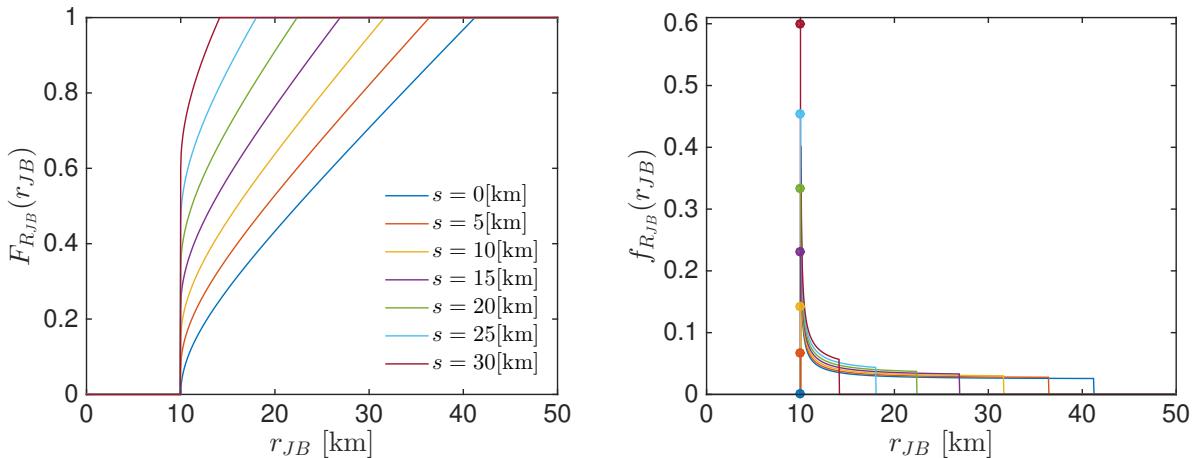


Figure 9: CDF and PDF of the Joyner-Boore distance for the given site. Mixed distribution, the stems at $R_{JB} = d$ indicate the discrete probability mass $P(R_{JB} = d) = s/(l - s)$.

Similarly to single random variables we can find the probability of falling in a given region, e.g. $x \in (l, u], y \in (\bar{l}, \bar{u}]$

$$P(l < X \leq u \cap \bar{l} < Y \leq \bar{u}) = \int_{\bar{l}}^{\bar{u}} \int_l^u f_{XY}(x, y) dx dy. \quad (49)$$

Moreover, $f_{XY}(x, y)$ must satisfy

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1. \quad (50)$$

Similarly to the single random variable the joint CDF can be written as

$$F_{XY}(x, y) = P(X \leq x \cap Y \leq y) = \int_{-\infty}^y \int_{-\infty}^x f_{XY}(x', y') dx' dy', \quad (51)$$

and it can be related to joint PDF by the following

$$f_{XY}(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}. \quad (52)$$

Moreover, we have $F_{XY}(x, -\infty) = F_{XY}(-\infty, y) = F_{XY}(-\infty, -\infty) = 0$, and $F_{XY}(\infty, \infty) = 1$.

Given the joint distribution of X and Y , one can obtain the the PDF of X , or Y , alone as follow

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy, \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx. \quad (53)$$

These operations are named marginalization, and $f_X(x)$ and $f_Y(y)$ are called the marginal distributions of X and Y . Figure 10-a shows this operation.

The above expressions can be extended to any number of random variables. For the sake of simplicity when dealing with multiple random variables we use the vectorial notations $\mathbf{X} = [X_1, \dots, X_N]^T$ (i.e. a random vector) and $\mathbf{x} = [x_1, \dots, x_N]^T$ (i.e a sample point). Given this notation, the joint PDF and CDF are written as $f_{\mathbf{X}}(\mathbf{x})$ and $F_{\mathbf{X}}(\mathbf{x})$. Then, the properties described for the bivariate PDF and CDF can be extended for the multivariate distribution function. In particular,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = 1; \quad (54)$$

moreover, given a vector partition, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$, where $\mathbf{X}_1 = [X_1, \dots, X_M]^T$ and $\mathbf{X}_2 = [X_{M+1}, \dots, X_N]^T$ we can obtain the PDF of \mathbf{X}_1 , or \mathbf{X}_2 , alone as follow:

$$f_{\mathbf{X}_1}(\mathbf{x}_1) = \int_{-\infty}^{\infty} \cdots \int_{N-M+1}^{\infty} f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2, \text{ and } f(\mathbf{x}_2) = \int_{-\infty}^{\infty} \cdots \int_M^{\infty} f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1, \quad (55)$$

4.1 Conditional Distribution and Statistical Independence

For two continuous random variables, X and Y , we define the conditional distribution of X given Y as

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{P(x < X \leq x + dx | y < Y \leq y + dy)}{dx} \\ &= \frac{P(x < X \leq x + dx \cap y < Y \leq y + dy)}{P(y < Y \leq y + dy)} \frac{1}{dx} \\ &= \frac{f(x, y)}{f(y)}. \end{aligned} \quad (56)$$

Figure 10-b shows this concept. One can observe that the joint PDF $f_{XY}(x, y)$ can be expressed as a product of one conditional and one marginal PDF, i.e.

$$f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x). \quad (57)$$

Two random variables X and Y are statistically independent if

$$f_{X|Y}(x|y) = f_X(x), \text{ or } f_{Y|X}(y|x) = f_Y(y). \quad (58)$$

It follows that for two statistically independent random variables knowledge of the marginals is sufficient to describe the joint probability distribution, i.e. $f_{XY}(x, y) = f_X(x)f_Y(y)$. Conditional distribution for multiple random variables are defined by generalizing the concepts introduced in the previous Section, e.g. the conditional distribution of \mathbf{X}_2 given $\mathbf{X}_1 = \mathbf{x}_1$ is

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2|\mathbf{x}_1) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}_1}(\mathbf{x}_1)}, \quad (59)$$

which leads to the following rule

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2|\mathbf{x}_1)f_{\mathbf{X}_1}(\mathbf{x}_1), \quad (60)$$

and more in general

$$f_{\mathbf{X}}(\mathbf{x}) = f(\mathbf{x}_N|\mathbf{x}_1, \dots, \mathbf{x}_{N-1})f(\mathbf{x}_{N-1}|x_1, \dots, x_{N-2}) \dots f(x_2|x_1)f(x_1), \quad (61)$$

The set of random variables \mathbf{X} are jointTLY statistically independent if

$$f_{\mathbf{X}}(\mathbf{x}) = f(x_1)f(x_2)\dots f(x_N), \text{ for } x_1, x_2, \dots, x_N. \quad (62)$$

Please note that in (61) and (62) for clarity of notation we drop the conventional subscript $\cdot_{\mathbf{X}}$. The reader should consider it implicit.

Common pitfalls: it is not uncommon to relate (wrongly) joint statistically independence to pairwise independence. It is important to note that pairwise independence DOES NOT guarantee independence as specified by the above condition.

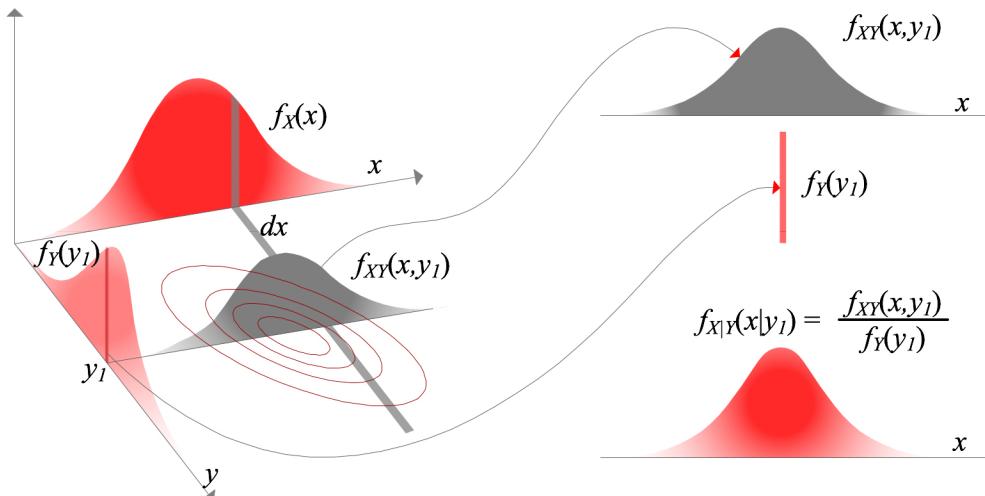


Figure 10: CDF and PDF of the epicentral distance for the given site

Example V

The Californian Memorial Stadium, Figure 11-a, is the stadium of the University of California, Berkeley football team. The stadium is a US historic landmark, opened in 1923, which currently seats circa 63,000 people. The Hayward fault crosses the infrastructure, Figure 11-b. The stadium has been recently retrofitted, by splitting it into two components to allow a relative permanent displacement in case of an earthquake. In the following Example, we compute the probability that the stadium will have a relative permanent displacement. We assume that there will be such displacement in the case the rupture will cross (even partially) the Stadium. We denote this event as E_s . Moreover, we denote with l_1 the distance from the left end of the Hayward fault to the left end of the Memorial Stadium, with $l_2 = l - l_1$, with w the width of the stadium, and with s the unknown rupture length. Given this, we consider the rupture length as a random phenomenon, and therefore we define the random variable S . Following the same geometrical considerations of Example -III and -IV, we can first define the conditional probability distribution for a given length s , $P(E_s|S = s)$, and then using the total probability theorem to compute $P(E_s)$ as

$$P(E_s) = \int_s P(E_s|S = s)f_S(s)ds.$$

In this Example V, we compute the conditional probability density $P(E|S = s)$; then, in Chapter 3 we will define $f_S(s)$ and compute $P(E_s)$. Following Figure (12), we can identify three cases: Case-I:

$$P(E|S = s) = \frac{w + s}{l_1 + l_2 - s}, \quad 0 < s \leq l_1,$$

Case-II:

$$P(E|S = s) = \frac{w + l_1}{l_1 + l_2 - s}, \quad l_1 < s \leq l_2 - w,$$

Case-III:

$$P(E|S = s) = 1, \quad s > l_2 - w.$$

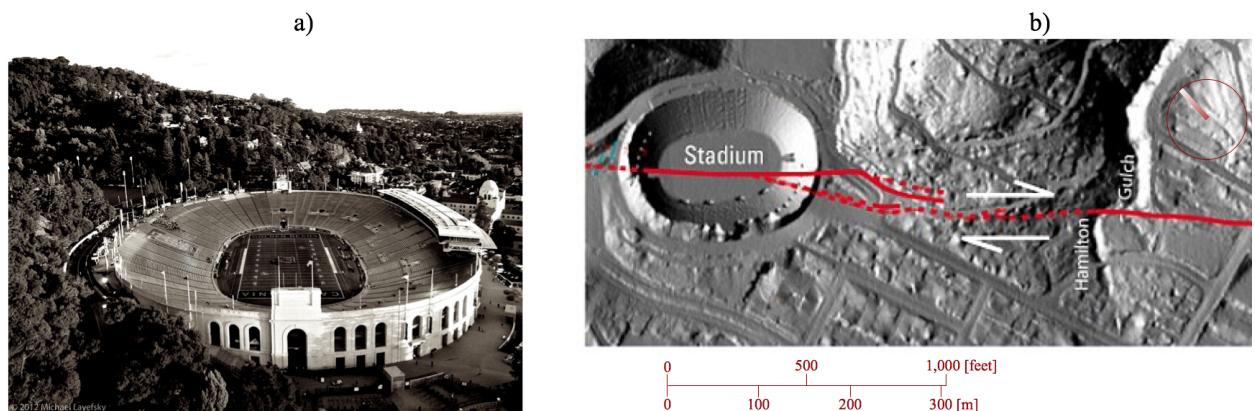


Figure 11: a- Memorial Stadium, Source: flickr, Michael Layefsky. b- Hayward fault crossing the stadium; Source: USGS

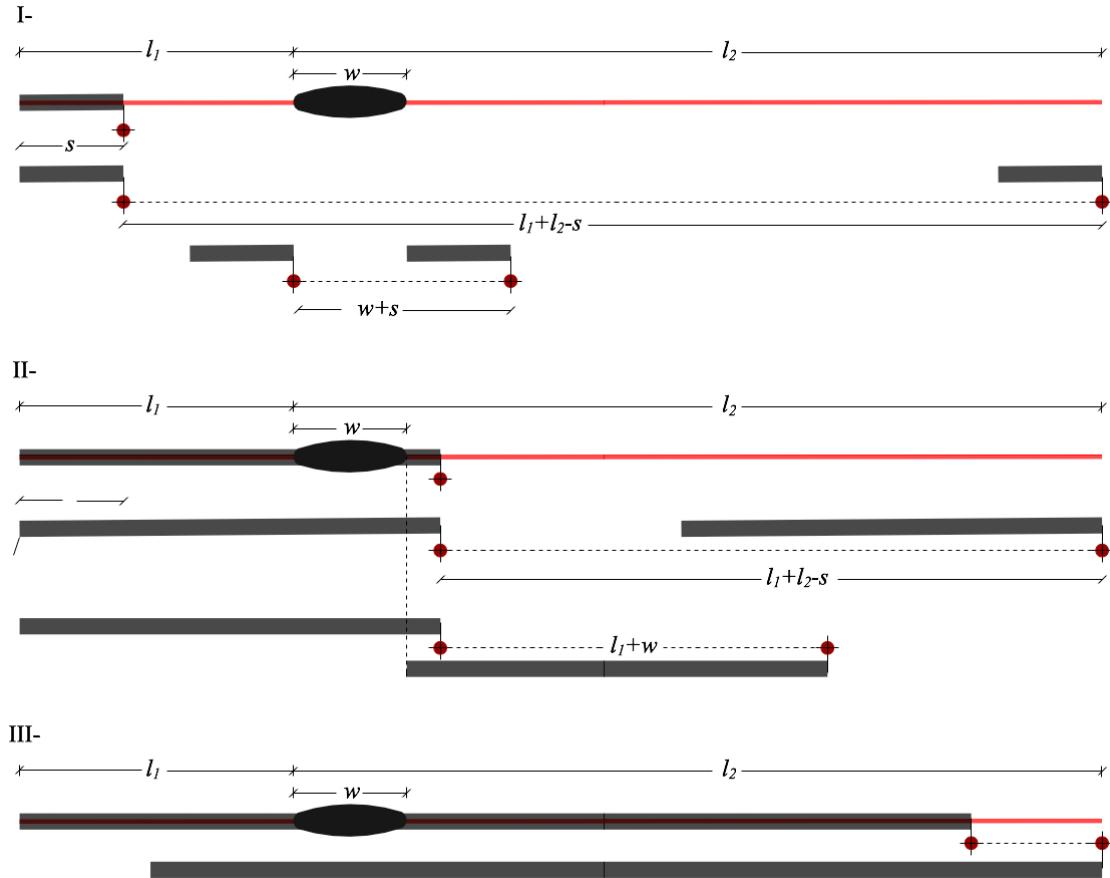


Figure 12: Example 5, case I- $0 < s \leq l_1$, case II- $l_1 < s \leq l_2 - w$, case III- $l_2 - w < s$

5 Partial Description of Random Variables, Moments and Expectation

Given a random variable, X , the complete information about this variable is given by its PDF. However, oftentimes we would like to have summary values which partially describe X . The most common partial descriptions of a random variable are: mean, variance (or standard deviation), coefficient of variation, median and mode. The mean is defined as

$$\mu_x = \int_{-\infty}^{\infty} x f_X(x) dx. \quad (63)$$

The variance is defined as

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_X(x) dx, \quad (64)$$

and the standard deviation as $\sigma_x = \sqrt{\sigma_x^2}$. The coefficient of variation (c.o.v) is defined as

$$\delta_x = \frac{\sigma_x}{|\mu_x|}, \text{ for } \mu_x \neq 0. \quad (65)$$

The median of a distribution, \tilde{x} , is defined as the “middle” point of the distribution, i.e.

$$\int_{-\infty}^{\tilde{x}} f_X(x) dx = 0.5. \quad (66)$$

The mode of the distribution, is defined as the point/s (or the interval) with the highest probability density

$$\text{mode} = \arg \max_x f_X(x). \quad (67)$$

5.1 Expectation

Consider a function $g(X)$ of random variable X . The mathematical expectation of $g(X)$ is defined by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \quad (68)$$

Given a vector of random variables \mathbf{X} and a multivariate function $g(\mathbf{X})$ the mathematical expectation is simply

$$E[g(\mathbf{X})] = \int_{-\infty}^{\infty} \underset{N \text{ folds}}{g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})} d\mathbf{x}. \quad (69)$$

Observe that the integral operation is linear; thus, given two functions $g_1(\mathbf{X})$ and $g_2(\mathbf{X})$, one can write

$$E[g_1(\mathbf{X}) + g_2(\mathbf{X})] = E[g_1(\mathbf{X})] + E[g_2(\mathbf{X})] \quad (70)$$

5.2 Moments of Random Variable

Consider the function $g(X) = X^k$ for a random variable X , where $k \in \mathbb{N}$. Then, the expectation

$$E[X^k] = \int_{-\infty}^{\infty} x^k f_X(x) dx \quad (71)$$

is known also as the k -moment of X . In particular, the first moment $E[X] = \mu_x$ describes the barycenter of the distribution, the second moment $E[X^2]$ is a measure of the dispersion, and the third and fourth moments are useful in computing measures of skewness and flatness of the distribution.

Consider $g(\mathbf{X}) = (X - \mu_x)^k$; then, the expectation

$$E[X^k] = \int_{-\infty}^{\infty} (x - \mu_x)^k f_X(x) dx \quad (72)$$

is the k -th central moment of X . For $k = 2$ we have $E[(X - \mu_x)^2] = \text{Var}[X] = \sigma^2$ is the central moment of inertia of the distribution. One can use the linear property of expectation to write

$$E[(X - \mu_x)^2] = E[X^2 - 2\mu_x X + \mu_x^2] = E[X^2] - E^2[X] \quad (73)$$

Continuing our analogy with mechanics, we observe that (73) is the equivalent of the parallel axis theorem. Similarly, the third central moment can be written as

$$E[(X - \mu_x)^3] = E[X^3] - 3E[X^2]E[X] + 2E^3[X] \quad (74)$$

Observe that $(X - \mu_x)^3$ retains the sign of the deviation from the mean; it follows that the third moment provides a good measure of the non-symmetry of the distribution. Usually, this is expressed by the coefficient of skewness defined by

$$\gamma_1 = \frac{E[(X - \mu_x)^3]}{\sigma^3}. \quad (75)$$

One can observe that for $\gamma_1 = 0$ the distribution is symmetric, whereas $\gamma_1 > 0$ indicates skewness of the distribution to the right, and $\gamma_1 < 0$ indicates skewness to the left. Finally, the coefficient of kurtosis is written as

$$\gamma_2 = \frac{E[(X - \mu_x)^4]}{\sigma^4}. \quad (76)$$

Kurtosis derives from Greek *κυρτός* which means “curved”; in fact, it is a measure of “flatness” of the distribution shape.

Consider the function $g(X) = \exp(\omega X)$ of random variable X . The mathematical expectation

$$E[\exp(\omega X)] = \int_{-\infty}^{\infty} \exp(\omega x) f_X(x) dx = M(\omega), \quad (77)$$

is named moment generating function. Note that this is essentially the same as the definition of the Laplace transform of a function $f(x)$, except that we are using ω instead of $-\omega$. The (77) is a convenient expression since

$$\frac{d^k M(\omega)}{d\omega^k} = \int_{-\infty}^{\infty} x^k \exp(\omega x) f_X(x) dx, \quad (78)$$

and

$$\left. \frac{d^k M(\omega)}{d\omega^k} \right|_{\omega=0} = \int_{-\infty}^{\infty} x^k f_X(x) dx = E[X^k]. \quad (79)$$

Similar expression can be derived in case of a random vector \mathbf{X} . In this case given $\boldsymbol{\omega} = [\omega_1, \dots, \omega_N]$, one can write

$$M(\boldsymbol{\omega}) = E[\exp(\boldsymbol{\omega} \mathbf{X})] = \int_{-\infty}^{\infty} \exp(\boldsymbol{\omega} \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \quad (80)$$

and

$$\frac{\partial^{k_1+k_2+\dots+k_N}}{\partial \omega_1^{k_1} \partial \omega_2^{k_2} \dots \partial \omega_N^{k_N}} M(\boldsymbol{\omega}) = E[X_1^{k_1}, X_2^{k_2}, \dots, X_N^{k_N}]. \quad (81)$$

5.3 Joint Moments of Random Variables

Given the random vector X , the joint moment of order $k_1 + \dots + k_N$ is defined as the expectation

$$E[X_1^{k_1} X_2^{k_2} \dots X_N^{k_N}] = \int_{-\infty}^{\infty} x_1^{k_1} x_2^{k_2} \dots x_N^{k_N} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \quad (82)$$

The joint central moment of order $k_1 + \dots + k_N$ is defined by

$$E[(X_1 - \mu_1)^{k_1}(X_2 - \mu_2)^{k_2} \dots (X_N - \mu_N)^{k_N}] = \int_{-\infty}^{\infty} (x_1 - \mu_{x_1})^{k_1} (x_2 - \mu_{x_2})^{k_2} \dots (x_N - \mu_{x_N})^{k_N} f(\mathbf{x}) d\mathbf{x}, \quad (83)$$

N folds

where $\mu_n = E[X_n]$ is the mean of X_n . Among these moments the most important are the lowest pair-wise central moments, i.e.,

$$E[(X_n - \mu_n)(X_m - \mu_m)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_n - \mu_n)(x_m - \mu_m) f_X(x_n, x_m) dx_n dx_m. \quad (84)$$

The (84) is known as covariance; therefore, we can write

$$\text{Cov}[X_n, X_m] = E[(X_n - \mu_n)(X_m - \mu_m)] \quad (85)$$

$$= E[X_n X_m] - E[X_n] \mu_m - \mu_n E[X_m] + E[X_n] E[X_m] \quad (86)$$

$$= E[X_n X_m] - E[X_n] E[X_m]; \quad (87)$$

moreover, if $X_m \perp\!\!\!\perp X_n \Rightarrow \text{Cov}[X_n, X_m] = 0$ since $E[XY] = E[X]E[Y]$. An important property of the covariance is that its absolute value is bounded by square root of the product of variances, i.e.

$$|\text{Cov}[X_n X_m]| \leq \sqrt{\text{Var}[X_n] \text{Var}[X_m]}. \quad (88)$$

The correlation coefficient between random variables X_m and X_n is defined as

$$\rho_{mn} = \frac{\text{Cov}[X_m X_n]}{\sigma_{X_m} \sigma_{X_n}}. \quad (89)$$

Observe that $-1 \leq \rho \leq 1$. This coefficient provides a dimensionless measure of linear dependence between the two variables. It is easy to show that when $|\rho| = 1$ there is perfect linear dependence.

Common pitfalls: it is not uncommon among student to reverse the following condition $X \perp\!\!\!\perp Y \Rightarrow \rho_{XY} = 0$. Please, NEVER do this. In fact $\rho_{XY} = 0$ or equivalently $\text{Cov}[XY] = 0$ indicates ONLY that there is not linear dependence. For example two random variables can have nonlinear dependence which leads to $\rho = 0$ but this does not mean that they are statistically independent! For example, let's investigate the correlation between the kinetic energy, E_K , of a particle of mass m and its velocity V . Suppose that the velocity is a random variable that has a symmetric distribution (i.e. $E[V^3] = 0$) with 0 mean (i.e. $E[V]=0$). It follows that the kinetic energy is a random variable since $E_k = 1/2mV^2$. Clearly V and K are perfectly correlated, i.e. knowledge of V is sufficient to know deterministically K . It follows that V and K are statistically dependent. However, $\text{Cov}[VK] = mE[VK] - E[V]E[K] = 1/2mE[V^3] = 0$. Thus, keep in mind $\rho_{XY} = 0 \not\Rightarrow X \perp\!\!\!\perp Y!!$

6 The Normal Distribution

The Gaussian (from Carl Friedrich Gauss⁶) normal distribution is a probability density function for a random variable that appears frequently in engineering applications.

The probability density function is

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty \leq x \leq \infty, \quad (90)$$

where μ is the location parameter (mean), and σ the scale parameter (standard deviation). A standard normal random variable, Z , is a normal random variable with zero mean and standard deviation one. It follows that its PDF is:

$$f_Z(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right], \quad -\infty \leq z \leq \infty, \quad (91)$$

and CDF

$$F_Z(z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z'^2\right] dz', \quad -\infty \leq z \leq \infty. \quad (92)$$

Observe that there is no close form solution for the CDF of the normal distribution. Moreover given the symmetry of the distribution $\Phi(-z) = 1 - \Phi(z)$. Since there is no close form solution the CDF is evaluated numerically and its values are reported in the normal distribution table, Figure 13. Alternatively, nowadays, any software can give directly the values of CDF.

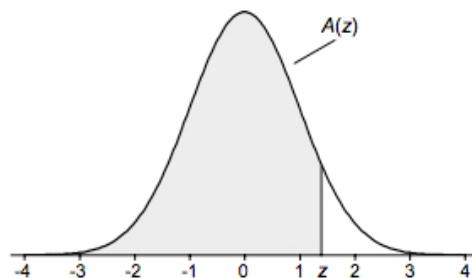
To use the normal distribution table with a classical normal random variable X , we should implement the following transformation:

$$Z = \frac{X - \mu}{\sigma}; \quad (93)$$

then,

$$P(X < x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right). \quad (94)$$

⁶Johann Carl Friedrich Gauss (*1777 †1855) was probably the greatest mathematician of modern history. Gauss had an unprecedented and extraordinary influence in several fields of mathematics and science. Oftentimes, he is remembered as the *Princeps mathematicorum* (the primer mathematician)



$A(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

Figure 13: Standard Normal CDF



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



LECTURE NOTES

Fundamentals of Probability and Statistics

Chapter 1: Lecture 3

Spring 2018

Written by marco broccardo
IBK, ETHZ

7 Probability Models for Random Processes

A random process is a probabilistic model that describes the evolution of a random phenomenon over time. Roughly speaking, a random process can be viewed as a sequence of random variables. We start this chapter by assuming discrete events; then, we extend the results for continuous time.

7.1 The Bernoulli process

The Bernoulli process can be thought of as a sequence of independent experiments, $X_1, X_2, \dots, X_n, \dots$, which having two possible outcomes (e.g. fail, not fail). Each experiment is named Bernoulli trial. The probability of success of an experiment is p (with $0 < p < 1$), (e.g. $P(X_n = 1) = p$ and $P(X_n = 0) = 1 - p$), and its mean and expectation are given by

$$E[X_n] = p \times 1 + (1 - p) \times 0 = p, \quad (95)$$

$$\text{Var}[X_n] = E[X_n^2] - E[X_n]^2 = p - p^2 = p(1 - p). \quad (96)$$

As mentioned above, a random process can be described as a collection of random variables. In Section 4, we have seen that in the presence of multiple random variables the complete probabilistic description of the problem is given in terms of the joint distribution. In the Bernoulli process, since each random variable is independent, the joint distribution can be written as $P_{\mathbf{X}}(x_1, x_2, \dots, x_n, \dots) = p_{X_1}(x_1)p_{X_2}(x_2)\dots p_{X_n}(x_n)\dots$. An alternative way to think about the Bernoulli process is to imagine that every infinite sequence of Bernoulli trials is a sample point of a sample space S . In this view, every single infinite sequence of Bernoulli trials has probability 0 (since $0 < p < 1$). It follows that the sample space must be continuous.

Problem Can you think what might be a good representation of the sample space S ?

Note that because of the assumption of independence, knowledge of the outcome of an experiment or of a sequence of experiments does not affect the outcome of future experiments. Oftentimes, this property is called memoryless property.

When facing with Bernoulli process there are three main probabilistic questions we might ask:

- i. For a given amount of experiments, how many will succeed?
- ii. How many experiments should we perform before observing the first success?
- iii. How many experiments should we perform before observing the second (or third, or fourth...) success?
- i. The first question is analogous to the Example II. In particular given n experiments, the probability that x out of n will succeed are given by the binomial distribution with parameters p and n , ($\text{bin}(p, n)$), i.e.

$$p_{\mathbf{X}}(x; p, n) = \binom{n}{x} p^x (1 - p)^{n-x}, \text{ for } x = 0, 1, 2, \dots, n. \quad (97)$$

The mean and the variance of the binomial distribution are given by

$$E[X] = np, \quad (98)$$

$$\text{Var}[X] = np(1 - p). \quad (99)$$

ii. To answer the second question, we define the number of performed experiments before we observe a success as the inter-arrival time, T_1 . Given this definition, a sequence of outcomes up to T_1 is of the following type $0, 0, \dots, 0, 1$. Then, given the assumption of independence, the probability of such sequence is

$$p_{T_1}(t; p) = P(T_1 = t) = (1 - p)^{t-1} p, \text{ for } t = 1, 2, \dots \quad (100)$$

Eq.(100) is known as the geometric distribution, $geom(p)$. The mean and variance of the geometric distribution are given as

$$E[T_1] = \frac{1}{p}, \quad (101)$$

$$\text{Var}[T_1] = \frac{1-p}{p^2}. \quad (102)$$

iii. To answer the third question, we observe that once the first success occurs, the next string of $0, 0, \dots, 0, 1$ is statistically independent from the past. It follows that the inter-arrival time $T_{1,2}$ is again a geometrical random variable with parameter p . More generally, any inter-arrival time between two sequential successful events is a geometric distribution with parameter p . Then, the arrival of the k th event is simply the sum of k statistically independent geometrical random variables, i.e. $T_k = T_1 + T_{1,2} + \dots + T_{(k-1),k}$. Alternatively, the distribution of T_k can be derived as follows

$$p_{T_k}(t; p) = P(T_k = t) = P(k-1 \text{ experiments succeeded before } t) \times P(\text{the } t\text{th experiment is successful}) \quad (103)$$

which leads to

$$\begin{aligned} p_{T_n}(t; p, k) &= \binom{t-1}{k-1} p^{k-1} (1-p)^{(t-k)} \times p, \\ &= \binom{t-1}{k-1} p^k (1-p)^{(t-k)} \text{ for } t = k, k+1, \dots \end{aligned} \quad (104)$$

The (104) is known as negative binomial distribution of order k , $negbin(p, k)$. Expectation and variance are given as

$$E[T_k] = \frac{k}{p}, \quad (105)$$

$$\text{Var}[T_k] = k \frac{1-p}{p^2}. \quad (106)$$

7.1.1 Bernoulli process with random selection

Suppose that you have a Bernoulli process composed of a series of experiments with probability p of success. Whenever there is a success, then you perform an additional experiment. This additional experiment has probability of success q , and it is statistically independent from the previous experiments. If the the second experiment succeed, then you keep the sample otherwise you discard it, Figure 14. If you observe the derived process, you might recognize that it is a Bernoulli process with probability pq . On the other hand, the process of discarded experiments is also a Bernoulli process, where each sample has a probability $p(1 - q)$.

7.1.2 Combining Bernoulli processes

Suppose you have two independent Bernoulli processes with parameters p and q and you would like to combine the two processes into one single process, Figure 15. A success on the merged process is recorded if there is a success in at least one of the two original processes. The probability that there is at least a success for every single experiment is $p + q - pq$. Notice that we are not accounting for the number of success at each sequential experiment, but only if there was at least one success.

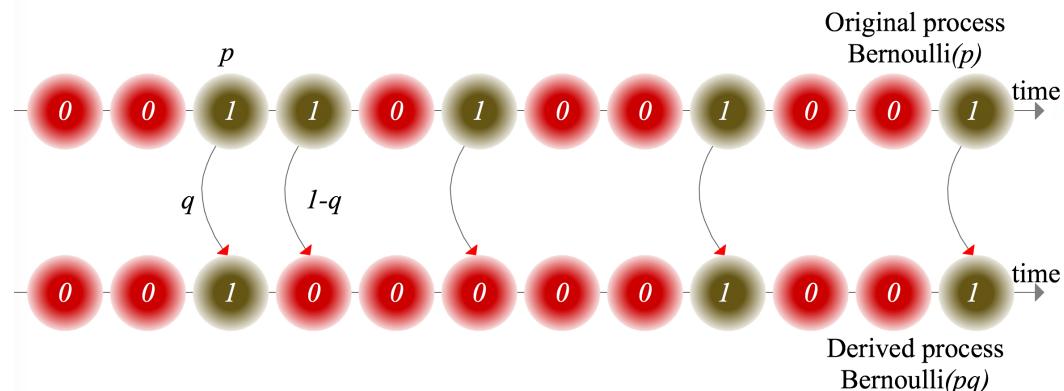


Figure 14: Bernoulli process with random selection

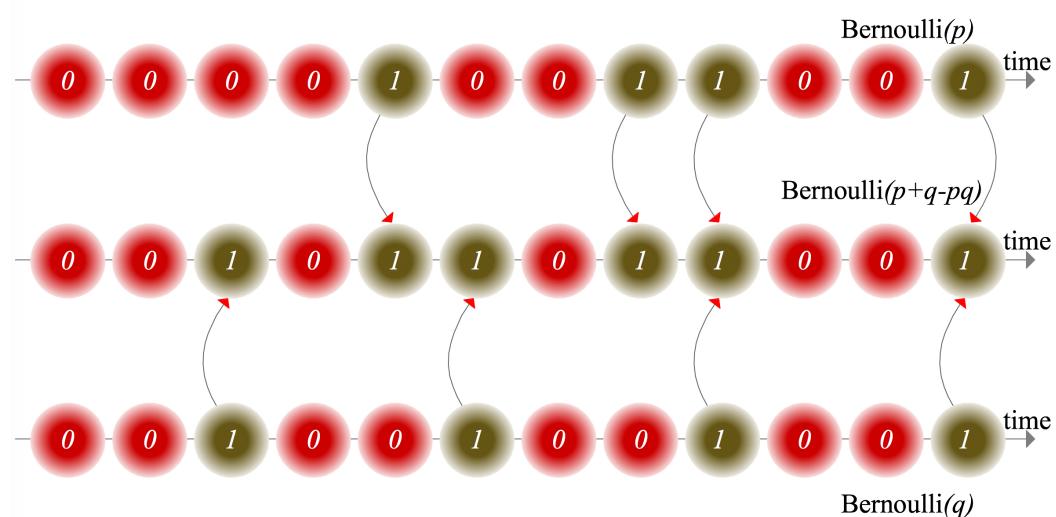


Figure 15: Combining Bernoulli processes

7.2 The Homogeneous Poisson Process, HPP

The Poisson⁷ process is a continuous time version of the Bernoulli process. Before examining the process in details, let's focus on the following problem.

Suppose that for a given experimental campaign n is large and p is small so that the mean np has a finite reasonable value. In this case, every string of sequential experiments has a large number of 0s and few 1s. If we make these experiments so fast that to the limit they become instantaneous experiments; then, we can record the exact time when an experiment succeed. Hence, we would like to use a continuous variable for the time, rather than discrete. For a given interval, we can address this problem by letting n growing to infinity, while p decreases so that for a fixed interval t , $np = \Lambda$. Thus, given a time interval t we can obtain the number of instantaneous success by finding the limit of the binomial distribution, i.e.

$$\begin{aligned}
p_X(x; \Lambda) &= \lim_{n \rightarrow \infty} \binom{n}{x} \left(\frac{\Lambda}{n}\right)^x \left(1 - \frac{\Lambda}{n}\right)^{n-x}, \\
&= \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\Lambda}{n}\right)^x \left(1 - \frac{\Lambda}{n}\right)^{(n-x)}, \\
&= \lim_{n \rightarrow \infty} \underbrace{\frac{n(n-1)\dots(n-x+1)}{n^x}}_{1 \text{ for } n \rightarrow \infty} \underbrace{\frac{\Lambda^x}{x!} \left(1 - \frac{\Lambda}{n}\right)^{-x}}_{1 \text{ for } n \rightarrow \infty} \underbrace{\left(1 - \frac{\Lambda}{n}\right)^n}_{e^{-\Lambda} \text{ for } n \rightarrow \infty}, \\
&= \frac{\Lambda^x}{x!} e^{-\Lambda}
\end{aligned} \tag{107}$$

The (107) is known as the Poisson distribution. Observe that Λ was fixed for a given duration of the experimental campaign. Then, we can define the mean *rate* of events as $\lambda = \Lambda/t$, where t is the duration of the experimental campaign.

A Poisson process builds on the following three properties:

Independence: the number of events for a given interval is independent from any other disjoint interval.

Stationarity (Homogeneity): the probability of x success for a given t , is the same for any interval of length t . This is equivalent of assuming constant λ . We later relax this condition by defining the non-homogeneous Poisson process.

Single event occurrence: given an infinitesimal time dt the probability of 0 event is $P(0, dt) = e^{-\lambda dt} = 1 - \lambda dt + \mathcal{O}(dt^2)$, the probability of one event is equal to $P(1, dt) = \lambda dt e^{-\lambda dt} = \lambda dt + \mathcal{O}(dt^2)$, where $\mathcal{O}(dt^2)$ are high order terms of dt . It follows that every single instant of time is a random variable which corresponds to a Bernoulli trial with probability of success (and expectation) $p = \lambda dt$. Then, λ is the expected number of success per unit time.

Like we did for the Bernoulli process, we might want to answer to the following questions:

⁷Siméon Denis Poisson (*1781 †1840) was a French mathematician and physicist. He accomplished many great results including the correction of the Laplace's second order partial differential equation for potential. It is also remembered as strenuous opponent of the wave theory of light; however, he was proven to be wrong by the civil engineer Augustin-Jean Fresnel (*1788 †1827). The famous experiment that brought shame upon Poisson and confirmed the Fresnel's waves theory of light is known as the Arago (*1786 †1853) spot.

- i. For a given time interval t , how many successful instantaneous events can we observe (equivalently, how many arrivals can we observe)?
- ii. How much time should we wait to observe the first success (equivalently, what is the first arrival time)?
- iii. How much time should we wait to observe the second (or third, or fourth...) success (equivalently, what is the k th arrival time)?
- i. Provided the introduction of this subsection, it should be easy to recognize that given a time t , the probability of x successes is the Poisson distribution, $Poi(\lambda, t)$, here rewritten as

$$p_X(x; \lambda, t) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad (108)$$

with mean and variance given as

$$E[X] = \lambda t, \quad (109)$$

$$\text{Var}[X] = \lambda t. \quad (110)$$

Common pitfalls: oftentimes students are confused whether the Poisson distribution is a discrete or continuous distribution. It is absolutely a DISCRETE distribution. Observe that the time t is a parameter of the distribution.

- ii. The second question can be answered in two way. The first one is to take the limit of the geometric distribution as n goes to infinity and $p = \Lambda/n$. The second approach is similar to the one we used to derive the geometric distribution. We follow the second approach. By definition

$$f_{T_1}(t) = \frac{P(0 \text{ events in } [0, t] \cap 1 \text{ event in } t + dt)}{dt}; \quad (111)$$

then,

$$\begin{aligned} f_{T_1}(t; \lambda) &= \frac{(e^{-\lambda t})(\lambda dt)}{dt}, \\ &= \lambda e^{-\lambda t}, \text{ for } t \geq 0. \end{aligned} \quad (112)$$

The (112) is known as the exponential distribution, $exp(\lambda)$. Mean and variance are given as follows

$$E[T_1] = \frac{1}{\lambda}, \quad (113)$$

$$\text{Var}[T_1] = \frac{1}{\lambda^2}. \quad (114)$$

- iii. The third question can be answered in three possible ways. The first one is by taking the limit of the negative binomial as n goes to infinity, and $p = \Lambda/n$. The second approach follows the same reasoning of the previous example. The third one is based on sum of statistically independent exponential random variables. In this notes, we follow the second method. By definition

$$f_{T_k}(t) = \frac{P(k-1 \text{ events in } [0, t] \cap 1 \text{ event in } t + dt)}{dt}; \quad (115)$$

then,

$$\begin{aligned} f_{T_k}(t; \lambda, k) &= \frac{1}{dt} \frac{(\lambda t)^{(k-1)} e^{-\lambda t}}{(k-1)!} (\lambda dt), \\ &= \frac{t^{k-1} \lambda^k e^{-\lambda t}}{(k-1)!}. \end{aligned} \quad (116)$$

The (116) is known as Erlang⁸ distribution which is simply a special case of the gamma distribution, $\Gamma(\lambda, k)$ with $k = 1, 2, \dots$. Mean and variance are given as

$$\begin{aligned} E[T_k] &= \frac{k}{\lambda}, \\ \text{Var}[T_k] &= \frac{k}{\lambda^2}. \end{aligned} \quad (117)$$

We have seen that the Poisson process can be seen as the continuous counterpart of the discrete Bernoulli process. In particular, we have seen that the geometric distribution is memoryless, i.e. the first arrivals does not affect the next sequence of experiments. The exponential distribution has a similar property. It follows that the inter-arrival time $T_{(k-1)-k}$ is exponential. The fact that at every instant of time the process renews itself is oftentimes counterintuitive. In particular, if for a given time t nothing has happen, it must $T_1 > t$. Then, the remaining time $T - t$ is again another exponential distribution, with the same parameter λ . This can be proved mathematically as follows

$$\begin{aligned} P(T > t + \tau | T > t) &= \frac{P(T > t + \tau, T > t)}{P(T > t)}, \\ &= \frac{P(T > t + \tau)}{e^{-\lambda t}}, \\ &= \frac{e^{-\lambda(t+\tau)}}{e^{-\lambda t}} = e^{-\lambda\tau}. \end{aligned} \quad (118)$$

It follows that the sum of statistically independent exponential distributions is the Erlang distribution.

7.2.1 Poisson process with random selection

Similar to the Bernoulli process, we start from a Poisson process with rate λ and select events as follows: at each event of the original process we perform an additional experiment and we keep the event with probability p and we discard it with probability $1 - p$, Figure 16. The resulting process is a Poisson process with rate λp . To confirm this statement, we should verify the three assumptions of the Poisson process. The independence property is satisfied since each additional experiment with probability p is statistically independent. The time stationarity holds because λ and p are constant in time. Finally, for a given dt we have for the Poisson process with random selection $P(0) = 1 - (\lambda dt)p$ and $P(1) = (\lambda dt)p$; hence, the third property is satisfied.

Example VI

Suppose that the occurrence of earthquakes in the Hayward fault is modeled with a Poisson process with rate of $M \geq 5.5$ equal to $\lambda_{hf} = 0.15$. Suppose that given an earthquake of $M \geq 5.5$

⁸Agner Krarup Erlang (*1878 †1929) was a Danish statistician, mathematician, and engineer. He is known as the father of transportation engineering.

occurred, the probability of observing a peak ground acceleration (PGA) greater than $0.1[g]$ is 0.1, i.e. $p = P(PGA \geq 0.1[g] | M \geq 5.5) = 0.1$.

Problem V

- What is the probability that at least one earthquake of $PGA \geq 0.1[g]$ will hit the site in Oakland in the next 20 years?

Suppose that the probability of exceeding the $PGA > 0.1[g]$ is given conditional to R (epicentral distance) and M magnitude, i.e. $P(PGA \geq 0.1[g] | M = m, R = r)$.

Problem VI

- Write the joint probability distribution of PGA , R , and M , i.e. $f_{PGA,M,R}(pga, m, r)$.
- Find the marginal probability of the PGA .
- Describe the Poisson process of the earthquakes with $PGA \geq 0.1[g]$.

7.2.2 Combining Poisson processes

Suppose you have two independent Poisson processes with rate λ_1 and λ_2 . The merged process is another Poisson process with rate $\lambda_1 + \lambda_2$, Figure 17. To verify this statement we should verify that the three assumptions of the Poisson process are satisfied. The independence property is satisfied since different intervals are independent within the two processes (and the two processes are independent) so it must be for the combined. The time stationarity holds because the rate $\lambda_1 + \lambda_2$ is stationary on time. Finally, for a given dt we have for the combined process

$$P(0 \text{ events in } dt) = (1 - \lambda_1 dt)(1 - \lambda_2 dt) = 1 - (\lambda_1 + \lambda_2)dt + \mathcal{O}(dt^2), \quad (119)$$

$$P(1 \text{ event in } dt) = \lambda_1 dt + \lambda_2 dt - \lambda_1 \lambda_2 dt^2 = (\lambda_1 + \lambda_2)dt + \mathcal{O}(dt^2); \quad (120)$$

hence, the third property is satisfied.

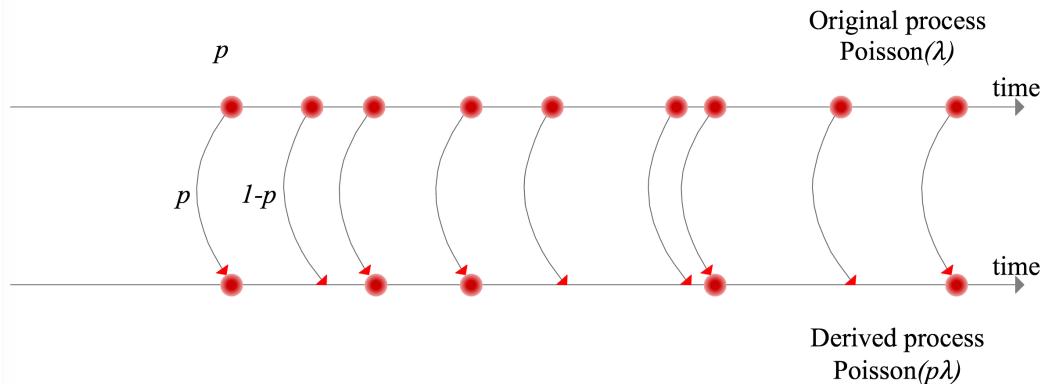


Figure 16: Poisson with random selection

Example VII

Suppose that the occurrence of earthquakes in the Hayward fault and St Andrea fault can be modeled as two statistically independent Poisson processes with rate of occurrence of $M \geq 6.7$ equal respectively to $\lambda_{hf} = 0.011$ and $\lambda_{af} = 0.019$.

Problem VII

- i What is the probability of having at least one earthquake of Magnitude 6.7 or greater in our location in South Oakland?
- ii Given that an earthquake occurred what is the probability that it is from St Andrea fault?

Suppose that an earthquake $M = 7$ occurred, and 4 out of the 6 buildings of Example II are heavily damaged. Assume (somewhat not very realistically) that the recovery time follow an exponential distribution, and the recovery of the 4 buildings are statistically independent.

Problem VIII

- i What is the expected time until the last building is restored?

Table 3 summarize the distributions for the Bernoulli and Poisson processes

	Bernoulli Process	Poisson Process
PMF of successes or arrivals	$bin(p, n)$, Equation (97)	$Poi(\lambda, t)$, Equation (108)
Interarrival time PDF	$geom(p)$, Equation (100)	$exp(\lambda)$, Equation (112)
k th arrival time	$negbin(p, k)$, Equation (104)	$\Gamma(\lambda, k)$, Equation (116)

Table 3: Summary Bernoulli and Poisson processes

8 Non-Homogeneous Poisson Process, NHPP

In this section, we derive a generalized form of the Poisson process where the mean rate of events are changing with time. A non-homogeneous Poisson process is defined by the following three assumptions

Independence: the number of events for a given interval is independent from any other disjoint interval.

Infinitesimal Poisson assumption: given a small interval dt , the probability of events in the interval $[t, t + dt]$ can be assumed as a Poisson distribution with mean number of events $d\Lambda(t) = \lambda(t)dt$.

Single event occurrence: given a time t and an infinitesimal time dt the probability of 0 events in the time interval $[t, t + dt]$ is $P(0, [t, t + dt]) = e^{-\lambda(t)dt} = 1 - \lambda(t)dt + \mathcal{O}(dt^2)$, the probability of one event is equal to $P(1, [t + dt]) = \lambda(t)te^{-\lambda(t)dt} = \lambda(t)dt + \mathcal{O}(dt^2)$,

where $\mathcal{O}(dt^2)$ are high order terms of dt . It follows that every single instant of time is a random variable which corresponds to a Bernoulli trial with probability of success (and expectation) $p = \lambda(t)dt$. It follows that $\lambda(t)$ is the instantaneous expected number of success for $[t, t + dt]$.

Denote with $p(x, t)$ the probability of x events in the interval $[0, t]$, then given the above assumption we have

$$p(x, t + dt) = p(x, t)p(0, [t, t + dt]) + p(x - 1, t)p(1, [t, t + dt]), \quad (121)$$

$$= p(x, t)(1 - \lambda(t)dt) + p(x - 1, t)\lambda(t)dt, \quad (122)$$

which leads to the following differential equation

$$\begin{aligned} \frac{p(x, t + dt) - p(x, t)}{dt} &= -\lambda(t)[p(x, t) - p(x - 1, t)], \\ \frac{dp(x, t)}{dt} &= -\lambda(t)[p(x, t) - p(x - 1, t)]. \end{aligned} \quad (123)$$

with initial conditions $p(0, 0) = 1$ and $p(x, 0) = 0$. The (123) can be solved recursively, for a fixed x . In particular for $x = 0$ we have $p(-1, 0) = 0$

$$\begin{aligned} \frac{dp(0, t)}{p(0, t)} &= -\lambda(t)dt, \\ \ln(p(0, t)) &= - \int_0^t \lambda(t)dt, \\ p(0, t) &= \exp(-\Lambda(t)), \end{aligned} \quad (124)$$

where

$$\Lambda(t) = \int_0^t \lambda(t)dt. \quad (125)$$

Finally we can verify by substitution that

$$p_X(x; \lambda(t), t) = \frac{\Lambda(t)^x}{x!} e^{-\Lambda(t)}. \quad (126)$$

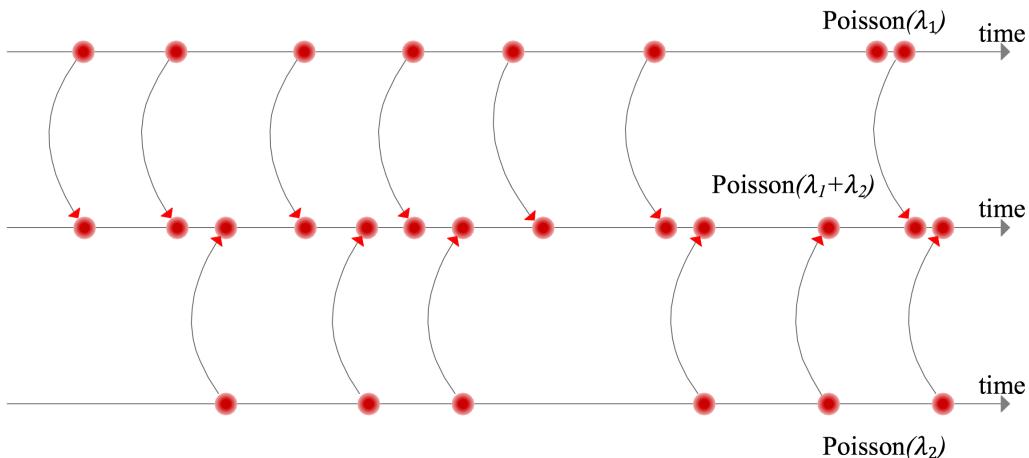


Figure 17: Combining two Poisson processes

Following the same reasoning of Section 7.2, we can derive the distribution of the first arrival as follow

$$f_{T_1}(t) = \frac{P(0 \text{ events in } [0, t] \cap 1 \text{ event in } t + dt)}{dt}; \quad (127)$$

then,

$$\begin{aligned} f_{T_1}(t; \lambda) &= \frac{(e^{-\Lambda(t)})(\lambda(t)dt)}{dt}, \\ &= \lambda(t)e^{-\Lambda(t)} \text{ for } t \geq 0. \end{aligned} \quad (128)$$

The probability distribution of the k th event is then

$$f_{T_k}(t) = \frac{P(k-1 \text{ events in } [0, t] \cap 1 \text{ event in } t + dt)}{dt}; \quad (129)$$

then,

$$\begin{aligned} f_{T_k}(t; \lambda, k) &= \frac{1}{dt} \frac{\Lambda(t)^{(k-1)} e^{-\Lambda(t)}}{(k-1)!} (\lambda(t)dt), \\ &= \frac{\Lambda(t)^{k-1} \lambda(t) e^{-\Lambda(t)}}{(k-1)!}. \end{aligned} \quad (130)$$

Problem IX

- i Find the Probability distribution of the inter-arrival time. Hint: you should use the total probability theorem
- ii Assume a rate function of the following type

$$\lambda(t; \alpha, \beta) = \frac{\alpha}{\beta} \left(\frac{t}{\beta} \right)^{\alpha-1} \quad (131)$$

Derive the distribution of T_1 .



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



LECTURE NOTES

Fundamentals of Probabilistic Seismic Hazard Analysis (PSHA)

Chapter 2: Lecture 4

Spring 2018

Written by simona esposito
marco broccardo
IBK, ETHZ

“Every earthquake contains a surprise”
Clarence Allen, 1952

Abstract: These lecture notes present the main aspects of Probabilistic Seismic Hazard Analysis (PSHA). PSHA is the basis for many earthquake design codes practices, and for seismic risk analysis of civil systems. The study of seismic hazard requires an understanding of the various processes by which earthquakes occur and their effects on ground motion. Therefore, in the first part of this note, we provide a brief introduction to the terminology used to describe the complex phenomena of earthquakes. Then, each of the steps needed to perform a site specific PSHA is presented and discussed.

1 Introduction

Seismic hazard analysis is one of the main components for the evaluation of the seismic risk at an earthquake prone site. In particular, the seismic hazard can be defined as the exceedance (or occurrence) probability for a given ground motion intensity measure threshold, site and time interval. Seismic risk is the exceedance (or occurrence) probability for a given loss threshold, site and time interval.

Reliable estimates of a seismic hazard are paramount to minimize loss of life, property damage and social and economic disruption. Further, seismic hazard analysis is the base-ground for seismic-structural design, financial-insurance policies, emergency-response policies, and post-earthquake recovery.

One of the main uses of hazard analysis is within the context of seismic design of buildings and infrastructures. These results are usually incorporated into building codes for residential buildings; while for critical facilities, hazard analysis are performed ad hoc.

The approaches to seismic hazard assessment can be grouped into two broad categories: deterministic and probabilistic.

A Deterministic Seismic Hazard Analysis (DSHA) consists in the analysis of a particular seismic scenario. Generally, the output of the analysis is the intensity level of the selected hazard scenario. In particular, the scenario consists of the postulated occurrence of an earthquake of a specific size occurring at a given location. A typical DSHA (shown in Figure 1) can be split in four stages:

- i Identification and characterization of all earthquake sources capable of producing significant ground motion at the site. Source characterization includes definition of the earthquake source's geometry (source zone) and earthquake potential.
- ii Selection of a source to site distance parameter for each source zone. Usually the shortest distance between the source zone and the site of interest is selected. The distance can be expressed in different ways (e.g. epicentral or hypocentral) depending on the measure of distance of the predictive equation used in the following step.
- iii Selection of the controlling earthquake (i.e. the earthquake that is expected to produce the strongest level of shaking) generally expressed in terms of ground motion parameters at the site. The selection is made comparing the levels of shaking produced by earthquakes (identified in step i) assumed to occur at the distances identified in steps ii. The controlling earthquake is described in terms of its size (i.e. magnitude) and distance from the site.
- iv The hazard at the site is formally defined, usually in terms of the ground motions produced at the site by the controlling earthquakes. Its characteristics are usually described by one or more ground motion parameters.

The DSHA provides a straightforward framework for the evaluation of the worst-case ground motions. This is a useful procedure when applied to critical facilities (e.g. nuclear power plant), where failure could have catastrophic effects. However, it does not provide any information regarding the likelihood of occurrence of the controlling earthquakes, the level of shaking that maybe expected during the lifetime of a particular structure, and the uncertainty involved in all the steps of the procedure. Finally, it is important to recognize that at the step i (but not only), DSHA involves critical subjective decisions.

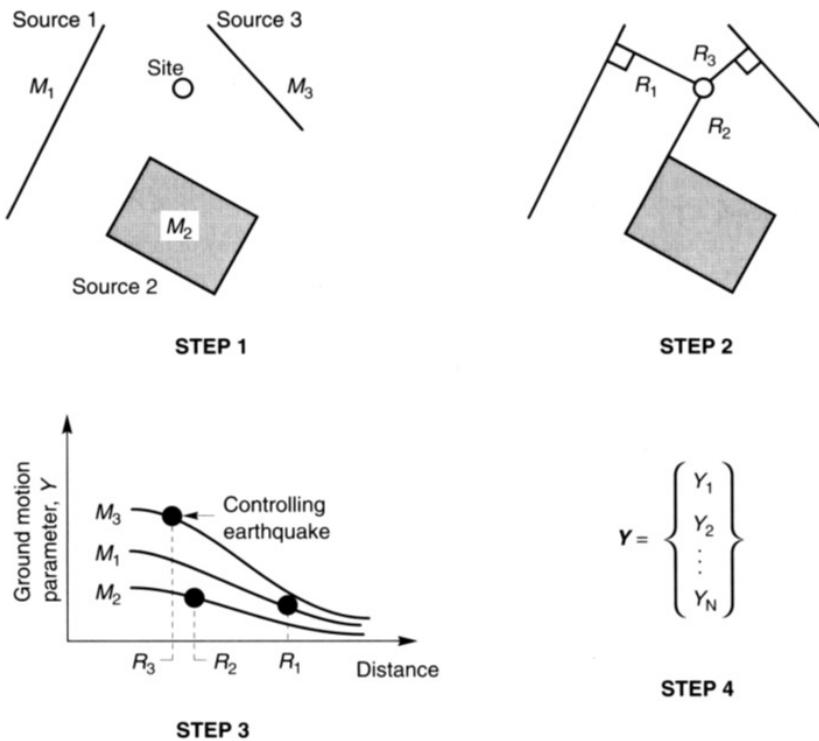


Figure 1: Four steps of a DSHA analysis (Kramer, 1996).

Probabilistic seismic hazard analysis (PSHA) provides an uncertainty quantification framework to address the hazard analysis. PSHA is an essential tool for current earthquake engineering practices and performance-based earthquake engineering (PBEE) design methodology. The goal of a PSHA is to evaluate the exceedance (or occurrence) probability of a given ground motion intensity measure threshold at a considered site and in a considered time interval. These intensity measures are then related to facilities damages, economic and social losses.

The numerical/analytical approach to PSHA was first formalized by Cornell (1968). The most comprehensive treatment to date is the SSHAC (1997) report, which covers many important procedural issues that will not be discussed here. The SSHAC report represents the best source of additional information for anyone conducting a PSHA.

The PSHA can also be described in four steps similarly to the steps of a DSHA procedure as illustrated in Figure 2:

- The first step is identical to DSHA, except that the probability distribution of potential earthquake locations within the source must also be characterized. These distributions are then combined with the source geometry to obtain the corresponding probability of source to site distance.
- The temporal distribution of earthquake recurrence and size distribution (magnitude) must be characterized for each source zone. A recurrence relationship specifies the average rate at which an earthquake of some size will be exceeded.
- The ground motion produced at the site by earthquakes of any possible size occurring at any possible point in each source zone must be determined with use of predictive equations.

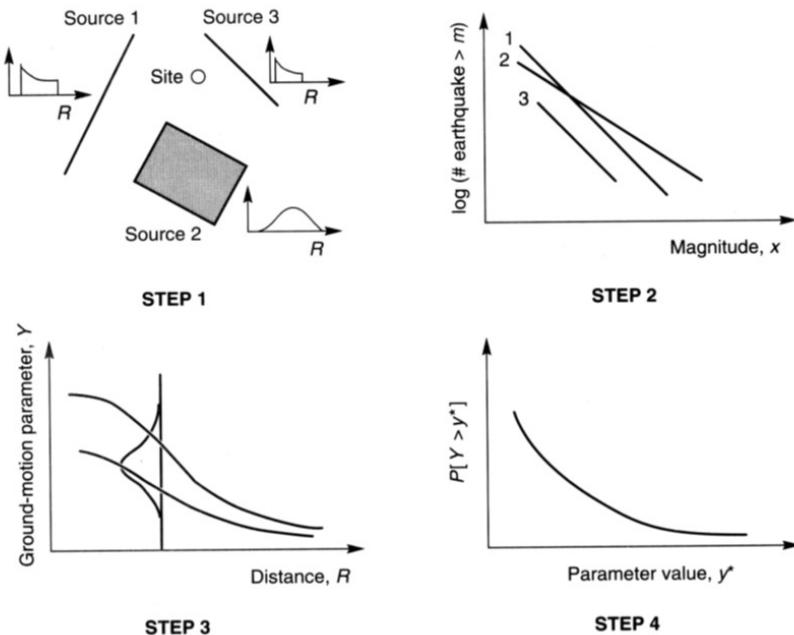


Figure 2: Four steps of a PSHA analysis (Kramer, 1996).

- iv In the last step the uncertainties in earthquake location, earthquake size, and ground motion are combined to obtain the probability that the ground motion parameter will be exceeded during a particular time period.

The selection of the most appropriate approach depends on several factors including the seismic environment, the structures and the aim of the project. However, “the more quantitative the decision to be made, the more appropriate is probabilistic hazard and risk assessment” (Mc Guire, 2004).

PART I
Seismology in a Nutshell

2 Earthquakes

Earthquakes are the result of a sudden release of energy in the Earth's crust that creates elastic waves: the Earth reacts as an elastic solid and seismic waves are propagated to all parts of the Earth following paths through the body of the Earth itself and around its surface.

Earthquakes may be classified as natural or anthropogenic according to the nature of the source. Anthropogenic earthquakes are induced by human activity, such as technological operations involving extraction and/or injection of fluids in underground rocks. Natural earthquakes are caused by natural processes in the Earth and their nature can be volcanic or tectonic.

Earthquakes produced by stress changes in solid rock due to the injection or withdrawal of magma are called volcano earthquakes. These earthquakes can cause land to subside and can produce large ground cracks. These earthquakes can occur as rock is moving to fill in spaces where magma is no longer present. Volcano earthquakes don't indicate that the volcano will be erupting but can occur at any time. In this type of earthquakes the direct cause is an induced effect of the geo-dynamic process. In tectonic earthquakes the direct cause is the geodynamic movement itself.

Earthquakes occurring at boundaries of tectonic plates are called interplate earthquakes, while the less frequent events that occur along faults in the normally stable interior of the lithospheric plates are called intraplate earthquakes. Intraplate earthquakes often occur at the location of ancient failed rifts, because such old structures may present a weakness in the crust where it can easily slip to accommodate regional tectonic strain. These earthquakes are not well understood, and the hazard associated may be difficult to quantify.

Most tectonic earthquakes are causally related to compressional or tensional stresses built up at the margins of the lithospheric plates. These occur when rocks in the earth's crust break due to geological forces created by movement of plates. Boundaries have different names depending on how the two plates are moving in relationship to each other (Figure 3):

- i Crashing: convergent (subduction zone) boundaries.
- ii Pulling apart: divergent (or spreading ridge) boundaries.
- iii Sideswiping: transform fault boundaries.

Plate movement causes the buildup of tremendous quantities of energy in the rock. When the rock's rupture strength is exceeded, the stored energy is suddenly released producing vibrations that travel through the rock, leading to earthquakes.

2.1 Faults

In some regions, plate boundaries are easy to identify, while in others they may consist in smaller plates or microplates trapped between larger plates. Locally the movement between two portions of the crust will occur in new or preexisting offsets in the geological structure of the crust known as *fault*.

Faults are planar rock fractures which show evidence of relative movement. Earthquakes are caused by energy release during rapid slippage along faults. The largest examples are at tectonic plate boundaries, but many faults occur far from active plate boundaries. Since faults usually do not consist of a single, clean fracture, the term fault zone is used when referring to the zone

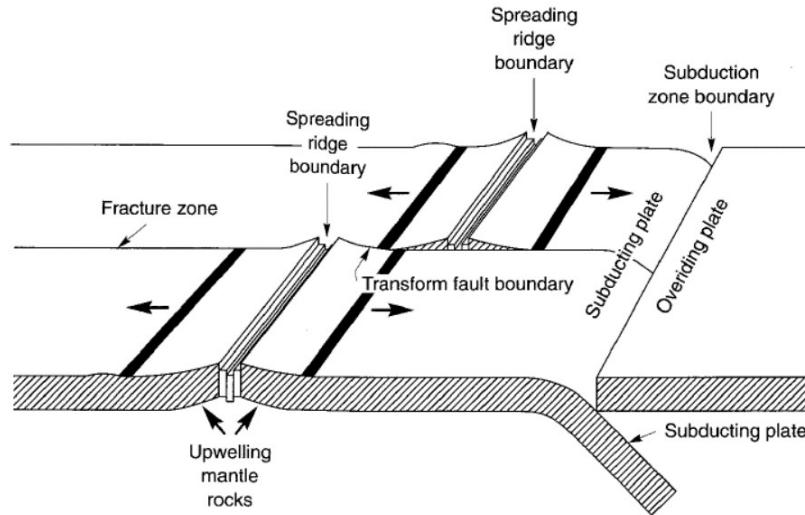


Figure 3: Plate boundaries (Kramer, 1996).

of complex deformation associated with the fault plane. The fault plane is then defined as the surface where the movement taken place within the fault.

The two sides of a fault are called the *hanging wall* and *footwall*. The hanging wall occurs above the fault. The footwall occurs below the fault.

The geometry of the fault plane is described by its strike and dip (Figure 4). The strike is the horizontal line produced by the intersection of the fault plane and a horizontal plane. The strike azimuth describes the orientation of the fault with respect to the north. The dip angle is the angle between the fault plane and the horizontal plane measured perpendicular to the strike.

The *sense of slip*, defined by the relative movements of geological features present on either side of the fault plane, defines the type of fault. Three groups are identified (5):

- i *Dip-slip fault*: where the main sense of movement (or slip) on the fault plane is vertical.
- ii *Strike-slip (or transform) fault*: where the main sense of slip is horizontal the fault is known as a fault.
- iii *Oblique-slip fault*: has significant components of both strike and dip slip.

Dip-slip faults include both normal and reverse. A *normal fault* occurs when the crust is in extension. The hanging wall moves downwards relative to the footwall. A *reverse fault* is the opposite of a normal fault: the hanging wall moves up relative to the footwall. Reverse faults are indicative of shortening of the crust. The dip angle of a reverse fault is relatively steep, greater than 45° . A *thrust fault* has the same sense of motion as a reverse fault, but with the dip of the fault plane at less than 45° .

In the strike-slip faults, the fault surface is usually near vertical and the footwall moves either left or right or laterally with very small vertical motion. The San Andreas fault is a remarkable example of a strike-slip fault. It marks the boundary between the North American and Pacific Plates in California.

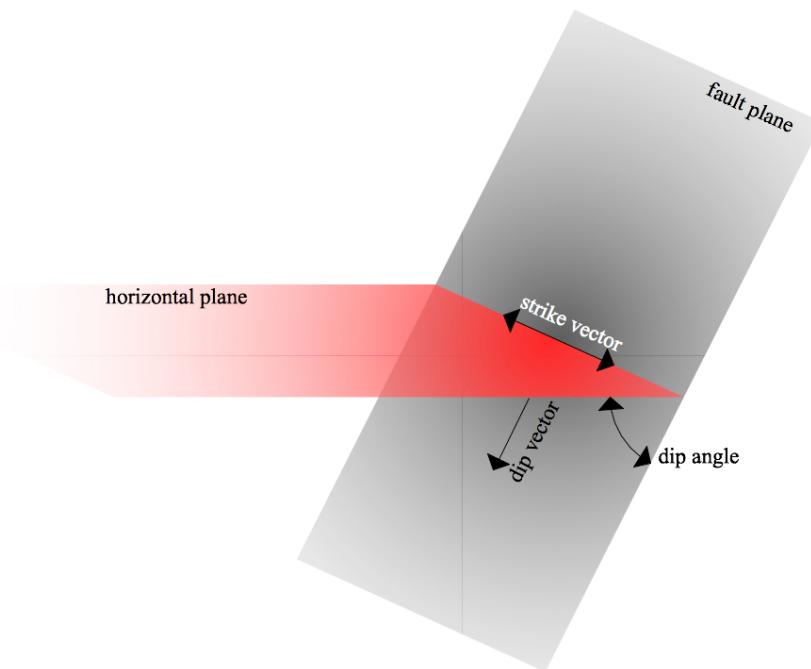


Figure 4: Fault plane orientation. Modified from Kramer, 1996.

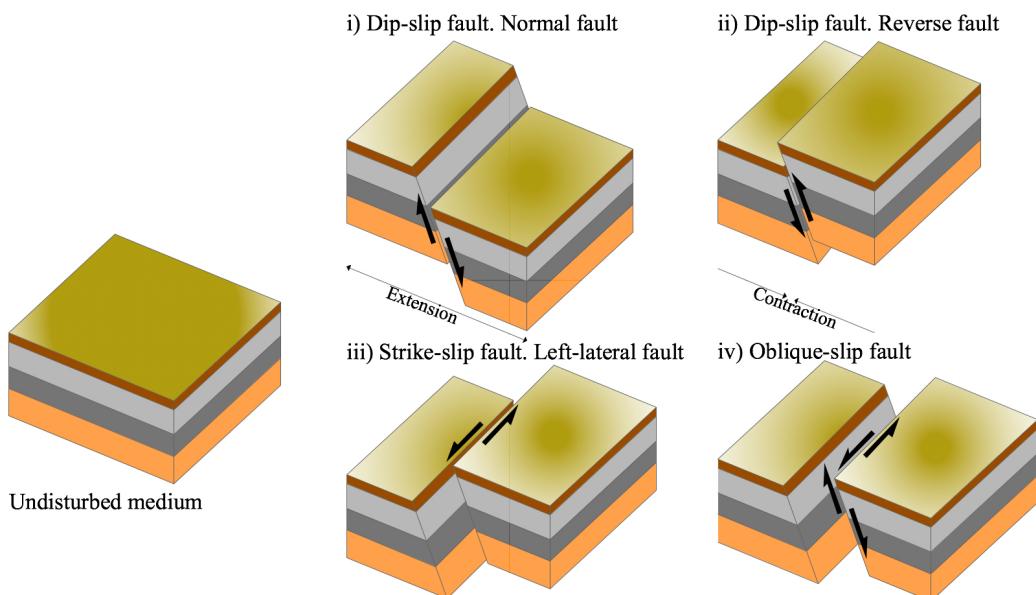


Figure 5: Type of faults

2.2 Seismic waves propagation

When an earthquake occurs, seismic waves radiate away from the source to the ground surface and travel rapidly through the earth's crust. Reaching the surface, seismic waves produce shaking which strength and duration depends on the size and location of the earthquake and on the characteristics of the site.

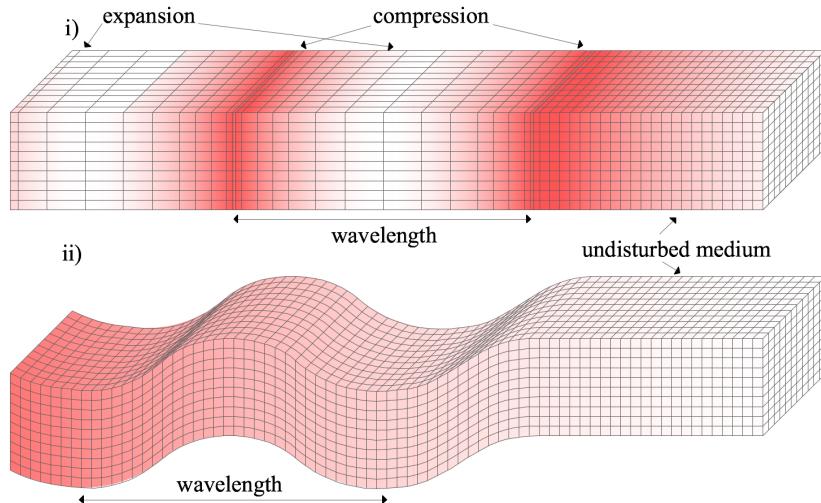


Figure 6: Deformations produced by body waves: i) P-waves ii) S-V waves.

The ground shaking felt at a given location will be made up of a combination of *body waves* and *surface waves*.

Body waves propagate through earth and they are generated by seismic faulting while surface waves travel along the ground surface and in most of cases are generated by the reflection and refraction of body waves. Body waves (Figure 6) include compressional waves (P-waves) where the ground moves parallel to the direction of propagation, and shear waves (S-waves) where the ground moves perpendicular to the direction of propagation; moreover depending on the direction of particle movement, S-waves can be divided in SV (vertical plane movement) and SH (horizontal plane movement).

Surface waves (*Rayleigh* and *Love waves*, Figure 7), instead, are more complex: for Rayleigh waves the particle motion traces a retrograde ellipse in a vertical plane with the horizontal component of motion being parallel to the direction of propagation; for the Love waves, the particle motion is along a horizontal line perpendicular to the direction of propagation.

The amplitude of ground motion reduces with distance from the source of seismic energy release. This is due to a combination of geometric attenuation, which accounts for the spread of the wave front as it moves away from the source, and anelastic attenuation, which is caused by material damping. In the immediate locality of the fault rupture, body waves will dominate the motion while ground motion at large distances to the source is generally dominated by surface waves because of the geometric attenuation is different for the two types of waves: assuming that the earthquake rupture zone may be represented as a point source and R is the distance from the rupture zone, the amplitude of body waves decreases in proportion to $1/R$, while the amplitude of surface waves decreases in proportion to $1/\sqrt{R}$.

2.3 Site effect

Although seismic waves travel through the rock for the majority of their trip from the source to the ground surface, the final part of the trip is through soil, which characteristics may influence the nature of shaking at the ground surface. The soil tends to act as a “filter” to seismic waves attenuating or amplifying the motion. Since soil conditions over short distances, levels of ground shakings may vary dramatically also within a small area. Generally site effects represent local

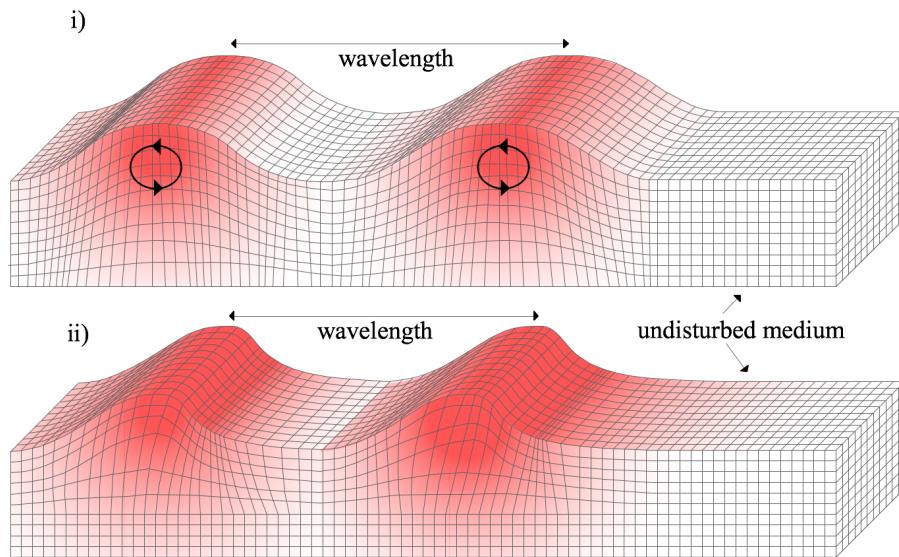


Figure 7: Deformations produced by surface waves: a) Rayleigh waves b) Love waves.

ground response effects, basin effects and the influence of surface topography on ground motion. Local ground response refers to the influence of relatively shallow geologic materials on (nearly) vertically propagating body waves.

The term basin effects refers to the influence of two- or three-dimensional sedimentary basin structures on ground motions, including critical body wave reflections and surface wave generation at basin edges.

Site effects due to surface topography (i.e., topographic effects) can amplify the ground shaking that would otherwise be expected on level ground along ridges or near the tops of slopes.

2.4 Location of earthquakes

Earthquakes result from rupture of the rock along a fault. The point at which rupture begins and first seismic waves originate is called the hypocenter (or focus) of the earthquake (Figure 8). From this point, located at some focal depth (or hypocentral depth) the earthquake spreads across the fault at velocities of 2 to 3 km/s. The point on the ground surface above the focus is called epicenter.

The distance between a site and the hypocenter is called hypocentral distance, while the distance on the ground surface between the site and the epicenter is known as the epicentral distance.

Earthquakes can be divided into three categories according to the hypocenter depth:

- i *Shallow*: maximum depth of 60 km.
- ii *Intermediate*: with a depth varying from 60 to 300 km.
- iii *Deep*: with a depth varying from 300 to 650 km.

The location of an earthquake is usually specified in terms of the location of its epicenter. Preliminary location of epicenter is based on the differential wave-arrival times measurements

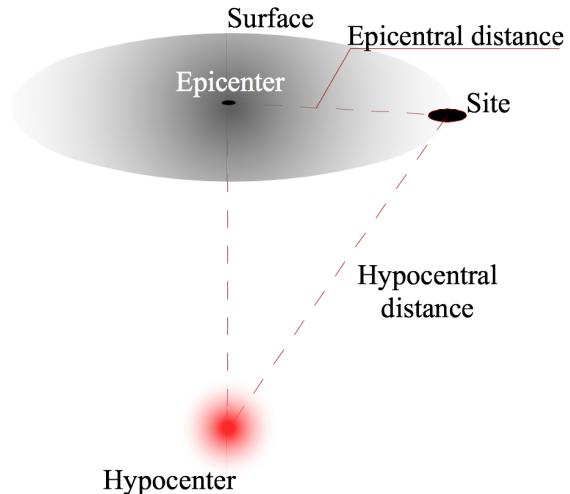


Figure 8: Earthquake location.

at a set of at least three seismographs. In particular it is based on the different arrival times of P - and S -waves. Since P -waves are faster than S -waves, they will arrive first at a given seismograph. The distance between the seismograph and the epicenter of the earthquake (d) will depend on the difference in arrival time of P -and S -waves (Δt_{p-s}) and the difference between the P - and S -wave velocities (λ_p and λ_s) according to:

$$d = \frac{\Delta t_{p-s}}{1/\lambda_s - 1/\lambda_p} \quad (1)$$

At any single seismograph it is possible to determine only the epicentral distance but not the direction of the epicenter. Plotting a circle of radius equal to the epicentral distance for a set of three seismograph, the location of the epicenter will correspond to the intersection of the three circles.

2.5 Size of earthquakes and magnitude scales

The oldest measure of the size of an earthquake is represented by the earthquake intensity. The intensity is a qualitative description of the effects of the earthquake at a particular location as evidenced by observed damage and human reactions. The increasing diffusion of modern instrumentations made necessary an absolute and instrumental scale of the earthquake size called earthquake magnitude. Most of the scales are based on measured ground motion parameters.

The fundamental step was done by Charles F. Richter in 1935 who used a Wood-Anderson torsional seismometer to define a magnitude scale for shallow, local earthquakes in California.

The *local magnitude* or *Richter magnitude* M_L is defined as the difference between the logarithm (base 10) of the maximum trace amplitude ($\log A$) (in millimeters) recorded at a known distance and the value of the “Zero Event” ($\log A_0$) associated to the same distance:

$$M_L = \log A - \log A_0 \quad (2)$$

The “Zero Event” (and the corresponding curve) is the event producing a 0.001 mm peak amplitude if recorded by Wood-Anderson instrument 100 km far from the epicenter of the

event. Usually for each seismic station, two signals are available recorded in two orthogonal horizontal directions (EW and NS): estimated local magnitude is the mean of values computed in each direction. Similarly, if more stations recorded the same event, local magnitude is the mean of values computed by each station.

Example I Evaluate the local magnitude of an event with a recorded peak amplitude of $A=45$ mm, located 100 km from the epicenter.

$$M_L = \log 45 - \log 0.001 = 1.66 - (-3.0) = 4.66$$

The local magnitude is the best known magnitude scale but it is not always the most appropriate scale for the description of the earthquake size. In order to extend the magnitude scale to more distant and deeper events, other magnitude scales based on the amplitude of a particular wave have been introduced. In particular two alternative magnitude scales M_S and m_b which refer to surface wave and volume (body) waves respectively, were introduced.

The *surface wave magnitude* (M_S) is based on the Rayleigh waves with a period of about 20 seconds which are not really influenced by the crust characteristics. It is obtained from:

$$M_S = \log A - 1.66 \log D + 2.0 \quad (3)$$

where A is the maximum ground displacement in micrometers and D is the epicentral distance of the seismometer measured in degrees. This magnitude is commonly used to describe the size of shallow (< 70 km focal depth) distant (Distance > 1000 km) moderate to large earthquakes.

For deeper events, volume waves have to be used because superficial waves are often too small to permit reliable evaluation of the surface wave magnitude. The magnitude scale with refers to volume P -waves (compressional and longitudinal in nature) which are not strongly influenced by the focal depth is the body wave magnitude, expressed as:

$$m_b = \log \left(\frac{A}{T} \right) + 0.01D + 5.9 \quad (4)$$

where A in this case is the P -wave amplitude in micrometers and T is the period of the P -wave. Previous described scales are empirical quantities based on measured ground motion parameters. However ground-shaking characteristics do not necessarily increase at the same rate of the amount of energy released during an earthquake. For large earthquakes, the energy recorded at one location does not continue to increase as the earthquake rupture area increases. This phenomenon is referred to as *saturation*. M_L and m_b saturate at magnitudes of 6 to 7 and M_S saturates at around 8. The only magnitude scale that is not subjected to saturation is the moment magnitude.

The *moment magnitude* (M_W) is a measure of earthquake size that can be related to physical parameters of an earthquake and it does not saturate because of seismograph limitations. It is based on the seismic moment which is a direct measure of factors that produce rupture along the fault. It is given by:

$$M_W = \frac{2}{3} \log(m_{min}) - 10.7 \quad (5)$$

where m_{min} is the seismic moment in dyne-cm, which is a function of soil shear modulus (μ), rupture area (A) and average displacement ($\overline{\Delta u}$) on A ,

$$m_{min} = \mu A \overline{\Delta u} \quad (6)$$

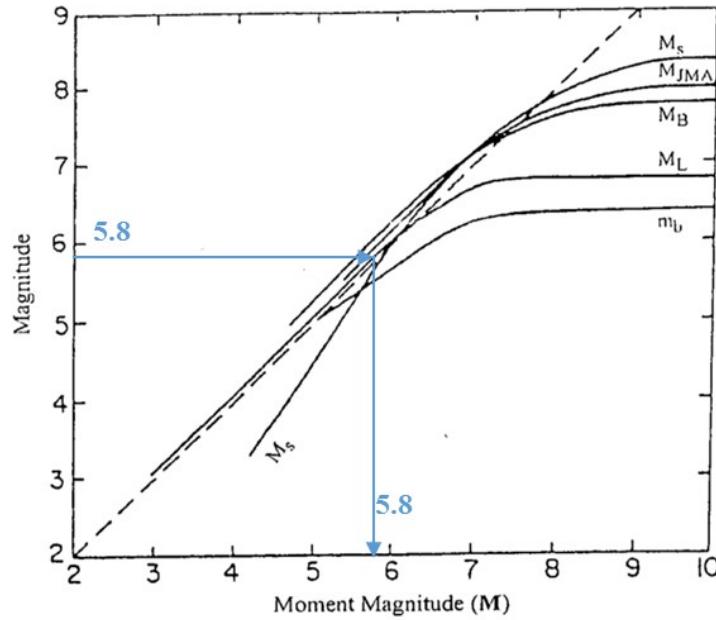


Figure 9: Relation between Moment Magnitude (M_W) and other Magnitude scales, adapted from McGuire (1994). M_{JMA} is the magnitude reported by the Japan Meteorology Agency. Note that $M_W = M_L$ for $M_W < 6.2$ and $M_W = M_S$ for $6.2 < M_W < 8.0$.

Fault type	Relationship	$\sigma_{\log L,A,D}$
Strike slip	$\log L = 0.74M_W - 3.55$	0.23
Reverse	$\log L = 0.63M_W - 2.86$	0.20
Normal	$\log L = 0.50M_W - 2.01$	0.21
All	$\log L = 0.69M_W - 3.22$	0.22
Strike slip	$\log A = 0.90M_W - 3.42$	0.22
Reverse	$\log A = 0.98M_W - 3.99$	0.26
Normal	$\log A = 0.82M_W - 2.87$	0.22
All	$\log A = 0.91M_W - 3.49$	0.24
Strike slip	$\log D = 1.03M_W - 7.03$	0.34
Reverse	$\log D = 0.29M_W - 1.84$	0.42
Normal	$\log D = 0.89M_W - 5.90$	0.38
All	$\log D = 0.82M_W - 5.46$	0.42

Table 1: Empirical Relationships between, surface rupture length (L), rupture area (A), maximum surface displacement (D).

Figure 9 summarizes and compares some of these magnitude scales and the saturation level of each scale.

The previous equation shows that seismic intensity is somehow correlated to the rupture dimensions. Analyzing historical earthquake data, Wells and Coppersmith (1994) provided empirical relationship between moment magnitude and all the geometrical parameters of the rupture (Table 1).

Example II Compute the probability that a moment magnitude 7.0 earthquake on the San Andreas Fault would cause a rupture area larger than 600 [km²].

Solution: San Andrea is a strike-slip movement. From Table 1, the mean rupture area is:

$$\log A = 0.9M_W - 3.42 = 6.3 - 3.42 = 2.88$$

from which

$$A = 10^{2.88} = 758.6 \text{ [km}^2\text{]}$$

The standard normal variate for a 400 [km²] area is

$$\begin{aligned} z &= \frac{\log 600 - \log 758.6}{0.22} = -0.46 \\ P(A > 600) &= 1 - P(A \leq 600) = 1 - \Phi(z) = 1 - 0.32 = 0.68 \end{aligned}$$

2.6 Ground motion intensity measures

The magnitude measures the intensity of a seismic event and it is a source parameter because it depends only on the characteristics of the rupture generating the event. On the contrary, it is also important to define some specific parameters able to identify the ground motion intensity in terms of effects in specific sites. Instrumental recordings of the strong ground movements during earthquakes are the clearest and most comprehensive definition of the actions against which structures and lifelines must be designed. Accelerographs record the acceleration of the ground as a function of time, while seismographs record the displacement or velocity of the ground. Double integration of the accelerogram provides the velocity and displacement time-histories to be recovered as well. Ground motion parameters (also called Intensity Measures, *IMs*) describe the most important characteristics of strong ground motion in compact and quantitative form. Many parameters have been proposed to characterize:

- i Amplitude (i.e. how large is the shaking?).
- ii The frequency content (what frequencies are particularly prevalent in the ground motion?).
- iii The duration (i.e. how long does the strong shaking last?) of strong ground motions.

Some intensity measures describe only one of these characteristics, while others may reflect two or three.

The most commonly used measure of the amplitude of a particular ground motion is the horizontal *peak ground acceleration* (PGA). The PGA for a given component of motion is the largest (absolute) value of horizontal acceleration obtained from the accelerogram of that component. Horizontal PGAs have commonly been used to describe ground motions because of their natural relationship to inertial forces; indeed, the largest dynamic forces induced in certain types of structures (i.e., very stiff structures) are closely related to the PGA.

The *peak ground velocity* (PGV) is another useful parameter for characterization of ground motion amplitude. PGV represents the maximum absolute value of velocity at the ground. The velocity is less sensitive to the higher-frequency components of the ground motion, therefore PGV is more likely than the PGA to characterize ground motion amplitude accurately at intermediate frequencies.

The *peak ground displacement* (PGD) is the maximum absolute value of displacement at the ground and it is generally associated with the lower frequency components of an earthquake motion. PGD is less commonly used as IM respect to PGV and PGA due to signal processing errors in the filtering and integration of accelerograms and due to long-period noise.

Loads produced by earthquakes are complex and characterized by components of motion that span a broad range of frequencies. The frequency content of ground motion describes how the amplitude of a ground motion is distributed among different frequencies. Since the dynamic response of civil structures is very sensitive to the frequency at which they are loaded, alternative IMs based on the frequency content of strong motion have been proposed to characterize the ground motion. Among them, the most used in earthquake engineering practice is the *response spectrum* defined as the maximum response (in terms of displacement, velocity or acceleration) of an elastic single degree-of-freedom (SDoF) system to a particular input motion as a function of the natural frequency (or natural period) and damping ratio of the SDoF system.

Duration of strong motion is also important for many physical processes that are sensitive to the number of load that occur during an earthquake (e.g. the degradation of stiffness and strength of certain old types of structures). The duration of strong ground motion is related to the time required for release of accumulated strain energy by rupture along the fault. As consequence, the duration increase with increasing earthquake magnitude. Due to the ground noise, a seismograph always receives little shakes that doesn't come from any seismic activity: wind, sea waves, road traffic and industrial activity are the principal causes for noise. As consequence some conventional measures of seismic duration have been introduced. Among them the *Bracketed duration* and the *Trifunac and Brady duration* are the most used. The *Bracketed duration* is defined as the time between the first and last exceedance of a threshold acceleration (usually 0.05 g). *Trifunac and Brady duration* depends on the amount of energy released in the site during the earthquake: it is the time interval between the 5% of the energy and the 95% of total released energy.

The preceding IMs are related primarily to only one of the three characteristics, i.e. the amplitude, frequency content, or duration of the ground motion. Since all of them are important, intensity measures that reflect more than one characteristic are very useful. An example is represented by the Arias Intensity. The *Arias Intensity* I_A is a descriptor of the energy of the earthquake and it is analytically defined as:

$$I_A = \frac{\pi}{2g} \int_0^{\infty} \ddot{u}_g(t) dt \quad (7)$$

where g is the acceleration due to gravity and $\ddot{u}_g(t)$ is the acceleration-time history in units of g .

I_A has the dimension of velocity. Since it is obtained by integration over the entire duration, its value is independent of the method used to define the duration of strong motion. This IM is well correlated to the damage level in soil and geotechnical structures.

2.6.1 Prediction of strong ground motion

Empirically based estimates of ground motion parameters are the oldest estimates in seismic hazard analysis, dating from the 1960s. Ground motion prediction equations (GMPEs) provide probabilistic distribution of the chosen IM (the predicted variable characterizing the level of shaking) conditional on a set of explanatory variables. GMPEs are obtained by regression of recorded data from historical events and they are popular for regions where many data are

available. GMPEs may change with time as additional strong motion data become available. Most predictive relationships are updated in the literature every 3 to 5 years or after the occurrence of a major event. In stable continental regions where low seismicity rates allow to collect only limited data, theoretical approaches are generally used instead. In such cases, synthetic time histories are simulated in order to enlarge the strong motion database on which regression techniques can then be applied.

The most used explanatory variables are the magnitude, the source-to-site distance and coefficients to take into account for style of faulting, wave propagation path, and/or local site conditions. The typical form is expressed in the following equation:

$$\log IM = \overline{\log IM}(M, R, \theta) + \varepsilon \quad (8)$$

$\overline{\log IM}(M, R, \theta)$ is the mean of the logs conditional on parameters such as magnitude (M), source-to-site distance (R), and others θ ; the difference between the observed and the predicted ground motion is the ground motion residual that represents the unexplained part of the model.

The functional form is usually selected to reflect the mechanics of the ground motion process as closely as possible. Common forms are based on the following observations (Kramer, 1996):

- i Peak values of strong motion parameters are approximately log normally distributed.
- ii Earthquake magnitude is typically defined as the logarithm of recorded peak amplitude.
- iii Body wave (P - and S -wave) amplitudes decrease according to $1/R$ and surface wave (primarily Rayleigh wave) amplitudes according to $1/\sqrt{R}$.
- iv The area over which fault rupture occurs increases with increasing earthquake magnitude. As a result, some of the waves that produce strong motion at a site arrive from a distance, R , and some arrive from greater distances. The effective distance, therefore, is greater than R by an amount that increases with increasing magnitude.
- v Some of the energy carried by stress waves is absorbed by the materials they travel through (material damping). This material damping causes ground friction amplitudes to decrease exponentially with R .
- vi Ground motion parameters may be influenced by source characteristics (e.g. type of fault) or site characteristics.

Therefore, a typical GMPE may have the following form:

$$\underbrace{\log IM}_{\text{i}} = \underbrace{C_1 + C_2 M + C_3 M^{C_4}}_{\text{ii}} + \underbrace{C_5 \log[R + C_6 \exp(C, M)]^{1/2}}_{\text{iii}} + \underbrace{C_8 R + f(\text{source}) + f(\text{site})}_{\text{v}} + \varepsilon \quad (9)$$

The residual term ε is assumed normally distributed with zero mean and standard deviation $\sigma_{\log IM}$ that describes uncertainty in the value of the ground motion parameters given by the predictive relationship (all the effects which are not accounted for in the chosen functional form).

In the most recent GMPEs the residual is expressed as the sum of two components: an inter-event term, which is constant for each earthquake (common for all sites) and represents average source effects not explicitly appearing in the model covariates, and an intra-event term representing the site-to-site variability of the ground motion parameter. Therefore for a particular

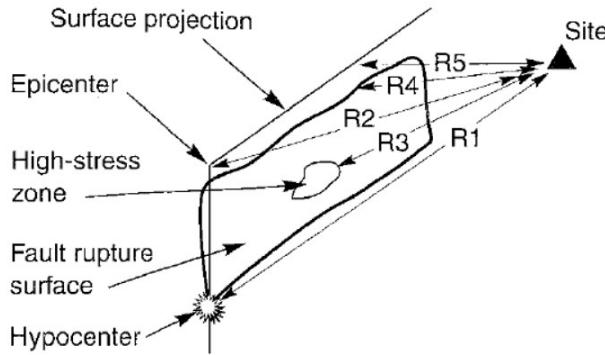


Figure 10: Distance metrics used in GMPEs (Kramer, 1996).

site p and earthquake j the logs of ground-motion intensities and related heterogeneity may be expressed as in the following equation:

$$\log IM_{pj} = \overline{\log IM_{pj}}(M, R, \theta) + \eta_j + \varepsilon_{pj}, \quad (10)$$

where η_j denotes the inter-event residual, which is a constant term for all sites in a given earthquake and represents a systematic deviation from the mean of the specific seismic event (i.e. average source effects not explicitly appearing in the model covariates); and ε_{pj} is the intra-event variability of ground motion. Residuals ε_{pj} and η_j are usually assumed to be *independent random variables*, normally distributed with zero mean and standard deviation σ_{intra} and σ_{inter} , respectively. Then, $\log IM_{pj}$ is modeled as a normal random variable with mean $\overline{\log IM_{pj}}(M, R, \theta)$ and standard deviation σ_t , where $\sigma_t = \sqrt{\sigma_{intra}^2 + \sigma_{inter}^2}$.

When using any GMPE, it is very important to know how parameters such as magnitude and source to site distance are defined. Concerning the distance, it is important to point out the different metrics used in the ground motion models (Figure 10): R1 and R2 are the hypocentral and epicentral distance, which are the easiest distance to determine after an earthquake. R3 is the distance to the zone of highest energy release. Since rupture of this zone is likely to produce the peak ground motion amplitudes, it represents the best distance measure for peak amplitude predictive relationships. Unfortunately, its location is difficult to determine after an earthquake and nearly impossible to predict before an earthquake. R4 is the closest distance to the zone of rupture and R5 is the closest distance to the surface projection of the fault rupture (also known as Joyner-Boore distance).

References

- Cornell, C.A. (1968). Engineering seismic risk analysis, Bull. Seism. Soc. Am., 58, 1583-1606.
- Grotzinger, J.P. and Jordan, T.H. (2014). Understanding Earth. W.H. Freeman and Company. New York, NY 10010.
- Kramer, S.L. (1996) Geotechnical earthquake engineering. Prentice Hall, Upper Saddle River, N.J.
- McGuire, R.K. (2004) Seismic Hazard and Risk Analysis, Earthquake Engineering Research Institute Publication, Report MNO-10, Oakland, CA, USA
- SHAAC, Senior Seismic Hazard Analysis Committee. (1997) Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts, US Nuclear Regulatory Commission report CR-6372, Washington DC.

Wells, D.L. and Coppersmith, K.J. (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. Bull. Seism. Soc. Am., 84, 974-1002.

PART II
Probabilistic Seismic Hazard Analysis

3 Probabilistic Hazard Model

Probabilistic Seismic Hazard Analysis (PSHA) evaluates the exceedance (or occurrence) probability of a given ground motion intensity measure threshold at given site and time interval.

This method has many applications in the field of earthquake engineering, including the design or retrofitting of critical facilities. More recently, results of PSHA have also been used for the determination of earthquake insurance coverage of private homes and businesses. PSHA provides a framework in which uncertainties are quantified, and combined in a rational manner to provide a comprehensive description of the seismic hazard. These uncertainties typically include magnitude size, earthquake location, soil condition, and rate of occurrence of earthquakes.

In detail, the calculation of seismic hazard is based on the Total Probability Theorem:

$$P[E] = \int_s P(E|S = s) f_S(s) ds, \quad (11)$$

where $P[E]$ represents the probability that the event E occurs, $P(E|S)$ is the conditional probability of the event E given the occurrence of the event S and $f_S(s)$ is the PDF of S , being S a continuous random variable.

In the context of seismic hazard, S is described by the magnitude M and the source to site distance R , and E is the event of overcoming an intensity measure level im at a given site. Given N_s seismic sources (e.g. multiple faults or multiple area sources), the (11) for a given source n becomes

$$P(IM > im) = \int_s P(IM > im|M = m, R = r) f_{R|M}^{(n)}(r|m) f_M^{(n)}(m) dr dm. \quad (12)$$

The (12) expresses the probability that a fixed value of ground motion im is exceeded at a given site, given the occurrence of *random* earthquake from the seismic source n . Observe that in case R is the epicentral or hypocentral distance R and M are statistically independent, i.e. $f_{R,M}^{(n)}(r,m) = f_R^{(n)}(r)f_M^{(n)}(m)$.

The general procedure is outlined by Cornell (1968) and it is based on four main steps:

- i **Earthquake Source Characterization.** Identification and classification of the N_s sources. Each source can be represented as area source, fault source, or, rarely, point sources, depending upon the geological nature of the sources and available data, Section 3.1. Definition of $f_{r|m}^{(n)}(r|m)$ or $f_r^{(n)}(r)$ (in case $R \perp M$). The distribution of $R|M$ or simply R depends on the definition of the distance the seismicity source, and whether considering or not the rupture size, Section 3.1.
- ii **Earthquake size.** Definition of $f_M^{(n)}(m)$ for each source n , based on magnitude recurrence relationship, Section 3.2.
- iii **Ground motion estimation.** Definition of $P(IM > im|M = m, R = r)$. These are empirical regression models named *ground motion prediction equations* (GMPE), Section 3.3. For a given magnitude and distance they define the probability of exceedance of a given intensity measure level im .
- iv **Hazard computation.** Solution of the integral 12 for all the N_s sources. Section 3.4.

This procedure is based on the assumption that earthquakes form a stochastic process. Because the earthquake can be considered instantaneous event with respect to the time periods of engineering interest, counting processes are the stochastic way to model earthquake occurrence which is a fundamental term of PSHA. Most applications of PSHA are based on the assumption that the earthquake process is memoryless, that is, there is no memory of the time, size and location of preceding events. This assumption is typically made by defining the occurrence of earthquakes as Homogenous Poisson Process (HPP) characterized by an exponential distribution of earthquake recurrence intervals (further details about this assumption will be provided in the following).

3.1 Earthquake source characterization

The first step required for PSHA is represented by the identification and the characterization of seismic sources that can affect the site of interest. The identification of seismic source zones is based upon the interpretation of geological, geophysical, and seismological data.

There are two general types (McGuire, 2004):

- i **Fault sources:** are faults or zones for which the tectonic features causing earthquakes have been identified. These are usually individual faults, but they may be zones comprising multiple faults or regions of faulting if surface evidence is lacking but the faults are suspected from evidence (e.g. seismicity patterns). Although originally only modeled as linear sources, most fault source models now have multi-planar features and ruptures are assumed to be distributed over the entire fault plane.
- ii **Area sources:** areas within for which future seismicity is assumed to have distributions of source properties and locations of energy release that do not vary in time and space. These are regions defined by polygons within which the seismicity is assumed uniform in terms of type and distribution of earthquakes.

In the context of PSHA, sources may be similar to or somewhat different than, the actual source, depending on the relative geometry of the source and site of interest and on the quality of the information about the sources (Kramer, 1996). In Figure 11 three examples of different source geometries are shown. The relative short fault shown in Figure 11a can be modelled as a point source since the distance between any point along its length and the site of interest is nearly constant. The fault plane shown in Figure 11b is characterized by a depth that is sufficiently small that the fault can be simplified as a linear source. The available data of the source shown in Figure 11c are not sufficient to determine the actual geometry, so it is represented as a volumetric source.

3.1.1 Distance

For a given earthquake source, it is generally assumed that earthquakes will occur with equal probability at any location on the source (i.e. uniformly distributed hypocenter location). Given that locations are uniformly distributed, the geometry of the source is used to identify the probability distribution of source-to-site distances.

Note that several distance definitions are used in PSHA (See Section 2.6.1 and Figure 10). Epicentral and hypocentral distances only consider the location of the rupture initiation; some others instead need to account for the fact that ruptures occur over a plane rather than a single point in space. The choice of distance definition will depend upon the required input to the GMPE. In case the estimate of fault rupture dimensions is needed, two methods can be applied: estimating the dimensions based directly on the size of the fault rupture plane or by basing the estimate on the size of the aftershock zone.

In the following examples we will derive the analytical distance distributions for simple geometric sources. Note that only the epicentral distance is considered here for simplicity. In the more general cases, (e.g.

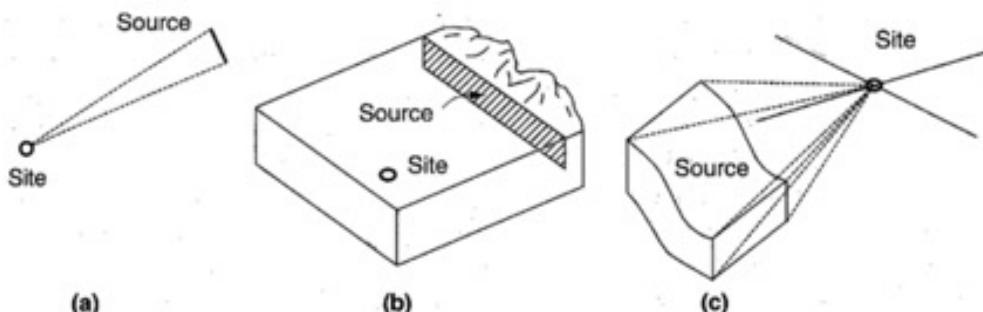


Figure 11: Example of different source geometries (Kramer, 1996).

complex geometries) distance distributions are more complicate to derive analytically and sometimes it is required numerical simulations.

Example III-a. Point source.

The problem is completely deterministic since the distance r is fixed, Figure 12-a). Then, $f_R(r') = \delta(r - r')$.

Example III-b. Linear source.

For the linear source in Figure 12-b), the probability that an earthquake occurs on the small segment of the fault between $L = l$ and $L = l + dl$ is the same as the probability that it occurs between $R = r$ and $R = r + dr$, i.e. $f(l)dl = f(r)dr$.

As $l^2 = r^2 - r_{min}^2$, and considering that the earthquakes are assumed to uniformly distributed over the length of the fault L_f , $f(l) = l/L_f$, the CDF of R is

$$F(r) = P(R \leq r) = \frac{\text{length of the fault within distance } r}{\text{total length of the fault}} \quad (13)$$

which can be written as

$$\begin{aligned} F(r) &= 0, \text{ for } r < r_{min} \\ &= \frac{\sqrt{r^2 - r_{min}^2}}{L_f}, \text{ for } r_{min} \leq r \leq r_{max} \\ &= 1 \text{ for } r > r_{max}. \end{aligned} \quad (14)$$

The PDF is obtained by differentiation of the (14) as

$$\begin{aligned} f(r) &= 0, \text{ for } r < r_{min} \\ &= \frac{r}{L_f \sqrt{r^2 - r_{min}^2}}, \text{ for } r_{min} \leq r \leq r_{max} \\ &= 0 \text{ for } r > r_{max}. \end{aligned} \quad (15)$$

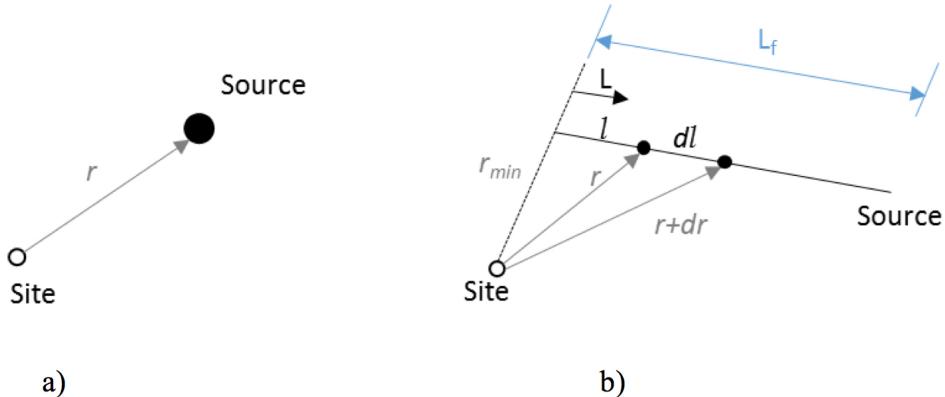


Figure 12: Example of a point source (a) and a linear source (b).

Example III-b. Area source.

This example is also the solution of the Problem *iii* in Section 3.1.2 of Fundamental of Probabilities. The site is located in an area source shown in Figure 13. Earthquakes are equally likely to occur anywhere in the area within K km from the site.

The probabilistic distribution of an epicenter being located at a distance of less than r is then equal to

$$F(r) = P(R \leq r) = \frac{\text{area of the circle with radius } r}{\text{area of the circle with radius } K} \quad (16)$$

which can be written as

$$\begin{aligned} F(r) &= \frac{r^2}{K^2}, \text{ for } 0 < r \leq K \\ &= 1 \text{ for } r \geq K. \end{aligned} \quad (17)$$

The PDF is obtained by differentiation of the (17) as

$$\begin{aligned} f(r) &= \frac{r}{2K^2}, \text{ for } 0 \leq r < K \\ &= 0 \text{ for } r \geq K. \end{aligned} \quad (18)$$

3.2 Earthquake size

Once the source characterization is completed, we define the distribution of the $f_M(m)$. This is usually based on a recurrence model describing “*the chance of an earthquake of a given size occurring anywhere inside the source during a specified period of time*”. A basic assumption of PSHA is that the recurrence law obtained from past seismicity is appropriate for the prediction of future earthquakes.

In many applications the exponential probability distribution is used to represent the relative frequency of different earthquake magnitudes(McGuire, 2004). In particular, the most used recurrence law model is the one proposed by Gutenberg and Richter (1954). The *Gutenberg-Richter law (G-R law)* expresses the relationship between the magnitude and rate of cumulative number of earthquakes in any given region:

$$\log \lambda(m) = a - bm, \quad (19)$$

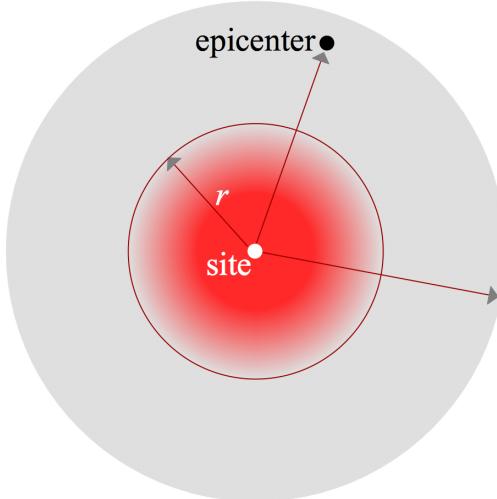


Figure 13: Example of an area source.

where \log is the logarithm base 10, $\lambda(m)$ is the mean annual rate of exceedance of magnitude m , and a and b are constants. The a value indicates the overall rate of earthquakes of the source, and the b value indicates the relative ratio of small and large magnitudes (a typical value of b is 1). As b increases, the number of larger magnitude events decreases compared to those of smaller magnitudes. Equation (20) is also expressed as

$$\lambda(m) = 10^{a-bm} = \lambda_0 \exp(-\beta m) = \exp(\alpha - \beta m), \quad (20)$$

where $\lambda_0 = 10^a$ represents the mean yearly number of earthquakes with $m \geq 0$, $\alpha = \ln(10)a$ and $\beta = \ln(10)b$.

The general formulation of the G-R law covers an infinite range of magnitudes, but often a lower and an upper-bound, m_{min} and m_{max} respectively, are used resulting in a truncated exponential distribution for magnitude frequency, named bounded G-R law (14).

The minimum magnitude is generally linked to the minimum magnitude which is believed to produce damages to the structures. It is usually set at values from about 4.0 to 5.0.

The truncation at m_{max} may arise because the magnitude scale saturates (see Section 2.5) or because the seismic zone cannot physically generate magnitudes above this value. It follows that the selection of m_{max} is complicate since it is generally estimated by using geologic evidence, geophysical data, analogies to similar tectonic regimes, or other methods (McGuire, 2004). The m_{max} can be estimated in several ways; the simplest is to take the maximum value historically observed and conservatively slightly increase this value, for example by 0.2 units. More advanced methods estimate m_{max} from the earthquake catalogue of the source considered or from the geometry of the source (e.g. Wells and Coppersmith, 1994, Table 1).

The CDF of the bounded GR-law considering only the lower bound can be obtained as follow

$$\begin{aligned} F_M(m) &= P(M \leq m | M > m_{min}) \\ &= \frac{\lambda_{min} - \lambda(m)}{\lambda_{min}}, \\ &= \frac{10^{a-bm_{min}} - 10^{a-bm}}{10^{a-bm_{min}}} \\ &= 1 - 10^{-b(m-m_{min})} \\ &= 1 - \exp[-\beta(m - m_{min})], \quad m > m_{min} \end{aligned} \quad (21)$$

and the PDF by differentiation as

$$f(m) = \frac{d}{dm} F(m) = b \ln(10) 10^{-b(m-m_{min})}, \quad (22)$$

$$= \beta \exp[-\beta(m - m_{min})], \quad m > m_{min}. \quad (23)$$

It follows that the mean annual rate can be written as

$$\lambda(m) = \lambda_{min} 10^{-b(m-m_{min})} = \lambda_{min} \exp[-\beta(m - m_{min})], \quad (24)$$

where $\lambda_{min} = 10^{a-bm_{min}} = \exp(\alpha - \beta m_{min})$ represents the mean yearly number of earthquakes with $m \geq m_{min}$. If the upper bound is also considered, then the (21) becomes

$$F_M(m) = \frac{1 - 10^{-b(m-m_{min})}}{1 - 10^{-b(m_{max}-m_{min})}}, \quad (25)$$

$$= \frac{1 - \exp[-\beta(m - m_{min})]}{1 - \exp[-\beta(m_{max} - m_{min})]}, \quad m_{min} \leq m \leq m_{max}, \quad (26)$$

with PDF as

$$f_M(m) = \frac{b \ln(10) 10^{-b(m-m_{min})}}{1 - 10^{-b(m_{max}-m_{min})}}, \quad (27)$$

$$= \frac{\beta \exp[-\beta(m - m_{min})]}{1 - \exp[-\beta(m_{max} - m_{min})]}, \quad m_{min} \leq m \leq m_{max}, \quad (28)$$

and mean annual rate of exceedance as

$$\lambda(m) = \lambda_{min} \frac{10^{-b(m-m_{min})} - 10^{-b(m_{max}-m_{min})}}{1 - 10^{-b(m_{max}-m_{min})}}, \quad (29)$$

$$= \lambda_{min} \frac{\exp[-\beta(m - m_{min})] - \exp[-\beta(m_{max} - m_{min})]}{1 - \exp[-\beta(m_{max} - m_{min})]}, \quad m_{min} \leq m \leq m_{max}. \quad (30)$$

Important considerations

- The use of the G-R law is not restricted to the use of magnitude as descriptor of the earthquake size. As explained in Section 2.5, intensity has also been used.
- Given a seismic source, the constants a and b (equation (20)) are usually estimated using statistical analysis of historical data (preinstrumental and instrumental seismicity) with additional constraining data provided by geological evidence. Historical data are usually provided by seismic catalogues, which summarized all important information about past events: e.g., date, hypocenter and epicenter location, earthquake size, fault mechanism, etc. The use of both instrumental and preinstrumental events implies a seismic catalogue that may contain both magnitude (based on different scales, see Section 2.5) and intensity data. Then, there is a need for conversion between the measures. This is expressed as: first by statistical and approximated correlation between macroseismic damage measures (i.e. intensity scales); second by a quantitative classification of earthquake intensity and between different magnitude scales. As an example, here it is reported a relationship between surface magnitude and epicentral intensity (a qualitative scale for damage estimation at the epicentral area) estimated for the Parametric Catalogue of Italian Earthquakes (Catalogo Parametrico dei Terremoti Italiani - CPTI): $M_s = 0.561I_0 + 0.937$. Note that whatever magnitude scale is chosen for the conversion, it must be consistent with the method chosen for estimating earthquake ground motion (See Section 2.6 and 3.4).
- Another important aspect to consider for the estimation of the G-R parameters is the *completeness* of the seismic catalogue in terms of intensity/magnitude and time intervals in which a certain intensity/magnitude range is likely to be completely reported (i.e. it is possible to assume that all the earthquakes of the considered magnitude are recorded in the catalogue). Geometry and

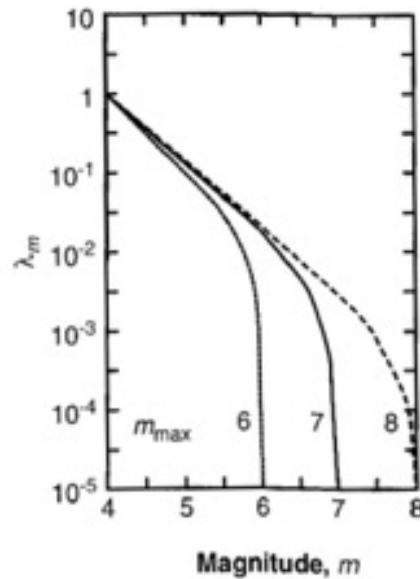


Figure 14: Bounded G-R law considering $m_{min} = 4$ and different m_{max} (Kramer, 1996).

coverage of the seismic network, or malfunction of seismic stations are the main responsible of the incompleteness of instrumental data. Analysis of catalogue completeness can be done via several methods such expert analysis of historical data and sources or visual procedures.

- Records of historical seismicity may be distorted by the presence of dependents events represented by the group of foreshock and aftershock which occur respectively before and after the mainshock in a variable time and space window. Since the PSHA is aimed at evaluating the seismic hazard associated to the main events (i.e. the mainshock), dependent events must be removed from the seismic catalogue and their effects accounted in separate analyses. One of the first analytical model for removing group of foreshock and aftershock was proposed by Gardner and Knopoff (1974) allowing the identification of spatial and temporal window of clustering around the mainshock as a function of mainshock magnitude.

3.2.1 Alternative magnitude distributions

Several magnitude distributions other than the truncated exponential distribution are available for modeling earthquake magnitudes as the Characteristic Magnitude Distribution. This distribution is used when continuous distributions encompassing all magnitudes are not appropriate (McGuire, 2004). Geological data may indicate that the characteristic earthquakes occur more frequently than would be implied by extrapolation of the GR-law from high exceedance rate (low magnitude) to low exceedance rates (high magnitude). This may be the case of individual sources (i.e. single faults) that usually generate earthquakes of similar size (i.e. characteristic earthquakes) at or near their maximum magnitude. The result is a more complex recurrence law that is governed by historical seismicity data at low magnitudes and geological data at high magnitudes (Kramer, 1996).

In most of the cases a continuous exponential distribution may be adequate for events up to, say, $m = 6-7$. Large earthquakes (e.g. $m = 7.5 - 8$) may occur with a characteristic magnitude whose frequency of occurrence is higher than obtained extrapolating from the smaller magnitude earthquakes. In this case, two truncated exponential distributions could model the magnitude occurrence of future events: a distribution between $m_{min} = 5$ and $m_{max} = 7$ with a standard β value (e.g. $\beta = 2.3$), and a separate distribution between $m_{min} = 7.5$ and $m_{max} = 8$ with $\beta = 0$ representing the equal likelihood of a characteristic event magnitude in that range. Seismicity data can be used then to estimate the rate for the first distribution while the rate for characteristic events can be estimated with geological evidence.

One form of the characteristic magnitude distribution is illustrated in Figure 15 developed by Young and Coppersmith (1985). This generalized recurrence law combines an exponential magnitude distribution at lower magnitudes with an uniform distribution in the proximity of the characteristic earthquake. The figure also shows a comparison with the bounded G-R law assuming the same upper magnitude bound value and slip rate.

Important considerations

- To adopt an alternative magnitude distribution the only requirement is to use the appropriate PDF. All the other steps of the PSHA remain identical.
- Evaluation of which model is most appropriate for a given source is hampered by brevity of historical and /or instrumental records. The seismicity records of the last decades for the major sources of the southern California suggest that while the available data were not sufficient to disprove the G-R model, the characteristic model better represented the observed distribution of magnitudes.

3.3 Ground motion estimation

Once the distribution of potential earthquakes magnitudes and locations has been identified, we can evaluate the ground motion at the site. There are two basic steps to estimate the ground motion at the site.

- i Identification of the important characteristics of the ground motion (i.e. the intensity measures, see Section 2.6). The selection of the IM of interest (amplitude, frequency and/or duration-based) depends on the element at risk under consideration, the effects being considered in the analysis and approach that is followed for the derivation of fragility curves. Several concepts and quantities are commonly used to assist in identifying the optimum intensity measure for the required purpose. These are defined as practicality, effectiveness, efficiency, sufficiency and robustness (Cornell et al., 2002).
- ii Estimation of the probability distribution of the selected IM must be estimated as a function of predictor variables such as the earthquake source properties, the relative location of the earthquake respect to the site, and the soil conditions.

Ground motion prediction equations (GMPEs) are usually adopted to evaluate the probability that a particular IM exceeds a certain value, im , for a given earthquake $M = m$, occurring at a given distance, $R = r$, (as illustrated graphically in Figure 16). In probabilistic terms, this is written as $P[IM > im|R = r, M = m] = 1 - F_{IM|Rm}(im|r, m)$.

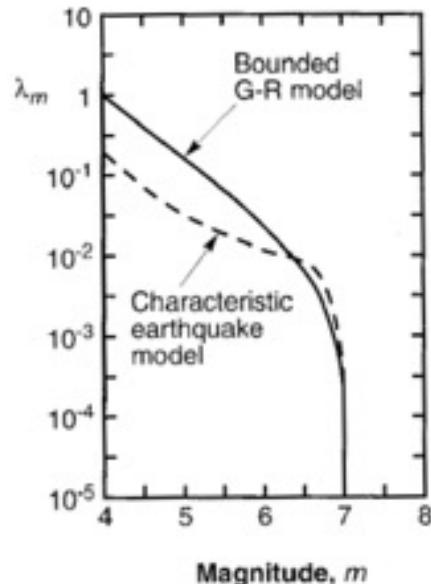


Figure 15: Comparison of recurrence laws from Characteristic and Bounded G-R magnitude models (Kramer, 1996).

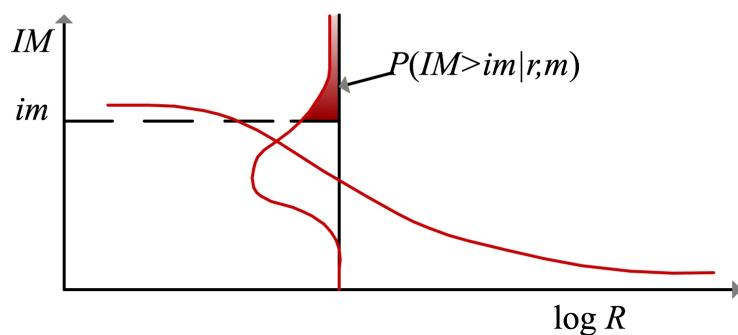


Figure 16: Schematic illustration of conditional probability of exceeding a particular value of a ground motion parameter for a given magnitude and distance.

In general, the conditional distribution of the ground motion intensity measure, i.e. $F(im|r, m)$ is assumed log normal. It follows that the logarithm of the conditional intensity measure is normally distributed. Then,

$$P[IM > im|R = r, M = m] = \int_{im}^{\infty} f_{IM|Rm}(u)du = \frac{1}{\sigma_T \sqrt{2\pi}} \int_{im}^{\infty} \exp \left[-\frac{(\ln u - \bar{\ln IM})^2}{2\sigma_T^2} \right] du \quad (31)$$

where $\bar{\ln IM}$ and σ_T are the mean value and the standard deviation of $\ln Y$ (See Section 2.6.1).

Important considerations

- Over decades of development, the prediction models have become complex, consisting of many parameters (see Section 2.6.1). More specifically, many factors related to source and site characteristics are considered. One of the most fundamental is the tectonic region in which the source is located. Typically, GMPEs are developed independently for each region (active, subduction and stable continental zones). Another important factor is the faulting mechanism that influences in particular amplitude and frequency content. Effects of local site conditions are represented in many forms, ranging from a simple constant to more complex terms that try to characterize non-linearity in ground response. Many recent works analyzed also the importance of near-fault effects (typically defined within 20-60 km of fault rupture), concluding that ground motion at near source sites are sensitive to the effect of the “rupture directivity” that affect duration and long period energy.
- As mentioned in Section 2.6, empirical or theoretical methods can be applied to estimate the ground motion. Stochastic methods are used to generate ground motions when there is insufficient amount of ground motion recordings to develop empirically-based equations. These methods are commonly used in stable tectonic regions and for high frequency motions characterized by a large magnitude and short source-to-site distance. Standard stochastic methods are based on the assumption that the far-field shear wave energy generated by an earthquake source can be represented as a band-limited white Gaussian noise random process.

3.4 Hazard computation

3.4.1 Seismic hazard curve

The seismic hazard curve is a function representing the annual frequency of exceeding various levels of ground shaking (i.e. the IM) at a specific site. The curve is obtained by integration of the first three steps over all possible magnitudes and earthquakes locations.

Seismic hazard curves are obtained for individual sources and, then, combined to express the aggregate hazard at a particular site. The aggregation is given after selecting a stochastic model for the occurrence of earthquakes in time. It is common to use the HPP model (Section 6.2 on Fundamentals of Probability and the following Section 3.4.2). For each source the rate of exceedance of a given im is derived as Poisson with random selection (Section 6.2.1 Fundamentals of Probability). In particular, given a rate of minimum earthquake of a source n , i.e. $\lambda_{min}^{(n)}$, the rate of the event $IM > im$, $\lambda(im)$, is simply given by $\lambda^{(n)}(im) = \lambda_{min}^{(n)} P(IM > im | M \geq m_{min})$, where $P(IM > im | M \geq m_{min})$ is obtained by (12). Observe that here, for the sake of simplicity in the notation, $P(IM > im | M \geq m_{min})$ is equivalent to $P(IM > im)$.

Given N_s statistically independent sources, the aggregate hazard curve is derived by merging all the N_s Poisson processes. Merging N_s statistically independent Poisson processes (Section 6.2.2 Fundamentals of Probability) is straightforward since it involves only the sum of the N_s different rates. Then, we can write

$$\lambda(im) = \sum_{n=1}^{N_s} \lambda_{min}^{(n)} \left[\int_{r_{min}}^{r_{max}} \int_{m_{min}}^{m_{max}} P(IM > im | M = m, R = r) f_{R|M}^{(n)}(r|m) f_M^{(n)}(m) dr dm \right], \quad (32)$$

or in case of $M \perp R$

$$\lambda(im) = \sum_{n=1}^{N_s} \lambda_{min}^{(n)} \left[\int_{r_{min}}^{r_{max}} \int_{m_{min}}^{m_{max}} P(IM > im | M = m, R = r) f_R^{(n)}(r) f_M^{(n)}(m) dr dm \right]. \quad (33)$$

As the individual components of equations (32) and (33) are complicated, the integral can be evaluated analytically only for some simple situations (Cornell, 1968). Then, numerical integration is required. Among a variety of different techniques, the most simple and used approach is to divide the possible ranges of magnitude and distance into N_m and N_r segments, respectively. It follows that (33), for example, can be estimated as

$$\lambda(im) \approx \sum_{n=1}^{N_s} \sum_{m=1}^{N_r} \sum_{l=1}^{N_m} \lambda_{min}^{(n)} P(IM > im | M^{(n)} = m_l, R^{(n)} = r_m) f_R^{(n)}(r_m) f_M^{(n)}(m_l) \Delta r \Delta m, \quad (34)$$

$$\approx \sum_{n=1}^{N_s} \sum_{m=1}^{N_r} \sum_{l=1}^{N_m} \lambda_{min}^{(n)} P(IM > im | M^{(n)} = m_l, R^{(n)} = r_m) P(M^{(n)} = m_l) P(R^{(n)} = r_l). \quad (35)$$

3.4.2 Earthquake Occurrence Model

Most applications of PSHA are based on the assumption that the earthquake process is memory-less, that is, there is no memory of the time, size and location of preceding events. Moreover, earthquake can be considered instantaneous event with respect to the time periods of engineering interest. We have seen that these assumptions are typically encompassed in a HPP process with rate λ (number of events in the unit time, that in the ordinary seismic application is the year) and exponential inter-arrival intervals (Section 6.2 on Fundamentals of Probability).

In particular, in the context of seismic hazard, HPPs possess the following properties:

- i The probability of more than one occurrence during a very short time interval is negligible.
- ii The number of earthquakes which occur in disjoint time intervals are independent (independent increments).
- iii The number of earthquake events in any interval of length t is Poisson distributed, $P(N(t+s) - N(s) = n) = \frac{[\lambda(m)t]^n}{n!} \exp[-\lambda(m)t]$, with mean equal to $\lambda(m)t$ (stationary increments), where $\lambda(m)$ is the rate of earthquake with magnitude greater than m .

We have seen that the exponential inter-arrival time distribution has mean equal to $E[T] = 1/\lambda(m)$. This physically represents the mean time in which the next event is expected. Given this, the mean of the exponential is also defined as return period, i.e. $T_R = 1/\lambda(m)$. As a consequence of the memoryless property, T_R is a constant value during the whole observed time interval (50 years or one day after a seismic event, Poisson T_R is always equal to $1/\lambda(m)$). In other words all the information about the history of past earthquakes has no influence on the distribution of expected future events.

In traditional PSHA, the occurrence probability of *at least one* event in the observed time interval is of concern. It can be computed as:

$$P(N \geq 1, t) = 1 - P(N = 0, t) = 1 - \exp[-\lambda(m)t] \quad (36)$$

Thus relationship between occurrence probability of at least one event and return period is:

$$\begin{aligned} P(N \geq 1, t) &= 1 - \exp\left(-\frac{1}{T_R} t\right) \rightarrow \\ T_R &= -\frac{t}{\ln[1 - P(N \geq 1)]}. \end{aligned} \quad (37)$$

For example, the 10% occurrence probability of at least an event in a considered period of $t = 50$ [years], corresponds to a return period of $T_R \approx 475$ [years].

Once the Poisson process is assumed, earthquake occurrence is completely defined by the mean annual number of earthquake occurrence $\lambda(m)$. As mentioned in the previous section, this allows two operations

- i Poisson with random selection (known also as “filtering a Poisson process”). Section 6.2.1 Fundamentals of Probability
- ii Combining Poisson processes (known also as “summing or merging Poisson processes”). Section 6.2.2 Fundamentals of Probability

Given the seismic source n the first operation is used to compute the rate of $IM > im$, i.e.

$$\lambda^{(n)}(im) = \lambda_{min}^{(n)} P(IM > im | M > m_{min}) = \lambda_{min}^{(n)} P(IM > im) \quad (38)$$

$$= \lambda_{min}^{(n)} \left[\int_{r_{min}}^{r_{max}} \int_{m_{min}}^{m_{max}} P(IM > im | M = m, R = r) f_{R|M}^{(n)}(r|m) f_M^{(n)}(m) dr dm \right] \quad (39)$$

The second operation is used to aggregate all the N_s sources, i.e.

$$\lambda(im) = \sum_{n=1}^{N_s} \lambda_{min}^{(n)} P(IM > im) \quad (40)$$

to obtain either the (32) or the (33).

Finally, the aggregate probability of at least one exceedance of im in a period t years is equal to:

$$P(\text{at least one event } IM > im, t) = F_t(im) = 1 - \exp[-\lambda(im)t]. \quad (41)$$

Given a time $t = 1[\text{year}]$ and a specific site, the (41) represents the annual exceedance probability for earthquake of $IM > im$. The (41) is known also as the hazard curve. In the case $t = 1[\text{year}]$ and for small value of λ_{im} the frequencies can be regarded as probabilities, since $1 - \exp[-\lambda(im)t] \approx \lambda(im)$.

Introducing T_R , we obtain

$$\begin{aligned} F_t(im) &= 1 - \exp\left[-\frac{1}{T_R}t\right] \rightarrow \\ T_R &= -\frac{t}{\ln[1 - F_t(im)]}. \end{aligned} \quad (42)$$

3.5 Uncertainty in PSHA

The definition and treatment of uncertainties represents an important aspect of PSHA. This involves identifying the intrinsic randomness of the modelled phenomenon, defined as *aleatory uncertainty* as well as the uncertainty related to our limited knowledge and data of the modelled phenomenon, defined as *epistemic uncertainty*.

The distinction between these two types of uncertainties is fundamental to understand where uncertainty originates and how it may be appropriately handled in hazard calculations.

The aleatory variability in a seismic hazard analysis is included directly in the calculations, specifically through the probability distributions of the parameters, and thus it directly influences the results. Epistemic uncertainty is related to the subjective decisions that are made as part of the process of carrying out the analysis, e.g.:

- i The definition of the geometry of the sources.
- ii The assumptions regarding the completeness of the seismic catalogue
- iii The definition of m_{max}
- iv The selection of the GMPE

Since epistemic uncertainty is characterized by the use of alternative models (i.e. alternative probability distributions) it is not considered directly in the hazard calculations but rather is treated by developing alternative hazard curves.

A common way to handle epistemic uncertainty is through the use of *Logic Trees*. This methods allows the use of different models, each of which is assigned a weighting factor that represents the likelihood of that model being correct. Representing current scientific judgment on the merit of the alternative models, the weights are based on data collected from analogous regions, simplified physical models, and empirical observations. The use of the logic tree is to calculate the hazard following every branch. The result of each analysis is finally weighted by the relative likelihood of its combination of branches, taken as the product of all the weightings along all the branches.

The logic tree shown in Figure 17 was used in the construction of the seismic hazard map of Italy (Gruppo di Lavoro, 2004). It considers the uncertainties related to the completeness of the catalogue of earthquakes, the method for calculating seismicity rates, and to the attenuation pattern. All these alternative options lead to a logic tree with $2 \times 2 \times 4 = 16$ branches.

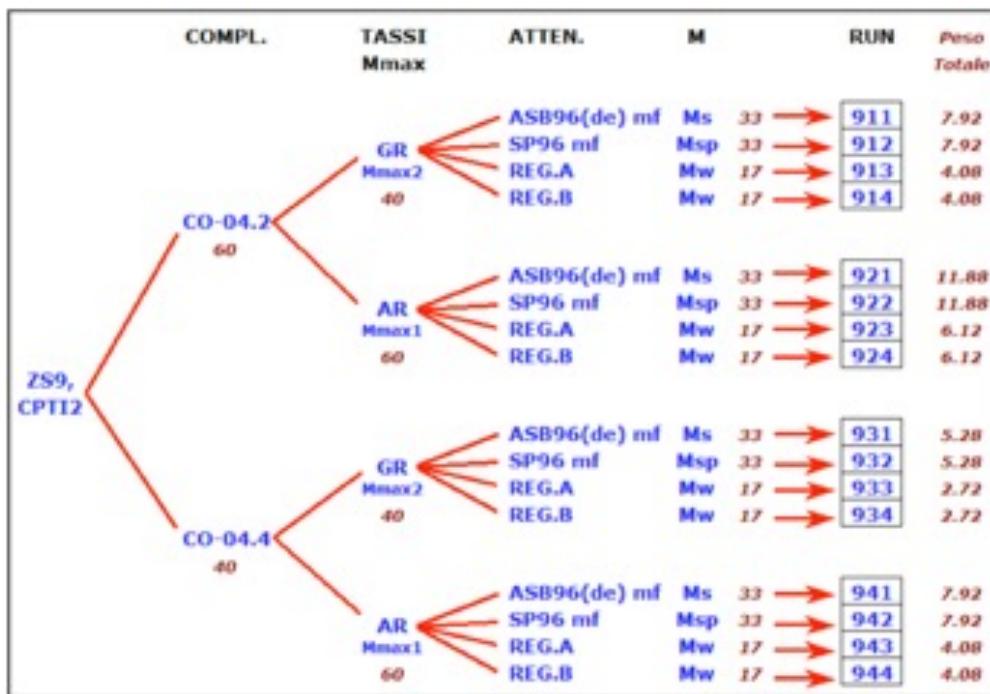


Figure 17: The logic tree used for the computation of the Italian seismic hazard map (Gruppo di Lavoro, 2004).

References

- Baker J. W. (2008). An Introduction to Probabilistic Seismic Hazard Analysis (PSHA), White Paper, Version 1.3, 72 pp.
- Cornell, C.A. (1968). Engineering seismic risk analysis, Bull. Seism. Soc. Am., 58, 1583-1606.
- Cornell CA, Jalayer F, Hamburger RO, Foutch DA (2002) Probabilistic basis for 2000 SAC/FEMA steel moment frame guidelines. J. Struct. Eng., 1284, 26?533.
- Gardner J.K. and Knopoff, L. (1974) Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian? Bull. Seism. Soc. Am., 64, 1363-1367.
- Gruppo di Lavoro (2004) Redazione della mappa di pericolosità sismica prevista dall'Ordinanza PCM 3274 del 20 marzo 2003. Rapporto conclusivo per il Dipartimento della Protezione Civile, INGV, Milano - Roma, 65 pp. + 5 App.
- Gutenberg, B. and Richter C.F.(1942) Earthquake Magnitude, Intensity, Energy and Acceleration, Bull. Seism. Soc. Am., 32, 163-191.
- Kramer, S.L. (1996) Geotechnical earthquake engineering. Prentice Hall, Upper Saddle River, N.J.
- McGuire, R.K. (2004) Seismic Hazard and Risk Analysis, Earthquake Engineering Research Institute Publication, Report MNO-10, Oakland, CA, USA
- Wells, D.L. and Coppersmith, K.J. (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. Bull. Seism. Soc. Am., 84, 974-1002.
- Youngs R.R. and Coppersmith K.J.(1985) Implications of fault slip rates and earthquake recurrence models to probabilistic seismic hazard assessments. Bull. Seism. Soc. Am., 75, 939-964.



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



LECTURE NOTES

Fundamentals of Probabilistic Seismic Hazard Analysis (PSHA)

Chapter 2: Lecture 5

Spring 2018

Written by simona esposito
marco broccardo
IBK, ETHZ

PART III
Principles of Disaggregation and Ground Motions
Selection

4 Introduction

In the Section we outline two key elements for selecting ground motions for time history dynamic analysis of a given structure. In fact, one pillar for dynamic analysis is the input selection, which is, oftentimes, a recorded ground motion. Selection of the proper set of ground motions for a given region and level of hazard is still an open research field and source of debate. In this section, we first outline the definition of the Uniform Hazard Spectrum (UHS), followed by the concept of disaggregation, and lastly, we introduce the most used practice for ground motions selection. These concepts are central steps for correctly representing the hazard of a given region.

5 Presentation of PSHA results

Results of a PSHA analysis can be presented in different ways. The basic output of a PSHA analysis for an individual site is the annual rate of exceedance of a particular IM expressed either as rate of exceedance or return period.

As an example, Figure 18 shows the annual frequency of exceedance of PGA (unit of g), $\lambda(pga)$, for three cities in Italy (Campobasso, Napoli, and Bari) provided by the Istituto Nazionale di Geofisica e Vulcanologia (INGV). In particular, the figure shows the distribution of the 50th percentile (median map, which is the reference map for every probability of exceedance) and the distribution of the 16th and 84th percentiles. These curves are the result of the use of the Logic Tree outlined in Figure 17 for the quantification of the epistemic uncertainties. A seismic hazard map shows the variation in seismic

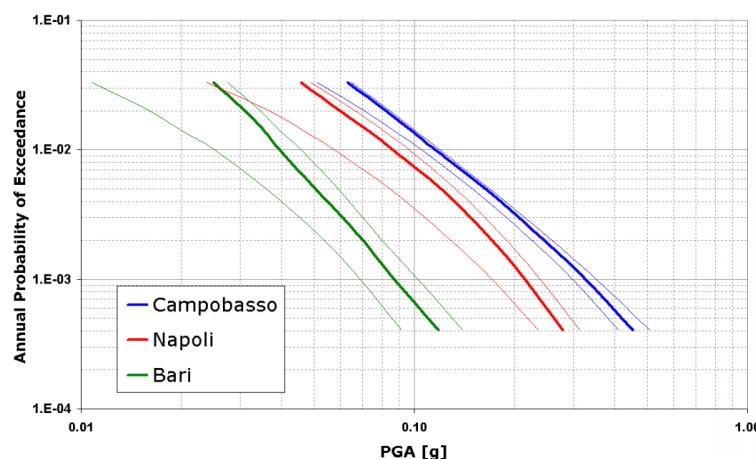


Figure 18: Seismic hazard curves for three sites in Italy. Tick lines represent median values, thin lines represent the 16th and 84th percentile of the epistemic uncertainties. (Source: <http://esse1.mi.ingv.it/data/D2.pdf>, Meletti and Montaldo, 2007).

hazard over a particular region or country. A hazard map is produced by performing hazard analysis at a large number of sites within the region under study (the task of LAB 1). A hazard curve is evaluated for each site and then the IM level determined at each point for the desired probability of exceedance and for a specific time interval is used to produce the map. Hazard contours are then drawn through the resulting values at the nodes to obtain IM level curve. Seismic hazard maps are available in literature for most areas of the world. They are usually updated as new earthquakes occur and new data are available in order to reflect the improved knowledge. As an example, Figure 19 shows the 475-year hazard map in terms of PGA (unit of [g]) for the Italian territory provided by INGV.

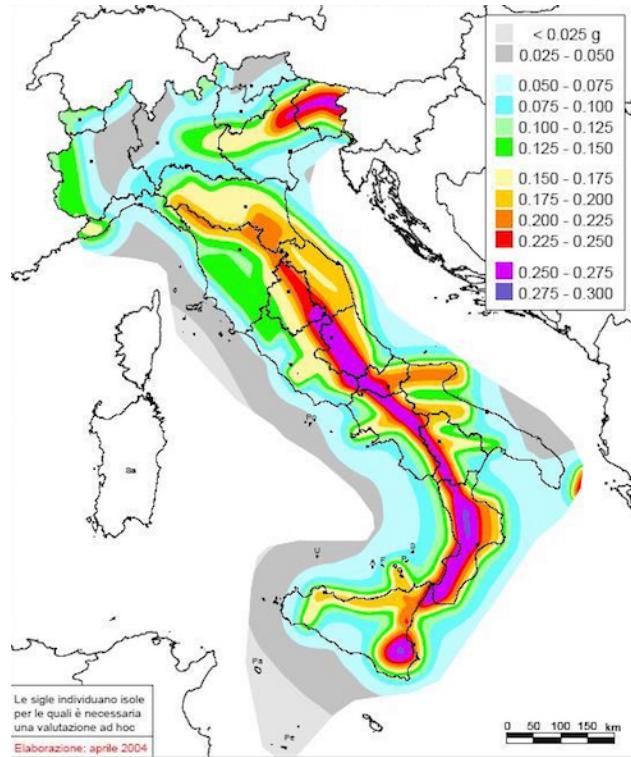


Figure 19: Seismic hazard map of the Italian national territory in terms of PGA (unit of g) considering a return period of 475 years (source: <http://esse1.mi.ingv.it>).

6 Uniform Hazard Spectrum

Response spectrum represents one of the most widely used representation of seismic actions in earthquake-resistant design and in response analysis of buildings. There are two different ways to obtain a response spectrum.

The simplest approach, is to use the output of a PSHA analysis in terms of PGA to anchor a standard spectral shape (usually the spectrum defined by the code) at zero period, selected according to a specific soil type. The limitation of this approach is that it does not account for the variation of the shape of the spectrum with earthquake's features such as magnitude and distance. As consequence the return period of the ordinates of the spectrum is known only at a period of zero (i.e. for the PGA) while at other response periods the return period will be longer or shorter. Further, the shape of the spectrum does not change with the hazard level but only with the soil type. Eurocode 8 (CEN, 2003) tried to overcome this problem by using two spectral shapes, one for regions affected by relatively low magnitude earthquakes, the other for areas with larger earthquakes.

The Uniform Hazard Spectrum (UHS) instead, is characterized by a uniform hazard level for all the ordinates of the spectrum. The UHS is developed by first performing the PSHA many times using period-dependent GMPEs for response spectral ordinates. Then, for each period the spectral amplitude corresponding to a target rate of exceedance is identified. Those spectral values are then plotted versus the structural periods. In this manner the spectrum represents the same level of hazard across the entire range of periods. The process of constructing a UHS is illustrated by Figure 20. Figure 21 shows, instead, the UHS of the same site considered in Figure 20 for different probabilities of exceedance in 50 years (source, INGV). This spectrum is called a uniform hazard spectrum because every ordinate is characterized by an equal rate of being exceeded. However it is apparent, from their computational procedures, that UHS accounts for the occurrence of many different seismic events and then each value of the spectrum may have come from a different event. Then, it is clear that UHS cannot have a shape similar to any actual recorded signal. If, for example, the seismicity of a site is characterized by nearby sources of moderate

magnitude earthquakes and distant sources with larger events, it is possible that different portions of the UHS will actually correspond to different type of earthquakes. This characteristic has important implications for the cases in which it is necessary to represent the seismic actions in terms of acceleration time-histories.

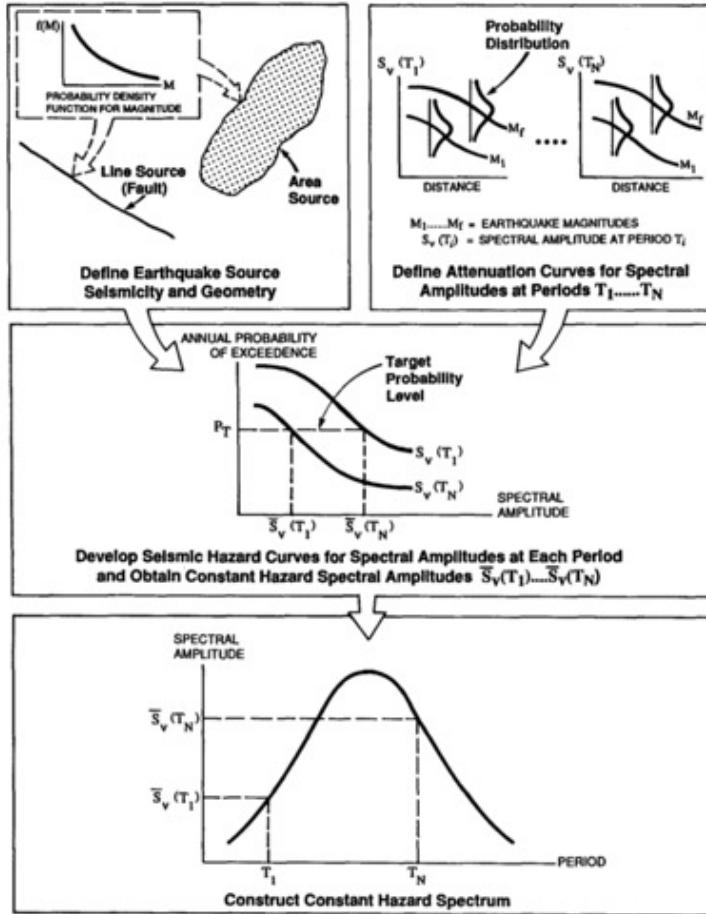


Figure 20: Illustration of the process of constructing a UHS as the output from a PSHA (EERI, 1989).

7 Disaggregation

Seismic hazard analysis leads to the hazard curve expressed in Eq. (32), (33) (Lecture Note 4) and shown in Figure 18. In some cases, it can be necessary to know the relative contributions to the hazard from different values of the random components of the problem, i.e. the magnitude, M , the source-to-site distance, R , and sometimes, ε , the measure of the deviation of the ground motion from the value predicted by the GMPE (Lecture Note 4, Eq. (8)).

These quantities may be used, for example, to select ground motion records for response analysis. In this cases, the required analytical instrument is called *disaggregation* (or deaggregation) of seismic hazard (Bazzurro and Cornell, 1999). Disaggregation is a procedure which allows identification of the hazard contribution of each $\{M, R, \varepsilon\}$ vector conditional to the exceedance of the hazard itself. The analytical result of disaggregation is the joint probability density function of $\{M, R, \varepsilon\}$ conditional to the exceedance

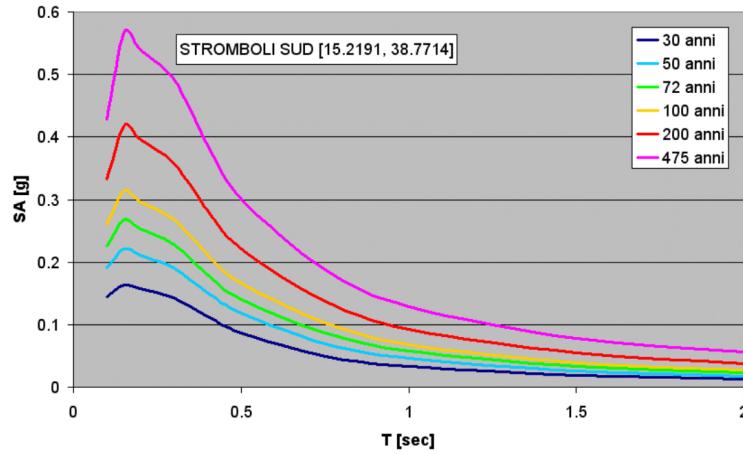


Figure 21: UHS for a site in Italy for different probabilities of exceedance in 50 years (source: <http://esse1.mi.ingv.it>, Montaldo and Meletti, 2007).

of an IM threshold (im^*), as expressed in equation (43),

$$f_{M,R,\epsilon}(m, r, \epsilon | IM > im^*) = \frac{\sum_{n=1}^{N_s} \lambda_{min}^{(n)} P[IM > im^* | m, r, \epsilon] f_{M,R,\epsilon}(m, r, \epsilon)}{\lambda(im^*)}, \quad (43)$$

where N_s is the number of seismic sources relevant for the hazard at the site, $f_{M,R,\epsilon}(m, r, \epsilon)$ is the joint PDF of $\{M, R, \epsilon\}$ and $\lambda(im^*)$ is the hazard for im^* . Practically, the disaggregation of the hazard from all N_s sources is obtained by accumulating in each $\{M, R, \epsilon\}$ bin the contribution to the global hazard, during the numerical integration of Eq. 32, (i.e. equations 34-35, Lecture Note 5).

Equation (43) computes the relative contribution to the hazard originated by a specific characterization of the seismicity. When epistemic uncertainty is of concern, in principle, it is possible to disaggregate the hazard from each considered seismicity model (i.e. for each brunch of a Logic Tree). In realistic cases, this could be impractical, due to the very large number of cases considered. Then, the hazard values usually considered correspond to the 50th percentile of the hazard distribution obtained by using a specific logic tree. Thus, the disaggregation is performed using the inputs along the logic tree path that provided hazard values closest to the reference 50th percentile hazard.

As an example, Figure 22 shows the disaggregation of median PGA for a site in Italy (Spallarossa and Barani, 2007), in terms of magnitude and distance, with a probability of exceedance of 10% in 50 years, provided by the Istituto Nazionale di Geofisica e Vulcanologia (INGV). Note that in this case, the disaggregation has been applied using the inputs along the logic tree path that provided hazard values closest to the reference 50th percentile hazard.

7.1 Magnitude disaggregation: an example

In case of magnitude disaggregation, we might be interested to calculate the probability of having an earthquake magnitude equal to m , given that the ground motion $IM > im^*$ has occurred, i.e.:

$$\begin{aligned} P(M = m | IM > im^*) &= \\ &= \frac{\lambda(M > im^* | M = m) P(M = m)}{\lambda(im^*)} \\ &= \frac{\sum_{n=1}^{N_s} \sum_{m=1}^{N_r} \lambda_{min}^{(n)} P[IM > im^* | M^{(n)} = m^{(n)}, R^{(n)} = r_m^{(n)}] P(M^{(n)} = m^{(n)}) P(R^{(n)} = r^{(n)})}{\lambda(im^*)} \end{aligned} \quad (44)$$

This is practically equal to the rate of earthquakes with $IM > im^*$ and $M = m$, divided by the rate of all earthquakes with $IM > im^*$, i.e. $\lambda(im^*)$.

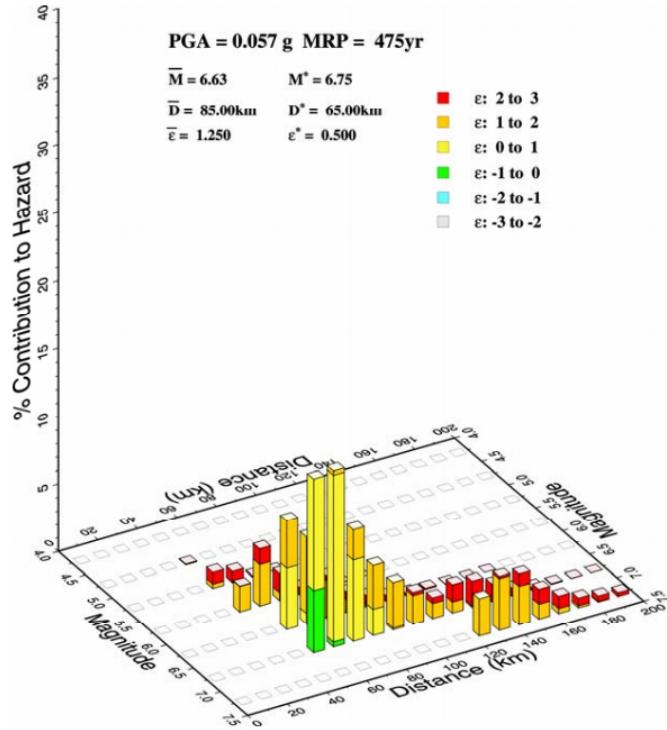


Figure 22: Disaggregation of median PGA with probability of exceedance of 10% in 50 years (source: <http://esse1.mi.ingv.it/>, Spallarossa and Barani, 2007).

Example

(Adapted from Baker). Compute the relative contributions of two linear faults to $\lambda(PGA) > 1[g]$. Fault 1 produces earthquakes with magnitude $M = 6.5$ and source to site distance (closest distance) $r_1 = 10[\text{km}]$ from the site; and has an annual occurrence rate of 0.01; Fault 2 produces earthquakes with magnitude $M = 7.5$ and source to site distance (closest distance) $r_2 = 20[\text{km}]$ from the site, and it has annual occurrence rate of 0.002.. Events will rupture all the fault (the distance is a constant). The used GMPE is (Cornell et al., 1979):

$$\begin{aligned} \overline{\ln PGA} &= -0.152 + 0.859M - 1.803 \ln(R + 25) \\ \sigma &= 0.57 \end{aligned} \quad (45)$$

From Eq. (33) (hazard integral) considering $\lambda_{min}^{(1)} = 0.01$ and $\lambda_{min}^{(2)} = 0.002$, we obtain

$$\lambda(PGA > 1[g]) = 0.01P[PGA > 1[g]|6.5, 10] + 0.002P[PGA > 1[g]|7.5, 20], \quad (46)$$

where

$$P[PGA > 1[g]|6.5, 10] = 1 - \Phi\left(\frac{\ln 1 - \ln(0.3758)}{0.57}\right) = 1 - \Phi(1.72) = 0.043 \quad (47)$$

$$P[PGA > 1[g]|7.5, 20] = 1 - \Phi\left(\frac{\ln 1 - \ln(0.5639)}{0.57}\right) = 1 - \Phi(1.01) = 0.158. \quad (48)$$

Finally $\lambda(PGA > 1[g]) = 0.01(0.043) + 0.002(0.158) = 0.000746$

Magnitude Disaggregation

The relative contributions of the two linear faults to exceedance of $PGA = 1[g]$ can be computed as following:

$$P[M = 6.5 | PGA > 1[g]] = \frac{0.01P[PGA > 1[g]|6.5, 10]}{\lambda(PGA > 1[g])} = 0.58 \quad (49)$$

$$P[M = 7.5 | PGA > 1[g]] = \frac{0.002P[PGA > 1[g]|7.5, 20]}{\lambda(PGA > 1[g])} = 0.42. \quad (50)$$

Then the more active but smaller source (Fault 1) makes a greater contribution to exceedance of the $PGA = 1[g]$.

7.2 Design earthquakes and engineering applications

Historically, mean values of M , R , and ε of the disaggregation, $\{\bar{M}, \bar{R}, \bar{\varepsilon}\}$ has been the most popular contender for the role of defining the dominant event, i.e. the design earthquake. Today modal values are preferred because respect to the mean values, they corresponds to a realistic scenario. A Design Earthquake (DE), defined as mode of the disaggregation joint PDF (equation (43)), is the vector $\{M^*, R^*, \varepsilon^*\}$ that gives the largest contribution to the hazard (i.e. to the exceedance of the $IM = im$ threshold corresponding to the considered return period).

As an example, Figure 23 shows maps of DEs in terms of mean $(\bar{M}, \bar{R}, \bar{\varepsilon})$ and modal $(M^*, R^*, \varepsilon^*)$ values as result of the disaggregation of the median PGA with probability of exceedance of 10% in 50 years computed for the Italian territory (source INGV).

In some cases, the use of single statistics, such as the mean or the mode is not sufficient to describe the characteristics of the ground motions that most likely to threaten the site. For example, disaggregation of the joint PDF may show more than a single mode that significantly contribute to the hazard. In this case a second mode, i.e. a second DE is defined as the second relative maximum of $f_{M,R,\varepsilon}(m, r, \varepsilon | IM > im^*)$ distribution.

Design earthquakes represent nowadays the preliminary criterion for record selection for assessing the nonlinear demand of structures. The current state of practices is based on first disaggregating the seismic hazard at the site for the level of spectral acceleration (at the first mode period of the structure) at a specified probability (e.g. 10% in 50 years). Then records are selected to match within tolerable limits the mean or modal value of the M and R (from a disaggregation analysis) and other source, path and site characteristics such as, style of faulting, site conditions etc. Finally the selected records are scaled to match, in some average sense, a target spectrum (e.g. code spectrum or UHS).

Another possible use of design earthquakes is the possibility to build hazard curves for secondary intensity measures conditional to design value of primary intensity measure for which a hazard curve is available by national authorities. In fact, in case a secondary IM is required to improve the estimation of the structural response and hazard curves are not available for this IM , conditional hazard maps can be built computing the probabilistic distribution for the secondary IM conditional to the design value of the primary IM . An example of this application is reported in Iervolino et al., 2011.

For further details about disaggregation procedures and applications, the reader is addressed to (Bazzurro and Cornell, 1999 and Iervolino et al., 2011).

8 Record Selection for Dynamic Response Analysis

The most complete analysis for the safety level of existing civil structures requires nonlinear dynamic analysis. This type of analysis requires a detailed modeling of the structure and a proper selection and scaling of seismic input. The main steps for record selection and scaling are summarized below and shown in Figure 25 (Danciu and Fah, 2016):

1. Define the target spectrum for the site and the limit state of interest
 - Usual target spectrum are: design-code spectrum, UHS, and Conditional Spectrum (Baker, 2010).

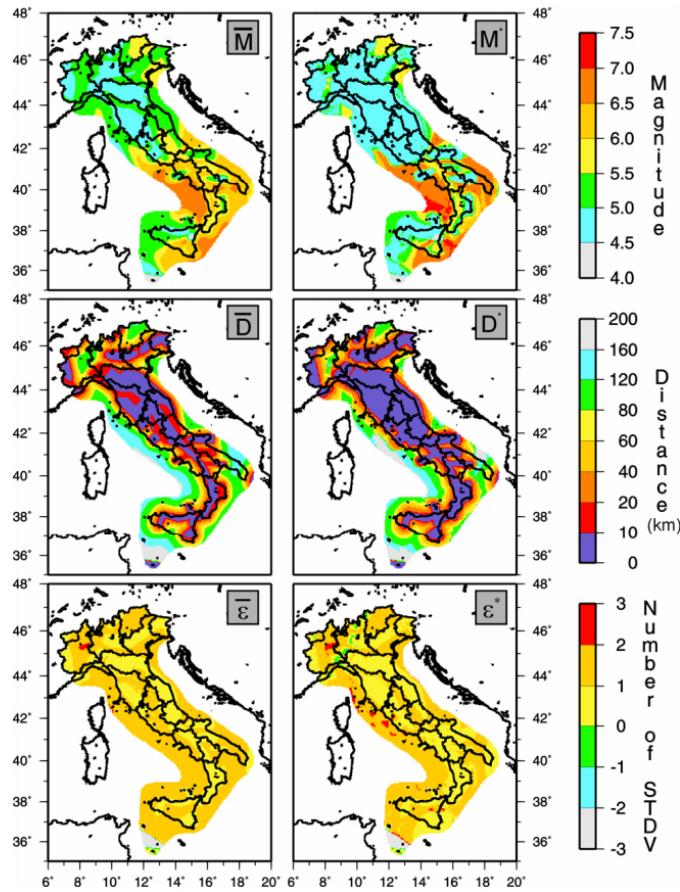


Figure 23: Mean (\bar{M} , \bar{R} , $\bar{\varepsilon}$) and modal (M^* , R^* , ε^*) values of the disaggregation of median PGA with probability of exceedance of 10% in 50 years (source: <http://esse1.mi.ingv.it/>, Spallarossa and Barani, 2007).

2. Assessment of the dominant earthquake characteristics (e.g. magnitude, distance etc.) which dominate the hazard for the spectral ordinate of interest.
 - From seismic hazard disaggregation (if available), otherwise from experience.
3. Choose a final reduced subset of ground motions matching the target spectrum T^* (or over a range of periods)
 - Ground motion signals can be real (better), synthetic, or artificial. We suggest to use artificial earthquake (i.e., the one derived by stochastic process) to augments a given subset rather than be used in substitution of real ground motions.
4. Modify the ground motion so that its response spectrum more closely matches the design spectrum (this is a conventional but controversial practice).
 - Amplitude based, or response spectrum (frequency-domain or by time-domain) modification methods.

8.1 Selection

Generally, the signals that can be used for the seismic structural analysis are of three types: artificial waveforms; simulated accelerograms (synthetic); and real (natural) records. Many codes worldwide, such

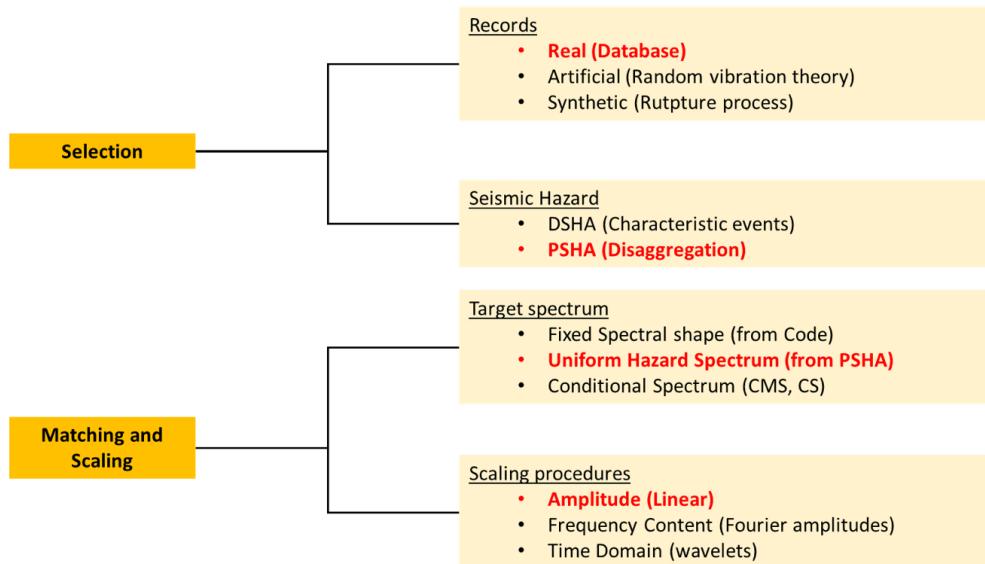


Figure 24: Overview of the main steps of record selection and matching for dynamic seismic analysis of structures. (source: Danciu and Fah, 2016).

as Eurocode 8 (EC8), allows the use of all the three types of signals listed above. Artificial signals are usually obtained generating a Power Spectral Density (PSD) from the code response spectrum, and deriving signals compatible to that. Simulated accelerograms are obtained via modelling of the seismological source and may account for path and site effects. The last type are ground-motion records from real events. There are many on-line, user friendly databases of strong-motion recording; however an issue regarding the use of these signals, whose spectra are generally non-smoothed, is the selection of a set compatible with a code-specified spectrum. The basic criteria for selecting ground motions are formulated on basis of earthquake source parameters of a design scenario defined either deterministically (through DSHA) or probabilistically (through PSHA). Design scenario is usually defined in terms of : a) source parameters, i.e. magnitude range of significant event(s), focal depth, style of faulting, directivity; b) Path: distance range of the site from the causative fault(s), fault azimuth, geometrical spreading and attenuation; c) Site: surface geology (generally described by average shear-wave velocity), topography.

Probabilistic methods require hazard disaggregation, i.e seismic hazard is first disaggregated at the site (Bazzurro and Cornell, 1999), by causative magnitude (M) and distance (R), for the level of spectral acceleration (at the first mode period of the structure, T^*) at a specified probability (e.g. 10% chance of exceedance in 50 years). The records are then chosen to match within tolerable limits the mean or modal value of the M and R and site conditions, i.e., the expected value or most likely value of these characteristics given that exceedance.

8.2 Matching

The selected set of records has then to match a predefined target response spectrum, which represents the link between the structure and the ground motion. The target spectrum can be defined by i) a design spectrum as specified by seismic design code provisions, ii) uniform hazard spectrum resulting from seismic hazard assessment and more recently by iii) a conditional mean spectrum. Design code spectrum is the simplest target spectrum and it is provided by seismic design codes and reference standards. Design spectra are generally determined by smoothing or enveloping multiple earthquake response spectra. Thus, the design spectra are typically conservative as they envelope spectral ordinates of equal probabilities of exceedance. The UHS, as introduced above, is generated by a PSHA and represents the response spectra for a specified probability of exceedance (or mean return period). During the last years, the ground motion epsilon, ε , has been used as an indicator of the spectral shape, which lead to the development

of the Conditional Mean Spectrum (CMS) as target response spectrum. The CMS is a spectrum that considers ε at a single conditioning period and the correlation of spectral ordinates at different periods to compute the conditional mean and standard deviation of the spectral ordinates at all other periods, given the conditioning period. The conditioning period is usually chosen as the fundamental (1st mode) period of the structure, although other periods could be used.

8.3 Scaling

When the number of eligible records selected is not reasonable to adequately evaluate the structural response, scaling methods are applied¹. The most common scaling procedures are amplitude-matching and spectrum-matching techniques. Spectrum matching is done either by modifying spectral ordinates in the frequency domain (FD) or adding (subtracting) wavelets in the time domain (TD). In both ways, the spectral shape is modified to match the target spectrum. The amplitude-matching procedure (Figure 8) linearly scales the amplitude of the ground motions to match the target spectrum. To this aim, a scale factor (SF) has to be determined to scale the ground motions to the design response values. The match

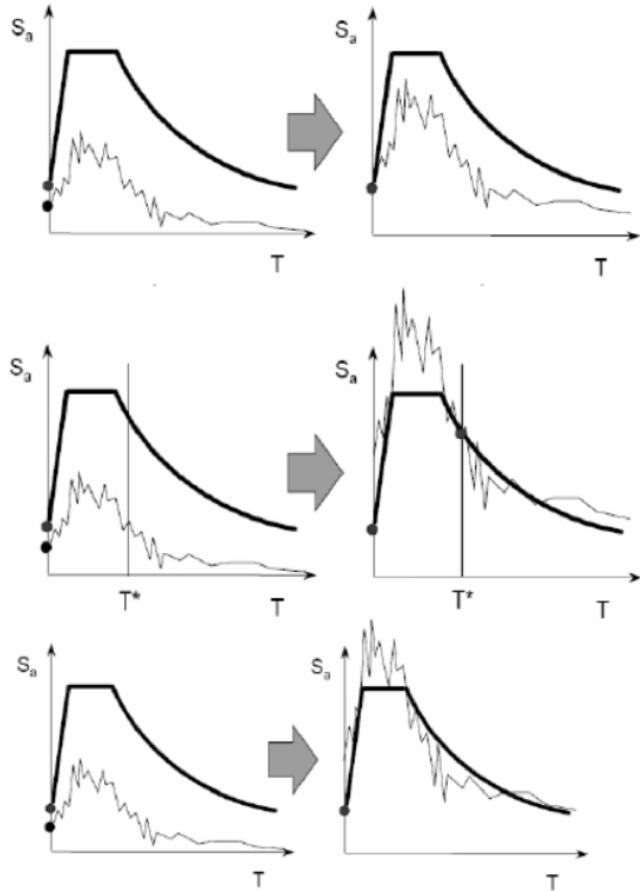


Figure 25: Amplitude-based scaling: a) based on PGA; b) based on T^* ; c) based over a range of periods $T_1 - T_2$. (Source)

can be based on (a) PGA, (b) the first fundamental period of the structure T^* or (c) minimizing the

¹Alternatively artificial ground motions can be used

misfit (least square error) over a range of periods $T_1 - T_2$, as expressed in the following equation:

$$r^2 = \int_{T_1}^{T_2} [S_a(T)_{design} - SFAa(T)_{unscaled}]^2 dt \quad (51)$$

$$r^2 = \sum_{i=1}^N [S_a(T_i)_{design} - SFAa(T_i)_{unscaled}]^2 \quad (52)$$

$$SF = \frac{\sum_{i=1}^N [S_a(T_i)_{design} - SFAa(T_i)_{unscaled}]^2}{\sum_{i=1}^N [SFAa(T_i)_{unscaled}]^2} \quad (53)$$

References

- Baker, J. W. (2008). An introduction to probabilistic seismic hazard analysis. Report for the US Nuclear Regulatory Commission, page Version, 1.
- Baker, J. W. (2010). Conditional mean spectrum: Tool for ground-motion selection. *Journal of Structural Engineering*, 137(3), 322-331.
- Bazzurro, P., Cornell, C.A. (1999) Disaggregation of Seismic Hazard, *Bulletin of Seismological Society of America*, 89: 501–520.
- CEN, European Committee for Standardisation TC250/SC8/ (2003) Eurocode 8: Design Provisions for Earthquake Resistance of Structures, Part 1.1: General rules, seismic actions and rules for buildings, PrEN1998-1.
- EERI (1989). The basics of seismic risk analysis. *Earthquake Spectra* 5, 675-702.
- Iervolino, I., Chioccarelli, E., Convertito, V. (2011). Engineering Design Earthquakes from Multi-modal Hazard Disaggregation, *Soil Dynamics and Earthquake Engineering*, 31(9): 1212–1231.
- Meletti C., Montaldo V., 2007. Deliverable D2 Valutazioni di ag (16mo, 50mo e 84mo percentile) con le seguenti probabilità di superamento in 50 anni: 81%, 63%, 50%, 39%, 30%, 22%, 5%, 2%, rispettivamente corrispondenti a periodi di ritorno di 30, 50, 72, 100, 140, 200, 1000 e 2500 anni. <http://esse1.mi.ingv.it/data/D2.pdf>).
- Montaldo V., Meletti C., 2007. Valutazione del valore della ordinata spettrale a 1sec e ad altri periodi di interesse ingegneristico. Progetto DPC-INGV S1, Deliverable D3, <http://esse1.mi.ingv.it/d3.html>.
- Spallarossa D., Barani S., 2007. Disaggregazione della pericolosità sismica in termini di M-R-?. Progetto DPC-INGV S1, Deliverable D14, <http://esse1.mi.ingv.it/d14.html>.



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



LECTURE NOTES

The PEER framework & Fundamental of Logistic Regression

Chapter 3: Lecture 7

Fall 2016

Written by marco broccardo
IBK, ETHZ

1 Introduction

Performance-Based Earthquake Engineering (PBEE) is a probabilistic framework, which supports designers and stakeholders in assessing design, evaluation, and planning of civil systems. The utmost goal of the PBEE is to replace the classical load-and-resistance-factor design (LRFD), which is still the most used practice in earthquake engineering. While the LEFD evaluates performances of civil systems primarily with respect to failure probability of individual structural components, the PBEE is a holistic framework that assesses performances at a system level with respect to different decision variables such as monetary losses, fatalities. The first efforts to standardize the PBEE framework is contained in the SEAOC's Vision 2000 and FEMA 273.

Two axes define the PEER framework, a discrete set of predefined system performance objectives, and a set (discrete or continuous) of seismic hazard level. The domain defined by the two axes is partitioned into two subdomains, the acceptable domain and its complement, the not acceptable. Within the acceptable domain, designer and stakeholders define the desired combination of performance objectives with respect to the hazard level and its likelihood, and the monetary cost of the system.

A schematic view of the PBEE framework is given in Figure 1. According to this scheme the system performance includes: fully operational, operational, life safety, and near collapse. The level of excitation are divided in frequent ($T_r = 43$ -years) return periods, occasional ($T_r = 72$ -years), rare ($T_r = 475$ -years), very rare ($T_r = 949$ -years). Under the Poissonian assumption, these correspond to events with 50% exceedance probability in 30 years, 50% probability in 50 years, 10% probability in 50 years, and 10% in 100 years.

In the earthquake engineering literature there are various attempts and analytical approaches to perform PBEE; in this class, we focus on the approach introduced by the Pacific Earthquake Engineering center (PEER).

2 The PEER-PBEE framework

The PEER-PBEE framework was initially introduced by [?] to establish a simple but yet meaningful tool for Performance-Based Earthquake Engineering. The framework is simply a statement of the total probability theorem for the yearly mean number of events of a selected decision variable. In specific, the so-named PEER formula was originally written as follow

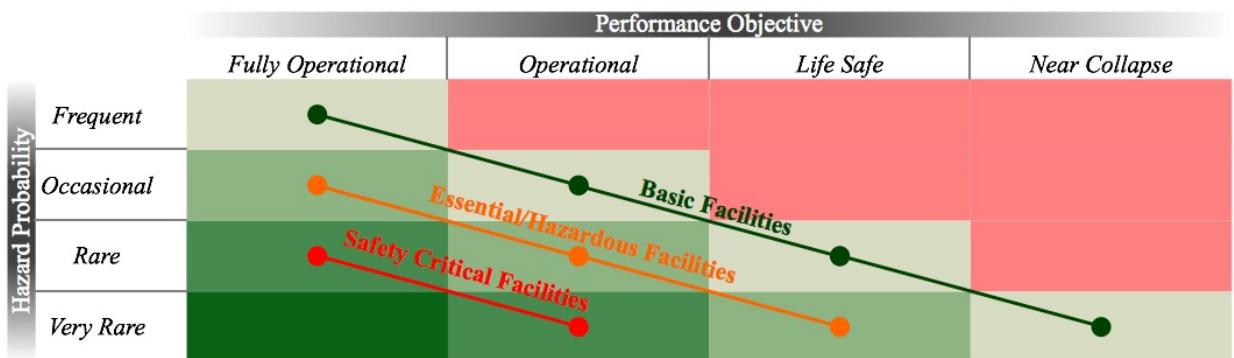


Figure 1: PBEE concept

$$\lambda(dv) = \int_d \int_{im} G(dv|d) |dG(d|im)| |d\lambda(im)|, \quad (1)$$

and later rewritten as

$$\lambda(dv) = \int_d \int_{edp} \int_{im} G(dv|d) |dG(d|edp)| |dG(edp|im)| |d\lambda(im)|, \quad (2)$$

where im is an intensity measure (e.g., peak ground acceleration, peak ground velocity, spectral acceleration, etc.), edp is an engineering demand parameter (e.g., interstorey drift), d is a damage measure (e.g., minor, medium, extensive, collapse), dv is a decision variable (e.g., monetary losses, fatalities, etc.), $\lambda(x)$ is the mean annual rate of events exceeding a given threshold for a given variable x , and $G(y|x) = P(Y \geq y|X = x)$ is the conditional complementary cumulative distribution function (CCDF). Note that, in general, d is a discrete variable so that \sum_d substitutes \int_d i.e.

$$\lambda(dv) = \sum_d \int_{edp} \int_{im} G(dv|d) P(d|edp) |dG(edp|im)| |d\lambda(im)| \quad (3)$$

Moreover, the variables im , edp , and d can be expressed in vector form; the integral, then, involves several folds.

$$\lambda(\mathbf{dv}) = \sum_d \int_{edp} \int_{im} G(\mathbf{dv}|d) P(\mathbf{d}|edp) |dG(edp|im)| |d\lambda(im)| \quad (4)$$

Often, the variable EDP is used directly to define damage states, i.e., no uncertainties are involved in determining the damage states given the observation of the EDP . In this case, the PEER framework is written in his original and most used form, i.e.

$$\lambda(dv) = \sum_d \int_{im} G(dv|d) P(d|im) |d\lambda(im)|, \quad (5)$$

where $P(d|im)$ is the so named fragility function.

One of the main advantages of the framework is that (3) and (5) build on four distinct stages: i. hazard analysis, ii. structural/fragility analysis, iii. damage analysis, and iv. loss analysis, each of which can be addressed by a different research group.

- i Hazard analysis. This analysis evaluates the mean annual rate of exceedance of the IM , $\lambda(im)$, at the facility site.
- ii Structural/fragility analysis. This analysis evaluates the probabilistic relationship between the selected IM and the EDP . In particular, the analyst creates a structural model of the system, where the input is derived from the hazard analysis and the output is computed via structural analysis. The structural model may or my not account for structural uncertainties.
- iii Damage analysis. This analysis evaluates the relationship between EDP and D . The relationship can be either purely deterministically, or probabilistically. In the first case a set of EDP thresholds ($[edp_1, edp_2, edp_3, edp_4]$) completely defines the damage states (e.g. $EDP \geq edp_1 \rightarrow D \geq d_1$, $EDP \geq edp_2 \rightarrow D \geq d_2$, $EDP \geq edp_3 \rightarrow D \geq d_3$, $EDP \geq edp_4 \rightarrow D \geq d_4$). In the second case, a statistical analysis is needed to characterize the probability distribution $P(D|EDP)$.

- iv Loss analysis. This analysis evaluates the probabilistic relationship between a given DV and a given damage state D . The decision variable measures the seismic performance of the facility in terms of the stakeholder interest.

Decision-making constitutes an additional layer of the framework. This layer is based on $G(dv)$ or $\lambda(dv)$. Observe that similar relationships to (2) can be written for each intermediate stage to produce edp mean rates, $\lambda(edp)$, and damage mean rates, $\lambda(d)$. Figure 2 illustrates all the steps of the PEER framework.

3 Underlying assumptions

The first underlying assumption in (2) is that the joint probability distribution of the variables, IM, EDP, D , and DV can be expressed with a Markovian structure Figure 2. This implies that EDP is defined only conditional to IM , D is defined only conditional to EDP , and DV is defined only conditional to D . It follows that $D \perp\!\!\!\perp IM|EDP = edp$, and $DV \perp\!\!\!\perp EDP, IM|D = d$. This is equivalent to require for each variable to be an efficient and sufficient predictor ???. In this context, efficient is that a variable is highly correlated with the following variable, and sufficient is that, conditioned to the previous variable, the current variable is not significantly correlated with the ancestors. The second implicit assumption is that the conditional distributions in (2) are identical for successive events. This implies that the facility has an instantaneous recovery to the original state, and does not deteriorate.

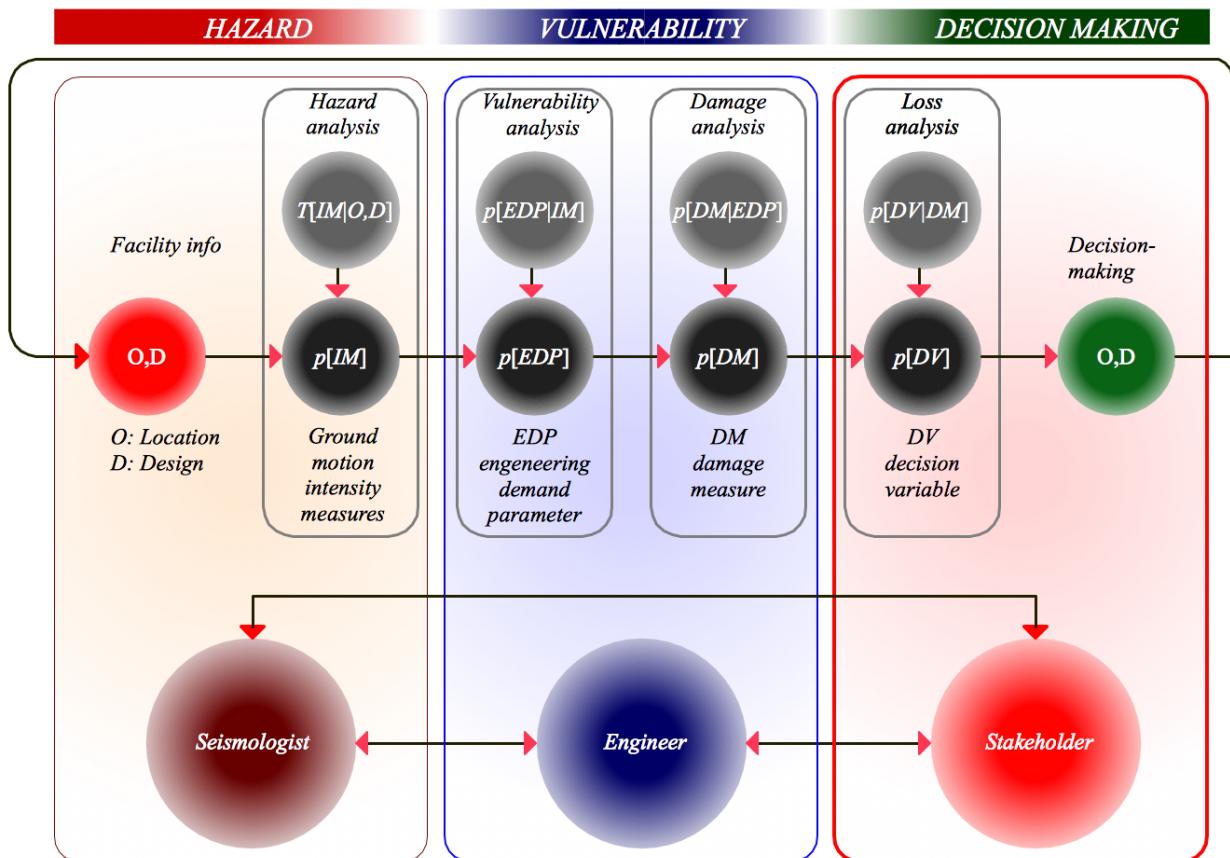


Figure 2: PEER framework concept

It is natural to think of the PEER framework as a natural extension of the PSHA analysis. However, particular caution must be used when using the PEER framework for the computation of the distribution of the extreme DV , D , or EDP for all earthquakes occurring during a specific period of time. In the PSHA analysis we have seen that given the hazard rate $\lambda(im)$ the CCDF for a given period of time was determined based on the homogeneous Poisson process as

$$P(\{IM > im\}, t) = 1 - P[0 \text{ events } \{IM > im\} \text{ in } t] \quad (6)$$

$$= 1 - \exp[-\lambda(im)t] \quad (7)$$

In the (10), there is implicit Poisson with random selection, i.e. $\lambda(im) = P(IM > im|M > m_{min})$, where λ_{min} is the rate of the minimum earthquake of interest, and the total probability theorem in the computation of $P(IM > im|M > m_{min})$. However, there is another important assumption implicit in Poisson, that is that the uncertainties are renewable, i.e. the Poisson process is a memoryless process. This is certainly true in case the uncertainties are statistically independent an assumption that has been proved to be sufficiently good in PSHA analysis¹. These assumptions make the Poisson process and in general all renewable processes ergodic. However, in (2) there are different type of uncertainties, which do not “renew” at each earthquake event, and therefore are not ergodic. Classical examples are the aleatory uncertainty of the system properties and the epistemic uncertainties related to model idealization. In particular, the gained knowledge on a civil system response after a seismic event modifies the nature and estimate of the uncertainties. It follows, that different stochastic processes with memory properties might be more suitable to tackle problems with non-renewable uncertainties. However, in practice, Eq. (10) is the most used formulation in Earthquake Engineering given its simplicity and because the error introduced is negligible for very low probabilities levels (typically for probabilities $< 10^{-3}$). The Eq. (10) is named the ergodic approximation of the probability of exceedence.

In the following example, we show in practice, the influence of nonergodic variables and solution to include them correctly in the framework.

3.1 Renewable vs not-renewable, exact solution vs ergodic approximation

Suppose we have structure for which

$$\{EDP|\Theta, IM = im\} = \Theta \times im + \varepsilon_{im} \quad (8)$$

where $\Theta = \theta$ represents the mean response of the structure for a given $IM = im$ (e.g. the stiffness of a structure), and ε_{im} is the variability due only to IM . If Θ is random variable, it reflects an epistemic uncertainty. In facts, these uncertainty are not renewable for a novel seismic event, and generally they can be reduced via structural monitoring. It follows that the event $\{EDP > edp|IM = im\} = \int_{\theta} \{EDP|\Theta = \theta, IM = im\} d\theta$ has a non-ergodic component for which the renewable process assumption does not work. However, the conditional event $\{EDP > edp|\Theta = \theta\}$ has only aleatory uncertainties for which Poisson process can be apply. Consequently, we can write

$$P(\{EDP > edp|\Theta = \theta\}, t) = 1 - P[0 \text{ events } \{EDP > im|\Theta = \theta\} \text{ in } t] \quad (9)$$

$$= 1 - \exp[-\lambda(edp|\theta)t] \quad (10)$$

¹This is true for the aleatory uncertainties, while epistemic uncertainties need a special treatment which is beyond the scope of the current notes

where $\lambda(edp|\theta) = \lambda_{min} \times P(EDP|\Theta = \theta)$, and

$$P(EDP|\Theta = \theta) = \int_{im} P(EDP|\Theta = \theta, IM = im) dG(im), \quad (11)$$

where $G(im) = \lambda(im)/\lambda(0)$. Finally we can use the total probability theorem to obtain $P(\{EDP > edp\}, t)$ as

$$P(\{EDP > edp, t\}) = \int_{\theta} (1 - \exp[-\lambda(edp|\theta)t]) |dG_{\theta}|, \quad (12)$$

$$= 1 - \int_{\theta} (\exp[-\lambda(edp|\theta)t]) |dG_{\theta}|. \quad (13)$$

On the other hand the $\lambda(edp)$ is simply obtained as

$$\lambda(edp) = \lambda_{min} \times \int_{\theta} P(EDP|\Theta = \theta) |dG_{\theta}| \quad (14)$$

If a fully ergodic assumption is used then

$$P(\{EDP > edp, t\}) \sim 1 - \exp[-\lambda(edpt)], \quad (15)$$

$$\sim 1 - \exp \left[- \left(\int_{\theta} \lambda(edp|\theta) |dG_{\theta}| \right) t \right]. \quad (16)$$

Observe that the approximate solution (16), which is used most of the time in practice, is different from the correct solution (13). However, if the rate instead of probability is used as the reference metric, then there is no approximation involved.

4 Classification Problem

Classification problems, are class of problems where the output of the analysis is not a quantitative value, but a *qualitative* value. Qualitative variables belong to an unordered categorical set \mathcal{C} , for example the state of a building after an earthquake $\in \{\text{no damage, light damage, heavy damage, collapse}\}$. In a classification problem, likewise in regression analysis, there is a “feature input” vector X , a qualitative (instead of quantitative) response Y which takes values in the set \mathcal{C} , and a classifier $C(X)$, which is a function that map the feature vector X into the possible values of Y , i.e. $C(X) \in \mathcal{C}$. However, oftentimes, our interest is in *measuring the probability* that X belongs to a particular outcome of Y rather than an *hard* classification. For example, in earthquake engineering it is more important to estimate the probability that a building is collapsed rather than a hard classification collapse or not collapse.

There are two main approaches to create a probabilistic classifier, the generative approach, and the discriminative approach. The generative approach is based on the definition of a joint probability model $p(y, x)$, from which $p(y|x)$ is derived by conditioning on x . On the other hand, the discriminative approach consist of fitting $p(y|x)$ directly. Seldom, classification problems are solved via regression analysis. Despite for the binary case this approach offers reasonable results, it is generally not recommendable in most of the classification problems. In the following sections, we introduce probably the most common and useful classification technique, namely *the logistic regression*, which is a linear discriminative model.

4.1 Logistic regression

Consider the simple example for which the response of a building after an earthquake falls into one of the two classes collapse, not collapse. Instead of modeling the response Y , the logistic regression models directly the probability that Y belongs to a particular category, i.e. $P(Y = y|X = x)$. To model a probability distribution, we should choose a function that gives output between 0 and 1 for any value of the feature vector X . Observe that any CDF meet this description; in the next chapter, for example, we will use the lognormal CDF to model the so named fragility functions. In logistic regression, the most common function is the *logit* function:

$$p(X) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \quad (17)$$

where $p(X) = P(Y = y|X = x)$. The name of “logistic” comes from the log-transformation (or logit trasformation) of the *odds*, i.e

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x, \quad (18)$$

where $\log\left(\frac{p}{1-p}\right)$ is the logit, and $\frac{p}{1-p}$ is the odds, which take values $[0, \infty)$. For example, odds close to 0 indicates low probability of collapse, while high odds indicate vey high probability of collapse. In particular, an odd of 1/9 indicate that in average one building of ten will collapse, conversely an odd 9 indicates that, in average, 9 building out of 10 will collapse. Odds are traditionally used in place of probabilities in sports bets. Observe that Eq. (18) is still a linear model, but it is modeling probabilities in a nonlinear scale; in particular, in logistic regression the *logit* is a linear function of the feature vector X . Likewise in linear regression, where β_1 gives the average change of Y associated with a unit increase on X , in logistic regression, $\frac{p}{1-p}$ gives the change in of the logit for an unit increase of X .

4.1.1 Coefficient estimations

The most suitable method to estimate the coefficient β_0 and β_1 is via Maximum Likelihood. In particular, we determine the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ by maximizing the likelihood function, that is the probability of observing a sequence of observations, given a set of parameters. In our guiding examples, the qualitative variable takes only two categories [collapse, not collapse] ([0, 1] in practice), and the probability of observing a collapse is simply $p(X)$, and not collapse $1 - p(X)$; it follows that the likelihood function can be written as

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{n:y_n=1} p(\beta_0, \beta_1; x_n) \prod_{n:y_n=0} (1 - p(\beta_0, \beta_1; x_n)), \quad (19)$$

where $p(\beta_0, \beta_1; x_n)$ is given by (17)², and the log-likelihood by

$$\log \mathcal{L}(\beta_0, \beta_1) = \sum_{n:y_n=1} p(\beta_0, \beta_1; x_n) + \sum_{n:y_n=0} (1 - p(\beta_0, \beta_1; x_n)). \quad (20)$$

The parameters β_0 and β_1 are then obtained as

$$[\hat{\beta}_0, \hat{\beta}_1] = \operatorname{argmax}_{[\beta_0, \beta_1]} [\log \mathcal{L}(\beta_0, \beta_1)]. \quad (21)$$

Given $\hat{\beta}_0$ and $\hat{\beta}_1$ it is a simple matter to estimate the probability of collapse or not collapse for a given feature vector X . Specifically, the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are plug in into (17) to obtain an estimate of $\hat{P}(Y|X)$. In the next lecture we will see much more in details the use of logistic regression in fragility computation models.

²In this case we made explicit the parameters β_0, β_1 .



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



LECTURE NOTES

Principle for fragility function computation

Chapter 3: Lecture 8

Spring 2018

Written by marco broccardo
IBK, ETHZ

1 Introduction

In this chapter, we outline the main concept behind fragility functions with a motivational example. Then, we review the central role of fragility functions in the context of the PEER framework, followed by a brief summary of the main classes of fragility function. Finally, we explore in details two of the main methods for computing fragilities based on time-history analysis of a given structural model.

2 Definition and motivational example

A seismic fragility function is defined as the conditional probability of an event (e.g. a defined damage state) given the observation of an intensity measure which describes the seismic event.

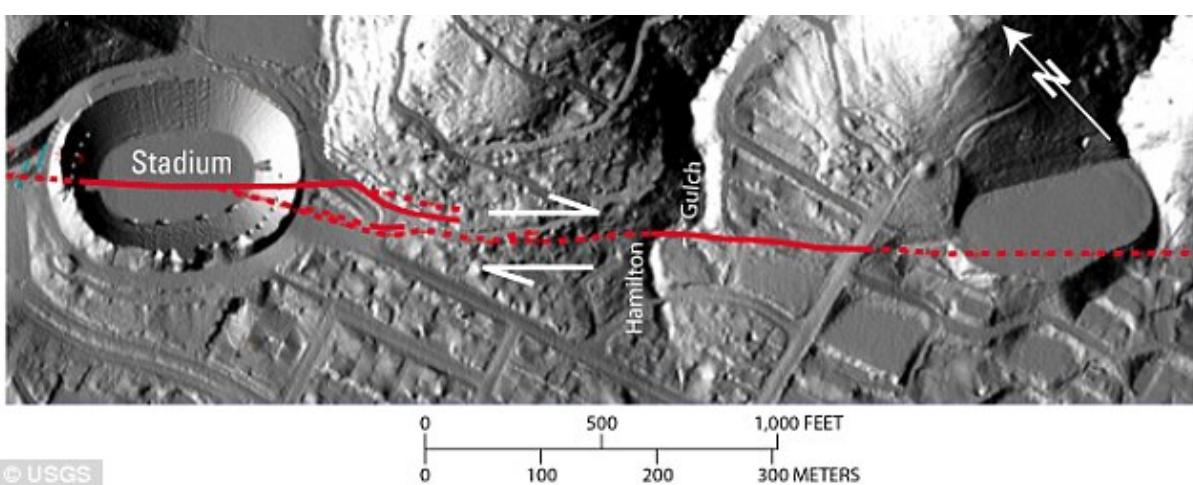
In Lecture 2, we start to investigate the problem of determining the probability that California Memorial Stadium, Figure 2, will split due to a seismic event. Here, we build on that example by calculating such probability, and by defining our first “fragility function”. Given the geometry



(a) Source: Michael Layefsky (via flickr)



(b) Source: USGS & Google maps



(c) source: USGS & Google maps

Figure 1: California Memorial Stadium: a) aerial view, b) Stadium with the crossing Hayward fault, c) Hayward fault

of the site Figure 2, a random variable S describing the size of the fault rupture, and the event D that the fault rupture is crossing the stadium, we have computed the following conditional probability distributions:

case 1 $0 < s < l_1$

$$P(D|S = s) = \frac{w + s}{l_1 + l_2 - s}, \quad (1)$$

case 2 $l_1 \leq s \leq l_2 - w$

$$P(D|S = s) = \frac{l_1 + w}{l_1 + l_2 - s}, \quad (2)$$

case 3 $s \geq l_2 - w$

$$P(D|S = s) = 1. \quad (3)$$

In the Lecture of PSHA analysis we have provided the Wells and Coppersmith (1994) relationships between moment magnitude M and rupture length here named S . In this case, since the Hayward fault is strike slip we write

$$G(S|M = m) = \hat{\Phi} \left(\frac{\log(s) - \mu(m)}{\sigma(m)} \right) \quad (4)$$

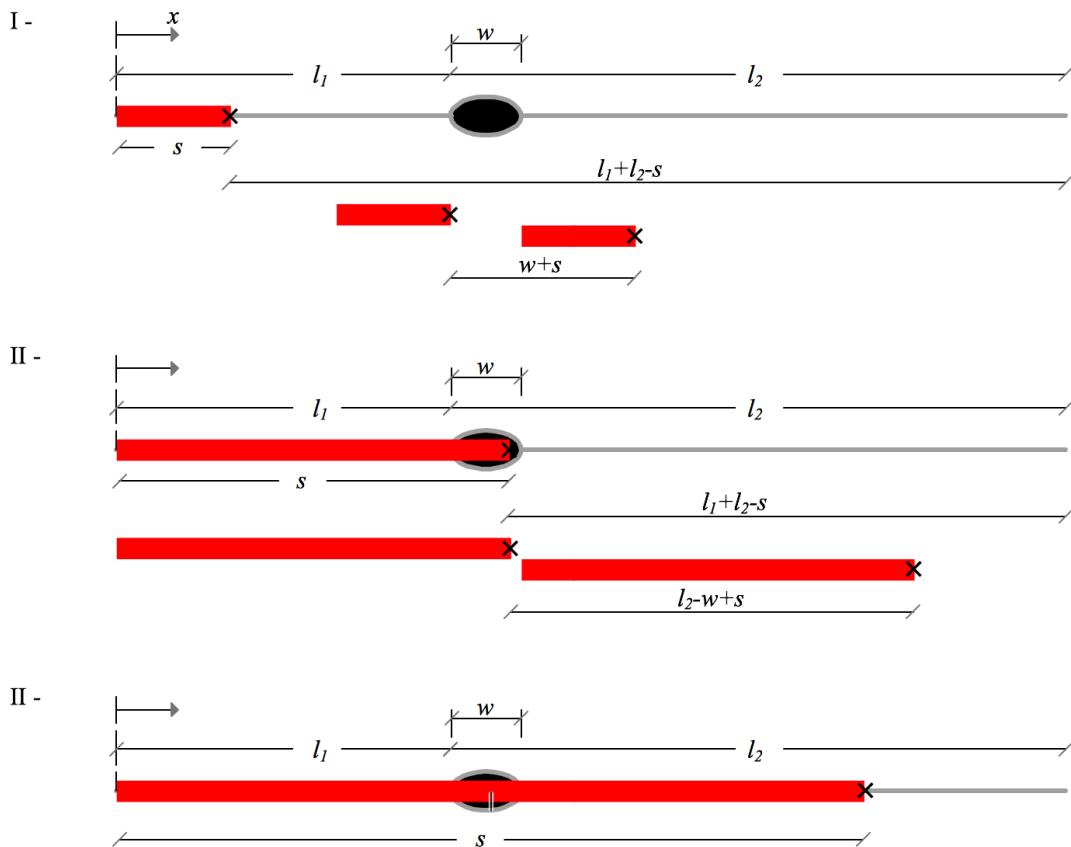


Figure 2: Configuration of the fault rupture system

where $\hat{\Phi}$ is the truncated standard normal CCDF,

$$\mu(m) = 0.74m - 3.55, \quad (5)$$

and

$$\begin{aligned} \sigma(m) &= \hat{\sigma}, \text{ for } \mu(m) \leq \log(l_1 + l_2) - 3\hat{\sigma} \\ \sigma(m) &= \frac{\log(l_1 + l_2) - \mu(m)}{3}, \text{ for } \mu(m) > \log(l_1 + l_2) - 3\hat{\sigma} \end{aligned} \quad (6)$$

The (6) has been introduced to take into account the boundary condition of the finite rupture length. Observe that for $\mu(m) = \log(l_1 + l_2)$ there is a deterministic relationship ($\sigma(m) \rightarrow 0$) between rupture length and maximum magnitude. In particular assuming that the maximum magnitude will occur when the rupture is equal to the total length of the fault, we obtain $m_{max} = 7.369$. Assuming a $M_{min} = 4$ we can define a truncated Gutenberg-Richter distribution with $\beta = 2.3$ and $\lambda_{min} = 0.15$.

Given the following information we can write

$$P(D) = \int_{m_{min}}^{m_{max}} \int_{s_{min}}^{l_1 + l_2} P(D|s) |dG_{S|M}(s|m)| |dG_M(m)| \quad (7)$$

$$= \int_{m_{min}}^{m_{max}} P(D|m) |dG_M(m)|. \quad (8)$$

For this case, observe that by selecting as intensity measure the moment magnitude, i.e. $IM = M$ and as damage state D ,

$$P(D|M = m) = \int_{s_{min}}^{l_1 + l_2} P(D|s) |dG(s|m)| \quad (9)$$

$P(D|M = m)$ is effectively our first derived fragility function. In particular, if we observe an event of a given magnitude we can estimate directly the probability of that given damage state.

Problem Compute the fragility function. What is the probability for the stadium of splitting in case $M_f = 4$? Which conclusions can you draw? What is the magnitude for which the stadium is surely damaged? List the possible use that you can make of this fragility function.

3 Fragility function in the PEER framework

The backbone of this course is the PEER framework. In particular, we would like to address the problem of determining probability of damages, or losses, given a specific hazard level, and a suitable structural analysis strategy.

The most common version of the PEER framework is given by

$$\lambda(dv) = \sum_d \int_{edp} \int_{im} G(dv|d) P(d|im) |d\lambda(im)|, \quad (10)$$

where the fragility function is formally defined as $P(d|im)$. One of the main advantage of the framework is that (10) builds on three distinct stages: i. hazard analysis, ii. structural/fragility analysis & damage analysis, and iii. loss analysis, each of which can be addressed by a different research group, or at different stages.

We have seen during Chapter 3 Lecture 7 the importance of clearly outline the different uncertainties between renewable (typically seismic hazard uncertainties) and not renewable (typical structural uncertainties). In this section, we focus on a brief discussion on the importance of the other two keys aspect of the fragility functions, which are: the choice of the intensity measure IM , and the definition of the damage state D .

3.1 Which intensity measure?

The choice of a suitable intensity measure is defined and classified in terms of efficiency, robustness, practicality, and effectiveness (Mackie and Stojadinovic 2003, 2005).

Efficiency refers to the total variability of an engineering demand parameter for a given IM . This is the most important quantitative measure for which an optimum IM can be obtained.

Robustness describes the efficiency $IM - EDP$ across different structures, and different fundamental period ranges.

Practicality refers to the fact that the IM should be correlated to known and easy identifiable engineering quantities. The practicality of an IM can be verified by computing the dependence of the structural response on the physical properties of the structures, i.e. energy, response of fundamental and higher modes, etc.

Sufficiency describes the validity of the assumption that the $EDP|IM$ is statistically independent of ground motion site source characteristics such as magnitude and distance. A sufficient IM is one that renders the structural demand measure conditionally independent of the earthquake scenario.

Effectiveness of an IM is determined by its ability to evaluate its relation with an engineering demand parameter (EDP) in closed form (Mackie and Stojadinovic 2003), so that the mean annual frequency of a given decision variable exceeding a given limiting value can be determined analytically.

3.2 How to define a damage state?

From our motivational example, it should be clear that the definition of damage states should suit the specific structural problem. Generally, the damage states are defined with categorical variables as follow

- D_0 No damage
- D_1 Slight, minor damage
- D_2 Moderate damage
- D_3 Extensive Damage
- $D_4 = C$ Collapse.

In this course, we follow the rule that damage states are deterministically defined by engineering demand parameters thresholds (i.e. $edp_1 \rightarrow D_1, \dots, edp_2 \rightarrow C$).

The art of modeling relies in associating each damage state to a specific engineering demand parameter or to a set of engineering demand parameters (for a series system). It must be said that it is also possible to define continuous damage states, e.g. damage indexes (usually between

0 and 1) and a probabilistic rather than deterministic relationship between EDP and D . These definition are beyond the topics of this course.

4 Class of fragilities

There are generally four macro-classes of fragility functions:

- i **Empirical.** Empirical fragility are obtained by fitting a function to observational data from past earthquake events or laboratory tests. Data are usually collected as ordered pairs of level of excitation and categorical variables of damage or collapse. Usually, these methods are based on pure statistical analysis (e.g. regression models).

Good: they are based on real data. It follows that they reflect several effects such as site effect, soil-structure interactions, structural uncertainties etc.

Bad: they are location specific since they reflect the data of a given area. Moreover, they are “time dependent” in the sense that they reflect the structural uncertainties of the exposed asset at the given time of the seismic event. It has been noted also that statistical relationships based on reconnaissance surveys are generally bias since the undamaged building are often not properly taken into account.

- ii **Analytical.** Analytical fragility functions are derived by defining an analytical, or numerical, structural model and analyzing its performance under different levels of the seismic hazard. There are two subclasses of analytical models:

(a) **Static.** The hazard is represented by a response spectrum, and the capacity model is defined either by an idealized structural behavior (typically bilinear) or by push-over analysis. The most popular static analytical method is the Capacity Spectrum method (CSM), which is used in HAZUS [] and RISK-UE methodologies for fragility computation. Generally, this method is used to compute fragility functions for specific class of buildings and they are computational not expensive. In this course we do not explore this method in details, since we decided to give preference to dynamic methods.

(b) **Dynamic.** The hazard is represented by a collection of ground motions time series, which reflects the hazard of a specific site. The structural-capacity model is defined by a numerical model which is, usually, based on finite element theory. Then, time history analysis is performed to determine the structural response. Post process of the output is used to determine the fragility function. There are several methods for computing fragility functions via dynamic time-history analysis. These methods deviates in two main aspect: i) on the selection and preprocess (e.g. scaling) of the ground motion time series, and ii) which statistical method to use to analyze the results. In this course, we examine first the most popular “dynamic” method to compute fragility which is named Incremental Dynamic Analysis (IDA); then we focus on the Multi-Stripes Analysis (MSA).

There is another subclass of fragility which is based on stochastic dynamic theory [?]. In this course, the the input is modeled as a stochastic process which reflects the properties of the seismic hazard. The fragility, then, can be computed both by simulations (i.e. similarly to any “dynamic” based fragility)[[?]] or by theory of stochastic dynamics [?]. Computation of fragility via these methods are beyond the purpose of this course.

The choice of which method to use (“static” or “dynamic”, and within these classes which subclasses) is based on the accuracy of the representation of the structural behavior and

the cost-efficiency of the analysis. As general rule, we suggest: first to understand the scale of the problem, second the level of detail of the analysis, finally the available resources for performing the analysis.

- iii **Expert opinion.** Expert opinion fragilities is created by polling one or more (independently) experts of the given structural asset. Usually, it is asked to guess or estimate the failure probability, or a range of failure probabilities, for a given hazard level.

Good: generally, they are useful tools for a fragility pre-assessment. They overcome the issue of extensive damage dataset, and they can be derived basically for classes of buildings.

Bad: They are solely based on expert experience and they can be potentially bias. When multiple expert are involved in the pool, usually these fragilities show a large variability which basically reflect the lack of uniformity in judging a complex natural phenomena such as a seismic event.

- iv **Hybrid** Hybrid fragility functions are obtained based on a combination of the different methods. The main advantage is that by combining alternative methods we can eliminate the singular negative consequences. For example, analytical fragility function can be modified and updated by integrating post-seismic outcomes, to reduce inherent model bias or to reduce the variance. Hybrid fragility functions are particularly suitable for Bayesian statistics, example of these are [?].

Between the different possible strategies for determining fragility functions, we decided to privilege the ones that requires time history analysis of a physical based model. In our view, this class of fragility are the most appropriate for a new or existing single civil facility, such as schools, hospitals, large buildings, bridges etc.

5 Dynamic based fragility functions

5.1 Incremental Dynamic Analysis

Incremental dynamic analysis IDA [?] is probably the most popular method to compute fragility functions via time-history analysis. Briefly stated IDA can be summarized as follows

- i Definition of a numerical model for the structure of interest. This is equivalent to define a capacity model $y(t) = \mathcal{M}[\ddot{x}_g(t|IM = im); \boldsymbol{\theta}_{\mathcal{M}}(t)]$ where $\mathcal{M}[\cdot; \boldsymbol{\theta}_{\mathcal{M}}(t)]$ is the the numerical model (e.g. a finite element model), $\boldsymbol{\theta}_{\mathcal{M}}(t)$ is a set of model parameters (e.g. masses, stiffnesses, damping, ductility parameters etc.), and $\ddot{x}_g(t|IM = im)$ is the input ground motion given a $IM = im$ of interest.
- ii Selection of a suitable IM given the structure or system of interest (e.g. PGA , SA , etc.).
- iii Selection of a suitable set of N ground motions for the location of interest (i.e definition of the seismic demand).
- iv Selection of an EPD of interest (e.g. interstorey drift, beam curvature, a combination of interstorey drifts).
- v Definition of damage limit states D via EPD thresholds (i.e. $[edp_1 \rightarrow D_1, edp_2 \rightarrow D_2, \dots, edp_4 \rightarrow C]$).

- vi Scale each ground motion based on the given IM until it reaches the collapse (i.e. run a time history analysis for each scaled ground motion).
- vii Save each $[EDP \rightarrow D]$ threshold- im_n pair for each ground motion.

For a given damage state, the output of the analysis is thus N intensity measures, which they represent the particular $IM_n = im_n$ at which the given ground motion reached the given damage state.

Given the N outcomes, it is common to assume a lognormal probability distribution for the random variable IM associated with the given damage state D . Then, the parameter of the lognormal can be estimated via the method of moments estimator (for exponential family distribution the method of moments and maximum likelihood are equivalent).

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N \ln(im_n), \quad (11)$$

$$\hat{\sigma} = \sqrt{\frac{1}{N} \sum_{n=1}^N (\ln(im_n) - \hat{\mu})^2}, \quad (12)$$

where in (12) $(N - 1)$ in place of N is used in case an unbiased estimator is desired. Then, the fragility function can be estimate via CDF of the log normal distribution as

$$P(D > d | IM = im) = \Phi\left(\frac{\ln(im) - \hat{\mu}}{\hat{\sigma}}\right), \quad (13)$$

where $\Phi(\cdot)$ is the CDF of the normal distribution. Assuming that the scaling is legitimate and well posed (we will come back to this concept), the previous set equations completely define the fragility of interest.

The method of moment to estimate the parameter of the lognormal distribution is equivalent to the method of maximum likelihood if the data are representative of the full distribution. One limitation of the IDA is that, oftentimes, the ground motions need to be scaled to large IM s. This causes several problems. First scaling the ground motions with large factors is not an accurate way to represent the ground motions for that particular intensity. Second, the computational cost can become quite large.

To overcome these issues, an upper limit \bar{IM} to the scaling is recommended. In this case the method of moments is not correct since the data are not representative of the full distribution. However, the method of maximum likelihood can still be applied. In particular, the data are divided into two groups. Data that causes collapse $n \in [1, \bar{N}]$ with $\bar{N} \leq N$, and data that do not cause collapse $n \in [\bar{N} + 1, N]$. For the first group of data the likelihood function under the assumption that the im_n are statistical independent is simply given by

$$\mathcal{L}(\mu, \sigma) \alpha \prod_{n=1}^{\bar{N}} \varphi\left(\frac{\ln(im_n) - \mu}{\sigma}\right) \quad (14)$$

where $\varphi(\cdot)$ is the standard normal PDF. For the second group of data the likelihood function is proportional to the probability that IM_n is grater than \bar{IM} ; then, under the assumption of independence of the im_n

$$\bar{\mathcal{L}}(\mu, \sigma) \alpha \prod_{n=\bar{N}+1}^N \left[1 - \Phi\left(\frac{\ln(\bar{IM}) - \mu}{\sigma}\right)\right] = \left[1 - \Phi\left(\frac{\ln(\bar{IM}) - \mu}{\sigma}\right)\right]^{N-\bar{N}}. \quad (15)$$

Finally the likelihood function for the entire set of data is given by

$$\mathcal{L}(\mu, \sigma) = \underline{\mathcal{L}}(\mu, \sigma) \bar{\mathcal{L}}(\mu, \sigma) = \prod_{n=1}^{\bar{N}} \varphi\left(\frac{\ln(im_n) - \mu}{\sigma}\right) \left[1 - \Phi\left(\frac{\ln(\bar{IM}) - \mu}{\sigma}\right)\right]^{N - \bar{N}}, \quad (16)$$

and the log likelihood as

$$\ln \mathcal{L}(\mu, \sigma) = \sum_{n=1}^{\bar{N}} \varphi\left(\frac{\ln(im_n) - \mu}{\sigma}\right) + (N - \bar{N}) \left[1 - \Phi\left(\frac{\ln(\bar{IM}) - \mu}{\sigma}\right)\right]. \quad (17)$$

Then, the estimate of the parameter is found by optimization as

$$[\hat{\mu}, \hat{\sigma}] = \operatorname{argmin}_{\mu, \sigma} [-\ln \mathcal{L}(\mu, \sigma)]. \quad (18)$$

Notice that in case $\bar{N} = N$ the solutions for (18) is (11) and (12). Observe that in general the likelihood function can be written also for a generic probability density function $f(x; \boldsymbol{\theta})$ (where $\boldsymbol{\theta}$ is set of parameter) as

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{n=1}^{\bar{N}} f(im_n; \boldsymbol{\theta}) [1 - F(\bar{IM}; \boldsymbol{\theta})]^{N - \bar{N}}, \quad (19)$$

or

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{\bar{N}} f(im_n; \boldsymbol{\theta}) + (N - \bar{N}) [1 - F(\bar{IM}; \boldsymbol{\theta})], \quad (20)$$

and solution of the parameter can be found

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} [-\ln \mathcal{L}(\boldsymbol{\theta})]. \quad (21)$$

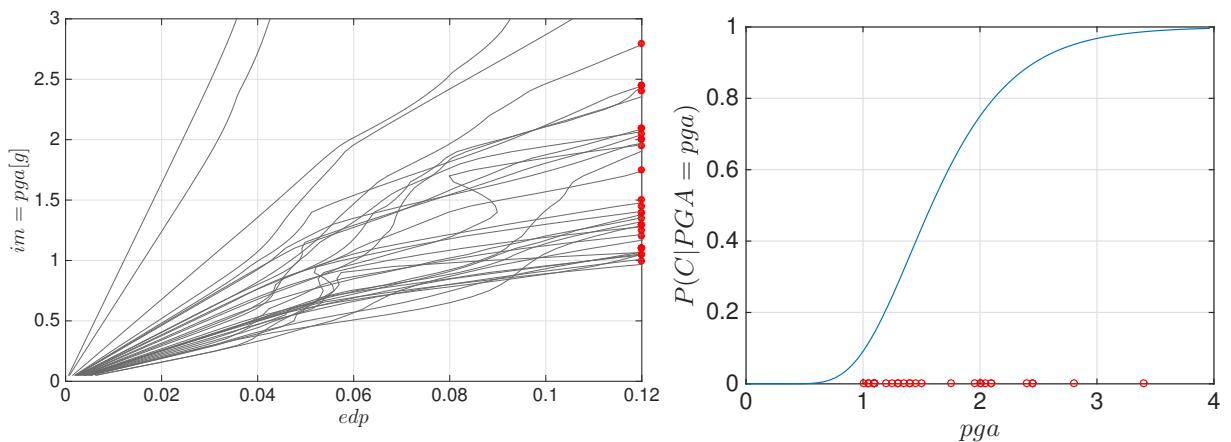


Figure 3: Fragility function computed via IDA

5.2 Multiple Stripes Analysis

There is profound assumption on the IDA analysis which is significantly arguable. The assumption is that ground motions can be scaled given an IM of reference. In essence we are modifying the recorded ground motions to infer the the output of the response. The operation of modifying data (which represent the true outcome of a natural process) rather than fitting a statistical model is undoubtedly not ideal, since it creates artificial time series which do not necessarily share the physical characteristics of the true records for that level of intensity. It follows that only “relative small” scaling procedure should be applied for a correct computation of the output.

An alternative to the IDA is the so named Multiple Stripes Analysis (MSA). The MSA method can be synthesize as follow:

- i Definition of a numerical model for the structure of interest, i.e. $y(t) = \mathcal{M}[\ddot{x}_g(t|IM = im); \boldsymbol{\theta}_{\mathcal{M}}(t)]$ where $\mathcal{M}[\cdot; \boldsymbol{\theta}_{\mathcal{M}}(t)]$ is the the numerical model (e.g. a finite element model), $\boldsymbol{\theta}_{\mathcal{M}}(t)$ is a set of model parameters, and $\ddot{x}_g(t|IM = im)$ is the input ground motion given a $IM = im$ of interest.
- ii Selection of a suitable IM given the structure or system of interest (e.g. PGA , SA , etc.).
- iii Selection of the Hazard curve for the given IM of interest and set of intensity measures associated with given return periods, e.g. $im \in [im_1(T_{R_1}), im_2(T_{R_2}), \dots, im_R(T_{R_M})]$.
- iv Selection of a suitable set of N_m ground motions, $\ddot{x}_g^{(n_m)}(t|IM = im_m)$, for each intensity measure im_m of interest. It follows that the total number of ground motions is $\sum_{m=1}^M N_m$ (the proper selection of ground motions is a topic of the next Lecture).
- v Selection of an EPD of interest based on the structural response $y(t)$. (e.g. if $y(t)$ is the interstorey drift; then, $EDP = \max(|y(t)|)$).
- vi Definition of damage limit states D via EPD thresholds (i.e. $[EDP_1 \rightarrow D_1, EDP_2 \rightarrow D_2, \dots, EDP_4 \rightarrow C]$).
- vii For each set of ground motions (based on a given intensity measure) perform time history analysis and compute the EDP of interest.
- viii For each $IM = im_n$ stripe of interest cluster the EDP according the defined damage state.

The suggested estimation of the fragility is based on maximum likelihood function. In particular for a given intensity level $IM = im_m$ and damage state threshold $EDP_i \rightarrow D_i$, the points are clustered between data that do not pass the threshold and data that do pass the threshold. By defining the number of *edps* passing the threshold D_i as \bar{N}_m , the probability of observing x events out N_m inputs given $IM = im_m$ is simply given by the binomial distribution (Lecture 2) as

$$P(\bar{N}_m \text{ collapses in } N_m \text{ ground motions}) = \binom{N_m}{\bar{N}_m} p_m^{\bar{N}_m} (1 - p_m)^{(N_m - \bar{N}_m)}, \quad (22)$$

where p_m is the probability that a given ground motion $\ddot{x}_g^{(n_m)}(t|IM = im_m)$ produce an $edp > edp_i$. Essentially, p_m is our fragility function, i.e. $p_m = P(D > d_i|IM = im_m)$. Assuming a general CDF for the fragility, e.g. $F(im; \boldsymbol{\theta})$, the (22) can be rewritten as

$$P(\bar{N}_m \text{ collapses in } N_m \text{ ground motions}) = \binom{N_m}{\bar{N}_m} F(im_m; \boldsymbol{\theta})^{\bar{N}_m} (1 - F(im_m; \boldsymbol{\theta}))^{(N_m - \bar{N}_m)}, \quad (23)$$

It follows that the likelihood function assuming that the events in the different stripes are statistical independent is proportional to

$$\mathcal{L}(\boldsymbol{\theta}) \propto \prod_{m=1}^M F(im_m; \boldsymbol{\theta})^{\bar{N}_m} (1 - F(im_m; \boldsymbol{\theta}))^{(N_m - \bar{N}_m)}, \quad (24)$$

or the loglikelihood as

$$\ln \mathcal{L}(\boldsymbol{\theta}) \propto \sum_{m=1}^M [\bar{N}_m \ln F(im_m; \boldsymbol{\theta}) + (N_m - \bar{N}_m)(1 - F(im_m; \boldsymbol{\theta}))], \quad (25)$$

from which the parameter $\boldsymbol{\theta}$ can be estimated with the (21).

A classical choice for the CDF is the lognormal distribution, it follows that (25), can be written as

$$\ln \mathcal{L}(\boldsymbol{\theta}) \propto \sum_{m=1}^M \left\{ \bar{N}_m \ln \Phi \left(\frac{\ln(im_m - \mu)}{\sigma} \right) + (N_m - \bar{N}_m) \ln \left[1 - \Phi \left(\frac{\ln(im_m - \mu)}{\sigma} \right) \right] \right\} \quad (26)$$

and solution is given by (18). Figure [?] shows the computation of lognormal fragility functions computed via MSA for two compelling structural systems.

Several authors [?] have observed that lognormal might not be the most suitable choice for $F(im; \boldsymbol{\theta})$. This obviously depends from the input ground motions and the structural model, and there is no reason why there should be an universal $F(im; \boldsymbol{\theta})$ which fits the model. A good practice should be to use different $F(im; \boldsymbol{\theta})$ to fit the output of the analysis and finally checking the $F(im; \boldsymbol{\theta})$ with the highest likelihood. Within the classical parametric function we suggest the followings:

- lognormal distribution. Reason: it is the most common used fragility. Moreover it can be shown that the (26) is equivalent to the Probit model, which is a particular form of generalized linear regression model.
- Logit function. Reason: in statistics the problem of classification of a binary variable (or categorical variables) given a continuous random variable such as IM is solved via logistic regression. The set up of the likelihood function is equivalent to (26) with $F(im; \boldsymbol{\theta})$ equal to the logit function, i.e.

$$F(im; im_0, k) = \frac{1}{1 + \exp[-k(im - im_0)]} \quad (27)$$

- Beta CDF. Reason: Beta distribution is a desirable function because it has compact support, moreover it is a conjugate prior for the binomial distribution and it is particular suitable for Bayesian updating,

$$F(im; \alpha, \beta) = \frac{B(im; \alpha, \beta)}{B(\alpha, \beta)}, \quad (28)$$

where $B(im; \alpha, \beta)$ is the Beta function, and $\boldsymbol{\theta} = [\alpha, \beta]$.

Observe also that the $F(im; \boldsymbol{\theta})$ can be also obtained via polynomial expansion [?], spectral decomposition [?], or cubic spline [?].

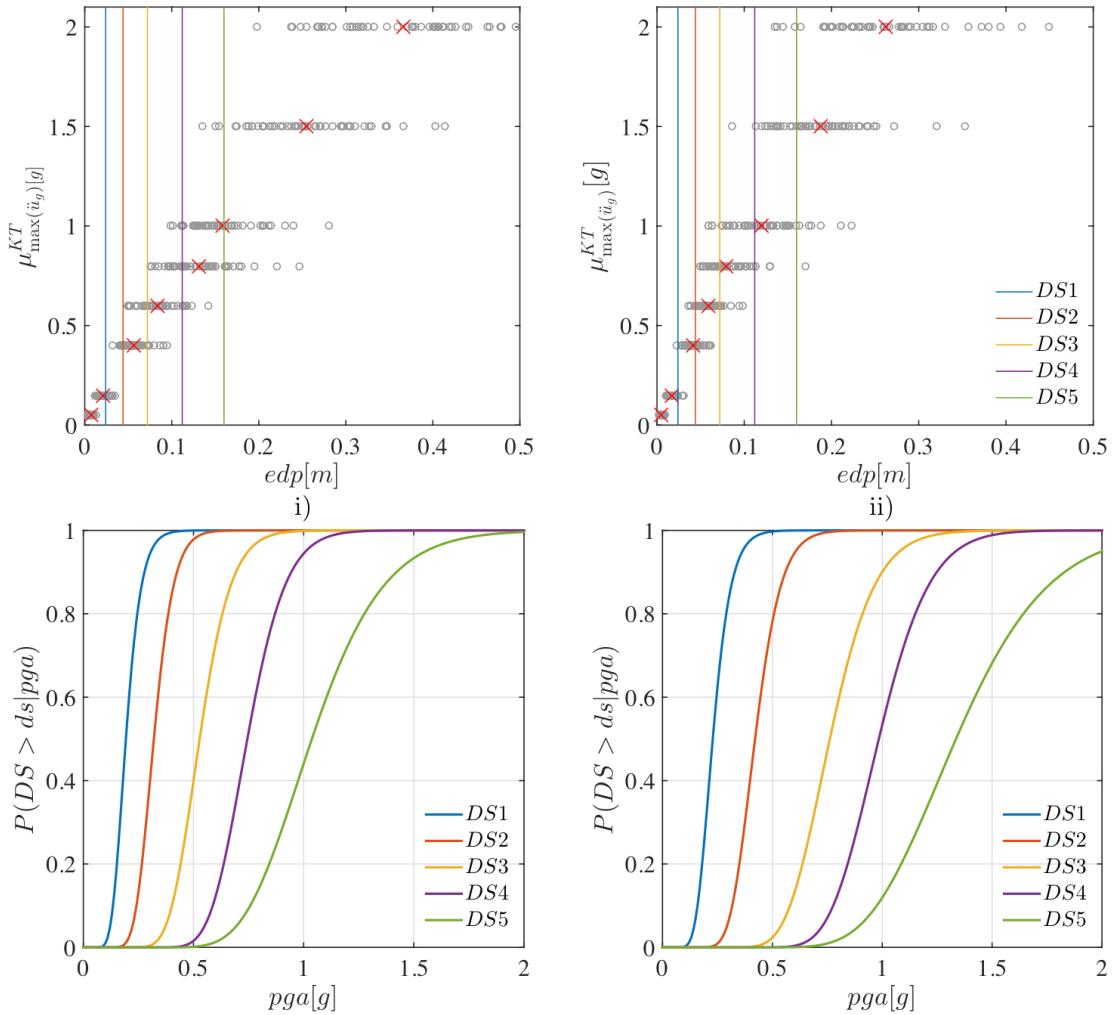


Figure 4: Fragility function computed via MSA for two compelling structural systems

An important observation is that any of the derived likelihood actually can be used also when the stripes are not defined, which is the case when the ground motion is not scaled. In this case, $N_m = 1$ and M the total number of ground motions. Moreover the IDA can be casted in this framework by defining each stripe associated with given scale factor.

The limitation of the method relies in the potential statistical dependence between different stripes, in case same ground motions are used for different stripes.