CISM COURSES AND LECTURES NO. 491
INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES

PROBABILISTIC METHODS IN GEOTECHNICAL ENGINEERING

EDITED BY

D. V. GRIFFITHS
GORDON A. FENTON

SpringerWienNewYork

SpringerWienNewYork

CISM COURSES AND LECTURES

The series presents lecture notes, monographs, edited works and proceedings in the field of Mechanics, Engineering, Computer Science and Applied Mathematics.
Purpose of the series is to make known in the international scientific and technical community results obtained in some of the activities organized by CISM, the International Centre for Mechanical Sciences.

INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES

COURSES AND LECTURES - No. 491

PROBABILISTIC METHODS IN GEOTECHNICAL ENGINEERING

EDITED BY

D. V. GRIFFITHS
COLORADO SCHOOL OF MINES, GOLDEN, USA

GORDON. A. FENTON
DALHOUSIE UNIVERSITY, HALIFAX, CANADA

SpringerWienNewYork

This volume contains 149 illustrations

Printed in Italy
SPIN 12048519

All contributions have been typeset by the authors.

ISBN 978-3-211-73365-3 SpringerWienNewYork

PREFACE

Soils and rocks are among the most variable of all engineering materials, and as such are highly amenable to a probabilistic treatment. The application of statistical and probabilistic concepts to geotechnical analysis is a rapidly growing area of interest for both academics and practitioners. The course is therefore aimed at students, researchers, and practitioners of geotechnical engineering who wish to keep abreast of developments in this evolving field of study. The course content and delivery will assume no more that an introductory understanding of probability and statistics on the part of the course participants.

The main objective of the course is to present a state-of-the-art training on probabilistic techniques applied to geotechnical engineering in relation to both theory and practice. The course will include:

(a) discussion of potential benefits of probabilistic approaches as opposed to the classical "Factor of Safety" methods, to review sources of uncertainty in geotechnical analysis and to introduce methods of LRFD and reliability concepts in Eurocode 7,

(b) review of relevant statistical theories needed to develop the methodologies and interpret the results of probabilistic analysis,

(c) examples of established probabilistic methods of analysis in geotechnical engineering, such as the First Order Second Moment (FOSM) method, the Point Estimate Method (PEM), the First and Second Order Reliability Methods (FORM/SORM) and Random Set (RS) theory.

(d) description of numerical methods of probabilistic analysis based on the finite element method, such as the Stochastic Finite Element Method (SFEM) and recent developments on the Random Finite Element Method (RFEM),

(e) practical examples and case histories of probabilistic applications in geotechnical engineering.

Gordon A. Fenton
D. V. Griffiths

CONTENTS

Review of Probability Theory, Random Variables, and Random Fields

Gordon A. Fenton* and D.V. Griffiths†

* Department of Engineering Mathematics, Dalhousie University, Canada
† Division of Engineering, Colorado School of Mines, U.S.A.

Abstract Regulatory bodies are increasingly asking geotechnical engineers to provide rational risk assessments to accompany their designs. In order to provide these assessments, practicing geotechnical engineers need a good understanding of both basic probability theory and the more sophisticated, but realistic, random field soil models. This chapter lays the groundwork for this understanding. Starting with the basics of probability, the reader is lead through the theory of random variables and random fields and how they can be used to realistically model spatially variable soils.

1 Event Probabilities

The probability of an event A, denoted by $\mathrm{P}[A]$, is a number satisfying

$$0 \le \mathrm{P}[A] \le 1$$

Also, we assume that

$$\mathrm{P}[\emptyset] = 0, \quad \mathrm{P}[S] = 1.$$

Probabilities can sometimes be obtained using the counting rules discussed in the previous section. For example, if an experiment can result in any one of N different but equally likely outcomes, and if exactly m of these outcomes correspond to event A, then the probability of event A is $\mathrm{P}[A] = m/N$.

1.1 Additive Rules

Often we must compute the probability of some event which is expressed in terms of other events. For example, if A is the event that the company A requests your services and B is the event that company B requests your services, then the event that at least one of the two companies request your services is $A \cup B$. The probability of this is given by the following relationship;

If A and B are any two events, then

$$\mathrm{P}[A \cup B] = \mathrm{P}[A] + \mathrm{P}[B] - \mathrm{P}[A \cap B] \tag{1}$$

This relationship can be illustrated by the following Venn diagram. The desired quantity, $\mathrm{P}[A \cup B]$, is the area of $A \cup B$ which is shaded. If the shaded area is computed as the sum of the area of A, $\mathrm{P}[A]$, plus the area of B, $\mathrm{P}[B]$, then the intersection area, $\mathrm{P}[A \cap B]$, has been added twice. It must then be removed once to obtain the correct probability.

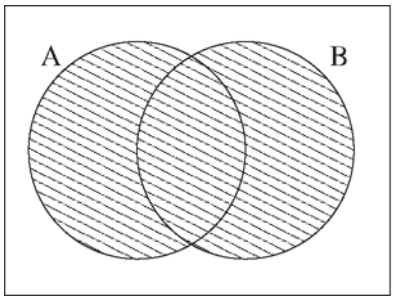

Figure 1. Venn diagram illustrating the union $A \cup B$.

Also,

If A and B are mutually exclusive, i.e. are disjoint and have no overlap, then

$$\mathrm{P}[A \cup B] = \mathrm{P}[A] + \mathrm{P}[B] \tag{2}$$

If $A_1, A_2, ..., A_n$ are mutually exclusive, then

$$\mathrm{P}[A_1 \cup ... \cup A_n] = \mathrm{P}[A_1] + ... + \mathrm{P}[A_n]. \tag{3}$$

Definition: We say that $A_1, A_2, ..., A_n$ is a **partition** of the sample space S if $A_1, A_2, ..., A_n$ are mutually exclusive and collectively exhaustive. Collectively exhaustive means that $A_1 \cup A_2 \cup ... \cup A_n = S$.
If $A_1, A_2, ..., A_n$ is a partition of the sample space S, then

$$\mathrm{P}[A_1 \cup ... \cup A_n] = \mathrm{P}[A_1] + ... + \mathrm{P}[A_n] = \mathrm{P}[S] = 1 \tag{4}$$

The above ideas can be extended to the union of more than two events. For example,

For any three events A, B, and C, we have

$$\begin{aligned}\mathrm{P}[A \cup B \cup C] =& \mathrm{P}[A] + \mathrm{P}[B] + \mathrm{P}[C] \\ &- \mathrm{P}[A \cap B] - \mathrm{P}[A \cap C] - \mathrm{P}[B \cap C] \\ &+ \mathrm{P}[A \cap B \cap C]\end{aligned} \tag{5}$$

This can be seen by drawing a Venn diagram and keeping track of the areas which must be added and removed in order to get $\mathrm{P}[A \cup B \cup C]$.

For the complementary events A and A^c, $\mathrm{P}[A] + \mathrm{P}[A^c] = 1$. This is often used to compute $\mathrm{P}[A^c] = 1 - \mathrm{P}[A]$.

2 Conditional Probability

The probability of an event is often affected by the occurrence of other events and/or the knowledge of information relevant to the event. Given two events, A and B, of an experiment, $\mathrm{P}[B \mid A]$ is called the conditional probability of B given that A has already occurred. It is defined by

$$\mathrm{P}[B \mid A] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[A]} \tag{6}$$

That is, if we are given that event A has occurred, then A becomes our sample space. The probability that B has also occurred within this new sample space will be the ratio of the "area" of B within A to the "area" of A.

Sometimes we know $\mathrm{P}[B \mid A]$ and wish to compute $\mathrm{P}[A \cap B]$. If the events A and B can both occur, then

$$\mathrm{P}[A \cap B] = \mathrm{P}[B \mid A]\,\mathrm{P}[A] \tag{7}$$

2.1 Total Probability

Sometimes we know the probability of an event in terms of the occurrence of other events and want to compute the *unconditional* probability of the event. For example, when we want to compute the *total* probability of failure of a bridge, we can start by computing a series of simpler problems such as

1) the probability of bridge failure given a maximum static load,
2) the probability of bridge failure given a maximum dynamic traffic load,
3) the probability of bridge failure given an earthquake,
4) the probability of bridge failure given a flood,

etc. The *Total Probability Theorem* can be used to combine the above probabilities into the unconditional probability of network failure. We need to know the above conditional probabilities along with the probabilities that the 'conditions' occur (e.g. the probability that the maximum static load will occur during the design life, etc.).

The Total Probability Theorem is stated generally as follows;

Total Probability Theorem:

If the events $B_1, B_2, ..., B_k$ constitute a partition of the sample space S (i.e. are disjoint but collectively exhaustive) then for any event A in S,

$$\mathrm{P}[A] = \sum_{i=1}^{k} \mathrm{P}[B_i \cap A] = \sum_{i=1}^{k} \mathrm{P}[A \mid B_i]\,\mathrm{P}[B_i] \tag{8}$$

2.2 Bayes' Theorem

Sometimes we want to improve an estimate of a probability in light of additional information. Bayes' Theorem allows us to do this. It arises from the observation that $\mathrm{P}[A \cap B]$ can be written in two ways;

$$\begin{aligned} \mathrm{P}[A \cap B] &= \mathrm{P}[A \mid B] \cdot \mathrm{P}[B] \\ &= \mathrm{P}[B \mid A] \cdot \mathrm{P}[A] \end{aligned} \tag{9}$$

which implies that $\mathrm{P}[B \mid A] \cdot \mathrm{P}[A] = \mathrm{P}[A \mid B] \cdot \mathrm{P}[B]$, or

$$\mathrm{P}[B \mid A] = \frac{\mathrm{P}[A \mid B] \cdot \mathrm{P}[B]}{\mathrm{P}[A]} \tag{10}$$

Bayes' Theorem is stated formally as follows.

Bayes' Theorem:

If the events $B_1, B_2, ..., B_k$ constitute a partition of the sample space S (i.e. are disjoint and collectively exhaustive) then for any event A of S, such that $\mathrm{P}[A] \neq 0$,

$$\mathrm{P}[B_j \mid A] = \frac{\mathrm{P}[B_j \cap A]}{\sum_{i=1}^{k} \mathrm{P}[B_i \cap A]} = \frac{\mathrm{P}[A \mid B_j]\,\mathrm{P}[B_j]}{\sum_{i=1}^{k} \mathrm{P}[A \mid B_i]\,\mathrm{P}[B_i]} = \frac{\mathrm{P}[A \mid B_j]\,\mathrm{P}[B_j]}{\mathrm{P}[A]} \tag{11}$$

for any $j = 1, 2, ..., k$.

Bayes' Theorem is useful for revising or updating probabilities as more data and information becomes available. In the previous example on piezocones, there was an *initial* probability that a piezocone would have been manufactured at plant A: $\mathrm{P}[A] = 0.5$. This probability is referred to as the **prior** probability of A. That is, in the absence of any other information, a piezocone chosen at random has a probability of having been manufactured at plant A of 0.5. However, if a piezocone, chosen at random, is found to be defective (so that there is now more information on the piezocone), then its probability that it was manufactured at plant A reduces from 0.5 to 0.294. This latter probability is referred to as the **posterior** probability of A. **Bayesian** updating of probabilities is a very powerful tool in engineering reliability-based design.

For problems involving conditional probabilities, event trees are usually the easiest way to proceed. However, event trees are not always easy to draw, and the purely mathematical approach is sometimes necessary. As an example of a tree which is not quite straightforward, see if you can draw the event tree and answer the questions in the following *Exercise*. Remember that you must set up the tree in such a way that you can fill in most of the probabilities on the branches. If you are left with too many empty branches and no other given information, you are likely to have confused the order of the events - try reorganizing your tree.

2.3 Problem-Solving Methodology

Solving real-life problems (i.e. 'word problems') is not always easy. It is often not perfectly clear what is meant by a worded question. Two things improve one's chances of successfully solving problems which are expressed using words: (a) a systematic approach, and (b) *practice*. It is practice that allows you to identify those aspects of the question that need further clarification, if any. Below, a few basic recommendations are outlined.

1) Solving a word problem generally involves the computation of some quantity. Clearly identify this quantity at the beginning of the problem solution. Before starting any computations, it is good practice to write out your concluding sentence first. This forces you to concentrate on the essentials.
2) In any problem involving the probability of events, you should
   a) **clearly define your events.** Use the following guidelines:
      i) Keep events as simple as possible.
      ii) if your event definition includes the words "*and*", "*or*", "*given*", "*if*", "*when*", etc., then **it is NOT a good event definition.** Break your event into two (or more, if required) events and use " $\cap$ ", " $\cup$ ", or "|"

operators to express what you had originally intended. The complement is also a helpful operator, see (iii).

iii) You do not need to define separate events for, for example, "*an accident occurs*" and "*an accident does not occur*". In fact, this will often lead to confusion. Simply define A to be one of the events and use A^c when you want to refer to the other. This may also give you some hints as to how to proceed since you know that $\mathrm{P}[A^c] = 1 - \mathrm{P}[A]$.

b) Once your events are defined, you need to go through the worded problem to extract the given numerical information. Write this information down in the form of probabilities of the events that you defined above. For example, $\mathrm{P}[A] = 0.23$, $\mathrm{P}[B \mid A] = 0.6$, etc. Note that the conditional probabilities, are often difficult to unravel. Phrases such as

'if ... occurs, the probability of ... doubles...'
'In the event that ... occurs, the probability of ... becomes 0.6'
'When ... occurs, the probability of ... becomes 0.43'
'Given that ... occurs, the probability of ... is 0.3'

all translate into a probability statement of the form $\mathrm{P}[A \mid B]$. In this case, you will likely be using one of the conditional probability relationship ($\mathrm{P}[A \cap B] = \mathrm{P}[B \mid A]\,\mathrm{P}[A]$), the Total Probability Theorem, or Bayes' Theorem.

c) Now review the worded problem again and write down the probability that the question is asking for in terms of the events defined above. Although the question may be in worded form, you should be writing down something like $\mathrm{P}[A \cap B]$ or $\mathrm{P}[B \mid A]$, etc. Make sure that you can express the desired probability in terms of the events you defined above. If you can't, then you need to revise your original event definitions.

d) Finally, use the rules of combining probabilities (e.g. probabilities of unions, intersections, Bayes' Theorem, etc) to compute the desired probability.

3 Random Variables and Probability Distributions

Although probability theory is based on the idea of events and associated set theory, it becomes very unwieldy to treat random events like 'time to failure' using explicit event definitions. One would conceivably have to define a separate event for each possible time of failure and so would soon run out of symbols for the various events. For this reason, and also because they allow the use of a wealth of mathematical tools, *random variables* are used to represent a suite of possible events. In addition, since most engineering problems are expressed in terms of numerical quantities, random variables are particularly appropriate.

Definition: Consider a sample space S consisting of a set of outcomes $\{s_1, s_2, \ldots\}$. If X is a function that assigns a real number $X(s)$ to every outcome $s \in S$, then X is a *random variable*. Random variables will be denoted with upper case letters.

Now what does this mean in plain English? Essentially a random variable is a means of identifying events in numerical terms. For example, if the outcome s_1 means that an apple was selected and s_2 means that an orange was selected, then $X(s_1)$ could be

set equal to 1 and $X(s_2)$ could be set equal to 0. $X > 0$ then means that an apple was selected. Now mathematics can be used on X, ie. if the fruit picking experiment is repeated n times and $x_1 = X_1(s)$ is the outcome of the first experiment, $x_2 = X_2(s)$ the outcome of the second, etc., then the total number of apples picked is $\sum_{i=1}^{n} x_i$. Note that mathematics could not be used on the actual outcomes themselves, e.g. picking an apple is a real event which knows nothing about mathematics nor can it be used in a mathematical expression without first mapping the event to a number.

For each outcome s, there is exactly one value of $x = X(s)$, but different values of s may lead to the same x.

The above discussion illustrates in a rather simple way one of the primary motivation for the use of random variables – simply so that mathematics can be used. One other thing might be noticed in the previous paragraph. After the 'experiment' has taken place and the outcome is known, it is referred to using the lower case, x_i. That is x_i has a known fixed value while X is unknown. In other words x is a realization of the random variable X. This is a rather subtle distinction, but it is important to remember that X is unknown. The most that we can say about X is to specify what its likelihoods of taking on certain values are – we cannot say exactly what the value of X is.

3.1 Discrete Random Variables

Discrete random variables are those that take on only discrete values $\{x_1, x_2, \ldots\}$, ie. have a countable number of outcomes. Note that countable just means that the outcomes can be numbered $1, 2, \ldots$, however there could still be an infinite number of them. For example, our experiment might be to count the number of soil tests performed before one yields a cohesion of 200 MPa. This is a discrete random variable since we outcome is one of $0, 1, \ldots$, but the number may be very large or even (in concept) infinite (implying that a soil sample with cohesion 200 MPa was never found).

Discrete Probability Distributions

As mentioned previously, we can never know for certain what the value of a random variable is (if we do measure it, it becomes a *realization* – presumably the next measurement is again uncertain until it is measured, and so on). The most that we can say about a random variable is what its probability is of assuming each of its possible values. The set of probabilities assigned to each possible value of X is called a **probability distribution.** The sum of these probabilities, over all possible values, must be 1.0.

Definition: The set of ordered pairs $(x, f_X(x))$ is the probability distribution of the discrete random variable X if, for each possible outcome x,

1) $0 \leq f_X(x) \leq 1$,
2) $\sum_{\text{all } x} f_X(x) = 1$,
3) $\mathrm{P}[X = x] = f_X(x)$

$f_X(x)$ is called the *probability mass function* of X. The subscript is used to indicate what random variable is being governed by the distribution. We shall see when we consider continuous random variables why we call this a probability 'mass' function.

Discrete Cumulative Distributions

An *equivalent* description of a random variable is the cumulative distribution function (CDF), which is defined as follows;

Definition: The cumulative distribution function, $F_X(x)$ of a discrete random variable X, with probability mass function $f_X(x)$, is defined by

$$F_X(x) = \mathrm{P}[X \leq x] = \sum_{t \leq x} f_X(t) \tag{12}$$

We say that this is equivalent to the probability mass function because one can be obtained from the other,

$$f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \tag{13}$$

3.2 Continuous Random Variables

Continuous random variables can take on an infinite number of possible outcomes – generally X takes values from the real line $\Re$. To illustrate the changes involved when we go from the discrete to the continuous case, consider the probability that a grain silo experiences a bearing capacity failure at *exactly* 4.3673458212... years from when it is installed. Clearly the probability that it fails at *exactly* that instant in time is essentially zero. In general the probability that it fails at any one instant in time is vanishingly small. In order to characterize probabilities for continuous random variables, we can't use probabilities directly (since they are all essentially zero) – we must use *relative likelihoods.* That is, we say that the probability that X lies in the small interval between x and $x + dx$ is $f_X(x)\,dx$, or

$$\mathrm{P}[x < X \leq x + dx] = f_X(x)\,dx \tag{14}$$

where $f_X(x)$ is now called the *probability density function* (pdf) of the random variable X. The word *density* is used because "density" must be multiplied by a length measure in order to get a "mass". Note that the above probability is vanishingly small because dx is vanishingly small. The function $f_X(x)$ is now the relative likelihood that X lies in a very small interval near x. Roughly speaking, we can think of this as $\mathrm{P}[X = x] = f_X(x)dx$.

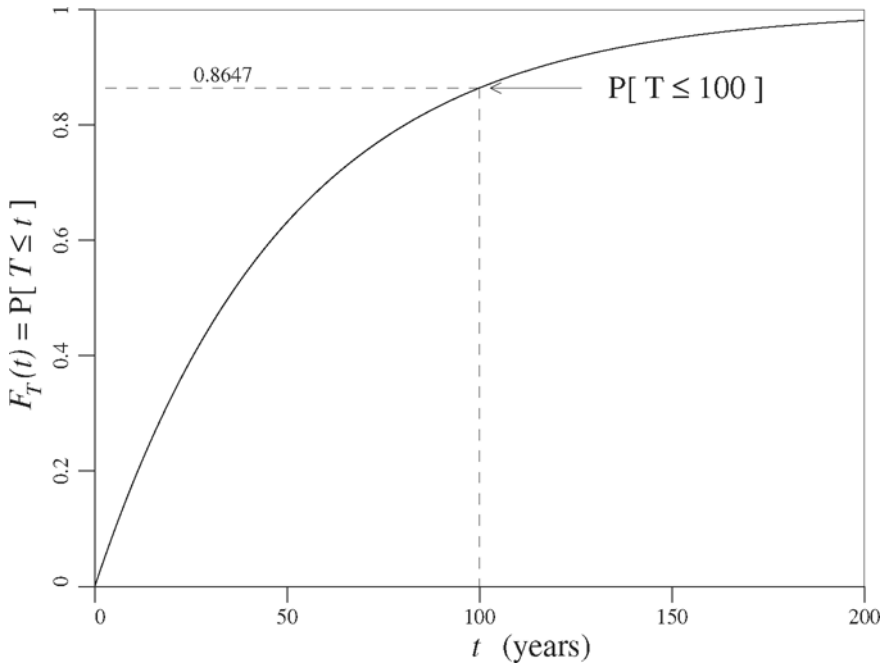

Figure 2. Cumulative distribution function for the exponential distribution.

Continuous Probability Distributions

Definition: The function $f_X(x)$ is a probability density function for the continuous random variable X, defined over the set of real numbers, if

1) $0 \le f_X(x) < \infty$, for all $-\infty < x < +\infty$,
2) $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (i.e. the area under the pdf is 1.0),
3) $\mathrm{P}[a < X < b] = \int_a^b f_X(x)dx$ (i.e. the area under $f_X(x)$ between a and b).

NOTE: it is important to recognize that, in the continuous case, $f_X(x)$ is no longer a probability. It has units of probability per unit length. In order to get probabilities, we have to find *areas* under the pdf, i.e. sum up values of $f_X(x)dx$.

Continuous Cumulative Distribution

The *cumulative distribution function* (cdf), for a continuous random variable is basically defined in the same way as it is for a discrete distribution.

Definition: The cumulative distribution function, $F_X(x)$, of a continuous random variable X having probability density function $f_X(x)$, is defined by the area under the density function to the left of x

$$F_X(x) = \mathrm{P}[X \leq x] = \int_{-\infty}^{x} f_X(t)dt \tag{15}$$

As in the discrete case, the cdf is equivalent to the pdf, in that one can be obtained from the other. It is simply another way of expressing the probabilities associcated with a random variable. Since the cdf is an integral of the pdf, the pdf can be obtained from the cdf as a derivative, ie.

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{16}$$

4 Measures of Central Tendency, Variability, and Association

A random variable is completely described, as well as can be, if its probability distribution is specified. However, we will never know the precise distribution of any natural phenomenon. Nature cares not at all about our mathematical models and the 'truth' is usually far more complex than we are able to represent. So we very often have to describe a random variable using less complete, but more easily estimated, measures. The most important of these measures are *central tendency* and *variability*. Even if the complete probability distribution is known, these quantities remain useful because they convey information about the properties of the random variable that are of first importance in practical applications. Also, the parameters of the distribution are often derived as functions of these quantities, or they may be the parameters themselves.

The most common measures of central tendency and variability are the *mean* and the *variance*, respectively. In engineering, the variability of a random quantity is often expressed using the dimensionless *coefficient of variation* which is the ratio of the *standard deviation* over the *mean*. Also, when one has two random variables, X and Y, it is frequently of interest to measure how strongly they are related (or associated) to one another. A typical measure of the strength of the relationship between two random variables is their *covariance*. As we shall see, covariance depends on the units of the random variables involved and their individual variabilities, and so a more intuitive measure of the strength of the relationship between two random variables is the *correlation coefficient*, which is both dimensionless and bounded. All of these characteristics will be covered in this section.

4.1 Mean

The *mean* is the most important characteristic of a random variable, in that it tells us about its central tendency. It is defined mathematically as follows;

Definition: Let X be a random variable with probability density function $f(x)$. The *mean* or *expected value* of X, denoted μ_X, is defined by

$$\mu_X = \mathrm{E}[X] = \sum_x xf(x) \qquad \text{if } X \text{ is discrete} \tag{17a}$$

$$\mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{if } X \text{ is continuous.} \tag{17b}$$

where the subscript on μ, when present, denotes what μ is the mean of.

Expectation

The notation $\mathrm{E}[X]$ refers to a mathematical operation called *expectation*. The expectation of any random variable is a sum of all possible values of the random variable weighted by the probability of each value occurring. For example, if X is a random variable with probability (mass or density) function, $f_X(x)$, then the expected value of the random variable $g(X)$, where g is any function of X, is

$$\mu_{g(X)} = \mathrm{E}[g(X)] = \sum_x g(x) f_X(x) \qquad \text{if } X \text{ is discrete}$$

$$\mu_{g(X)} = \mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \qquad \text{if } X \text{ is continuous.} \tag{18}$$

If we have a sample of observations , $x_1, x_2, \ldots, x_n$, of some population X, then the population mean, μ_X, is estimated by the *sample mean*, $\bar{x}$, defined as

$$\boxed{\text{Sample Mean:}} \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

4.2 Median

The median is another measure of central tendency. We shall denote the median as $\tilde{\mu}$. It is the point which divides the distribution into two equal halves. Most commonly, $\tilde{\mu}$ is found by solving

$$F_X(\tilde{\mu}) = \mathrm{P}[X \leq \tilde{\mu}] = 0.5$$

for $\tilde{\mu}$. For example, if $f_X(x) = \lambda e^{-\lambda x}$, then $F_X(x) = 1 - e^{-\lambda x}$, and we get

$$1 - e^{-\lambda \tilde{\mu}} = 0.5 \qquad \Longrightarrow \qquad \tilde{\mu}_X = -\frac{\ln(0.5)}{\lambda} = \frac{0.693}{\lambda}$$

While the mean is strongly affected by extremes in the distribution, the median is largely unaffected.

In general the mean and the median are not the same. If the distribution is positively skewed (or skewed right, which means a longer tail to the right than to the left), as are most soil properties, then the mean will be to the right of the median. Conversely, if the distribution is skewed left, then the mean will be to the left of the median. If the distribution is symmetric, then the mean and the median will coincide.

NOTE: the median is the point which divides the distribution in half.

If we have a sample of observations , $x_1, x_2, \ldots, x_n$, of some population X, then the population median, $\tilde{\mu}_X$, is estimated by the *sample median*, $\tilde{x}$. To define $\tilde{x}$, we must first order the observations from smallest to largest, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. When we have done so, the sample median is defined as

$$\boxed{\text{Sample Median}} \qquad \tilde{x} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{((n+1)/2)}\right) & \text{if } n \text{ is even} \end{cases}$$

4.3 Variance

The mean (expected value) or median of the random variable X tells where the probability distribution is "centered". The next most important characteristic of a random variable is whether the distribution is "wide", "narrow", or somewhere in between. This distribution "variability" is commonly measured by a quantity call the variance of X.

Definition: Let X be a random variable with probability (mass or density) function $f_X(x)$ and mean μ_X. The **variance**, σ_X^2, of X is defined by

$$\sigma_X^2 = \mathrm{Var}[X] = \mathrm{E}\left[(X-\mu_X)^2\right] = \sum_x (x-\mu_X)^2 f_X(x) \tag{19a}$$

$$\sigma_X^2 = \mathrm{Var}[X] = \mathrm{E}\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx \tag{19b}$$

for the discrete and continous cases, respectively.

The variance of the random variable X is sometimes more easily computed as

$$\sigma_X^2 = \mathrm{E}\left[X^2\right] - \mathrm{E}^2[X] = \mathrm{E}\left[X^2\right] - \mu_X^2 \tag{20}$$

The variance, σ_X^2, has units of X^2. The square root of the variance, σ_X, is called the *standard deviation* of X. Since the standard deviation has the same units as X, it is often preferable to report the standard deviation as a measure of variability.

Even though the standard deviation has the same units as the mean, it is often still not particularly informative. For example, a standard deviation of 1.0 may indicate significant variability when the mean is 1.0, but indicates virtually deterministic behaviour when the mean is one million. For example, an error of 1 m on a 1 m survey would be considered unacceptable, whereas an error of 1 m on a one thousand km survey might be considered quite accurate. A measure of variability which is both non-dimensional and delivers a relative sense of the magnitude of variability is the *coefficient of variation*, defined as

$$v = \frac{\sigma}{\mu} \tag{21}$$

Note that the coefficient of variation becomes undefined if the mean of X is zero. It is, however, quite popular as a way of expressing variability in engineering, particularly for material property and load variability, which generally have non-zero means.

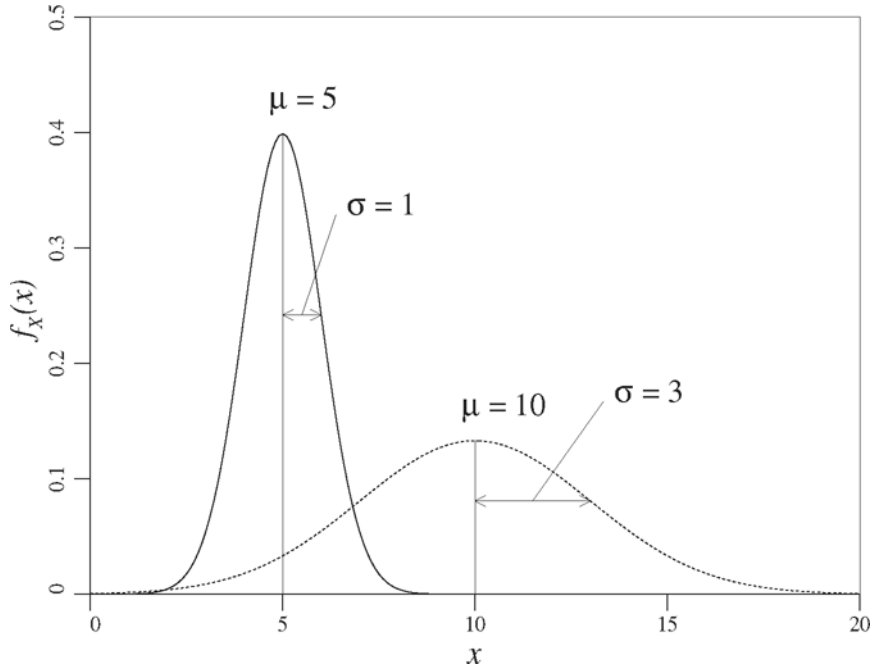

Figure 3. Two distributions illustrating how the position and shape changes with changes in mean and variance.

4.4 Covariance

Often one must consider more than one random variable at a time. For example, the two components of a drained soil's shear strength, $\tan(\phi')$ and c', will vary randomly from location to location in a soil. These two quantities can be modeled by two random variables, and since they may influence one another (or they may be jointly influenced by some other factor), they are characterized by a *bivariate distribution*.

Properties of the Bivariate Distribution:

Discrete:

$$f_{XY}(x,y) = \mathrm{P}[X = x \cap Y = y]$$

$$0 \le f_{XY}(x,y) \le 1$$

$$\sum_{\text{all } x} \sum_{\text{all } y} f_{XY}(x,y) = 1$$

Continuous:

$$f_{XY}(x,y)\,dx\,dy = \mathrm{P}[x < X \le x + dx \cap y < Y \le y + dy]$$

$$f_{XY}(x,y) \ge 0 \text{ for all } (x,y) \in \Re^2$$

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

$$\mathrm{P}[x_1 < X \le x_2 \cap y_1 < Y \le y_2] = \int_{y_1}^{y_2}\int_{x_1}^{x_2} f_{XY}(x,y)\,dx\,dy$$

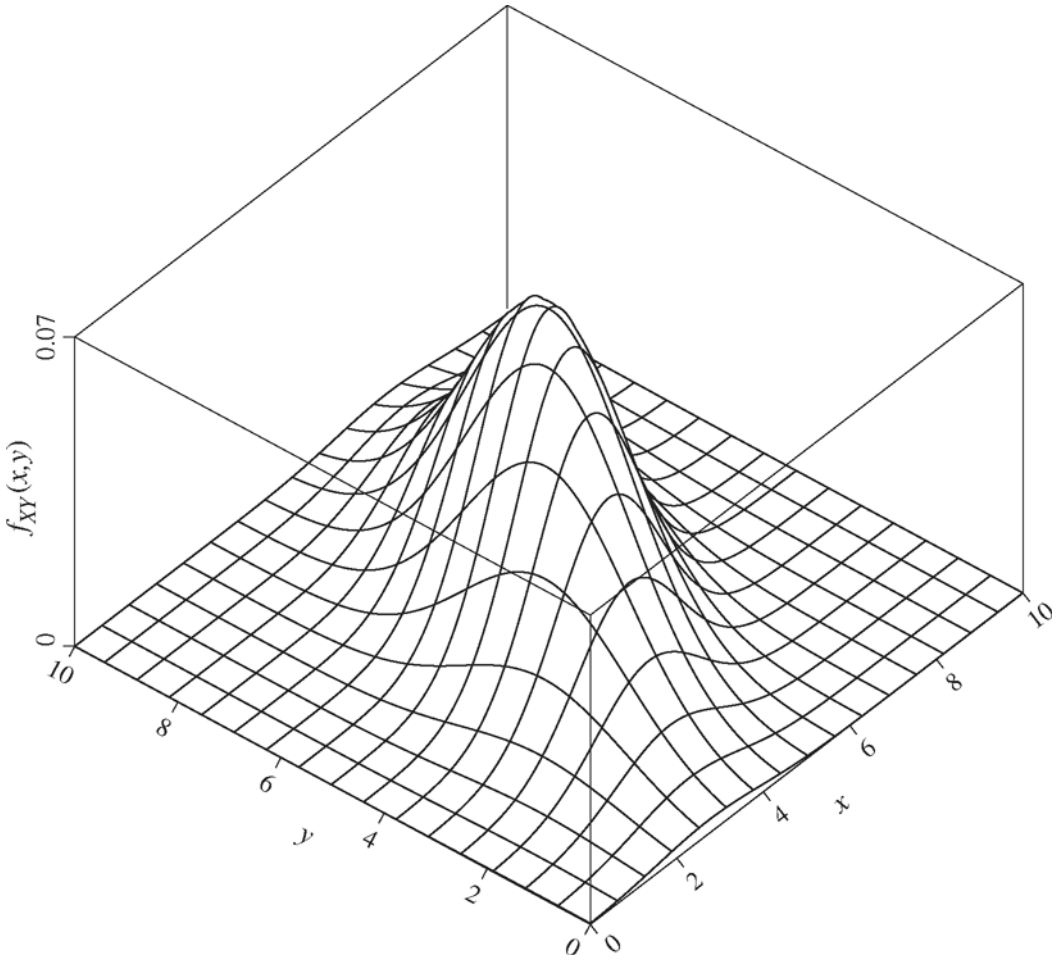

Figure 4. Example bivariate probability density function, $f_{XY}(x,y)$.

Definition: Let X and Y be random variables with joint probability distribution $f_{XY}(x,y)$. The *covariance* between X and Y is defined by

$$\begin{aligned} \mathrm{Cov}[X,Y] &= \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \sum_x \sum_y (x-\mu_x)(y-\mu_Y) f_{XY}(x,y), \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{XY}(x,y)\, dx\, dy, \end{aligned} \tag{22}$$

for the discrete and continuous cases, respectively.

The covariance between two random variables X and Y, having means μ_X and μ_Y, respectively, may also be computed as

$$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = \mathrm{E}[XY] - \mu_X\mu_Y \tag{23}$$

Although the covariance between two random variables does give information regarding the nature of the relationship, the magnitude of $\mathrm{Cov}[X,Y]$ does not indicate anything regarding the strength of the relationship. This is because $\mathrm{Cov}[X,Y]$ depends on the units and variability of X and Y. A quantity which is both normalized and non-dimensional is the correlation coefficient, to be discussed next.

4.5 Correlation Coefficient

Definition: Let X and Y be random variables with joint probability distribution $f_{XY}(x,y)$. The **correlation coefficient** between X and Y is defined to be

$$\rho_{XY} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X\sigma_Y}. \tag{24}$$

Figure 5 illustrates the effect that the correlation coefficient has on the shape of a bivariate probability density function, in this case for X and Y jointly normal. If $\rho_{XY} = 0$, then the contours form ovals with axes aligned with the cartesian axes (if the variances of X and Y are equal, then the ovals are circles). When $\rho_{XY} > 0$, the ovals become stretched and the major axis has a positive slope. What this means is that when Y is large, X will also tend to be large. For example, when $\rho_{XY} = 0.6$, as shown on the right plot of Figure 5, then when $Y = 8$, the most likely value X will take is around 7, since this is the peak of the distribution along the line $Y = 8$. Similarly, if $\rho_{XY} < 0$, then the ovals will be oriented so that the major axis has a negative slope. In this case, large values of Y will tend to give small values of X.

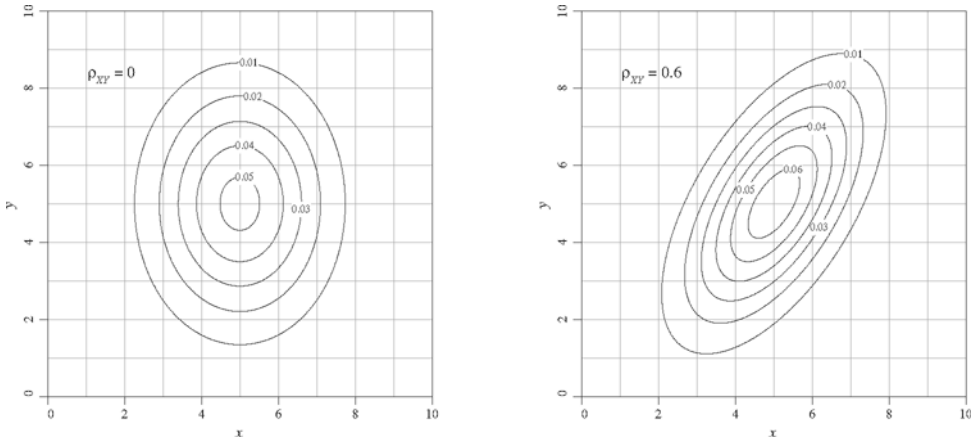

Figure 5. Effect of correlation coefficient, ρ_{XY}, on contours of a bivariate probability density function, $f_{XY}(x,y)$, having $\mu_X = \mu_Y = 5$ and $\sigma_X = 1.5$ and $\sigma_Y = 2.0$.

We can show that $-1 \leq \rho_{XY} \leq 1$ as follows: Consider two random variables X and Y having variances σ_X^2 and σ_Y^2, respectively, and correlation coefficient ρ_{XY}. Then

$$\begin{aligned} \text{Var}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] &= \frac{\sigma_X^2}{\sigma_X^2} + \frac{\sigma_Y^2}{\sigma_Y^2} + 2\frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y} \\ &= 2\Big[1 + \rho_{XY}\Big] \\ &\geq 0 \end{aligned}$$

which implies that $\rho_{XY} \geq -1$. Similarly,

$$\begin{aligned} \text{Var}\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right] &= \frac{\sigma_X^2}{\sigma_X^2} + \frac{\sigma_Y^2}{\sigma_Y^2} - 2\frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y} \\ &= 2\Big[1 - \rho_{XY}\Big] \\ &\geq 0 \end{aligned}$$

which implies that $\rho_{XY} \leq 1$. Taken together, these imply that $-1 \leq \rho_{XY} \leq 1$.

The correlation coefficient is a direct measure of the degree of *linear* dependence between X and Y. When the two variables are perfectly linearly related, ρ_{XY} will be either +1 or -1 (+1 if Y increases with X and -1 if Y decreases when X increases). When $|\rho_{XY}|$ is less that 1, the dependence between X and Y is not completely linear; however, there could still be a strong nonlinear dependence. If two random variables X and Y are independent, then their correlation coefficient will be 0. If the correlation coefficient between two random variables X and Y is 0, **it does not mean** that they are independent, only that they are uncorrelated. Independence is a much stronger statement than is $\rho_{XY} = 0$, since the latter only implies linear independence. For example, $Y = X^2$ may be linearly independent of X (this depends on the range of X), but clearly Y and X are completely (non-linearly) dependent.

5 Common Discrete Probability Distributions

Many engineered systems have the same statistical behaviour – we generally only need a handful of probability distributions to characterize most naturally occurring phenomena. In this section, the most common discrete distribution will be reviewed (the next section looks at the most comment continous distributions). These are the *Bernoulli family* of distributions, since they all derive from the first,

1) Bernoulli
2) Binomial
3) Geometric
4) Negative Binomial
5) Poisson
6) Exponential
7) Gamma

The Poisson, Exponential, and Gamma are the continuous-time analogs of the Binomial, Geometric, and Negative Binomial, respectively, arising when each instant in time is viewed as an independent Bernoulli trial. In this section we consider the *discrete* members

of the Bernoulli family, which are the first five members listed above, looking briefly at the main characteristics of each of these distributions and describing how they are most commonly used in practice.

For a more complete description of these distributions, the interested reader should consult an introductory textbook on probability and statistics, such as Law and Kelton (2000) or Devore (2003).

5.1 Bernoulli Trials

All of the discrete distributions considered in this section (and the first two in the next section) are derived from the idea of *Bernoulli Trials*. A Bernoulli trial is an experiment which has only two possible outcomes, *success* or *failure* (or [1, 0], or [true, false], or [$< 5, \geq 5$], etc). If a sequence of Bernoulli trials are mutually independent with constant (stationary) probability, p, of success, then the sequence is called a *Bernoulli Process*. There are many examples of Bernoulli processes: one might model the failures of earth dams using a Bernoulli process. The success or failure of each of a sequence of bids made by a company might be a Bernoulli process. The failure of piles to support the load applied on them might be a Bernoulli process if it can be assumed that the piles fail (or survive) independently and with constant probability. However, if the failure of one pile is dependent on the failure of adjacent piles, as might be the case if the soil structures are similar and load transfer takes place, the Bernoulli model may not be appropriate and a more complex, 'dependent', model may be required, e.g. random field modeling of the soil and finite element analysis of the structural response within a Monte Carlo simulation. Evidently, when we depart from satisfying the assumptions underlying the simple models, such as those required for the Bernoulli model, the required models rapidly become very much more complicated. In some cases, applying the simple model to the more complex problem will yield a ballpark estimate, or at least a bound on the probability, and so it may be appropriate to proceed with a Bernoulli model taking care to treat the results as approximate. The degree of approximation depends very much on the degree of dependence between 'trials' and the 'stationarity' of the probability of 'success', p.

If we let

$$X_j = \begin{cases} 1 & \text{if the } j^{th} \text{ trial results in a success,} \\ 0 & \text{if the } j^{th} \text{ trial results in a failure} \end{cases} \tag{25}$$

then the Bernoulli distribution, or probability mass function, is given by

$$\begin{aligned} \mathrm{P}[X_j = 1] &= p \\ \mathrm{P}[X_j = 0] &= 1 - p = q \end{aligned} \tag{26}$$

for all $j = 1, 2, \ldots$. Note that we commonly denote $1 - p$ as q for simplicity.

For a single Bernoulli trial the following results hold

$$\mathrm{E}[X_j] = \sum_{i=0}^{1} i \cdot \mathrm{P}[X_j = i] = 0(1-p) + 1(p) = p \tag{27a}$$

$$\mathrm{E}[X_j^2] = \sum_{i=0}^{1} i^2 \cdot \mathrm{P}[X_j = i] = 0^2(1-p) + 1^2(p) = p$$

$$\mathrm{Var}[X_j] = \mathrm{E}\left[X_j^2\right] - \mathrm{E}^2[X_j] = p - p^2 = pq \tag{27b}$$

For a sequence of trials, the assumption of independence between the trials means that

$$\mathrm{P}[X_1 = x_1 \cap X_2 = x_2 \cap \cdots X_n = x_n] = \mathrm{P}[X_1 = x_1]\,\mathrm{P}[X_2 = x_2]\cdots\mathrm{P}[X_n = x_n] \tag{28}$$

The Maximum Likelihood Estimate of p is just the average of the set of observations, $x_1, x_2, \ldots, x_n$, of X,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{29}$$

Notice that we use a hat to indicate that this is just an *estimate* of the true parameter p. Since the next set of observations will likely give a different value for $\hat{p}$, we see that $\hat{p}$ is actually a random variable itself, rather than the true population parameter, which is non-random. The mean and variance of the sequence of $\hat{p}$ can be found by considering the random $\hat{P}$,

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{30}$$

obtained *prior* to observing the results of our Bernoulli trials. We get

$$\begin{aligned}\mathrm{E}\left[\hat{P}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}(np) \\ &= p\end{aligned} \tag{31}$$

which means that the estimator given by Eq. (29) is *unbiased* (that is, the estimator is 'aimed' at its desired target on average).

The estimator variance is

$$\begin{aligned}\mathrm{Var}\left[\hat{P}\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{1}{n^2}(npq) \\ &= \frac{pq}{n}\end{aligned} \tag{32}$$

where we made use of the fact that the variance of a sum is the sum of the variances **if the random variables are uncorrelated.** We are assuming that, since this is a Bernoulli process, not only are the random variables uncorrelated, they are completely independent (the probability of one occurring is not affected by the probability of other occurrences).

Note that the estimator variance depends on the true value of p on the right-hand-side of Eq. (32). Since we are estimating p, we obviously don't know the true value. The solution is to use our estimate of p to estimate its variance, so that

$$\sigma^2_{\hat{p}} \simeq \frac{\hat{p}\hat{q}}{n} \tag{33}$$

Once we have determined the estimator variance, we can compute its *standard error*, which is commonly taken to be equal to the standard deviation and which gives an indication of how accurate our estimate is,

$$\sigma_{\hat{p}} \simeq \sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{34}$$

For example, if $\hat{p} = 0.01$, then we would prefer $\sigma_{\hat{p}}$ to be quite a bit smaller than 0.01 and we can adjust the number of observations, n, to achieve this goal.

Later in this book, we will be estimating the probability of failure, p_f, of various classic geotechnical problems using a technique called *Monte Carlo simulation*. The standard error given by Eq. (34) will allow us to estimate the accuracy of our failure probability estimates, assuming that each 'simulation' results in an independent failure/success 'trial'.

Applications

The classic Bernoulli trial is the toss of a coin, but many other experiments can lead to Bernoulli trials under the above conditions. Consider the following examples;

1) Soil anchors at a particular site have a 1% probability of pulling out. When an anchor is examined, it is classified as a *success* if it has not pulled out, or a *failure* if it has. This is a Bernoulli trial with $p = 0.99$ if the anchors fail independently and if the probability of success remains constant from trial to trial.
2) Suppose that each sample of soil at a site has a 10% chance of containing significant amounts of chromium. A sample is analyzed and classified as a *success* if it does not contain significant amounts of chromium, and a *failure* if it does. This is a Bernoulli trial with $p = 0.90$ if the samples are independent and if the probability of success remains constant from trial to trial.
3) A highway through a certain mountain range passes below a series of steep rock slopes. It is estimated that each rock slope has a 2% probability of failure (resulting in some amount of rock blocking the highway) over the next 10 year. If we define each rock slope as a trial which is a success if it does not fail in the next 10 years, then this can be modeled as a Bernoulli trial with $p = 0.98$ (assuming rock slopes fail independently – which might not be a good assumption if they generally fail due to earthquakes...).

5.2 Binomial Distribution

Let N_n be the number of successes in n Bernoulli trials, each with probability of success p. Then N_n follows a binomial distribution where

$$\mathrm{P}[N_n = k] = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n \tag{35}$$

The quantity $p^k q^{n-k}$ is the probability of obtaining k successes and $n-k$ failures in n trials and $\binom{n}{k}$ is the number of possible ways of arranging the k successes over the n trials.

For example, consider 8 trials, which can be represented as a series of 8 dashes:

____ ____ ____ ____ ____ ____ ____ ____ .

One possible realization of 3 successes in 8 trials might be:

F S F F S S F F

where successes are shown as S and failures as F. Another possible realization might be

S F F S F F F S

and so on. Clearly these involve 3 successes, which have probability p^3, and 5 failures, which have probability q^5. Combining these two probabilities with the fact that 3 successes in 8 trials can be arranged in $\binom{8}{3}$ different ways leads to

$$\mathrm{P}[N_8 = 3] = \binom{8}{3} p^3 q^{8-3}$$

which generalizes to the binomial distribution, for n trials and k successes, given above.

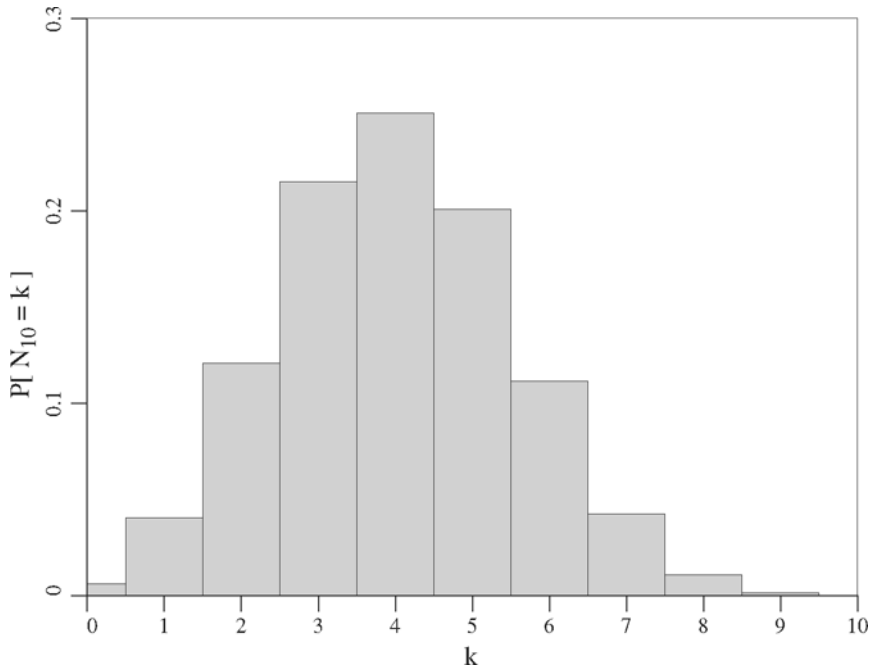

Figure 6. Binomial distribution for $n = 10$ and $p = 0.4$.

Properties:
In the following proofs, we make use of the binomial theorem, which states that

$$(\alpha+\beta)^{n}=\sum_{i=0}^{n}\binom{n}{i}\alpha^{i}\beta^{n-i}=\sum_{i=0}^{n}\frac{n!}{i!(n-i)!}\alpha^{i}\beta^{n-i} \tag{36}$$

The expected number of successes in n trials can be found directly from the definition of the discrete case expectation,

$$\begin{aligned}\mathrm{E}\left[N_{n}\right]&=\sum_{i=0}^{n}i\binom{n}{i}p^{i}q^{n-i}=\sum_{i=0}^{n}i\left(\frac{n!}{i!(n-i)!}\right)p^{i}q^{n-i}\\&=np\sum_{i=1}^{n}\frac{(n-1)!}{(i-1)!(n-i)!}p^{i-1}q^{n-i}=np\sum_{i=0}^{(n-1)}\frac{(n-1)!}{i!((n-1)-i)!}p^{i}q^{(n-1)-i}\\&=np(p+q)^{n-1}\\&=np\end{aligned} \tag{37}$$

since $p+q=1$.
Alternatively, we could write

$$\begin{aligned}\mathrm{E}\left[N_{n}\right]&=\mathrm{E}\left[X_{1}+X_{2}+\cdots+X_{n}\right]\\&=\mathrm{E}\left[X_{1}\right]+\mathrm{E}\left[X_{2}\right]+\cdots+\mathrm{E}\left[X_{n}\right]\\&=np\end{aligned}$$

where X_i is a Bernoulli random variable, having expectation p.

To find the variance of N_n, we first need to find

$$\begin{aligned}\mathrm{E}\left[N_{n}^{2}\right]&=\sum_{i=0}^{n}i^{2}\binom{n}{i}p^{i}q^{n-i}=\sum_{i=1}^{n}i^{2}\left(\frac{n!}{i!(n-i)!}\right)p^{i}q^{n-i}\\&=np\sum_{i=1}^{n}i\left(\frac{(n-1)!}{(i-1)!(n-i)!}\right)p^{i-1}q^{n-i}=np\sum_{i=0}^{n-1}(i+1)\left(\frac{(n-1)!}{i!(n-1-i)!}\right)p^{i}q^{n-1-i}\\&=np\left\{\sum_{i=0}^{n-1}i\left(\frac{(n-1)!}{i!(n-1-i)!}\right)p^{i}q^{n-1-i}+\sum_{i=0}^{n-1}\left(\frac{(n-1)!}{i!(n-1-i)!}\right)p^{i}q^{n-1-i}\right\}\\&=np\left\{(n-1)p+1\right\}\\&=npq+n^{2}p^{2}\end{aligned}$$

where for the first sum, we made use of the result given by Eq. 37. The variance is thus,

$$\mathrm{Var}\left[N_{n}\right]=\mathrm{E}\left[N_{n}^{2}\right]-\mathrm{E}^{2}[N_{n}]=npq+n^{2}p^{2}-n^{2}p^{2}=npq \tag{38}$$

The same result could have been obtained much more easily by considering the variance of a sum of independent random variables, since in this case, the variance of a sum is the sum of the variances;

$$\mathrm{Var}\,[N_n] = \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} VarX_i = npq$$

5.3 Geometric Distribution

Consider a Bernoulli process in which T_1 is the number of trials required to achieve the first success. Thus, if $T_1 = 3$, then we must have had 2 failures followed by a success (the value of T_1 fully prescribes the sequence of trials). This has probability

$$\mathrm{P}\,[T_1 = 3] = \mathrm{P}\,[\{\text{failure, failure, success}\}] = q^2 p$$

In general

$$\mathrm{P}\,[T_1 = k] = q^{k-1} p, \quad k = 1, 2, \ldots \tag{39}$$

Note that this is a valid probability mass function since

$$\sum_{k=1}^{\infty} q^{k-1} p = p \sum_{k=0}^{\infty} q^k = \frac{p}{1-q} = 1$$

where we used the fact that for any $|alpha| < 1$ (see, e.g., Gradshteyn and Ryzhik, 1980),

$$\sum_{k=0}^{\infty} \alpha^k = \frac{1}{1-\alpha} \tag{40}$$

As an example, in terms of the actual sequence of trials, the event that the first success occurs on the 8^{th} trial appears as

F F F F F F F S

That is, the single success always occurs on the last trial. If $T_1 = 8$, then we have had 7 failures, having probability q^7, and 1 success, having probability p. Thus

$$\mathrm{P}\,[T_1 = 8] = q^7 p$$

Generalizing this for $T_1 = k$ leads to the geometric distribution shown above.

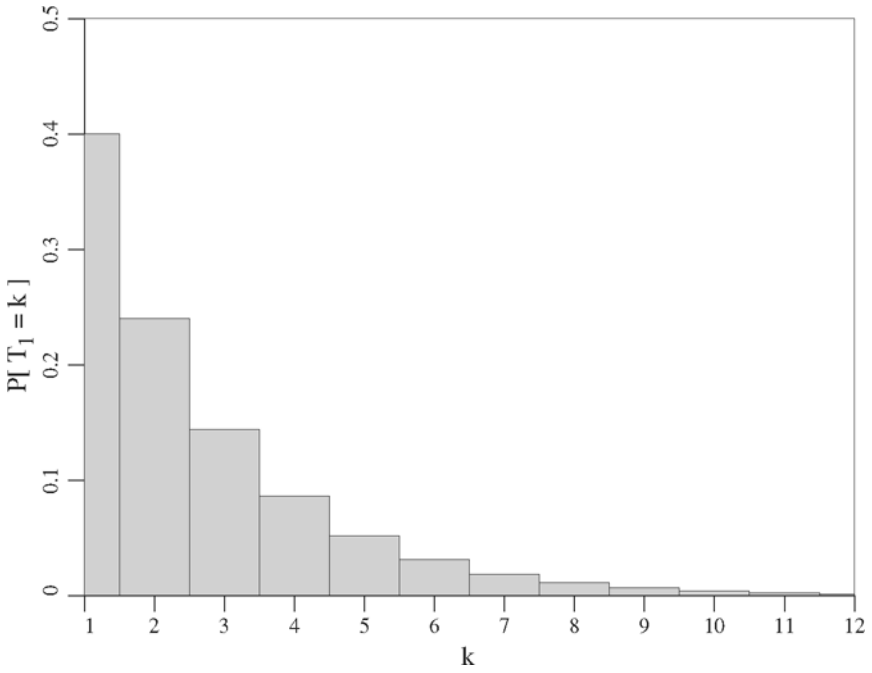

Figure 7. Geometric distribution for $p = 0.4$.

Because trials are assumed independent, the geometric distribution also models the number of trials *between* successes in a Bernoulli process. That is, suppose we observe the result of the Bernoulli process at trial number 1032. We will observe either a success or failure, but whichever is observed, it is now *known*. We can then ask a question such as: What is the probability that the next success occurs on trial 1040? To determine this, we start with trial 1032. Because we have observed it there is no uncertainty associated with trial 1032, so it does not enter into the probability problem. However, trials 1033, 1034, ..., 1040 are unknown. We are asking for the probability that trial 1040 is the first success after 1032. In order for this *event* to occur, trials 1033 to 1039 must be failures. Thus, the 8 trials, 1033 to 1040, must involve seven failures (q^7) followed by one success (p). The required probability is just the product

$$\mathrm{P}\,[T_1 = 8] = q^7 p$$

What this means is that the geometric distribution, by virtue of the independence between trials, is *memoryless*. It doesn't matter when you start looking at a Bernoulli process, the number of trials to the next 'success' is given by the geometric distribution (and is independent of the trial number).

Properties:
The mean of T_1, which is also sometimes refered to as the *return period* or the *mean recurrence time*, is determined as follows;

$$\begin{aligned} \mathrm{E}[T_1] &= \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} kq^{k-1} = p\frac{d}{dq}\sum_{k=1}^{\infty} q^k = p\frac{d}{dq}\left(\frac{q}{1-q}\right) \\ &= p\left(\frac{1}{(1-q)^2}\right) \\ &= \frac{1}{p} \end{aligned} \tag{41}$$

where we used Eq. 40 to evaluate the final sum above. We will use the second to last sum in the following proof.

The variance of T_1 is obtained from $\mathrm{Var}[T_1] = \mathrm{E}\left[T_1^2\right] - \mathrm{E}^2[T_1]$ as follows;

$$\begin{aligned} \mathrm{E}\left[T_1^2\right] &= \sum_{k=1}^{\infty} k^2 pq^{k-1} = p\sum_{k=1}^{\infty} k^2 q^{k-1} = p\frac{d}{dq}\sum_{k=1}^{\infty} kq^k \\ &= p\frac{d}{dq}\left(\frac{q}{(1-q)^2}\right) \\ &= \frac{1}{p} + \frac{2q}{p^2} \end{aligned}$$

thus

$$\begin{aligned} \mathrm{Var}[T_1] &= \mathrm{E}\left[T_1^2\right] - \mathrm{E}^2[T_1] \\ &= \frac{1}{p} + \frac{2q}{p^2} - \frac{1}{p^2} \\ &= \frac{q}{p^2} \end{aligned} \tag{42}$$

Aside: In engineering problems, we often reverse the meaning of 'success' and 'failure', and use the geometric distribution to model 'time' to failure, where 'time' is measured in discrete steps (trials).

5.4 Negative Binomial Distribution

Suppose we wish to know the number of trials ('time') in a Bernoulli process until the m^{th} success. Letting T_m be the number of trials until the m^{th} success, then

$$\mathrm{P}[T_m = k] = \binom{k-1}{m-1} p^m q^{k-m} \qquad \text{for } k = m, m+1, \ldots \tag{43}$$

which is the *negative binomial distribution*. Whereas a binomial distributed random variable is the number of successes in a fixed number of trials, a negative binomial distributed random variable is the number of trials for a fixed number of successes.

We note that the negative binomial is also often used to model the number of failures before the m^{th} success, which results in a somewhat different distribution. We prefer the interpretation that the negative binomial distribution governs the number of trials until the m^{th} success because it is a natural generalization of the geometric distribution and because it is then a discrete analog of the Gamma distribution considered in Section 1.6.2.

The name of the negative binomial distribution arises from the *negative binomial series*

$$(1-q)^{-m} = \sum_{k=m}^{\infty} \binom{k-1}{m-1} q^{k-m} \tag{44}$$

which converges for $|q| < 1$. This series can be used to show that the negative binomial distribution is a valid distribution, since

$$\sum_{k=m}^{\infty} \mathrm{P}\left[T_m = k\right] = \sum_{k=m}^{\infty} \binom{k-1}{m-1} p^m q^{k-m} = p^m \sum_{k=m}^{\infty} \binom{k-1}{m-1} q^{k-m} = p^m (1-q)^{-m} = 1 \tag{45}$$

as expected.

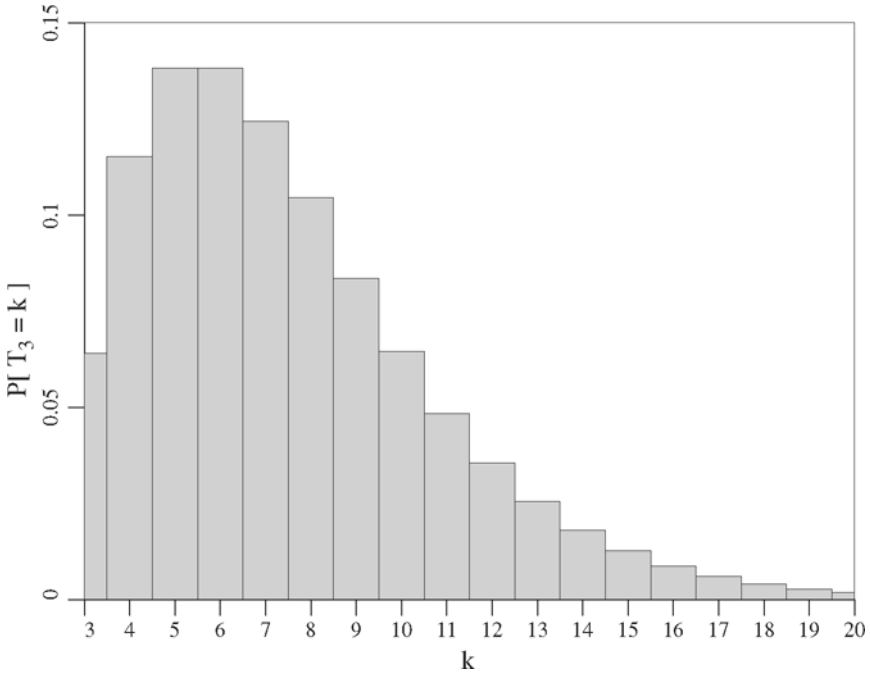

Figure 8. Negative binomial distribution for T_3 (ie. $m = 3$) and $p = 0.4$.

We see that the Geometric distribution is a special case of the Negative Binomial distribution with $m = 1$. The negative binomial distribution is often used to model 'time

to the m^{th} failure', where 'time' is measured in discrete steps, or trials. Consider one possible realization which has the 3rd success on the 8th trial;

F S S F F F F S

Another possible realization might be

F F F S F S F S

In both cases, the number of successes is 3, having probability p^3, and the number of failures is 5, having probability q^5. In terms of ordering, if $T_3 = 8$, then the 3rd success must occur on the 8'th trial (as shown above). Thus, the only other uncertainty is the ordering of the other two successes. This can occur in $\binom{7}{2}$ ways. The probability that the 3rd success occurs on the 8th trial is therefore given by

$$\mathrm{P}[T_3 = 8] = \binom{7}{2} p^3 q^5$$

Generalizing this for m successes and k trials leads to the negative binomial distribution shown above.

Properties:

Mean:

$$\begin{aligned}
\mathrm{E}[T_m] &= \sum_{j=m}^{\infty} j\mathrm{P}[T_m = j] = \sum_{j=m}^{\infty} j\binom{j-1}{m-1} p^m q^{j-m} = \sum_{j=m}^{\infty} j\left(\frac{(j-1)!}{(m-1)!(j-m)!}\right) p^m q^{j-m} \\
&= mp^m \sum_{j=m}^{\infty} \left(\frac{j!}{m!(j-m)!}\right) q^{j-m} \\
&= mp^m \left[1 + (m+1)q + \frac{(m+2)(m+1)}{2!}q^2 + \frac{(m+3)(m+2)(m+1)}{3!}q^3 + \cdots\right] \\
&= \frac{mp^m}{(1-q)^{m+1}} \\
&= \frac{m}{p}
\end{aligned} \tag{46}$$

which is just m times the mean of a single geometrically distributed random variable, T_1, as expected since the number of trials between successes follows a geometric distribution. In fact, this observation leads to the following alternative representation of T_m,

$$T_m = T_{1,1} + T_{1,2} + \cdots + T_{1,m} \tag{47}$$

where $T_{1,1}$ is the number of trials until the first success, $T_{1,2}$ is the number of trials after the first success until the second success, and so on. That is, the $T_{1,i}$ terms are just the 'times' between successes. Since all trials are independent, each of the $T_{1,i}$ terms will be independent geometrically distributed random variables, all having common probability of success, p. This leads to the following much simpler computation,

$$\mathrm{E}[T_m] = \mathrm{E}[T_{1,1}] + \mathrm{E}[T_{1,2}] + \cdots + \mathrm{E}[T_{1,m}] = \frac{m}{p} \tag{48}$$

since $\mathrm{E}[T_{1,i}] = 1/p$ for all $i = 1, 2, \ldots, m$. The mean in Figure 8 is $3/0.4 = 7.5$.
Variance:

To get the variance, $\mathrm{Var}[T_m]$, we'll again use Eq. 47. Due to independence of the $T_{1,i}$ terms, the variance of the sum is the sum of the variances,

$$\begin{aligned} \mathrm{Var}[T_m] &= \mathrm{Var}[T_{1,1}] + \mathrm{Var}[T_{1,2}] + \cdots + \mathrm{Var}[T_{1,m}] \\ &= m\mathrm{Var}[T_1] \\ &= \frac{mq}{p^2} \end{aligned} \tag{49}$$

which is just m times the variance of a single geometrically distributed random variable, T_1, as expected.

5.5 Poisson Distribution

The Poisson distribution governs many 'rate' dependent processes – for example, arrivals of vehicles at an intersection or number of points where a soil's cohesion exceeds some high threshold in a region. The Poisson is yet another distribution arising from the Bernoulli family and can be derived directly from the binomial distribution by letting each instant in time (or space) become an independent Bernoulli trial. For simplicity, we will talk about Poisson processes in time, but recognize that they can be equivalently applied over space, simply by replacing t with a distance (or area, volume, etc) measure.

For any non-zero time interval we have an infinite number of Bernoulli trials, since any time interval is made up of an infinite number of instants. Thus, the probability of success, p, in any one instant must go to zero – otherwise we would have an infinite number of successes in each time interval ($np \to \infty$ as $n \to \infty$). This means that we must abandon the probability of success, p, in favour of a *mean rate of success*, λ, which quantifies the mean number of successes per unit time.

The basic assumption on which the Poisson distribution rests is that each instant in time is a Bernoulli trial, with mean success (arrival) rate given by the parameter λ. This basic assumption leads to the following statements (which also define λ);

1) successes occur at random and at any point in time (or space),
2) the occurrence of a success in a given time (or space) interval is independent of successes occurring in all other disjoint intervals,
3) the probability of a success occurring in a small interval, Δt,is proportional to the size of Δt, ie., is $\lambda \Delta t$, where λ is the mean *rate* of occurrence.
4) for $\Delta t \to 0$, the probability of two or more successes in Δt is negligible (e.g. a Bernoulli trial can only have one success).

Now define N_t to be the number of successes ('arrivals' or 'occurrences') occurring in time t. If the above assumptions hold, then N_t is governed by the following distribution,

$$\mathrm{P}[N_t = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, 2, \ldots \tag{50}$$

where λ is the mean rate of occurrence (λ has units of 1/time).

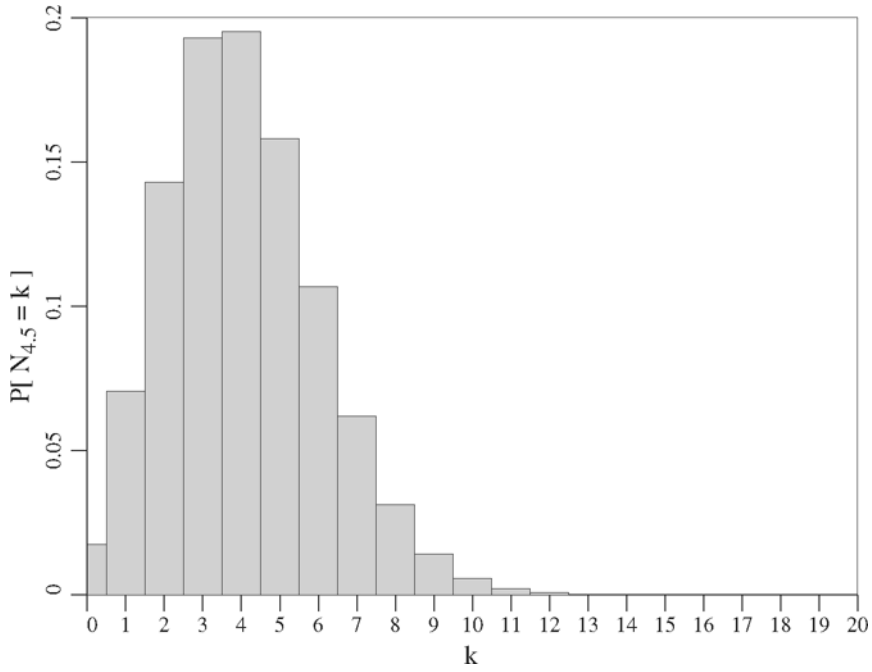

Figure 9. Poisson distribution for $t = 4.5$ and $\lambda = 0.9$.

Properties

Mean:

$$\begin{aligned} \mathrm{E}[N_t] &= \sum_{j=0}^{\infty} j \frac{(\lambda t)^j}{j!} e^{-\lambda t} = \lambda t e^{-\lambda t} \sum_{j=1}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} \\ &= \lambda t e^{-\lambda t} \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!} \\ &= \lambda t \end{aligned} \tag{51}$$

The mean of the distribution shown in Figure 9 is $\mathrm{E}[N_{4.5}] = 0.9(4.5) = 4.05$.

Variance:

$$\begin{aligned} \mathrm{E}\left[N_t^2\right] &= \sum_{j=0}^{\infty} j^2 \frac{(\lambda t)^j}{j!} e^{-\lambda t} = \lambda t e^{-\lambda t} \sum_{j=0}^{\infty} (j+1) \frac{(\lambda t)^j}{j!} \\ &= \lambda t e^{-\lambda t} \left[\sum_{j=0}^{\infty} j \frac{(\lambda t)^j}{j!} + \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!} \right] \\ &= (\lambda t)^2 + (\lambda t) \end{aligned}$$