3.11 Problems

TS

Problem 3.1 GRAIN Several physical processes can be modeled with a *random walk*: a process of interatively changing a quantity by some random amount. Infinitely many variations are possible, but common factors of variation include probability distribution, step size, dimensionality (e.g. one-dimensional, two-dimensional, etc.), and coordinate system. Graphical representations of these walks can be beautiful. Develop a computer program that generates random walks and corresponding graphics. Do it well and call it art because it is.

Problem 3.2 FREE Consider the defective spring problem from example 3.4. One way to improve the probability of a true positive test (i.e., the sensitivity) is to add a second test for which a positive event is called C. Again assuming that the sensitivity and specificity are equal for tests B and C, and that the sensitivity of test B is $P(B|A) = 0.995$ what is the required sensitivity for test C? Clearly state any assumptions.

4 Statistics

1M

Whereas probability theory is primarily focused on the relations among mathematical objects, statistics is concerned with making sense of the outcomes of observation (Skiena 2001). However, we frequently use statistical methods to **estimate** probabilistic models. For instance, we will learn how to estimate the standard deviation of a random process we have some reason to expect has a Gaussian probability distribution.

Statistics has applications in nearly every applied science and engineering discipline. Any time measurements are made, statistical analysis is how one makes sense of the results. For instance, determining a reasonable level of confidence in a measured parameter requires statistics.

A particularly hot topic nowadays is **machine learning**, which seems to be a field with applications that continue to expand. This field is fundamentally built on statistics.

A good introduction to statistics appears at the end of (Ash 2008). A more involved introduction is given by (Jaynes et al. 2003). The treatment by (Kreyszig 2011) is rather incomplete, as will be our own.

4.1 Populations, Samples, and Machine Learning

3E

An experiment's **population** is a complete collection of objects that we would like to study. These objects can be people, machines, processes, or anything else we would like to understand experimentally.

Of course, we typically can't measure *all* of the population. Instead, we take a subset of the population—called a **sample**—and infer the characteristics of the entire population from this sample.

However, this inference that the sample is somehow representative of the population assumes the sample size is sufficiently large and that the sampling is **random**. This means selection of the sample should be such that no one group within a population are systematically over- or under-represented in the sample.

Machine learning is a field that makes extensive use of measurements and statistical inference. In it, an algorithm is **trained** by exposure to sample data, which is called a **training set**. The variables measured are called **features**. Typically, a **predictive model** is developed that can be used to extrapolate from the data to a new situation. The methods of statistical analysis we introduce in this chapter are the foundation of most machine learning methods.

Example 4.1

Consider a robot, Pierre, with a particular gravitas and sense of style. He seeks the nicest pair of combat boots for wearing in the autumn rains. Pierre is to purchase the boots online via image recognition, and decides to gather data by visiting a hipster hangout one evening to train his style. For a negative contrast, he also watches footage of a white nationalist rally, focusing special attention on the boots of wearers of khakis and polos. Comment on Pierre's methods.

Pierre must identify *features* in the boots, such as color, heel-height, and stitching. Choosing two places to sample certainly enhances the *sample* or *training set*. Positive correlations can be sought with the first group in the sample and negative with the second. The choosing of "desirable" and "undesirable" sample groups is an example of *supervised learning*, which is to say the desirability of one group's boots and the undesirability of the other's is assumed to be known.