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.

4.2 Estimation of Sample Mean and Variance

4.2.1 Estimation and Sample Statistics

LU

The mean and variance definitions of section 3.7 and section 3.8 apply only to a random variable for which we have a theoretical probability distribution. Typically, it is not until after having performed many measurements of a random variable that we can assign a good distribution model. Until then, measurements can help us *estimate* aspects of the data. We usually start by estimating basic parameters such as *mean* and *variance* before estimating a probability distribution.

There are two key aspects to randomness in the measurement of a random variable. First, of course, there is the underlying randomness with its probability distribution, mean, standard deviation, etc., which we call the *population statistics*. Second, there is the *statistical variability* that is due to the fact that we are *estimating* the random variable's statistics—called its *sample statistics*—from some sample. Statistical variability is decreased with greater sample size and number of samples, whereas the underlying randomness of the random variable does not decrease. Instead, our estimates of its probability distribution and statistics improve.

4.2.2 Sample Mean, Variance, and Standard Deviation

The *arithmetic mean* or **sample mean** of a measurand with sample size N, represented by random variable X, is defined as

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

If the sample size is large, $\overline{x} \to m_X$ (the sample mean approaches the mean). The **population mean** is another name for the mean m_X, which is equal to

$$m_X = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Recall that the *definition* of the mean is $m_X = \mathrm{E}[x]$.

The **sample variance** of a measurand represented by random variable X is defined as

$$S_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

If the sample size is large, $S_X^2 \to \sigma_X^2$ (the sample variance approaches the variance). The **population variance** is another term for the variance σ_X^2, and can be expressed