

# 1. Introduction

---

## The Basic Statistical Model

As usual, our starting point is a **random experiment** with **probability measure**  $\mathbb{P}$  on a **sample space**. In the basic statistical model, we have an observable **random variable**  $\mathbf{X}$  (which we call the **data variable**) taking values in a set  $S$ . In general,  $\mathbf{X}$  can have quite a complicated structure. For example, if the experiment is to sample from a population and record various measurements of interest, then the outcome variable is

$$\mathbf{X} = (X_1, X_2, \dots, X_n)$$

where  $X_i$  is the vector of measurements for the  $i^{\text{th}}$  object. Here are some specific examples.

1. In the **M&M data set**, a sample of 30 bags of M&Ms were studied. For this study,  $X_i$  records the color counts for red, green, blue orange, yellow, and brown candies, and the net weight for the  $i^{\text{th}}$  bag.
2. In **Fisher's iris data set**, a sample of 150 irises were studied. For this study,  $X_i$  records the type, petal length, petal width, sepal length, and sepal width for the  $i^{\text{th}}$  iris.
3. In the **cicada data**, 104 cicadas were captured. For this study,  $X_i$  records the body weight, body length, wing width, wing length, gender, and species for the  $i^{\text{th}}$  cicada.

On the other hand, the hallmark of mathematical abstraction is the ability to gray out out the features that are not relevant at any particular time, to treat a complex structure as a single object. Thus, although  $\mathbf{X}$  may actually be a vector of vectors, the crucial fact at this point is that it is simply a random variable for an experiment.

There are two broad branches of statistics. The term **descriptive statistics** refers to methods for summarizing and displaying the observed data  $\mathbf{x}$ . The term **inferential statistics** refers to methods of drawing inferences about the distribution of  $\mathbf{X}$  from an observed value  $\mathbf{x}$ . Thus, in a sense, inferential statistics is the dual of probability. In probability, we try to *predict* the value of  $\mathbf{X}$  *assuming* knowledge of the distribution. In statistics, by contrast, we *observe* the value of  $\mathbf{X}$  and try to *infer* information about the underlying distribution..

The techniques of statistics have been enormously successful; these techniques are widely used in just about every subject that deals with quantification--the natural sciences, the social sciences, law, and medicine. On the other hand, statistics has a legalistic quality and a great deal of terminology that can make the subject a bit intimidating at first. In this section, we will discuss some of the basic definitions.

## Random Samples

The most common and important special case of this statistical model occurs when the observation variable

$$X = (X_1, X_2, \dots, X_n)$$

is a sequence of independent and identically distributed random variables. Again, in the standard sampling model,  $X_i$  is itself a vector of measurements for the  $i^{\text{th}}$  object in the sample, and thus, we think of  $(X_1, X_2, \dots, X_n)$  as independent copies of an underlying measurement vector  $X$ . In this case,  $(X_1, X_2, \dots, X_n)$  is said to be a **random sample** of size  $n$  from the distribution of  $X$ .

## Types of Variables

### Discrete and Continuous

Recall that a real variable is **continuous** if the possible values form an interval of real numbers. For example, the weight variable in the M&M data set, and the length and width variables in Fisher's iris data are continuous. In contrast, a **discrete variable** is one whose set of possible values forms a discrete set. For example, the counting variables in the M&M data set, the type variable in Fisher's iris data, and the denomination and suit variables in the card experiment are discrete. Continuous variables represent quantities that can, in theory, be measured to any degree of accuracy. In practice, of course, measuring devices have limited accuracy so data collected from a continuous variable is necessarily discrete. That is, there is only a finite (but perhaps very large) set of possible values that can actually be measured.

### Levels of Measurement

A real variable is also distinguished by its **level of measurement**, which determines the mathematical operations that make sense for the variable. **Qualitative variables** simply encode types, and thus no mathematical operations make sense, even if numbers are used for the encoding. Such variables have the **nominal level of measurement**. For example, the type variable in Fisher's iris data is qualitative. A variable for which only order is meaningful is said to have the **ordinal** level of measurement; differences are not meaningful even if numbers are used for the encoding. For example, in many card games, the suits are ranked, so the suit variable has the ordinal level of measurement. A quantitative variable for which differences, but not ratios are meaningful is said to have the **interval** level of measurement. Equivalently, a variable at this level has a relative zero value. Typical examples are temperature (in Fahrenheit or Centigrade) or time (clock or calendar). Finally, a quantitative variable for which ratios are meaningful is said to have the **ratio** level of measurement. A variable at this level has an absolute zero value. The count and weight variables in the M&M data set, and the length and width variables in Fisher's iris data are examples.

## Parameters and Statistics

### Parameters

The term **parameter** refers to a non-random variable in a model that, once chosen, remains constant. Almost all probability models are actually parametric families of models; that is, they are models governed by one or more parameters that can be adjusted to fit the random process being modeled. More technically, a parameter is a characteristic of the distribution of the observable variable  $X$ . As usual, we will take the general point of view and allow parameters to be vector valued.

1. Identify the parameters in each of the following:

- a. [Buffon's Coin Experiment](#)
- b. [Buffon's Needle Experiment](#)
- c. [The Binomial Experiment](#)

## Statistics

A **statistic**  $W = W(X)$  is a random variable that is an observable function of the outcome variable of the experiment. The term **observable** means that the function should not contain any unknown parameters. After all, we need to be able to compute the value of the statistic from the observed data. The crucial point is that a statistic is a random variable and hence, like all random variables, it has a probability distribution.. Ultimately, what we observe is a *value* of this random variable. As with the data  $X$ , a statistic  $W$  may have a complicated structure; typically,  $W$  is vector valued. Note that  $X$  itself is a statistic, the original observed data variable; all other statistics are derived from  $X$ .

Statistics  $U$  and  $V$  are **equivalent** if there exists a one-to-one function  $r$  from the range of  $U$  onto the range of  $V$  such that  $V = r(U)$ . Equivalent statistics give equivalent information, in terms of drawing inferences.

2. Show that statistics  $U$  and  $V$  are equivalent if and only if the following condition holds: for any  $x \in S$  and  $y \in S$ ,  $U(x) = U(y)$  if and only if  $V(x) = V(y)$ .

3. Show that equivalence really is an **equivalence relation** on the collection of statistics for a given random experiment. That is, if  $U$ ,  $V$ , and  $W$  are arbitrary statistics then

- a.  $U$  is equivalent to  $U$  (the **reflexive property**).
- b. If  $U$  is equivalent to  $V$  then  $V$  is equivalent to  $U$  (the **symmetric property**).
- c. If  $U$  is equivalent to  $V$  and  $V$  is equivalent to  $W$  then  $U$  is equivalent to  $W$  (the **transitive property**).

[Virtual Laboratories](#) > [6. Random Samples](#) > [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#)

[Contents](#) | [Applets](#) | [Data Sets](#) | [Biographies](#) | [External Resources](#) | [Key words](#) | [Feedback](#) | ©