Probability and Statistics

Arthur Ryman

November 20, 2016

Abstract

This article provides a precise formulation of the basic concepts of Probability and Statistics using the language of category theory and typed set theory. The main benefit of viewing the subject in terms of category theory is that it unifies and simplifies many concepts. For example, random variables and p-values are seen to be arrows in the category of probability spaces. The main benefit of using typed set theory is that it allows one to make explicit many aspects of the structure of probability spaces and thereby clarifies the subject. All formal definitions in this article are encoded using Z Notation and have been validated using the *f*UZZ type-checker.

1 Introduction

This goal of this article is to present a precise formulation of the the basic concepts of probability and statistics. The approach taken here differs in several ways from the approach found in many textbooks.

1.1 Embracing the Mathematical Essence of Probability and Statistics

The approach taken here is mathematical in the sense that all concepts are defined precisely using the language of mathematics. The reader is expected to be comfortable with the concept of a set, and know about the natural numbers, real numbers, and other important sets.

This use of precise mathematical language contrasts with the usual, more relaxed, approach taken in books aimed at practitioners, especially those in the life sciences, commerce, and other domains where the practitioners are not expected to have mathematical backgrounds.

For example, consider [4], a textbook used in courses on introductory data analysis. The textbook first acknowledges that statistics is a branch of mathematics and warns the reader that it can be overwhelming [4, p3]:

> Statistics is a branch of mathematics that has applications in almost every facet of our daily life. It is a new and unfamiliar language

> for most people, however, and, like any new language, statistics can seem overwhelming at first glance. But once the language of statistics is learned and understood, it provides a powerful tool for data analysis in many different fields of application.

The textbook authors then go on to say that the subject involves many new mathematical terms and concepts, but the student can safely learn how to apply them simply by following the numerical examples and common-sense arguments provided in the textbook [4, p5]:

> As you begin to study statistics, you will find that there are many new terms and concepts to be mastered. Since statistics is an applied branch of mathematics, many of these basic concepts are mathematical—developed and based on results from calculus or higher mathematics. However, you do not have to be able to derive results in order to apply them in a logical way. In this text, we use numerical examples and common-sense arguments to explain statistical concepts, rather than more complicated mathematical arguments.

Here the textbook authors seem to equate mathematical discourse with deriving results, aka proofs. However, there is much more to mathematics than that. Mathematics provides a language for identifying and precisely defining concepts. Mathematics allows one to build up a conceptual vocabulary in which new concepts are defined in terms of previously defined ones or, ultimately, in terms the foundational concepts of set theory. It is this conceptual vocabulary for Probability and Statistics that is somewhat lacking in most practitioner-oriented textbooks. In contrast, this article takes the position that precise definitions, rather than being pedantic, make the subject easier to understand.

1.2 The Category of Probability Spaces

One of the great organizing principles of mathematics is that rather than study mathematical objects in isolation it is more illuminating to study families of structurally similar objects and the mappings between them that preserve their structure. A structure-preserving mapping from one object to another is generically referred to as a *morphism* or, more simply an it arrow. Such a family of objects and arrows is called a *category*.

For example, linear algebra is the study of the category of vector spaces and linear transformations, group theory is the study of the category of groups and homomorphisms, and topology is the study of the category of topological spaces and continuous mappings.

Viewed through the lens of category theory, probability and statistics is the study of the category of *probability spaces* and *random variables*.

Many important concepts can be simply expressed in the language of category theory. As an illustration of this point in the category of probability spaces, consider the usual textbook definition of a *p-value*:

> A p-value is a function that assigns to a given measurement the probability of observing a measurement as or more extreme than the given value.

This definition is somewhat hard to grasp at first. In terms of category theory, a p-value is simply a morphism from a probability space to the uniform probability space on the unit interval. The uniform probability space on the unit interval is an important object in the category of probability spaces.

In general, morphisms to and from important objects are also important. One of the underlying assumptions of this article is that every important concept in probability and statistics has a natural and simple interpretation in terms of category theory.

Category theory also teaches us that mappings between categories are important objects. In fact, much of the motivation behind category theory came from the subject of *algebraic topology* which can be viewed as the study of mappings from the category of topological spaces to various algebraic categories. The recent development of *topological data analysis* should provide ample motivation for data scientists to understand the ideas behind topology and category theory.

1.3 Structures and Types

Many interesting mathematical objects have significant structure. For example, a topological space is a structure that consists of a set of points and a *topology* on that set of points. A topology is a collection of sets, referred to as *open sets*, that satisfy certain axioms. Normal mathematical discourse defines structure informally. In contrast, the approach taken here is to define structure using a formal language, namely *Z Notation* [2].

Using Z Notation, each structure is defined as a *type*. New types are constructed by combining previously constructed types using standard mathematical operations such as forming cartesian products and power sets. Z Notation is based on *typed set theory* which was introduced by Bertrand Russell as a way to avoid certain paradoxes in the foundations of set theory. Anyone with some computer programming experience should be very comfortable with types since they are closely related to the datatypes found in most programming languages.

The benefit of using a formal language to define types is that all definitions can be automatically type-checked. This ensures that all required concepts are explicitly defined and that all statements makes sense. Informally, a type-checker will make sure that you are not trying to add apples to oranges.

The LaTeX source for this article has been type-checked using the *f*UZZ type-checker [3] and is available in a GitHub repository [1].

2 Probability and Probability Spaces

2.1 Probability

Let $\mathbb{R}$ denote the set of all real numbers. A *probability* $p \in \mathbb{R}$ is a real number between 0 and 1 inclusive. The set of all probabilities is denoted by I, the unit interval on the real number line.

$$I == \{\, p : \mathbb{R} \mid 0 \leq p \leq 1 \,\}$$

There are two main interpretations of probability, namely as a *relative frequency* or as a *degree of belief*. The frequency interpretation applies to processes, such as flipping a coin, that can be repeated many times. In this case the probability of an outcome is the expected ratio of the number of times the outcome occurs to the total number of times the process is repeated as the number of repetitions becomes large. The degree of belief interpretation applies to situations, such as criminal trials, which cannot be repeated many times, and where a judgement has to be made based on evidence. Both interpretations are useful and we model both using the same type of object, namely as a *probability space*.

2.2 Probability Spaces

3 Measurable Spaces

A probability spaces is a type of measurable space. A measurable space $M = (X, A)$ is a set X together with a distinguished set A of subsets that are capable of being measured. The measurable sets are required to form a σ-algebra under the usual operators of set theory.

The axioms for a σ-algebra of subsets of X are as follows:

- The empty set is measurable: $\varnothing \in A$
- X is measurable: $X \in A$
- The complement of a measurable set is measurable: $\forall\, Y \in A, X \setminus Y \in A$
- The union of a countable number of measurable sets is measurable: $\forall f : N -> A, \cup_{i \in N} f(i) \in A$

The axiom about the union of a countable number of measurable sets being measurable is related to the definition of the sum of infinite series.

A consequence of these axioms is that the intersection of a countable number of measurable sets is also measurable.

Clearly, if X is any set then its power set 2^X is a σ-algebra.

If A and B are sigma algebras on X, then so their intersection $A \cap B$. Given any collection C of subsets of X that is contained in the sigma algebras A and B, then C is also contained in their intersection. Therefore, the intersection of all sigma algebras that contain C is also a sigma-algebra that contains C. It is the sigma-algebra generated by C.

When dealing with the real numbers, we are interested in the sigma algebra generated by the set of all closed, bounded intervals.

A measure space is a measurable space together with a measure. A measure if a function $m : A- > R^+$ that assigns to any measurable set, a non-negative real number, where we have extended the real numbers with positive infinity and defined addition in the obvious way.

A measure space satisfies the following axioms:

- The empty set has measure 0: $m(\varnothing) = 0$
- The measure of the union of a countable sequence of disjoint measurable sets is the sum of the measures of the sets in the sequence: $\forall f : N- > A \mid \forall i, j : N \mid i \neq j => A_i \cap A_j = \varnothing, m(\cup_i A_i) = \Sigma_i m(A_i)$

For example, if X is a countable set then m(Y) = #Y is a measure.

References

[1] Ryman, A. agryman/probability-and-statistics. source code repository, GitHub, 2016. `https://github.com/agryman/probability-and-statistics`.

[2] Spivey, M. *The Z Notation: a reference manual*. Prentice Hall, 2001. `https://spivey.oriel.ox.ac.uk/mike/zrm/index.html`.

[3] Spivey, M. The fuzz type-checker for Z. web page, Oxford University, 2008. `https://spivey.oriel.ox.ac.uk/mike/fuzz/`.

[4] William Mendenhall, I., Beaver, R. J., and Beaver, B. M. *Introduction to Probability and Statistics, 14th Edition*. Brooks/Cole CENGAGE Learning, 2013.