Probability and Statistics

Arthur Ryman

December 18, 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 whose 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, as an *arrow.* Such a family of objects and arrows is called a *category.*

For example, linear algebra is the study of vector spaces and their linear transformations, group theory is the study of groups and their homomorphisms, and topology is the study of topological spaces and their continuous mappings. Each of these collections of objects and arrows form a category.

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

Many important concepts can be expressed simply in the language of category theory. As an illustration of this point in the category of probability spaces, consider the following typical textbook definition of p-value [4, p351]:

> The *p-value* is the probability of observing a test statistic as extreme as or more extreme than the observed value, if in fact H_0 is true.

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 structure-preserving mappings between categories are important objects. A structure-preserving mapping from one category to another is called a *functor*. In fact, much of the motivation behind category theory came from the subject of algebraic topology which can be viewed as the study of functors from the category of topological spaces to various algebraic categories. The recent development of topological data analysis provides another 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. Stated informally, a type-checker makes 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

The concept of probability is useful for modelling situations in which we have uncertain knowledge. For example, when we flip a coin we are uncertain which side, heads or tails, will land face up. Imagine performing an experiment in which we flip the coin n times. Let n_h and n_t be the number of times that heads and tails occur. Clearly we have:

$$0 \leq n_h \leq n \tag{1}$$
$$0 \leq n_t \leq n \tag{2}$$
$$n_h + n_t = n \tag{3}$$

Let f_h and f_t be the relative frequencies of heads and tails:

$$f_h = n_h/n \tag{4}$$
$$f_t = n_t/n \tag{5}$$

The relative frequencies satisfy:

$$0 \leq f_h \leq 1 \tag{6}$$
$$0 \leq f_t \leq 1 \tag{7}$$
$$f_h + f_t = 1 \tag{8}$$

In the long run we expect the relative frequency of occurrence for each side to approach some limiting value, called the *probability* of occurrence of the side. This view of probability is called the *frequentist* interpretation. Let p_h and p_t denote the probability of occurrence of heads and tails:

$$\lim_{n\to\infty} f_h = p_h \tag{9}$$
$$\lim_{n\to\infty} f_t = p_t \tag{10}$$

These probabilities satisfy:

$$0 \leq p_h \leq 1 \tag{11}$$
$$0 \leq p_t \leq 1 \tag{12}$$
$$p_h + p_t = 1 \tag{13}$$

If the coin is fair then $p_h = p_t = 0.5$ since half the time the coin will land heads and half the time it will land tails.

In general, a probability is any real number between 0 and 1. Let $\mathbb{R}$ denote the set of all real numbers and let I denote the unit interval on the real number line.

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

I is therefore also the set of all probabilities.

Now let's formalize the coin flipping experiment. The outcome of a coin flip is either heads or tails. Let *Flip* denote the set of coin flip outcomes:

$$\mathit{Flip} ::= \mathit{Head} \mid \mathit{Tail}$$

We can model a coin by assigning a probability to each of the flip outcomes such that the total probability is 1. Such an assignment of probabilities to outcomes is called a *discrete probability distribution* or a *probability mass function*. Let *Coin* denote the set of all probability mass functions on the set of coin flips:

$$\mathit{Coin} == \{\, p : \mathit{Flip} \longrightarrow I \mid p(\mathit{Head}) + p(\mathit{Tail}) = 1 \,\}$$

For example, let *fairCoin* denote the probability assignments for a fair coin:

$$\begin{array}{|l} \mathit{fairCoin} : \mathit{Coin} \\ \hline \mathit{fairCoin}(\mathit{Head}) = \mathit{fairCoin}(\mathit{Tail}) = 0.5 \end{array}$$

A probability mass function is adequate to describe situations in which the set of outcomes is finite or countable since there is a well-defined mathematical meaning to adding up a countable set of numbers—that is the subject matter of the theory of convergent sequences and series. However, in many important situations the set of outcomes is uncountable, for example when an outcome is a measurement, such as weight or height, that can take on any real number in some interval. There is no well-defined mathematical meaning to adding up an uncountable set of non-zero numbers. Rather, the generalization of addition on uncountable sets is integration, which leads to the concept of a continuous probability distribution or, more generally, a probability measure.

2.2 Events

The first step towards a more general framework for probability is to focus on sets of outcomes, rather than individual outcomes. A set of outcomes is referred to as an *event*. Let X be an arbitrary, possibly uncountable, set of outcomes. The empty set is the *impossible event* since it contains no outcomes. Its probability is always 0. The set X is the *certain event* since it contains every outcome. Its probability is always 1. However, when it comes to uncountable sets of outcomes, some restrictions must be placed on what constitutes on event in order to be able to assign a probability to each event in a mathematically consistent way.

Let $\mathcal{E}$ be a collection of events on X.

$$\mathcal{E} \subseteq \mathbb{P}\, X \tag{14}$$

The following natural restrictions on $\mathcal{E}$ are sufficient to enable probabilities to be assigned in a mathematically consistent way.

A collection of events must always include the impossible event and the certain event since these always have a well-defined probability.

$$\varnothing \in \mathcal{E} \tag{15}$$

$$X \in \mathcal{E} \tag{16}$$

Suppose A is an event. Since A can be assigned a probability, so can its complement $X \setminus A$, namely 1 minus the probability of A.

$$A \in \mathcal{E} \Rightarrow X \setminus A \in \mathcal{E} \tag{17}$$

Finally, let $\mathcal{A} = \{A_1, A_2, \ldots\}$ be a countable set of events. The union and intersection of $\mathcal{A}$ must also be events. The restriction to countable sets of events is natural because the probability of the union of a collection of mutually disjoint events must be the sum of the probabilities of the individual events, but we can only assign a well-defined mathematical meaning to the sum of countable sequences of non-zero numbers.

$$\forall\, i \in \mathbb{N} \bullet A_i \in \mathcal{E} \Rightarrow \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{E} \tag{18}$$

$$\forall\, i \in \mathbb{N} \bullet A_i \in \mathcal{E} \Rightarrow \bigcap_{i \in \mathbb{N}} A_i \in \mathcal{E} \tag{19}$$

Note that there is no requirement that uncountable unions and intersections of events must also be events.

2.3 σ-Algebras

The above restrictions on events define a mathematical structure called a σ-algebra of sets on X. In general, there are many σ-algebras on a given set. Let *SigmaAlgebra*$[X]$ denote the set of all σ-algebras on the set X.

$$\begin{array}{|l}
[X] \\
\hline
\mathit{SigmaAlgebra} : \mathbb{P}(\mathbb{P}(\mathbb{P}\, X)) \\
\hline
\forall\, \mathcal{E} : \mathit{SigmaAlgebra} \bullet X \in \mathcal{E} \\
\forall\, \mathcal{E} : \mathit{SigmaAlgebra} \bullet \forall\, A : \mathcal{E} \bullet X \setminus A \in \mathcal{E} \\
\forall\, \mathcal{E} : \mathit{SigmaAlgebra} \bullet \forall\, A : \operatorname{seq} \mathcal{E} \bullet \bigcup(\operatorname{ran} A) \in \mathcal{E} \\
\hline
\end{array}$$

- X is in the algebra.
- If A is in the algebra then its complement is in the algebra.
- If $A_1, A_2, \ldots$ are in the algebra then their union is in the algebra.

The empty set is always in the algebra because it is the complement of X, which is always in the algebra.

$$\forall\, \mathcal{E} : \mathit{SigmaAlgebra}[X] \bullet \varnothing \in \mathcal{E} \tag{20}$$

The intersection of a countable collection of sets can be expressed as the complement of the union of a countable collection of complements of sets.

$$\begin{aligned} \bigcap(\operatorname{ran} A) &= X \setminus (X \setminus \bigcap(\operatorname{ran} A)) \\ &= X \setminus (X \setminus (A_1 \cap A_2 \cap \ldots)) \\ &= X \setminus ((X \setminus A_1) \cup (X \setminus A_2) \cup \ldots) \end{aligned} \tag{21}$$

Therefore, the intersection of a countable set of members of the algebra is also in the algebra.

$$\forall \mathcal{E} : \mathit{SigmaAlgebra}[X] \bullet \forall A : \operatorname{seq} \mathcal{E} \bullet \bigcap(\operatorname{ran} A) \in \mathcal{E} \tag{22}$$

It is easy to see that if X is any set then $\mathbb{P}\, X$, the set of all subsets of X, is a σ-algebra.

$$\mathbb{P}\, X \in \mathit{SigmaAlgebra}[X] \tag{23}$$

Similarly, the collection $\{\varnothing, X\}$ is a σ-algebra.

$$\{\varnothing, X\} \in \mathit{SigmaAlgebra}[X] \tag{24}$$

2.4 Measurable Spaces

A *measurable space* is a structure that consists of a set of points X and a σ-algebra $\mathcal{E}$ on X. The members of $\mathcal{E}$ are referred to as *measurable sets*. Let U be some universe from which the points of measurable spaces are drawn and let $\mathit{MeasurableSpace}[U]$ denote the set of all measurable spaces from that universe.

$$\begin{array}{|l} \hline \mathit{MeasurableSpace}[U] \\ \mathit{points} : \mathbb{P}\ U \\ \mathit{sets} : \mathbb{P}(\mathbb{P}\ U) \\ \hline \mathit{sets} \in \mathit{SigmaAlgebra}[\mathit{points}] \\ \hline \end{array}$$

- The measurable sets form a σ-algebra on the points of the space.

Category theory teaches us that important mathematical structures are connected to each other through structure-preserving maps. What are the structure-preserving maps between measurable spaces? To answer this question, note the similarity between measurable spaces and topological spaces. Both structures are defined by a set of points and a distinguished collection of subsets of points. In topological spaces, the distinguished collection of subsets are called the open sets and the structure-preserving maps are the continuous functions, namely those maps for which the inverse image of an open set is open. This definition applies directly to measurable spaces.

A *measurable map* from one measurable space to another is a function from the points of one to the other such that the inverse image of a measurable set is measurable. Let U and V be universes of points and let R and S be measurable spaces on U and V. Let *MeasurableMap*$[U, V]$ denote the set of all measurable maps from R to S. S is called the domain of the map and T is called the codomain.

$$\begin{array}{|l}
\textit{MeasurableMap}[U, V] \\
\hline
\textit{domain} : \textit{MeasurableSpace}[U] \\
\textit{codomain} : \textit{MeasurableSpace}[V] \\
\textit{map} : U \pfun V \\
\hline
\textit{map} \in \textit{domain.points} \longrightarrow \textit{codomain.points} \\
\forall A : \textit{codomain.sets} \bullet \textit{map}^{\sim} (\!| A |\!) \in \textit{domain.sets} \\
\hline
\end{array}$$

- The map sends the points of the domain to the point of the codomain.
- The inverse images of each measurable set in the codomain is a measurable set in the domain.

Category theory further teaches us that certain pairs of so-called *composable* maps can be composed to form other maps. Let R, S, and T be measurable spaces and let f and g be measurable maps from R to S and from S to T. Then g and f are composable. Let *ComposableMM* denote the set of all pairs (g, f) of measurable maps that are composable.

$$\begin{array}{|l}
[U, V, W] \\
\hline
\textit{ComposableMM} : \\
\quad \textit{MeasurableMap}[V, W] \leftrightarrow \textit{MeasurableMap}[U, V] \\
\hline
\textit{ComposableMM} = \\
\quad \{\, g : \textit{MeasurableMap}[V, W]; f : \textit{MeasurableMap}[U, V] \mid \\
\qquad g.\textit{domain} = f.\textit{codomain} \,\} \\
\hline
\end{array}$$

- The domain of g must be the codomain of f.

The composition h of g and f is the measurable map from R to T obtained by first applying f and then applying g. The composition is written $h = g \circ f$. Let *CompositionMM* denote the relationship between f, g, and h.

$$\begin{array}{|l}
\textit{CompositionMM}[U, V, W] \\
\hline
f : \textit{MeasurableMap}[U, V] \\
g : \textit{MeasurableMap}[V, W] \\
h : \textit{MeasurableMap}[U, W] \\
\hline
g.\textit{domain} = f.\textit{codomain} \\
h.\textit{domain} = f.\textit{domain} \\
h.\textit{codomain} = g.\textit{codomain} \\
h.\textit{map} = g.\textit{map} \circ f.\textit{map} \\
\hline
\end{array}$$

- (g, f) must be composable.
- h is uniquely determined by f and g.

Note that the composition $g \circ f$ is indeed a measurable map. Let A be a measurable set in T. Then $g^{\sim}(\!| A |\!)$ is a measurable set in S because g is measurable and $f^{\sim}(\!| (g^{\sim}(\!| A |\!)) |\!)$ is a measurable set in R because f is measurable. But $(g \circ f)^{\sim}(\!| A |\!) = f^{\sim}(\!| (g^{\sim}(\!| A |\!)) |\!)$ so $g \circ f$ is measurable.

Let *ComposeMM* denote the operation that maps each composable pair (g, f) to $h = g \circ f$.

$$
\begin{array}{|l}
[U, V, W] \\
\mathit{ComposeMM} : \\
\qquad \mathit{ComposableMM}[U, V, W] \longrightarrow \mathit{MeasurableMap}[U, W] \\
\hline
\mathit{ComposeMM} = \\
\qquad \{\, \mathit{CompositionMM}[U, V, W] \bullet (g, f) \mapsto h \,\} \\
\hline
\end{array}
$$

It is often useful to illustrate the relationships between objects and maps in a category using *commutative diagrams* like the one below. The nodes of the diagram are objects, e.g. measurable spaces, and the arrows between the nodes are maps, e.g. measurable maps. The diagram is commutative when the map that results from composing the maps along a path is independent of the path. The following diagram is commutative by definition of composition.

$$\begin{tikzcd}
R \arrow[r, "f"] \arrow[dr, "h=g\circ f"'] & S \arrow[d, "g"] \\
& T
\end{tikzcd} \tag{25}$$

There are two other properties that are required for a set of objects and maps to form a category, namely that the composition operation is associative and that each object has an associated identity map. It is easy to verify that composition of measurable maps is associative because it is based on composition of the underlying maps of the set of points of the measurable spaces, and that is associative Identity maps also exist for measurable spaces because the identity map of the underlying set of points of a measurable space is always itself a measurable map. The set of measurable spaces and maps therefore forms a category.

2.5 Probability Spaces

Let S be a measurable space with points X and measurable sets $\mathcal{E}$. A *probability measure* on S is a function $P : \mathcal{E} \longrightarrow I$ that maps each measurable set to a probability, and that satisfies the following conditions. The empty set has probability 0 and the whole set has probability 1.

$$P(\varnothing) = 0 \tag{26}$$

$$P(X) = 1 \tag{27}$$

Let $A_1, A_2, \ldots$ be a countable sequence of mutally disjoint measurable sets. The probability of their union is the sum of the probabilities of each set in the sequence.

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots \tag{28}$$

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.