CS1810 Math / Statistics Primer

CS181, Fall 2020

Last updated: September 12, 2020

1 Intro

As you will learn in this class, analyzing biological data relies heavily on statistics and probability. The goal of this guide is to provide a broad overview of the essential math topics that are required to understand the algorithms in CS181.

2 Common Statistical Terms

There are many ways to define and describe trends in a dataset $x = x_1, x_2, ..., x_n$. The mean, median, and mode are three important properties of datasets.

Definition: *Mean* - The mean represents the average of the data. Formally, the arithmetic mean of a sample $x_1, x_2, \ldots, x_n$ usually denoted as $\bar{x}$, is defined as the sum of the sampled values divided by the number of items in the sample:

$$\bar{x} = \frac{1}{n}(\sum_{i=1}^{n} x_i) = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Definition: *Median* - The median value of a dataset is thought of as the "middle value". The median value of a distribution is a datapoint that is separated from the minimum and maximum data point by the same number of entries.

Definition: *Mode* - The mode represents the data point that occurs most frequently.

A more general form of *mean* that is used in distributions of data is the **expected value**.

Definition: *Expected Value* - The expected value of a distribution is the sum (or integral for continuous data) of all possible outcomes, each multiplied by the probability of that outcome occurring. Denoted by angled brackets $\langle\rangle$, the expected value can be defined for both discrete and continuous distributions.

- *Discrete Distribution*: $\langle x \rangle = \sum xp(x)$
- *Continuous Distribution*: $\langle x \rangle = \int xp(x)$

Notice that if all outcomes are equally probable, then the expected value is simply the arithmetic mean as defined earlier. On the other hand, if the outcomes in x are not equally probable, then the expected value is essentially a *weighted* average, since it takes into account that some outcomes are more likely than others.

3 Probability

Probability theory is a mathematical framework that is used to describe the likelihood of a particular outcome of a given an experiment. There are a number of important applications of probability theory to biology. For example, in a simple model of a DNA, each position in the DNA sequence has four possible outcomes: A, T, C, and G. Experimental data could indicate, for example, that the probability of finding a C at particular position in a DNA sequence is p(C) = 0.22 (22%).

In general, probabilities are constrained by the following axioms:

- The probability of an event is greater or equal to zero: $p_i \geq 0$.
- The summed probability of all of N possible outcomes is equal to 1. $\sum_{i=1}^{N} p_i = 1$

The probability of two or more events taking place is called the **joint probability**.

- Two events are **independent** if the outcome of one event does not influence the outcome of the other event. For instance, a model of DNA as a random sequence of A, T, C, and G might assume that the bases at different positions of the DNA sequence are independent of each other. To calculate the joint probability of N independent events, take the product of their individual probabilities:
  $P(e_1, e_2, \ldots, e_N) = \prod_{i=1}^{N} P(e_i)$
- Two events are **dependent** if the outcome of one event gives information about the outcome of the other event. For instance, the expression of two genes might be tightly coupled if both are governed by the same transcription factor.

Conditional probability describes the likelihood of a particular event given what we know about the outcome of another event. Denoted as $P(A|B)$, this conditional probability expression is read as the probability of A *given* B. Note that if A and B are independent, then $P(A|B) = P(A)$ because the outcome of A does not depend on B.

We can use conditional probability to calculate the joint probability that two events take place.

$$P(A, B) = P(B|A) * P(A)$$

We would expect the joint probability of two dependent events to be the same regardless of the order of the events: $P(A, B) = P(B, A)$. Using the equation of conditional probability, we can expand the terms such that:

$$P(B|A)P(A) = P(A|B)P(B)$$

This equation can be re-arranged to one form of **Bayes' Theorem**, which is widely used in probability theory.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

4 Set Notation

In math, a **set** is a unordered collection of distinct elements. Below are common notations used to describe sets and set operations.

- Sets are defined within curly brackets containing all the elements of the set. Sets are usually defined with capital letters (A,B,C,...), whereas elements in the set are often written in lowercase (a_1, a_2, ...). For example, $C = \{c_1, c_2, c_3\} = \{1, 2, 3\}$
- a $\in A$ means a is a member of A.
- $|A|$, called **cardinality** of A, denotes the number of elements of A
- $A \subseteq B$ means A is a subset of B. This occurs if and only if every element of A is also an element of B.
- $A \cap B$ is the **intersection** of A and B, denoting the set containing elements that are in *both* A and B.
- $A \cup B$ is the **union** of A and B, denoting the set containing elements that are either in A *or* B.
- $A = B$ if and only if they have precisely the same elements.

Special sets:

- $\emptyset$ denotes the empty set (the set with no elements).
- $\mathbb{N}$ denotes set of natural numbers; i.e. {1,2,3...}.
- $\mathbb{Z}$ denotes set of integers; i.e. {...,-2,-1,0,1,2,...}.
- $\mathbb{Q}$ denotes set of rational numbers
- $\mathbb{R}$ denotes set of real numbers

5 Questions?

If you would like any additional help with any of the math in this primer or taught in the class, please feel free to come to office hours or ask a question on Piazza!

References / Additional Resources:

- Expected Value
- Bayes' Theorem
- Statistics in Computational Biology