# Implementations of Probability Theory

Independent Study Report

Andrew Simonson

Compiled on: October 16, 2024

# Contents

# 1 Objective

The educational focus of Implementations of Probability Theory surrounds the application of data models that produce non-deterministic insights through probabilistic methodology. By pursuing this study I hope to gain a deeper understanding of how to apply data in risk calculation for mitigation scenarios as they appear in real life, rather than the experimental lab conditions that enable algorithmic certainty.

In contrast to the path of black-box artificial intelligence and algorithms taught in **CSCI 335: Machine Learning**, this study is tailored to methods designed to produce confidence levels for uncertain events using certain terms, leveraging logical, traceable, and definite, calculations. Current course offerings in the realm of data science focus largely on the storing and management of data, and it is noted that the cluster of data science was until very recently under the branding of data management. Implementations of Probability Theory is intended to extend learnings in previous courses, notably **CSCI 420: Principles of Data Mining**, for more advanced algorithms used at the intersection of data and computing after the preprocessing stage.

After beginning this study the intended deliverable outline was determined to be technically implausible and has been replaced with demonstrations of applied algorithms. Taking inspiration from the retinal mosaic as displayed in **CSCI 431: Intro to Computer Vision** and discussion in **IGME 589: Computational Creativity and Algorithmic Art** on the appearance and nature of randomness in graphics, I hope to create a program that can determine the liklihood that randomly distributed colors on a hexagonal grid appear as they do in an image.

# 2 Units

---

## 2.1 Unit 1: Statistics Review

To ensure a strong statistical foundation for the future learnings in probabilistic models, the first objective was to create a document outlining and defining key topics that are prerequisites for probabilities in statistics or for understanding generic analytical models.

### 2.1.1 Random Variables

1. **Discrete Random Variables -** values are selected by chance from a countable (including countably infinite) list of distinct values
2. **Continuous Random Variables -** values are selected by chance with an uncountable number of values within its range

### 2.1.2 Sample Space

A sample space is the set of all possible outcomes of an instance. For a six-sided dice roll event, the die may land with 1 through 6 dots facing upwards, hence:

$$S = [1, 2, 3, 4, 5, 6] \quad \text{where } S \text{ is the sample space}$$

### 2.1.3 Probability Axioms

There are three probability axioms:

1. **Non-negativity**:

   $$P(A) \geq 0 \quad \text{for any event } A,\ P(A) \in \mathbb{R}$$

   No event can be less likely to occur than an impossible event ( $P(A) = 0$ ). P(A) is a real number. Paired with axiom 2 we can also conclude that $P(A) \leq 1$.

2. **Normalization**:

   $$P(S) = 1 \quad \text{where } S \text{ is the sample space}$$

   **Unit Measure -** All event probabilities in a sample space add up to 1. In essence, there is a 100% chance that one of the events in the sample space will occur.

3. **Additivity**:

$$P(A \cup B) = P(A) + P(B) \quad \text{if } A \cap B = \emptyset$$

A union between events that are mutually exclusive (events that cannot both happen for an instance) has a probability that is the sum of the associated event probabilities.

### 2.1.4 Expectations and Deviation

1. **Expectation -** The weighted average of the probabilities in the sample space

$$\sum^{S} P(A) * A = E \quad \text{where } E \text{ is the expected value}$$

2. **Variance -** The spread of possible values for a random variable, calculated as:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Where $N$ is the population size, $\mu$ is the population average, and $X$ is each value in the population.
For samples, variance is calculated with **Bessel's Correction**, which increases the variance to avoid overfitting the sample:

$$s^2 = \frac{\sum (X - \bar{x})^2}{n - 1}$$

3. **Standard Deviation -** The square root of the variance, giving a measure of the average distance of each data point from the mean in the same units as the data.

$$\sigma = \sqrt{V} \quad \text{where variance is } V$$

### 2.1.5 Probability Functions

Probability Functions map the likelihood of random variables to be a specific value.

### Probability Mass Functions

Probability Mass Functions (PMFs) map discrete random variables. For example, a six-sided die roll creates a uniform random PMF:

$$P(A) = \begin{cases} 1/6 & \text{if} \quad X = 1 \\ 1/6 & \quad\quad X = 2 \\ 1/6 & \quad\quad X = 3 \\ 1/6 & \quad\quad X = 4 \\ 1/6 & \quad\quad X = 5 \\ 1/6 & \quad\quad X = 6 \end{cases}$$

### Probability Density Functions

Probability Density Functions (PDFs) map continuous random variables. For example, this is a PDF where things happen.

$$P(A) = \begin{cases} X & \text{if} \quad 0 \leq X \leq .5 \\ -X + 1 & \quad\quad .5 < X \leq 1 \\ 0 & \quad\quad \textit{otherwise} \end{cases}$$

## 2.1.6 Limit Theorems

### Law of Large Numbers

The Law of Large Numbers states that as the number of independent random samples increases, the average of the samples' means will approach the true mean of the population.

$$\text{true average} \approx \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{as } n \to \infty$$

### Central Limit Theorem

The Central Limit Theorem states that the sampling distribution of a sample mean is a normal distribution even when the population distribution is not normal.

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow{d} N(0, 1)$$

Where $X_i$ is the sample mean, $N(0, 1)$ is a standard normal distribution, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

This is a challenging to understand solely as an equation. As an example, take a sample of two six-sided dice rolls and average their numbers. The more sample

averages taken, the more they will resemble a normal distribution where the majority of samples average around 3.

### 2.1.7 Confidence

Confidence is described using a confidence interval, which is a range of values that the true value is expected to be in, and its associated confidence level, which is a probability (expressed as a percentage) that the true value is in the confidence interval.

It is important to note that confidence levels, such as 95%, do not indicate that the real value is within 5% of the point estimate. The confidence level expresses the probability that the real value is in the range provided by the confidence interval.

At the highest level, calculating confidence intervals is simply the observed statistic (generally the mean) plus or minus the standard error.

To calculate standard error, kys.

### 2.1.8 Statistical Inference

Statistical Inference is any data analysis to draw conclusions from a sample to make assertions about the population. Methods include estimation via averages and confidence intervals, and hypothesis testing, which attempts to invalidate (never *validate*) a hypothesis.

## 2.2 Unit 2: Probabilistic Theories and Epistemology

When developing probabilistic models it is vital to use domain expertise to expose the product to the full range of external variables that would be expected of a model applied to the real world. Without an appropriate understanding of both the limitations in research procedures and the true value of the data collected, the integrity of the model becomes inherently compromised.

As data scientists, we are uniquely at risk of falling for this trap because it is hard to fully grasp domain expertise when the nature of data science in a business setting frequently means consulting for many separate projects with a collectively massive scope. Of equal consideration, it is also easy to assume that the sophistication of our tools overrides imperfections in the data, in spite of mantras like 'Garbage In, Garbage Out'.

In this unit I explored some common fallacies and assumptions held by analysts who may not fully grasp the content that they work with, nor the problems they intend to solve. This required extensive research that I found was best digested in the form of books whose chapters chronicle multiple examples of a given principle. As such, the reading was not confined to just the timeslot designated for this unit. Research started during the months leading up to the start of the semester[1] and have continued through the independent study. This structure was particularly helpful to pull me back and gain perspective of what my goal was when I was knee-deep in feature construction and model formulation.

### 2.2.1 Moral Hazards and The Bob Rubin Trade

Picking pennies in front of a steamroller. When studying the effectiveness of a model the scope of review must capture the entire range of the sample space. Discarding black swans that don't impact the client does not mean the results will not reflect on the client for an oversight. There is therefore a question of obligation for data scientists to include flags for significant events in reality that do not effect the proposed course of action to the client.

The 2009 recession, attributed to the collapse of the housing market bubble, is the most common example of a moral hazard because the displacement of risk from banks who were federally required to give subprime loans to the taxpayer meant that banks could profit from subprime loans but would not be harmed when the inevitable occurred. In popular media, the housing bubble bursting is attributed to the banks where those in the industry passed off the event as something that nobody could have forseen.[2] In reality, banks only ignored a probablistic eventuality because their

[1] Only research during the semester was logged in the timesheet

[2] For instance, in the 2015 movie *The Big Short*, only a few savvy traders who bothered to look into the details find that banks had, in their ignorance, built the bundled mortgages on an unstable foundation.

models did not need to account for such an event.

Most emphasize the problems with risk transferrence when creating models. For this study's purposes, the important learning is that probablistic models should not drop evaluations as soon as an event leaves the scope of the immediate client.

### 2.2.2 Ignoring Improbable Outliers with Outsized Impact

In machine learning it is common for algorithms to drop the most extreme (or a random selection of) datapoints to avoid overfitting and errors in data collection. One issue with the current implementation of this procedure is that it is often done blindly, ignorant of information that these outliers may relay. For instance, in a selection of 300 water samples from a stream, all but a few show a normal amount of oxygen in the stream. A citizen scientist may discount the remaining pockets as a statistical implausibility that is most likely indicative of a failure in sample testing and drop the most extreme 5% of datapoints. However, if these few pockets show a complete disruption of the dissolution process, the vast majority of aquatic life in the stream will eventually pass through these pockets without oxygen and die, resulting in an outsized impact from just a few sources.

Nassim Taleb in *Fooled By Randomness* describes this event with an analogy to Russian Roulette: If there was a 5/6 chance of winning a million dollars and a 1/6 chance of killing yourself, many people would at least hesitate before pulling the trigger. But what if the barrel is 10,000 rounds and it was only a 1/10,000 chance of harm? In this case, many less-than-rational actors use the game repeatedly to acquire wealth indefinitely, forgetting or even outright ignorant that eventually the unlikely, or, as the actor would see it, the unthinkable, happens and all of the gains are completely negated.

### 2.2.3 Fooled By Randomness

May justify its own subsection since the others acknowledge small probabilities whereas this is outright randomness.

### 2.2.4 Lindy Effect

"For the perishable, every additional day in its life translates into a shorter additional life expectancy. For the nonperishable, every additional day may imply a longer life expectancy." A tool that is proven is more likely to stand the test of time than a new tool replacing it since it is unproven. "The robustness of an item is proportional to its life!"

"Inaccurate science... is constantly being published. The Lindy-conscious consumer of scientific data will take seriously only information that has held up over a

period of time."[3]

### 2.2.5 Decision Theory

Decision theory is the study of how people make decisions with uncertain information. There are two main branches of decision theory:

#### Normative/Rational Decision Theory

This branch studies how people *should* make decisions. In problems with other actors, as in game theory, it is assumed that all other actors will also act with perfect rationality, allowing for precise calculation of the actions of all of the others and their expected utility to the agent.

#### Descriptive Decision Theory

This branch studies how people actually make decisions which includes factors such as psychological and emotional biases.

### 2.2.6 Info Gap Decisions

In info gap decision theory there is not enough information to assign probabilities to events and the goal is to select a course of action that is robust in the face of uncertainty. Where decision theory can predict expectations in irrationality to determine expected values, info gap decisions approximate the range of probabilities and weight them to estimate expected value. In essence, it applies probabilities to probabilities, adding an additional layer to insulate calculations from a lack of data or lack of understanding of a topic.

### 2.2.7 Methodology Considerations

Given I have taken 10134023 instances of the last 40 years, all of which Obama has been alive, I can say with a high degree of certainty that Obama is immortal.

An event never occurring in history does not discount its possiblity of occurring in the future. Similarly, events that may have been impossible in the past are not necessarily impossible in the future. Also, psychology. Someone who knows they are being studied will act differently than someone who isn't being studied so models will be inaccurate.

[3] `https://www.nytimes.com/2021/06/17/style/lindy.html`

## 2.3 Unit 3: Bayesian Statistics

This unit was deliberately separated from statistical review due to the percieved complexity of the topic and the magnitude of usage in recent data science breakthroughs. Bayes Theorem is a part of the cirriculum for both **MATH 351 - Probability and Statistics** and **CSCI 420 - Principles of Data Mining**. However, as both approached the topic from different perspectives and while neither solidified my personal confidence in its use, I chose to take extra time to learn this important topic in my own way.

It has been said that statistics does not come naturally to the human brain, hence statistics is, by mathematical standards, a young discipline. Resulting research on Bayesian statistics has led me to the conclusion that the opposite may be true - Bayes Theorem is quite intuitive, but its discipline has not had the time to crystallize best practices for instructing it. For instance, updating one's beliefs to compare probabilities with the number of documented occurrences is frequently used in philosophical discussion in the form of explanations that subsets with high liklihood of fufilling terms are valid classifications even when the subset size results in overall fufilled terms to be infrequently categorized as the proposed subset. Most people understand these expressions but, when shown a table and how to calculate those ratios, the content enters the realm of collegiate instruction.

### 2.3.1 Bayes Theorem

The equation for Bayes Theorem is as follows:

$$P(A|E) = \frac{P(A) * P(E|A)}{P(A) * P(E|A) + (1 - P(A)) * P(E|\neg A)}$$

This formula appears more complex as it is. The denominator, while directly translating to "The probability of A times the probability of event E occuring in A divided by the probability of A times the probability of event E occuring in A plus the probability of not A times the probability of E occuring in not A" can be more easily expressed simply as $P(E)$ or the probability of event E occuring.

By utilizing venacular more familiar to everyday life, Bayes Theorem can be translated into:

$$\text{P(occurence came from category)} = \frac{\text{\# of occurences from category}}{\text{total \# of occurences}}$$

Finally, this equation is updated to replace descriptions with technical terms:

$$\text{Posterior Probability} = \frac{\text{prior} * \text{likelihood}}{\text{Evidence}}$$

Even this equation can be misconstrued as a number of arrangements of ratios

involving total occurrences from a category or non-occurrences from outside of the category so as a final demonstration, the sample space will be visualized geometrically [4] as a 1 unit by 1 unit square.

### 2.3.2 Bayesian Updating

Bayesian Updating is another term that has been added to buzzword vocabulary to describe a process that isn't directly related to Bayesian Statistics but appears to have been rediscovered by academia through study of applied Bayes Theorem. In essence, Bayesian Updating simply states that observed occurrences should not override previous evidence and that it should instead be added to it in equal weight (equal value being a naive assumption). This evidence updating makes applications of Bayes Theory calculate posterior probabilities continuously as new information enters the system rather than a calculation that is only done once.

### 2.3.3 Bayesian Belief Networks

Bayesian Belief Networks are probablistic graphical models that preserve conditional dependence between random variables. In spite of its name, Bayesian Belief Networks do not necessarily apply Bayesian models, though they are a way to utilize Bayes Theorem for domains with greater complexity beyond a single posterior probability. In this type of network, edges are directed and the structure is utilized in a single direction. This is in contrast to undirected Hidden Markov Models that do not assume the order of aquisition of random variables.

[4] Concept credit to 3Blue1Brown on Youtube, this video is what finally clarified in my mind what the equation behind Bayes Theorem meant.
`https://www.youtube.com/watch?v=HZGCoVF3YvM`