Introduction

In our increasingly data-driven world in the 21^{st} century, statistics has moved from dusty textbooks to the forefront of scientific discovery, business strategy, political decision-making, and in many such areas. It is the language we use to understand uncertainty, make informed choices, and uncover hidden patterns in the vast amounts of information surrounding us. From the choices made by gamblers in the mid 17^{th} century to the sophisticated algorithms powering artificial intelligence today, the principles of statistics have been instrumental in human progress. It is a field that has continuously adapted and has evolved to meet the challenges of an ever-more complex world nowadays. Statistics fundamentally revolves around the concepts of population and sample. The total collection of all the elements that we are interested in is called a *population*. A subset of the population that will be studied in detail is called *sample*. To visualize this, I can make you understand this difference by the following Venn-diagram (See Fig. 1).

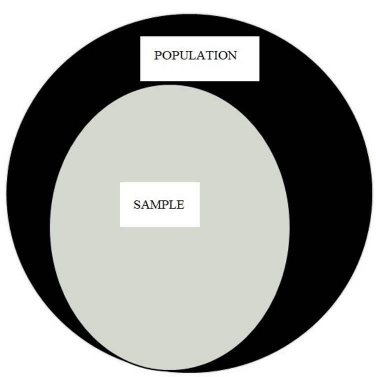

Figure 1: The outer circle represents population, and the inner section is the sample.

The black circle representing the population, and the gray area denoting the sample in Fig. 1 is my deliberate choice. Complete information about the population is unattainable due to time and cost constraints, size and diversity, accessibility and privacy etc., and the representativeness of a sample is subject to multiple factors like sample size, sampling method, selection bias, timing of data collection, measurement errors etc. (hence, the "gray zone" metaphor for uncertainty in Fig. 1). To illustrate these concepts, let me put forward my personal observations of female smoking habits in Sector - V, Kolkata (India). Pre-COVID (before 2020), I observed very few female corporate workers smoking, primarily in the evenings. However, post-COVID, there has been a significant increase in female smokers across various demographics, from students to working women and home-makers. This raises questions about the underlying causes:

> Is it a consequence of pandemic-related confinement and potential depression, the influence of over-the-top (OTT) media platforms like *Netflix*, *Prime Video* et al., or a perceived means of maintaining social status?

If I were to hypothesize that over 75% of females in Sector - V, are now prone to smoking post-COVID, how could we test this claim? Validating this requires rigorous data collection, a core task of social statistics that must be conducted ethically. The fundamental question that one can ask is that what empirical basis supports this specific figure of 75%? This 75% figure regarding female smoking habits is just an intuition.

In statistics, such a preliminary assumption is termed a *hypothesis*. To determine the validity of this hypothesis, we require sample data that is supposed to claim my intuitive hypothesis. This is why I intentionally marked the Population circle black in the Fig. 1, signifying the entire group we aim to understand. Carefully and ethically gathering data from a well-defined group is essential to ensure that the findings truly reflect the larger population. However, when investigating sensitive behaviors such as smoking habits

Figure 2: Smoking habits among females are rising in urban India (Photo Courtesy: Google).

in women, researchers often face challenges. Respondents may be reluctant to cooperate fully or may provide inaccurate information. This potential for unreliable data is precisely why I have depicted the inner section of Fig. 1 as a gray zone, which symbolizes potential gaps in data accuracy, not the sample itself. Consequently, understanding both the broader population and the limitations of the collected sample is of paramount importance in applied statistics.

A Brief History of Statistics: From Counting to the Age of Big Data:

The story of statistics is a long and fascinating journey, evolving from simple record-keeping to a sophisticated science that underpins decision-making in nearly every field imaginable. Statistics evolved from basic administrative record-keeping. Ancient civilizations like Babylonia, Egypt, and China conducted censuses as early as 3000 BCE for taxation and resource management. The Romans also extensively gathered data on population, land, and wealth for control. Medieval rulers, including Charlemagne, ordered land surveys [1]. The Domesday Book in England (1086) was a comprehensive survey of land ownership and resources. In 17^{th}-century England, *political arithmetic* (John Graunt's analysis of mortality records in Natural and Political Observations upon the Bills of Mortality, 1662) emerged as a foundational work for demography and early statistics. The term *statistics*, linked to *state*, began to develop, with Gottfried Achenwall credited with its formal introduction in the mid-18th century [2]. The 18^{th} century saw the formal development of probability theory, driven by games of chance due to the works of Mathematicians like Blaise Pascal, Pierre de Fermat and Christian Huygens. Jacob Bernoulli's *Ars Conjectandi (The Art of Conjecturing)* [3] was the first comprehensive work in this regard. The 19th-century pursuit of precise astronomical measurements spurred the independent development of the method of least squares by Gauss and Legendre. Gauss also introduced the foundational concept of normal (or Gaussian) distribution. The 19th century saw statistics extend to social phenomena. Adolphe Quetelet pioneered statistical methods in sociology, conceiving the *Average Man* [4].

Sir Francis Galton applied statistics to heredity, introducing correlation and regression. Karl Pearson, Galton's student, further developed these, established the first statistics department (1911), and introduced the *Chi-Squared Test* [5].

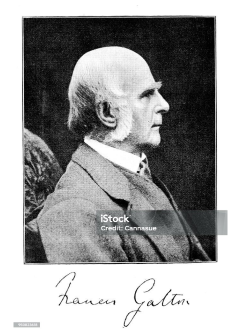

Figure 3: Photo Courtesy: istockphoto.com

The early 20^{th} century saw major theoretical strides in statistics. R.A. Fisher revolutionized experimental design, Analysis of Variance (ANOVA), and maximum likelihood estimation theory. Neyman and Pearson formalized hypothesis testing and introduced confidence intervals. Robust sampling techniques became key (with significant contributions from P.C. Mahalanobis, who founded the Indian Statistical Institute in Kolkata, India) [6]. The 20^{th} century witnessed widespread application across diverse fields and a statistical revolution with the advent of computers. The 21^{st} century is the era of *Big Data*. Statistics is now deeply linked with data science and machine learning, with its principles driving data analysis and prediction.

Probability theory and statistics:

Having introduced the issue of contemporary female smoking habits, the relevant population is female individuals. Theoretically, we can model this population using a *random variable* which is a quantity representing possible outcomes (e.g. "smokes" or "doesn't smoke"). To do this, we try to identify the probability distribution first, a mathematical function describing the likelihood of different outcomes of the random variable, that governs their smoking behaviour. In general, the theory of probability provides the theoretical framework for statistics. It deals with quantifying the likelihood that events occur based on a known underlying model or process. Statistics, on the other hand, uses observed data to make inferences about the unknown underlying processes or populations. Probability distributions are characterized by their parameters. Parameters are fixed, usually, unknown constants that define the shape, central tendency, and spread of a probability distribution. If these parameters are known, the distribution is fully specified. For example, the Binomial distribution models "yes/no" outcomes (like "Smoking"/"No Smoking") and is governed by the parameters p (probability of success) and n (number of trials). However, many real-world models involve unknown parameters. Imagine that my neighborhood records higher summer temperatures than other areas in Kolkata. Is this due to local construction or global warming? Without knowing parameters like the *true effect* of construction, we cannot assume a specific distribution. In such cases, a nonparametric approach, which does not rely on fixed parameters, may be appropriate. Thus, learning probability distributions is fundamental to developing statistical models of systems. Probability theory provides tools to quantify uncertainty, while statistics uses data to estimate unknown parameters or to test hypotheses. For instance, if we survey 100 women about smoking habits, the data helps us to estimate the parameter in question, even if we don't start with prior knowledge.

What is probability?

Following our earlier discussion, the question naturally arises: how should one learn probability theory? In my nineteen years of teaching, I've consistently posed a fundamental coin toss problem to my students: *what is the probability of getting heads when a single coin is tossed?* Almost without fail, they answer $\frac{1}{2}$. When I inquire about their reasoning, they invariably invoke the classical definition of probability – "*favorable outcomes divided by total outcomes*". However, their confidence wavers when I remind them that the coin's fairness was never established. The classical definition highlights a critical limitation, which assumes equally likely outcomes without any justification. To address such gaps, a rigorous and robust foundation for probability is needed, which is provided by the *Axiomatic Approach*. A more sturdy approach to learning probability theory involves understanding the three fundamental axioms upon which the entire field rests. Grasping these axioms requires familiarity with certain essential definitions such as *events* and *sample space*. The set of all possible outcomes of a random experiment[1] is called the *sample space* of the experiment, which I denote here by $\mathcal{S}$ and a subset of the sample space of a random experiment is called the *event*. So for a single coin toss, if $\mathcal{S} = \{H, T\}$, then we may consider $\mathcal{E} = \{H\}$ to be one of the events of the sample space $\mathcal{S}$. In his seminal 1933 work, Grundbegriffe der Wahrscheinlichkeitsrechnung (Foundations of the Theory of Probability), the renowned Russian mathematician Andrey Nikolaevich Kolmogorov established the foundational principles known as *Kolmogorov's Axioms of Probability*

[1] An experiment that can result in different outcomes, even though it is repeated in the same manner every time, is called a random experiment.

[5]. The axioms are as follows:

- **Non-negativity:** The probability of any event $\mathcal{E}$ is a non-negative real number, i.e. $P(\mathcal{E}) \geq 0$ for all events $\mathcal{E}$.
- **Unit measure:** The probability of the sample space $\mathcal{S}$ is always 1.
- **Countable additivity:** If $\mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_n, \cdots$ is a countable sequence of mutually exclusive events[2], by which we mean, $\mathcal{E}_i \cap \mathcal{E}_j = \phi$ for $i \neq j$, then the probability of their union is the sum of their individual probabilities i.e. $P(\cup_{i=1}^{\infty} \mathcal{E}_i) = \sum_{i=1}^{\infty} P(\mathcal{E}_i)$.

You take down any of the above axioms, the theory of probability will collapse. Given the established axioms, let us consider the probability of obtaining a head when a coin is tossed. Without assuming the coin is unbiased, the probability of heads, denoted by p actually can have any value in the interval $[0, 1]$. Consequently, the probability of tails, denoted by q, is $1 - p$. Empirically, as an experiment is repeatedly performed under identical conditions, the *relative frequency* of an event $\mathcal{E}$ (like getting heads in coin tosses) tends to approach a specific value as the number of trial increases. This limiting value, observed as the number of repetitions increases, defines the probability of $\mathcal{E}$. For instance, the proportion of heads in repeated coin flips approaches the probability of heads. Thus, for example, given 10000 flips of a fair coin resulting in 5011 heads, the relative frequency of heads is calculated as $\frac{5011}{10000} = 0.5011$. This long run stability is called the *frequentist definition of probability*. The classical definition of probability, which is the ratio of favourable outcomes to total outcomes, is a consequence of the axiomatic laws of probability under the assumption of equally likely outcomes.

[2] In coin toss, events $\{H\}$, $\{T\}$ are mutually exclusive events as occurrence of one event excludes the chance of occurrence of the other event.

If an experiment can result in any one of N different equally likely outcomes, and if n of these outcomes together constitute the event $\mathcal{E}$, then the probability of event $\mathcal{E}$ is $\frac{n}{N}$. Here we can denote individual outcomes by $O_1,\ O_2, \cdots, O_N$ having equal probability $\frac{1}{N}$ (and which is actually our hypothesis), then the event $\mathcal{E}$ can be thought of as the union of these O_is ($i = 1, 2, \cdots, n$), using the axiom of countable additivity, we can immediately write $P(\mathcal{E}) = \frac{1}{N} + \frac{1}{N} + \cdots + \frac{1}{N} = \frac{n}{N}$, thereby giving us the classical definition. The choice of the fraction $\frac{1}{N}$ and assigning it to each outcome is the limitation of the classical definition of probability.

The classical definition fails in real-world scenarios where outcomes are not equally likely. For instance, if a coin is biased, assuming $P(head) = \frac{1}{2}$ is incorrect; or if the sample space is infinitely large, the classical definition cannot be applied. So the classical definition is too restrictive - it relies entirely on the hypothesis of equally likelihood which often does not hold.

If you're finding the mathematical details overwhelming, a brief look at the history of probability might offer a welcome respite.

Historical Perspective:

The quantification of chance, or probability, represents a relatively recent intellectual achievement. For much of human history, events were perceived as governed by inscrutable forces, beyond the realm of rational understanding. It was only during the vibrant intellectual ferment of the early 17th century, amidst the waning embers of the Renaissance, that a burgeoning curiosity about the natural world and its underlying laws took hold. Within this milieu, the world of gambling, often a crucible for mathematical inquiry, played an unexpected role. A group of Italian gamblers, grappling with the intricacies of dice games, sought the wisdom of Galileo Galilei [7]. Despite his myriad scientific pursuits, Galileo recognized the inherent mathematical elegance in their questions, not only providing solutions but also composing a concise treatise on games of chance, a nascent exploration of probability. A similar narrative unfolded in France, centered on Chevalier de Mere, a gambler and amateur mathematician of considerable acumen. De Mere, seeking answers to more complex gaming dilemmas, turned to the prodigious Blaise Pascal. One particularly vexing problem, the *problem of the points*, concerning the fair division of stakes in an interrupted game, captivated Pascal's attention. In 1654, he initiated a correspondence with Pierre de Fermat, a mathematician of equal brilliance [8].

Figure 4: Blaise Pascal and Pierre de Fermat were renowned Mathematicians who contributed to the theory of chance (Photo Courtesy: Google).

This exchange, a testament to their intellectual prowess, transcended the immediate problem, laying the foundational framework for the analysis of numerous chance-based phenomena. Their celebrated correspondence, often hailed as the genesis of probability theory, ignited a fervent interest among Europe's leading mathematicians. The youthful Dutch prodigy, Christian Huygens, journeyed to Paris to engage in discourse on this burgeoning field, contributing to the rapid proliferation of research and development in this nascent domain of mathematical inquiry [5].

Theorems:

There is a fundamental difference between axioms and theorems. You can say in simpler terms that *axioms are the basic rules we accept to be true without proof, while theorems are the statements* *proven using these axioms.* From the axioms of probability, some theorems will instantly follow. One such theorem is about finding the probability of impossible events, for which we first have to understand what we mean by *impossible events*. In the theory of probability, an "impossible event" refers to an event that cannot occur under any circumstances within a given sample space. For example one cannot expect both head and tail from a single coin toss, be the coin is biased or unbiased (the only exception is when the coin is standing on its edge, which we avoid considering here). Such an event can be termed as impossible. Let me denote any impossible event by Φ. Obviously Φ is an event of any sample space considered[3]. Now $\mathcal{S}$, the sample space, can be expressed as the union of the Φ and itself, so that both being mutually exclusive, $P(\mathcal{S} \cup \Phi) = P(\mathcal{S}) + P(\Phi)$, implies that $P(\Phi) = 0$ (which directly follows from the axioms of *unit measure* and *countable additivity*.

It is interesting to note that, if you define an impossible event and puts its probability as one of the axioms, then the axiom of unit measure will automatically turn into a theorem. Some other theorems that instantly can be proved instantly using an axiomatic approach are as follows:

- If $\mathcal{E}_1$ and $\mathcal{E}_2$ are events in a sample space $\mathcal{S}$ and $\mathcal{E}_1$ is a sub-event of $\mathcal{E}_2$ (we write it as $\mathcal{E}_1 \subset \mathcal{E}_2$), then $P(\mathcal{E}_1) \leq P(\mathcal{E}_2)$.
- For any event $\mathcal{E}$, $0 \leq P(\mathcal{E}) \leq 1$.
- If $\mathcal{E}$ is any event and $\bar{\mathcal{E}}$ is the complementary event, then $P(\mathcal{E}) + P(\bar{\mathcal{E}}) = 1$.
- If $\mathcal{E}_1$ and $\mathcal{E}_2$ are any two events, then $P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \cap \mathcal{E}_2)$.

However, for any finite sequence of events (not necessarily mu-

[3]From set theoretic approach Φ being the null set is a subset of any universal set.

tually exclusive) $\mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_n$, one can show by induction that $\cup_{i=1}^{n} P(\mathcal{E}_i) \leq \sum_{i=1}^{n} P(\mathcal{E}_i)$ [9].

These theorems hold substantial importance within probability theory, and their proofs are readily established through the axiomatic foundation.

About Kolmogorov:

Andrey Nikolaevich Kolmogorov, a polymath of exceptional breadth, indelibly shaped the landscape of 20^{th}-century mathematics, most notably through his axiomatic foundation of modern probability theory in *Grundbegriffe der Wahrscheinlichkeitsrechnung* [9].

Born near Moscow in Tambov, 1903, Andrey Kolmogorov lost his mother, Maria Yakovlevna Kolmogorova, at birth and was raised by aunts in Tunoshna. His exiled revolutionary father, Nikolai Matveyevich Katayev, likely perished in the Russian Civil War. Kolmogorov's early talent emerged in his aunt Vera's village school, where, at five, he *edited* the mathematics section of *The Swallow of Spring*, publishing his first observation: the pattern in sums of odd numbers $1 = 1^2$; $1 + 3 = 2^2$; $1 + 3 + 5 = 3^2$ [10]. His intellectual pursuits, however, extended far beyond, encompassing seminal contributions to topology, intuitionistic logic, turbulence, classical mechanics, algorith-

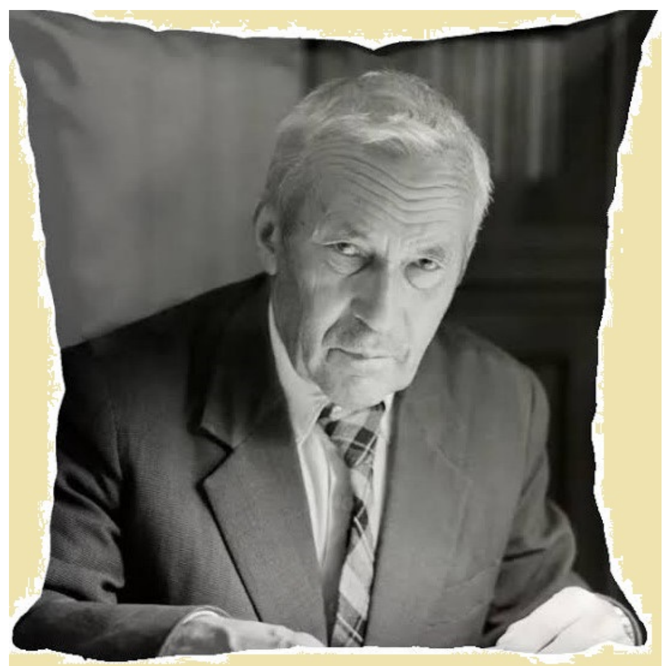

Figure 5: Andrey Nikolaevich Kolmogorov (25 April 1903 – 20 October 1987), a Soviet mathematician who played a central role in the creation of modern probability theory (Photo Courtesy: Google).

mic information theory, and computational complexity. From early forays into historical research on Novgorod landholding practices to groundbreaking work on Fourier series divergence, his precocity was evident. His 1930 European sojourn fostered collaborations with leading mathematicians, culminating in his influential *Über die analytischen Methoden der Wahrscheinlichkeitsrechnung*. Appointed professor at Moscow State University in 1931, he established its probability theory department in 1935, while also expanding the Lotka-Volterra ecological model. His 1938 work on stationary stochastic processes proved strategically vital during the Cold War, and his wartime contributions to Soviet artillery and aerial defence further demonstrated his applied genius. Alongside Sydney Chapman, he independently developed the eponymous Chapman-Kolmogorov equations, a cornerstone of stochastic process theory, solidifying his legacy as a titan of mathematical thought. In the latter part of his distinguished career, Andrey Nikolaevich Kolmogorov extended his global influence to the Indian subcontinent, notably visiting the Indian Statistical Institute in Kolkata, (India). During this period, he engaged with the vibrant community of Indian statisticians and mathematicians, fostering intellectual exchange and contributing to the development of probability theory and related fields within India. His presence served as a catalyst, inspiring researchers and students alike, and leaving an enduring legacy on the institute's academic pursuits, strengthening the collaborative ties between Soviet and Indian scientific communities.

Thought of the day:

Let me pause here, leaving the readers with a contemplative reflection:

While it is axiomatic that an impossible event possesses a probability of zero, one might mull over whether the converse holds true, i.e. whether probability of zero for an event invariably implies impossibility of that event !

References

[1] S. M. Stigler, *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1990.

[2] H. C. Darby and H. C. Darby, *Domesday England*. Cambridge University Press, 1986.

[3] I. Schneider, "Jakob bernoulli, ars conjectandi (1713)," in *Landmark Writings in Western Mathematics 1640-1940*, Elsevier, 2005, pp. 88–104.

[4] G. Eknoyan, *Adolphe quetelet (1796–1874)—the average man and indices of obesity*, 2008.

[5] S. M. Ross, *Introductory statistics*. Academic Press, 2017.

[6] C. Rao, "Pc mahalanobis.," *Current Science (00113891)*, vol. 99, no. 3, 2010.

[7] F. N. David, *Games, gods, and gambling: A history of probability and statistical ideas*. Courier Corporation, 1998.

[8] K. Binmore, "Blaise pascal versus pierre de fermat," in *Imaginary Philosophical Dialogues: between Sages down the Ages*, Springer, 2020, pp. 55–61.

[9] A. N. Kolmogorov, "Foundations of the theory of probability, chelsea pub," *Co., New York*, 1956.

[10] S.-S. Chern and F. Hirzebruch, *Wolf prize in mathematics*. World Scientific, 2000.