An Intuitive Introduction to Probability
University of Zurich

Instructor: Karl Schmedders

August 2021

1 Brief Introduction

Probability is a simple but a broad concept that we use several times a day right from checking the weather forecast before stepping out in the morning to checking the stocks prices before investing any.

Probability has several definitions and the most common one being the **Exact Definition of Probability**. By exact definition we mean the simple probability formula based problems like what is the probability of getting a 6 when we roll a dice etc which we have often heard since high school and college days. But in real life nobody cares what we get after rolling a dice, its close to useless.

However there is more to probability!

The second definition is the **Empirical Definition of Probability** or the definition according to **relative probabilities**. Here we look at past data or historical trends and try to predict the probability of occurrence of an event. For instance, I want to study in a particular college, enroll in a specific course and aim to get placed by the end of college. Now my probability of taking a decision whether or not I should enroll in that college will be based on the past years' placement records. Here although we do not have a crisp definition, this is a real life situation of probability and we have to deal with it in this broad domain.

The third definition of probability is the **Subjective Definition of Probability** which is used when we do not have any past data or trend to predict the occurrence. This probability is just based on an opinion or life experience which will influence the decision making. For example, what is the probability that I will finish my first Coursera course with full grades. In this case, we have no data as such to predict the outcome so it will be based on (say) the opinions of the instructors or my mentors. Another example is when product managers have to predict how their new product will be received by their audience.

2 Basic Definitions

- **Random Experiment**: Any experiment with some uncertain outcome. For e.g., rolling a die, predicting the weather etc. It is denoted by R.
- **Basic Outcomes**: Any outcome from a Random Experiment.
- **Sample Space, S**: Collection of all basic outcomes.
- **Event**: A subset of the sample space. (*denoted by any titlecase letter*)

Probability can never be negative. It is always between 0 and 1 or 0% to 100%. Growth rates however can be negative and even greater than 100.

Three axioms of probability theory (*also called Kolmogorov's axioms*) -

- The probability of an outcome in the sample space is 1, $P(S) = 1$
- For any event A, the probability of A is between 0 and 1, $0 \leq P(A) \leq 1$
- For disjoint events A and B, $P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$ — [*If two events have no elements in common those are called disjoint events.*]

Some additional rules:

- $P(A^c) = 1 - P(A)$ - Compliment rule
- $P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$ - This rule also holds when A and B are not disjoint.

Note: The probability axioms holds true for all the three definitions of probability.

3 Statistical Independence

When two events occur one after the other and the outcome of the first event in no way effects the outcome of the second, then such events are called independent events. For e.g. if we throw two dices then the outcome of the first doesn't effect the outcome of the second.

Imp: If $P(A)$ and $P(B)$ are two statistically independent events, then $P(A \cup B) = P(A)P(B)$.

4 Subjective Probabilities

For example, "Joe is a shy girl. She studies geography as a part of her UG degree and loves to draw and collect country flags and coins. She takes part in various painting competitions as a part of her hobby"

What do you think will be her job?

- A social leader
- A art teacher in college
- A professional geographer
- A social worker
- A professional geographer who is also a social worker
- A professional dancer

It is a subjective probability problem. We do not have any past data, however we can make assumptions based on who she is, like she might be a art teacher, a geographer or a social worker but definitely not a dancer or social leader. These seem to be reasonable predictions.

Since it is assumption-based we can make any predictions but if we look at this problem from a probabilistic mindset, it is always evident that $P(A \cap B) \leq P(A)$ or $P(A \cap B) \leq P(B)$ i.e., the probability of Joe being a professional geographer who is a also a social worker is less likely than Joe being either a professional geographer or a social worker.

Thus, When dealing with subjective probability we must keep note of this rule i.e., two events occurring simultaneously cannot be more likely than the individual events by themselves. We often make this mistake as humans. This is actually a famous fallacy called the **conjunction fallacy**. This work is very important in domains like behavioural economics, behavioural finance, etc.

5 Empirical Probabilities - Benford's Law

Empirical probability is predicting something with the help of the past data available with us. One of the most interesting example is that of predicting the first digit of a number. The first digit (most significant) of a number can be anything between 1 to 9 and it might seem that in all kinds of data, the candidates for the first digits (1-9) are equally likely but interestingly this is not the case in many real-life sets of numerical data. For instance, the probability of getting 1 as the first digit of a number differs from the probability of getting an 8.

This is Benford's Law. It states the probability empirically found of getting a particular digit as the first digit.

This is very widely seen in finance industries, stock markets, real estates, etc. Even this law is use to detect fraud cases in many domains like tax evaluation, etc.

Benford's Law

Frequency of "d" being the first digit,

$$P(d) = \log_{10}(1+1/d),\ d \in \{1,2,3,\ldots,9\}$$

d	1	2	3	4	5	6	7	8	9
P(d)	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Figure 1: Benford's Law

Imp: Note that Benford's Law is not followed in datasets that have a fixed maximum or minimum value and also when the data is manually assigned as in the case of phone numbers.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	County	State	Population	Leading Digit		digit	count	proportion	Benford						
2	Autauga County	Alabama	54571	5		1	953	0.30321	0.30103						
3	Baldwin County	Alabama	182265	1		2	594	0.18899	0.176091						
4	Barbour County	Alabama	27457	2		3	374	0.11899	0.124939						
5	Bibb County	Alabama	22915	2		4	308	0.09800	0.09691						
6	Blount County	Alabama	57322	5		5	213	0.06777	0.079181						
7	Bullock County	Alabama	10914	1		6	211	0.06713	0.066947						
8	Butler County	Alabama	20947	2		7	182	0.05791	0.057992						
9	Calhoun County	Alabama	118572	1		8	152	0.04836	0.051153						
10	Chambers County	Alabama	34215	3		9	156	0.04963	0.045757						
11	Cherokee County	Alabama	25989	2		Σ	3143	1	1						
12	Chilton County	Alabama	43643	4											
13	Choctaw County	Alabama	13859	1											
14	Clarke County	Alabama	25833	2		Source:		2010 U.S. Census Data							
15	Clay County	Alabama	13932	1				U.S. Census Bureau, Population Division							
16	Cleburne County	Alabama	14972	1				http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml							
17	Coffee County	Alabama	49948	4				(last accessed: August 30, 2016)							
18	Colbert County	Alabama	54428	5											
19	Conecuh County	Alabama	13228	1											
20	Coosa County	Alabama	11539	1											
21	Covington County	Alabama	37765	3											
22	Crenshaw County	Alabama	13906	1											

Figure 2: US census data from the year 2010 when the United States counted the number of people residing in its various counties.

	A	B	C	D	E	F	G	H	I	J	K	L
1	OCC_CODE	OCC_TITLE	TOT_EMP	Leading Digit		digit	count	proportion	Benford			
2	11-1011	Chief Executives	248760	2		1	271	0.33049	0.30103			
3	11-1021	General and Operations Managers	1973700	1		2	134	0.16341	0.176091			
4	11-1031	Legislators	55800	5		3	105	0.12805	0.124939			
5	11-2011	Advertising and Promotions Managers	28530	2		4	75	0.09146	0.09691			
6	11-2021	Marketing Managers	174010	1		5	62	0.07561	0.079181			
7	11-2022	Sales Managers	352220	3		6	61	0.07439	0.066947			
8	11-2031	Public Relations and Fundraising Managers	53730	5		7	45	0.05488	0.057992			
9	11-3011	Administrative Services Managers	269500	2		8	35	0.04268	0.051153			
10	11-3021	Computer and Information Systems Managers	319080	3		9	32	0.03902	0.045757			
11	11-3031	Financial Managers	499320	4		Σ	820	1	1			
12	11-3051	Industrial Production Managers	165340	1								
13	11-3061	Purchasing Managers	69620	6								
14	11-3071	Transportation, Storage, and Distribution Managers	102610	1		Source:		May 2013 National Occupational Employment and Wage Est				
15	11-3111	Compensation and Benefits Managers	17570	1				U.S. Bureau of Labor Statistics, Division of Occupational Emp				
16	11-3121	Human Resources Managers	110650	1				http://www.bls.gov/oes/current/oes_nat.htm#00-0000				
17	11-3131	Training and Development Managers	28340	2				(last accessed August 30, 2016)				
18	11-9013	Farmers, Ranchers, and Other Agricultural Managers	3770	3								
19	11-9021	Construction Managers	213720	2								
20	11-9031	Education Administrators, Preschool and Childcare Ce	47560	4								
21	11-9032	Education Administrators, Elementary and Secondary	226760	2								
22	11-9033	Education Administrators, Postsecondary	126340	1								
23	11-9039	Education Administrators, All Other	30880	3								

Figure 3: Employment in the United States from the Bureau of Labor Statistics across 820 job titles.