Probability and Statistics
MANUSCRIPT

William F. Barnes
[1]

November 9, 2024

Contents

Chapter 1

Probability and Statistics

1 Events and Probability

Probability theory is a branch of mathematics for studying systems with inherent randomness or uncertainty. It works closely with *statistics*, another branch of mathematics concerned with the organization and interpretation of data, along with *combinatorics*, a formal method of counting.

Stochastic Process

Systems that exhibit random or pattern-less behavior contain a *stochastic* component. Typical stochastic processes may include flipping a coin, drawing a card from a deck, rolling a dice, or playing darts while blindfolded. A *stochastic event* is any data generated by a stochastic process, and the set of all possible stochastic events is called the system's *sample space*.

1.1 Events

Elementary Events

Events that are considered *elementary* carry one 'unit' of information, loosely speaking. A coin landing on 'heads', or a dice landing on 3 qualify as elementary events.

Compound Events

Simple events that occur in groups are called *compound events*. Drawing a Queen of Hearts from a deck of cards carries two units information, and may be interpreted in several ways: 'draw a Queen AND a heart', or 'draw a Queen OR a Heart', or perhaps 'draw NOT a Diamond'. Such events are compound for this reason.

Compound Event Notation

Borrowing the familiar symbols from elementary logic, we denote the word 'AND' with the 'cap' symbol $\cap$, equivalent to multiplication ($\cdot$). Meanwhile, the word 'OR' uses the 'cup' symbol $\cup$, or sometimes just a plus sign ($+$). The 'NOT' operator is abbreviated by a dash above the symbol, as in 'NOT' $A = \bar{A}$. Any event that is infinitely improbable, impossible, or undefined is denoted by the 'Empty set' symbol, $\emptyset$. In summary:

$$
\begin{aligned}
A \text{ AND } B &= A \cdot B = A \cap B \\
A \text{ OR } B &= A + B = A \cup B \\
\text{NOT } A &= \bar{A} \\
\text{Empty set} &= \emptyset
\end{aligned}
$$

The logic of probabilistic analysis is the same as 'ordinary' logic. For instance, the philosophical axiom 'nothing can be and not be simultaneously' is contained in the statement:

$$A \cap \bar{A} = \emptyset$$

State

The *state* of a system, loosely defined, is any particular configuration of the variables used to describe that system. For instance, a snapshot of a chessboard contains the present state of the game. Any event taking place in a system usually changes its state. If the system is to evolve in time, as would a game of chess, then future states evolves from the present state according to some rules or model of evolution.

1.2 Probability

Statistical Probability

A stochastic process that iterates over a very large or infinite number of trials will produce data points randomly distributed among the space of all possible data points for that process. For all events of type A, the ratio of occurrences N_A over all N events is called the *statistical probability* of event A, defined as:

$$P(A) = \lim_{N\to\infty} \frac{N_A}{N} \tag{1.1}$$

$P(A)$ strictly has values between 0 and 1, inclusive.

Normalization Conditions

All other events B, C, etc., are represented by the symbol $\bar{A}$ ('NOT' A), and obey:

$$N_A + N_{\bar{A}} = N \tag{1.2}$$

$$P(\bar{A}) + P(A) = 1 \tag{1.3}$$

Classical Probability

A definition that skirts around the invocation of $N \to \infty$ is called the *classical probability.* For instance, it does not require an infinite number of rolls on a six-sided dice to know the chances of landing on 3 are one out of six, as this quality is built into the dice itself. Classical systems like dice or playing cards are most succinctly analyzed using classical probability.

Counting States

In probability and statistics, it's often necessary to know the total number of states available to a system, sometimes requiring rigorous combinatorial consideration.

Problem 1

The last four digits of a phone number have the format ABCD, where each letter represents any integer from 0 to 9, inclusive. What is the probability of randomly guessing the number 7766?

Answer: Right away, we know how to list all possible states of the password, starting from 0000 and ending at 9999 in numerical order. With $N = 1000$ passwords, the probability of randomly choosing the correct password $N_A = 7766$ is:

$$P(7766) = \frac{1}{N} = \frac{1}{10000}$$

Problem 2

A bank account password has format ABCD, where each letter represents any integer from 0 to 3, inclusive. What is the probability of randomly guessing the password?

Answer: All two-digit arrangements solved by AB are contained in:

$$\begin{aligned} \omega = &00, 01, 02, 03, \\ &10, 11, 12, 13, \\ &20, 21, 22, 23, \\ &30, 31, 32, 33 \end{aligned}$$

From here, observe that all four-digit arrangements are contained on an $\omega \times \omega$ grid having $N = 16^2 = 256$ total members, or

$$P(N_A) = \frac{1}{N} = \frac{1}{256} \,.$$

1.3 Mutually Exclusive Events

A pair of *mutually exclusive events* A and B are those that cannot occur simultaneously. Their coincidence can only belong to the *empty set* as

$$A \cap B = \emptyset \,.$$

If two events are mutually exclusive, the probability of either event occurring is the sum of the individual probabilities:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) \tag{1.4}$$

Problem 3

Calculate the probability of rolling a 3 or a 4 on a six-sided dice.

Answer: As mutually exclusive events, we simply have

$$P = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \,.$$

Problem 4

Calculate the probability that a random three-card hand drawn from a 52-card deck contains the Queen of Hearts. Answer:

$$P = \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{3}{52}$$

1.4 Non-Exclusivity

Non-mutually exclusive events are those that cause 'double counting' in $P(A \cup B)$, and are adjusted by subtracting the probability that both occur:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.5}$$

Or, in street terms, the above reads:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Problem 5

From a 52-card deck, calculate the probability of drawing a Heart or a Face card, or one that is both. Answer:

$$P = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52}$$

Problem 6

A class of 30 students is in session. 16 are studying French, and 21 are studying Spanish. Choosing a student at random, find the probability that they:

- study French
- study Spanish
- study French and Spanish
- study only French
- study only Spanish
- study French or Spanish

Answer: Denote F for French and S for Spanish. Then the easy ones can be listed off:

$$P(F) = 16/30$$
$$P(S) = 21/30$$

If the number of 'multilingual' students studying both French and Spanish is denoted M, then

$$(16 - M) + M + (21 - M) = 30$$

must hold, telling us $M = 7$, or

$$P(F \cap S) = 7/30\,.$$

With M known, the the number of students studying just one subject can be written:

$$P(\text{French only}) = P(F) - P(F \cap S) = 9/30$$
$$P(\text{Spanish only}) = P(S) - P(F \cap S) = 14/30$$

Finally, the number of students studying French or Spanish should equal the total, which is the sum of those studying French only, Spanish only, or both. The probability should equal one:

$$P(F \cup S) = \frac{9}{30} + \frac{14}{30} + \frac{7}{30} = 1$$

1.5 Independent Events

Two events A and B that occur simultaneously as the compound event $A \cap B$ are *independent* if not causally connected.

In general, the statistical probability for the compound event $A \cap B$ reads

$$P(A \cap B) = \lim_{N \to \infty} \frac{1}{N} N_{A \cap B}\,,$$

where in the $N \to \infty$, limit the quantity $N_{A \cap B}$ becomes $N_A \cdot P(B)$. We deduce that, for independent events, the compound probability is the product of the individual probabilities:

$$P(A \cap B) = P(A)\,P(B) \tag{1.6}$$

Problem 7

Calculate the probability of two fair coin tosses each landing on 'tails'. Answer:

$$P(T \cap T) = P(T) \cdot P(T) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Problem 8

From a 52-card deck, what is the probability of randomly drawing (i) a Queen, (ii) a Heart, (iii) the Queen of Hearts? Answer:

$$P(Q) = 1/4$$
$$P(H) = 1/13$$
$$P(Q \cap H) = P(Q) \cdot P(H) = \frac{1}{4} \cdot \frac{1}{13} = \frac{1}{52}$$

Coworker Problem

Two people are neighbors and travel to the same job. Person A owns car A, which has a 70% chance of starting in the morning, and a 30% chance of stalling (not starting). Person B owns car B with an 80% chance of starting. If one or both cars start, both people arrive at work. If neither car starts, they both miss work. In a span of 100 workdays, how many days are missed?

Each morning, one of four things happen:

$$\begin{aligned} P_1 &= \text{A starts, B starts} \\ P_2 &= \text{A starts, B stalls} \\ P_3 &= \text{A stalls, B starts} \\ P_4 &= \text{A stalls, B stalls} \end{aligned}$$

As independent events, we further have

$$\begin{aligned} P_1 &= (0.7)\,(0.8) = 0.56 \\ P_2 &= (0.7)\,(1-0.8) = 0.14 \\ P_3 &= (1-0.7)\,(0.8) = 0.24 \\ P_4 &= (1-0.7)\,(1-0.8) = 0.06\ , \end{aligned}$$

which passes the sanity check

$$\sum_{j=1}^{4} P_j = 1\ .$$

The answer to the question is the combined probability of at least one car starting. For this, we simply have

$$P = P_1 + P_2 + P_3 = 0.56 + 0.14 + 0.24 = 0.94\ ,$$

or 94%. Six days are missed of every hundred.

1.6 Conditional Probability

In contrast to independent events, systems may bear a notion of 'dependent events', meaning that event B can occur only if event A occurs. This is called a *conditional probability*, denoted $P(B|A)$, enunciated 'B given A'. By definition, the probability of event B occurring given condition A is

$$P(B|A) = \lim_{N\to\infty} \frac{1}{N_A} N_{A\cap B}\ .$$

The term $N_{A\cap B}$ is the number of events B that occur given event A, which shows up again in the equation for $P(A\cap B)$:

$$P(A\cap B) = \lim_{N\to\infty} \frac{N_{A\cap B}}{N}\ .$$

Divide the two equations and simplify to derive the statement of conditional probability:

$$P(A\cap B) = P(B|A)\,P(A) \tag{1.7}$$

Note that the above generalizes the case of independent events, for if events A and B are independent, this result reduces to $P(A\cap B) = P(A)\,P(B)$ again.

<u>Problem 9</u>

In a 52-card deck, calculate the probability that the first three cards are Kings. Answer:

$$\begin{aligned} P(KKK) &= P(K)\,P(K|K)\,P(K|\,(K|K)) \\ &= \frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50} \approx 0.000181 \end{aligned}$$

<u>Problem 10</u>

In a 52-card deck, calculate the probability that the first three cards are KQJ, in that order, with mixed suits allowed. Answer:

$$\begin{aligned} P(KQJ) &= P(K)\,P(Q|K)\,P(J|\,(Q|K)) \\ &= \frac{4}{52}\cdot\frac{4}{51}\cdot\frac{4}{50} \approx 0.000483 \end{aligned}$$

<u>Problem 11</u>

Suppose a pair of six-sided dice are rolled, landing on faces X and Y, respectively. What is the probability that $X = 2$ given that $X + Y \le 5$?

Answer: Of the 36 possible outcomes for a pair of dice, 10 of them obey $X + Y \le 5$. Listing these, find that only $(2,1)$, $(2,2)$, $(2,3)$ satisfy $X = 2$, a total of three outcomes. The final answer is the ratio of these counts:

$$P\left(X = 2 \middle| X + Y \le 5\right) = \frac{3}{10}$$

Inversion Trick

Making use of the normalization condition

$$P\left(\bar{A}\right) + P(A) = 1\ ,$$

it often helps in problem solving to use the negated event $\bar{A}$ as the working variable.

<u>Problem 12</u>

Calculate the probability that a random three-card hand drawn from a 52-card deck contains the Queen of Hearts.

Answer: For this, define the event $\bar{A}$ of *not* drawing the Queen of Hearts and use the inversion trick as follows:

$$\bar{A} = \bar{A}_1 \cdot \bar{A}_2 \cdot \bar{A}_3$$

$$P\left(\bar{A}_1 \cap \bar{A}_2\right) = P\left(\bar{A}_1\right) P\left(\bar{A}_2 | \bar{A}_1\right) = \frac{51}{52} \cdot \frac{50}{51} = \frac{50}{52}$$

$$P\left(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3\right) = P\left(\bar{A}_1 \cap \bar{A}_2\right) P\left(\bar{A}_3 | \bar{A}_1 \cap \bar{A}_2\right) = \frac{50}{52} \cdot \frac{49}{50} = \frac{49}{52}$$

$$P(A) = 1 - P\left(\bar{A}\right) = 1 - \frac{49}{52} = \frac{3}{52}$$

Radioactive Decay

An unstable atom is one that expels energy by ejecting a subatomic particle or photon. Having no internal time-keeping mechanism, an unstable atom is entirely 'unaware' of its absolute age, and the its decay occurs at a random moment after becoming unstable.

Supposing the observation of an unstable atom begins at $t = 0$, the conditional probability of the atom decaying in a small time window Δt after time $t > 0$ is

$$P^{\text{decay}}_{\Delta t/t} = \tau^{-1} \Delta t \,,$$

where τ^{-1} is a constant related to (but not precisely equal to) the statistical half-life of the element, defined such that $\Delta t \ll \tau$.

The probability of the atom being 'still alive' in the interval Δt is

$$P^{\text{alive}}_{\Delta t/t} = 1 - \Delta t / \tau \,.$$

Decompose the entire 'alive' state into a product of conditional probabilities by slicing the time t into n identical copies of the short interval Δt as:

$$P^{\text{alive}}(t) = P^{\text{alive}}_{\Delta t/t_1} \cdot P^{\text{alive}}_{\Delta t/t_2} \cdots P^{\text{alive}}_{\Delta t/t_n} = \left(1 - \frac{t}{n\tau}\right)^n$$

Letting $n \to \infty$ permits use of the identity

$$\lim_{n\to\infty} \left(1 + \frac{A}{n}\right)^n = e^A \,.$$

It follows that the probability that a single unstable atom will still be 'alive' obeys

$$P(t) = e^{-t/\tau} \,. \tag{1.8}$$

One can work out the so-called half life $\tau_{1/2}$ of the atom by inquiring when $P(t)$ reduces to $1/2$.

Missing Face Problem

A six-sided dice is rigged to keep rolling if it lands on 2. Prove that the statistical probability of rolling a 3 is $1/5$.

Denote B as the event 2, and denote A as event 3. With a single roll, the probabilities of A or B occurring are easy to write down:

$$p_1(A) = 1/6$$
$$p_1(B) = 1/6$$

Of course, event B is unstable and induces a re-roll, which has a $1/6$ chance of generating event A again, and the same chance for event B:

$$p_2(A|B) = (1/6)(1/6) = (1/6)^2$$
$$p_2(B|B) = (1/6)(1/6) = (1/6)^2$$

With event $B|B$ comes another re-roll, and we stack on the conditional probabilities as

$$p_3(A|B|B) = (1/6)^3$$
$$p_3(B|B|B) = (1/6)^3 \,,$$

and the pattern is clear.

It follows that A could occur after any number of rolls, or potentially occur after an infinite string of B-events, and the total probability for A occurring is

$$P(A) = p_1(A) + p_2(A|B) + p_3(A|B|B) + p_4(A|B|B|B) + \cdots \,,$$

simplifying to:

$$P(A) = \frac{1}{6} \cdot \left(1 + \left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^3 + \cdots\right)$$

In the infinite limit, the geometric series in parentheses converges to $6/5$. The probability of event A occurring is therefore:

$$P(A) = \frac{1}{6} \cdot \frac{6}{5} = \frac{1}{5}$$

1.7 Bayes' Theorem

For two events A and B, recall that the statement of conditional probability reads

$$P(A \cap B) = P(B|A) P(A) \,,$$

which gives the likelihood events A and B simultaneously occurring. It's equally valid to write the statement with A and B swapped, giving a complimentary statement for 'A given B':

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

Now, since $A \cap B$ is logically equivalent to $B \cap A$, we immediately know $P(A \cap B) = P(B \cap A)$, allowing the two conditional equations to be combined to arrive at *Bayes' theorem*:

$$P(B|A) = \frac{P(A|B) P(B)}{P(A)} \tag{1.9}$$

Laptop Repair Shop

Source: CMPSCI240 UMass Amherst 2013

Problem 13

You work in a laptop repair shop. 80% of laptops brought in have been dropped, 15% of laptops have had a drink spilled on them, and 5% of laptops have a variety of other problems. A customer drops off a laptop and doesn't tell you what happened to it. You notice the laptop is emitting a slight coffee-like smell. Based on your knowledge of broken laptops, you estimate that 20% of dropped laptops have a slight coffee-like smell, 65% of laptops that have had something spilled on them have a slight coffee-like smell, and 5% of laptops that have some other problem have a slight coffee-like smell.

Provide labels for the events described in the problem. Find the probability that the laptop had something spilled on it given that is has a slight coffee-like smell.

Part 2:

After closer inspection, you note that the laptop has no cracks on the case. Based on your knowledge of broken laptops, you estimate that 80% of dropped laptops have cracked cases, 11% of laptops that have had something spilled on them have cracked cases, and 9% of laptops that have some other problem have cracked cases.

If the probability that a laptop smells like coffee and the probability that a laptop has a cracked case are conditionally independent of each other given the cause of the damage (drop, spill, or other), what is the probability that the laptop had something spilled on it if it has a slight coffee-like smell and no cracks in the case?

Solution 13

Denote D for 'drop', S for 'spill', and O for 'other'. Let the letter F denote 'slight coffee-like smell'. The information provided in the problem may be written:

$$\begin{aligned} P\left(D\right) &= .80 \\ P\left(S\right) &= .15 \\ P\left(O\right) &= .05 \end{aligned}$$

$$\begin{aligned} P\left(F|D\right) &= .20 \\ P\left(F|S\right) &= .65 \\ P\left(F|O\right) &= .05 \end{aligned}$$

We further deduce:

$$\begin{aligned} P\left(F\right) = P\left(F|D\right)P\left(D\right) + P\left(F|S\right)P\left(S\right) \\ + P\left(F|O\right)P\left(O\right) = .26 \end{aligned}$$

To answer the question, we need to compute $P\left(S|F\right)$, which is the inversion of $P\left(F|S\right)$. Applying Bayes' theorem, we easily find:

$$P\left(S|F\right) = \frac{P\left(S\right)P\left(F|S\right)}{P\left(F\right)} = .375$$

Part 2:

The problem asks us to evaluate $P\left(S|F \cap Z^C\right)$, where Z denotes 'crack' and Z^C denotes 'no crack,' using the information

$$\begin{aligned} P\left(Z|D\right) &= .80 \\ P\left(Z|S\right) &= .11 \\ P\left(Z|O\right) &= .09\,, \end{aligned}$$

or equivalently:

$$\begin{aligned} P\left(Z^C|D\right) &= 1 - .80 = .20 \\ P\left(Z^C|S\right) &= 1 - .11 = .89 \\ P\left(Z^C|O\right) &= 1 - .09 = .91 \end{aligned}$$

Proceed by applying Bayes' theorem directly to write

$$P\left(S|F \cap Z^C\right) = \frac{P\left(F \cap Z^C|S\right)P\left(S\right)}{P\left(F \cap Z^C\right)}\,.$$

Due to the independence between F and Z^C, the term

$$P\left(F \cap Z^C|S\right)$$

decouples into

$$P\left(F|S\right)P\left(Z^C|S\right)\,.$$

The denominator term $P\left(F \cap Z^C\right)$ can be recast as a sum that factors in a similar way:

$$\begin{aligned} P\left(F \cap Z^C\right) &= \sum_{i=D,S,O} P\left(F \cap Z^C|X_i\right)P\left(X_i\right) \\ &= \sum_{i=D,S,O} P\left(F|X_i\right)P\left(Z^C|X_i\right)P\left(X_i\right) \end{aligned}$$

Evaluating the final answer is now straightforward:

$$P\left(S|F \cap Z^C\right) = .7169$$

2 Applications

2.1 Copernicus Method

Source: Futility Closet
`https://futilitycloset.com/`

Princeton astrophysicist J. Richard Gott was visiting the Berlin Wall in 1969 *when a curious thought occurred to him.* His visit occurred at a random moment t years after the wall was created. Dividing the total lifespan T of the wall into four equal intervals, Gott reasoned there is a 50% chance that t lands within the middle two quarters of the wall's lifespan. Built in 1961, the wall was $t = 8$ years old at the time of his visit.

With this setup, we see at minimum that t is one quarter of T. At the other extreme, t could be three quarters of T. We therefore write

$$T_{\text{max}} = 4t \qquad T_{\text{min}} = t \times \frac{4}{3}$$

to establish upper and lower estimates of the wall's lifespan. Inserting $t = 8$ years, we find

$$T_{\text{max}} = 32\,\text{yr} \qquad T_{\text{min}} \approx 10.67\,\text{yr}\ ,$$

which, when these are added to 1961, produces the pair of results

$$\text{Year}_{\text{max}} \approx 1993 \qquad \text{Year}_{\text{min}} \approx 1972\ .$$

That is, Gott found a 50% chance that the Berlin Wall would fall between the years 1972 and 1993. The wall came down in 1989.

Generalization

Now generalize the above method using N intervals instead of four. Doing so, we begin with

$$T_{max} = Nt \qquad T_{min} = \frac{Nt}{N-1}\ .$$

Of course, the window defined by $T_{max} - T_{min}$ no longer corresponds to a probability of 50%, but must be adjusted to

$$p\,(N) = \frac{N-2}{N}\ ,$$

or

$$N\,(p) = \frac{2}{1-p}\ ,$$

which must be an integer.

An interesting exercise is one that calculates $\Delta T = T_{max} - T_{min}$ and expresses the result all in terms of $p\,(N)$. Figuring this out, one finds

$$\Delta T = \frac{4tp}{1-p^2}\ ,$$

which can be inverted via the quadratic formula:

$$p\,(\Delta T) = \frac{-2t}{\Delta T} + \sqrt{\left(\frac{2t}{\Delta T}\right)^2 + 1}$$

<u>Problem 14</u>

Suppose you encounter a man for the first time in the 42^{nd} year of his life. Determine the upper and lower bounds of the interval in which he has a 33% chance of expiring (in total years). Repeat for 66% and 75%. Answer:

$$\begin{aligned} T_{max(33\%)} &= 3 \times 42 = 126 \\ T_{min(33\%)} &= \frac{3}{2} \times 42 = 63 \end{aligned}$$

$$\begin{aligned} T_{max(66\%)} &= 6 \times 42 = 252 \\ T_{min(66\%)} &= \frac{6}{5} \times 42 = 50.4 \end{aligned}$$

$$\begin{aligned} T_{max(75\%)} &= 8 \times 42 = 336 \\ T_{min(75\%)} &= \frac{8}{7} \times 42 = 48 \end{aligned}$$

2.2 Dice Stacking

Suppose you are given a black box containing a pair of three-sided dice, abbreviated $2d3$. Denoting the outcome of a given dice as 1, 2, 3, we can list the 'internal outcomes' if the whole box is rolled:

$$\omega = 11, 12, 21, 13, 22, 31, 23, 32, 33$$

As a black box, there is no way to look inside. Instead, the box contains a small creature who announces the sum of the digits after a roll. Summing pairs of digits in ω, we write the corresponding black box outcomes as

$$\sigma = 2, 3, 3, 4, 4, 4, 5, 5, 6\ ,$$

having five unique members.

In a more visual notation, the above is equivalent to:

$\sigma_2 =$ ⚀⚀
$\sigma_3 =$ ⚀⚁, ⚁⚀
$\sigma_4 =$ ⚀⚂, ⚁⚁, ⚂⚀
$\sigma_5 =$ ⚁⚂, ⚂⚁
$\sigma_6 =$ ⚂⚂

Effective d5?

The question on hand is, can we use the black box as an effective five-sided dice? There are five unique members in σ, but the density of outcomes is not uniform.

From the information contained in σ, we can work toward an effective $d5$ so long as the nonuniform distribution problem can be dealt with. Proceed by listing the allowed outcomes divided by the respective density, i.e.

$$\gamma = \frac{1}{1}(2), \frac{1}{2}(3), \frac{1}{3}(4), \frac{1}{2}(5), \frac{1}{1}(6) \ .$$

Multiply by $3 \cdot 2$ to get rid of all denominators:

$$\gamma = 6(2), 3(3), 2(4), 3(5), 6(6)$$

Explicitly, γ is a list with twenty items:

$$\gamma = 2,2,2,2,2,2,3,3,3,4,4,5,5,5,6,6,6,6,6,6$$

Outcomes that are under-represented in σ are over-represented in γ, and vice-versa.

The d5 Gamma Ray

The item γ is the 'gamma-array', or 'gamma ray' for short. Taking any gamma ray γ, assign the items in γ to an index via $\gamma(k)$, where k is an integer between 1 and 20, inclusive.

In the above example, one can see there is nothing particularly special about the arrangement of items 2, 3, 4, etc. in the array. These are listed by group in ascending order for mere convenience.

Since the intent is to build a uniformly-behaving $d5$ dice, it's prudent to conceive of the set $\{\gamma\}$ of *all* possible reshuffles of the items in γ. Of course, one wouldn't attempt writing down the full set $\{\gamma\}$, or stepping through its members in any systematic way. It suffices instead to have a function that randomly shuffles an existing gamma ray to produce another.

The d5 Algorithm

An effective $d5$ is achieved with the following steps:

1. Choose a random gamma ray from $\{\gamma\}$ and let $k = 1$.
2. Let x equal the random sum of a $2d3$ roll:
$$x = \texttt{dice}(3) + \texttt{dice}(3)$$
3. If $x = \gamma(k)$, record $x - 1$ as a valid result.
4. Let $k \to k + 1$.
5. If $k > 20$ then Goto 1.
6. Goto 2.

3 Combinatorics

3.1 Arrangements

The number A_n of all *arrangements* of n distinguishable (non-repeated) elements is equal to the factorial of the total number of elements:

$$A_n = n!$$

If the set of n elements contains any number m identical members, the number of arrangements overcounts by a factor of m-factorial, which must be divided out:

$$A_n^m = \frac{n!}{m!} \tag{1.10}$$

<u>Problem 15</u>

Consider the set of twelve elements *ACEFGILMNTUX*. What is the probability that a random arrangement of the elements will spell out *MAGNETICFLUX*? Answer:

$$P(MAGNETICFLUX) = \frac{1}{12!}$$

<u>Problem 16</u>

Consider the word *FLUXELECTRIC*. What is the probability that a random arrangement of the letters will spell out *ELECTRICFLUX*? Answer:

$$P(ELECTRICFLUX) = \frac{2!2!2!}{12!}$$

3.2 Permutations

For a set of width n, partition each of the $n!$ arrangements into two bins such that one bin contains the first m elements in the arrangement, and the other bin contains the remaining $n - m$ elements. For each of the m elements in the first bin, the unused elements in the second bin are subject to $(n - m)!$ arrangements. Dividing out this factor yields the *permutation* number:

$$P_n^m = \frac{n!}{(n-m)!} \tag{1.11}$$

Qualitatively, the permutation number tells how many ways there are to choose m unique elements from a set of n total elements.

<u>Problem 17</u>

A door keypad is unlocked by a code of four different integers between 0 to 9, inclusive. The same integer cannot be used twice. How any possible passwords are there?

Answer: For a $k = 4$ digit password drawing (and consuming) from $N = 10$ integers, observe that N of them are available for the first digit A. $N - 1$ of the digits are available for the second digit B, and so on, with the k^{th} digit selecting from $N - k + 1$ unused integers. In general, we can intuitively write

$$P_N^k = N(N-1)\dots(N-k+1) = \frac{N!}{(N-k)!} ,$$

which builds the permutation formula:

$$P_{10}^4 = \frac{10!}{(10-4)!} = \frac{10!}{6!} = 5040$$

<u>Problem 18</u>

In a 52-card deck, calculate the probability that the first three cards are Kings. (This is a repeat of an earlier problem.)

Answer: The total number of ways to draw any three cards from 52 is

$$P_{52}^3 = \frac{52!}{(52-3)!} = 52 \cdot 52 \cdot 50 .$$

Meanwhile, the number of ways to draw any three Kings from four total Kings is

$$P_4^3 = \frac{4!}{(4-3)!} = 4 \cdot 3 \cdot 2 .$$

The ratio of these is the answer:

$$P(KKK) = \frac{P_4^3}{P_{52}^3} = \frac{4 \cdot 3 \cdot 2}{52 \cdot 51 \cdot 50} \approx 0.000181$$

<u>Problem 19</u>

In a 52-card deck, calculate the probability that the first three cards are KQJ, in that order, with mixed suits allowed. (This is a repeat of an earlier problem.)

Answer: Consider the three cards KQJ in that order. Listing off all ways this could occur, we see there are 4^3 possibilities in total, as each card has four suits to choose from. The total number of ways to draw any three cards from 52 is

$$P_{52}^3 = \frac{52!}{(52-3)!} = 52 \cdot 52 \cdot 50 .$$

The ratio of these is the answer:

$$P(KQJ) = \frac{4^3}{52 \cdot 51 \cdot 50} \approx 0.000483$$

Birthday Problem

Consider a room populated by N people. What is the probability that any two people were born on the same day? (Ignore leap year.)

Begin with the trivial case $N = 2$, in where there is a 1/365 chance of a common birthday:

$$P(2) = \frac{1}{365} = 1 - \frac{364}{365}$$

The result is written in the form $1 - X$ so we may focus on X, the probability of *no* common birthday.

A third person entering the system, making $N = 3$, has $365 - 2 = 363$ available days to avoid a common birthday. The probability becomes

$$P(3) = 1 - \frac{364}{365} \cdot \frac{363}{365} ,$$

and the pattern becomes clear. For total population N, the probability that some pair of people share a birthday ought to be:

$$\begin{aligned} P(N) &= 1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{(365-N+1)}{365} \\ &= 1 - \frac{365!}{365^N (365-N)!} \end{aligned}$$

On the right, note that X has been expressed as a recursion of conditional probabilities:

$$X(n|n-1) = \frac{365-(n-1)}{365}$$

$$X(N) = \prod_{n=2}^{N} X(n|n-1)$$

The same result can be written directly by the permutation formula. Choosing N people of 365, we have

$$P_{365}^N = \frac{365!}{(365-N)!} .$$

Meanwhile, the number of ways to assign birthdays to N people is 365^N without worrying about sharing. The ratio of these is the same $X(N)$ calculated above:

$$X(N) = \frac{1}{365^N} P_{365}^N = \frac{365!}{365^N (365-N)!} .$$

Following is a list of various populations N with their corresponding $P(N)$:

N	P(N)
5	2.71%
10	11.7%
20	41.1%
23	50.7%
30	70.6%
50	97.0%

Remarkably, the population need only be 23 in order for there to be a 50% chance that any two people share a birthday.

3.3 Combinations

Extending the derivation of the permutation formula, it may happen that the precise order of elements in the 'm' bin does not matter, meaning the list of permutations is overpopulated by a factor of $m!$. Dividing out this factor, we attain the number of *combinations* in the system:

$$C_n^m = \frac{n!}{m!\,(n-m)!}$$

The numbers C_n^m are none other than the binomial coefficients:

$$C_n^m = \binom{n}{m}$$

Problem 20

From a 52-card deck, a five-card hand is drawn at random. How many five-card hands are possible? Answer:

$$C_{52}^5 = \frac{52!}{5!\,(52-5)!}$$

Problem 21

From a 52-card deck, calculate the probability of drawing a royal flush (A-K-Q-J-10) in any order in any one suit. Answer:

$$P\left(RF\right) = \frac{4}{C_{52}^5} = \frac{1}{649740} \approx 0.00000154$$

Lottery Game

In a lottery game, the winning numbers are five non-repeating integers between 1 and 75, inclusive, along with one bonus integer between 1 and 15, inclusive. Guessing the five winning numbers at random, let us calculate the probability $P = (n, b)$ that n of the guessed numbers match the winning numbers, with or without the bonus b also being correctly guessed.

As an application of combinatoric analysis, it follows that there are $C_{75}^5 = 17,259,390$ ways to guess the winning five numbers, and $C_{15}^1 = 15$ choices for the bonus number. Following are the probabilities of guessing partial winning numbers, with and without the bonus.

$$P\left(5,1\right) = \frac{1}{C_{75}^5 \cdot C_{15}^1} = \frac{1}{258,890,850}$$
$$P\left(4,1\right) = \frac{C_5^4 \cdot C_{70}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{739,688}$$
$$P\left(3,1\right) = \frac{C_5^3 \cdot C_{70}^2}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{10,720}$$

$$P\left(2,1\right) = \frac{C_5^2 \cdot C_{70}^3}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{473}$$
$$P\left(1,1\right) = \frac{C_5^1 \cdot C_{70}^4}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{56}$$
$$P\left(0,1\right) = \frac{C_5^0 \cdot C_{70}^5}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{21}$$

$$P\left(5,0\right) = \frac{C_5^5 \cdot C_{70}^0 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{18,492,204}$$
$$P\left(4,0\right) = \frac{C_5^4 \cdot C_{70}^1 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{52,835}$$
$$P\left(3,0\right) = \frac{C_5^3 \cdot C_{70}^2 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{766}$$

$$P\left(2,0\right) = \frac{C_5^2 \cdot C_{70}^3 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{34}$$
$$P\left(1,0\right) = \frac{C_5^1 \cdot C_{70}^4 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{4}$$
$$P\left(0,0\right) = \frac{C_5^0 \cdot C_{70}^5 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{2}{3}$$

4 Variables and Data

4.1 Normalization

For an event A, the probability $P\left(A\right)$ of the event occurring has a trivial yet important relationship to $P\left(\overline{A}\right)$, via the normalization condition

$$1 = P\left(A\right) + P\left(\overline{A}\right) \; .$$

In words, normalization means there is a 100% chance that event A either occurs or does not occur.

For n repeated events, also called trials, A is replaced by A_k, where the index k tracks the event number such that $1 \leq k \leq n$. Using this notation, we lump $\overline{A}$ and all subsequent $\overline{A}_k$ into the coefficients A_k to write a more general normalization condition:

$$1 = \sum_{k=1}^{n} P\left(A_k\right) \tag{1.12}$$

4.2 Statistical Average

Expanding out the normalization condition above, we have a sequence with n terms on the right:

$$1 = P(A_1) + P(A_2) + \cdots + P(A_n)$$

By multiplying A_k into each respective $P(A_k)$ term, the equation becomes the *statistical average*, or *weighted average* $\langle A \rangle$ of the events A_k. To denote this, we write:

$$\langle A \rangle = A_1 \cdot P(A_1) + A_2 \cdot P(A_2) + \cdots + A_n \cdot P(A_n)$$

In summation notation, the above result reads:

$$\langle A \rangle = \sum_{k=1}^{n} A_k \cdot P(A_k) \tag{1.13}$$

4.3 Expectation Value

A function f that depends on any event A_k can also be averaged using this apparatus. Generalizing the above, we can easily write an equation for the *expectation value* of f:

$$\langle f \rangle = \sum_{k=1}^{n} f(A_k) \cdot P(A_k) \tag{1.14}$$

With the above, we may also calculate $\left\langle f^2 \right\rangle$ without hesitation:

$$\left\langle f^2 \right\rangle = \sum_{k=1}^{n} (f(A_k))^2 \cdot P(A_k) \tag{1.15}$$

<u>Problem 22</u>

A six-sided dice that chooses a random number $1 \leq A_k \leq 6$ is tossed in succession to produce $n \gg 1$ events. Calculate the average outcome. Answer:

$$\langle A \rangle = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{21}{6} = 3.5$$

<u>Problem 23</u>

A six-sided dice that is missing the 2-face but has an extra 4-face is tossed in succession to produce $n \gg 1$ events. Calculate the average outcome. Answer:

$$\langle A \rangle = \frac{1}{6} + \frac{0}{6} + \frac{3}{6} + \frac{4 \cdot 2}{6} + \frac{5}{6} + \frac{6}{6} = \frac{23}{6} = 3.83$$

4.4 Standard Deviation

Further insight into f can be gained by inserting $(f(A_k) - \langle f \rangle)^2$ as the argument in Equation (1.15). By doing so, and then taking the square root of the entire result, we arrive at the *standard deviation* in the system:

$$\sigma_f = \sqrt{\sum_{k=1}^{n} (f(A_k) - \langle f \rangle)^2 \cdot P(A_k)} \tag{1.16}$$

Using only the definitions above, it's easy to show that the standard deviation is equivalent to

$$\begin{aligned} \sigma_f &= \sqrt{\langle f^2 \rangle - 2\langle f \rangle \langle f \rangle + \langle f \rangle^2} \\ \sigma_f &= \sqrt{\langle f^2 \rangle - \langle f \rangle^2}\,. \end{aligned} \tag{1.17}$$

4.5 Mean, Median, Mode

Mean

For a given set of numerical data, a synonym for the statistical average is the *mean*.

Median

The *median* in a sorted numerical set is the number directly in the 'middle' of the list. If the set has an odd number of members, simply pick the middle number. If the set has an even number of members, the median is the average of the two numbers competing for the middle slot. Note that the median is not the same as the midpoint value.

Mode

The *mode* is the member or set of members that occurs most often in a given set. For example, the set

$$\{1, 2, 2, 2, 3, 3, 4, 4, 4\}$$

has two modes, namely 2 and 4. On the other hand, the set

$$\{2, 4, 6, 8, 10\}$$

has no mode.

<u>Problem 24</u>

Suppose:

- x is the average of m and 9
- y is the average of $2m$ and 15
- z is the average of $3m$ and 18

What is the average of x, y, z in terms of m? Answer: $m + 7$

Index