NCEA Level 2 Mathematics
22. Probability and Risk

Basic probability

Suppose we are performing an experiment where the outcome is uncertain. Some examples of such experiments include:-

- Flipping a coin.
- Measuring rainfall.
- Determining the sex of a newborn child.
- Checking the outcome of a sporting event.

The set of possible outcomes of an experiment is called the *sample space*. The sample space of the first experiment above, that of flipping a coin, is

$$S_{\text{coin flip}} = \{H, T\}.$$

The *probability* of a given outcome is the proportion of times that we would expect an experiment to give a particular outcome if we run the experiment multiple times. We write $P(\text{outcome})$ for the probability of a particular outcome.

For example, $P(H) = 0.5$ because, if we were to flip a coin many times, we would expect half the tosses to result in an outcome of H.

A probability must lie between 0 and 1 inclusive (because it makes no sense for an outcome to happen in 200% of the experimental runs).

There are a couple of ways to determine the probability of something. We either find a probability experimentally, by doing an experiment a bunch of times, or we do it theoretically, by counting outcomes.

- To determine a probability experimentally — based on empirical evidence:
$$P(A) = \frac{\text{number of times } A \text{ happened}}{\text{number of times the experiment ran}}.$$
- If there are n ways for an outcome A to occur, and each way is equally likely, then
$$P(A) = \frac{\text{number of ways } A \text{ could happen}}{\text{number of ways any outcome could happen}}.$$

We have some rules to calculate with probabilities. If A and B are two possible outcomes, I will write $P(A \text{ or } B)$ for the probability that A *or* B happens,* and $P(A \text{ and } B)$ for the probability that A *and* B happens.

If the total sample space consists of n outcomes, $S = \{A_1, A_2, ..., A_n\}$, then

$$P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_n) = 1.$$

In other words, there is a probability of 1 — absolute certainty — that there will be an outcome. Either heads or tails must be flipped.†

Two outcomes A and B are said to be *mutually exclusive* if they can never happen at the same time; in other words, if $P(A \text{ and } B) = 0$. To take the example of flipping a coin, heads and tails are mutually exclusive outcomes.

---

*By 'A or B happens', I mean that one of A, B, or both A and B happens.

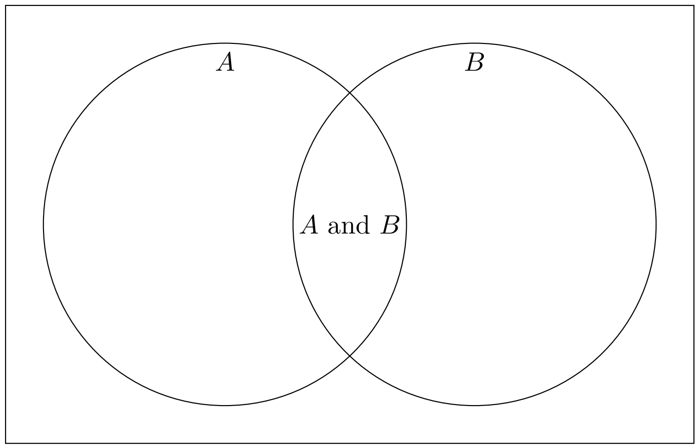
†Yes, it is possible for a coin to land on its rim. I am assuming an infinitely thin coin, or maybe a counter.

If two events A and B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$: the number of times that A or B happens is the sum of the number of times that A or B happens.

Suppose A is an outcome. Then 'not A' is an outcome as well. Further, A and not A are mutually exclusive; and they make up an entire sample space for our experiment because if we do the experiment, then either A is the outcome or it isn't. Hence $P(A) + P(\text{not } A) = 1$, and thus $P(\text{not } A) = 1 - P(A)$. To save ~~chalk~~ typing, I will write $\neg A$ for 'not A'.

Now, suppose two events A and B are *not* mutually exclusive: so it is possible for A and B to be an outcome of an experiment. We want to show that $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

Proof 1: via Venn diagram. If we draw a Venn diagram showing the $P(A \text{ and } B)$, we end up with:



So we see that $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. □

Proof 2: algebraic. First, see that A or B = A or (B and $\neg A$), where these events are mutually exclusive. (Either A happens — which could include A and B happening — or B happens but not A). We can apply the rule from above: $P(A \text{ or } B) = P(A) + P(B \text{ and } \neg A)$.

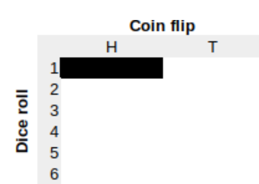
Now, note that $B = (B \text{ and } A)$ or $(B \text{ and } \neg A)$ (B happens whenever B and A both happen, or when B happens without A), where again the events are mutually exclusive. Hence $P(B) = P(A \text{ and } B) + P(B \text{ and } \neg A)$. So we have both of the following:

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

From the second, $P(B \text{ and } \neg A) = P(B) - P(A \text{ and } B)$; and substituting into the first, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. □

Two outcomes A and B are called *independent* if the the occurrence or nonoccurrence of one does not affect the likelihood of occurrence or nonoccurence of the other. For example, the events 'it will rain tomorrow in Brazil' and 'the sun will rise at 5:30 am in New Zealand' are independent.[‡]

Suppose I flip a coin and roll a die. By drawing a table (see below) and counting the outcomes, I find that there is a 1 in 12 chance ($P = \frac{1}{12}$) that I flip a heads and roll a 6.



[‡]One hopes. It is possible to imagine scenarios where they turn out not to be independent — for example, suppose a crazed supervillain from Dunedin decides he will blow up the country if it rains in Brazil tomorrow.

Further experimentation should convince you that, if A and B are independent, then $P(A \text{ and } B) = P(A)P(B)$.

If two events are *not* independent, we can't apply this rule. We can, however, discuss the probability that A occurs given that we know B has happened; we write $P(A|B)$, read *the probability of A given B*, for this.

Formally, what we want to know is the proportion of times that A happens, out of all the times that B happens; writing this down, we have

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

Instead of our sample space being the entire set of possibilities, we restrict our sample space to just being the set of times B has occurred, and then we calculate this probability in the usual way.

It is often easier to work out relative probabilities using either a table or a tree diagram.

Example (NZQA, 2012)**.** BigGen power company keeps track of complaints made to the company by its customers.

- 12% of customers have made a complaint in the last 12 months.
- If a customer made a complaint, there was a 0.7 chance that they left BigGen.
- The percentage of customers leaving BigGen over the last 12 months was 10%.

Calculate the following:

1. The proportion of customers that complained and left BigGen.
2. The number of customers that would be expected to complain, but stay with BigGen, given that there were 250,000 customers in total at the start of the year.
3. The probability that a customer who didn't complain left BigGen.
4. If a customer left BigGen, the probability that they complained.

Solution. Consider (1) first. The probability that a customer complained and left can be calculated directly, since

$$P(\text{complained and left}) = P(\text{left}|\text{complained}) \cdot P(\text{complained}) = 0.7 \cdot 0.12 = 0.084.$$

On the other hand, consider the tree diagram drawn below; the probability $P(\text{complained and left})$ can be calculated from it by just multiplying down the branch.

For (2), we need the probability $P(\text{complained and not left})$; we can fill in the probability $P(\text{not left}|\text{complained})$, as it will be $1 - P(\text{left}|\text{complained}) = 1 - 0.7 = 0.3$. Then $P(\text{complained and not left}) = 0.3 \cdot 0.12 = 0.036$.

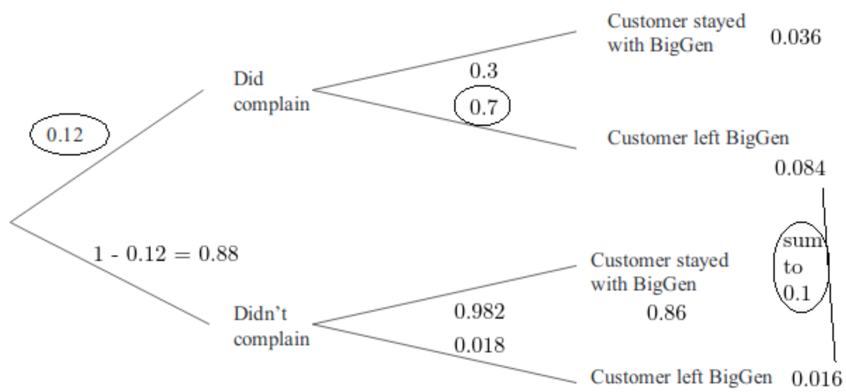
For (3), we need $P(\text{left}|\text{not complained})$. We know that $P(\text{left}) = 0.1$ (it was given); and

$$P(\text{left and not complained}) + P(\text{left and complained}) = P(\text{left})$$

so $P(\text{left and not complained}) = 0.1 - 0.084 = 0.016$; then

$$P(\text{left}|\text{not complained}) = 0.016/P(\text{not complained}) = 0.016/0.88 = 0.018.$$

Finally, for (4) we want $P(\text{complained}|\text{left})$. This is given by $P(\text{complained and left})/P(\text{left})$; filling this in, $P(\text{complained}|\text{left}) = 0.084/0.1 = 0.84$.



All this information is summarised in the tree diagram above; the circled pieces of information are the ones given, and they are sufficient to reproduce the rest. Cover up the example above, and attempt to reproduce the tree diagram yourself using only the information given and the probability laws.

Note that $P(A|B) \neq P(B|A)$ in general: the probability that someone complained given that they left was 0.84, while the probability that someone left given that they complained was 0.7.

Relative risk

Suppose we are conducting a medical trial, and we want to examine the differences in outcomes between two groups. For example, suppose we survey 1000 people and we split them up into two groups according to whether they smoke. We find that 200 people are smokers, and 800 are non-smokers. We then track them for the rest of their lives, and find that twenty of the smokers developed lung cancer, and four of the non-smokers did.

	Smoker	**Non-smoker**	**Total**
Cancer	20	4	24
Non-cancer	180	796	976
Total	200	800	1000

We can't directly compare the actual numbers of cancer sufferers, 4 versus 20, to show that lung cancer and smoking are linked because the sizes of the cohorts are different: it makes no sense to say that smokers are five times (20/4) more likely to get cancer on the basis of these figures. To solve this problem, we want to calculate the risks (probabilities) of lung cancer in the presence of smoking versus non-smoking, and then compare these.

We learn that, given a person is a smoker, they have a $20/200 = 0.1$ probability of developing lung cancer. On the other hand, if a person does not smoke, they have a $4/800 = 0.005$ probability of developing lung cancer. This means that the *relative risk* of lung cancer in the presence of smoking is $0.1/0.005 = 20$: smokers are twenty times more likely to get lung cancer than non-smokers.

The 2017 NCEA L2 probability exam cites another example; the following table shows data from a sample of 2500 New Zealanders aged between 15 and 24:

	Obese	**Non-obese**	**Total**
Male	222	983	1205
Female	285	1010	1295
Total	507	1993	2500

The risk of being obese, given that an individual in the sample is male, is $222/1205 = 0.184$; the risk of being obese given that an individual is female is $285/1295 = 0.220$. Thus the relative risk is $0.220/0.184 = 1.196$: if an individual from the sample is female, they are almost 20% more likely to be obese than a male from the sample.

Reports like this often generate sensational headlines in the media, and so it is important to understand how they are calculated and the assumptions made.

Questions

Basic probability

1. A newspaper reports the probability that a certain airline's flights arrive on time.
   (a) Only one of these numbers could represent this probability; which one, and why?

   $$-6.86, -0.686, 0.686, \text{or } 6.86$$

   (b) Was the probability based on equally likely outcomes, a long-run set of observed outcomes, or subjective assessment?

2. Suppose a fair 12-sided die is rolled. What is the probability that:
   (a) An even number is rolled.
   (b) A number divisible by three is rolled.
   (c) A number greater than (but not equal to) seven is rolled.
   (d) The number 12 is rolled.
   (e) The number 14 is rolled.
   (f) On two consecutive (independent) rolls, a six and then a five are rolled.
   (g) On two consecutive rolls, a five is not rolled either time.

3. A factory makes three models of car, and each car can be one of three colours. The following table provides some information about the cars manufactured in a particular week.

	Silly Sedan	**Horrible Hatchback**	**Vast Van**	**Total**
Yellow	7			23
Black		16		34
Green	3	8	2	13
Total	20		14	

   (a) Complete the table.
   (b) Suppose a random car is chosen for inspection. What is the probability that the car is either a yellow car, or a sedan, or both?

4. In the ANA online Health and Safety Survey of 2001, several thousand nurses reported the usual length of the shift they worked at their main job. What is the probability that one of these nurses worked a 12-hour shift?

Less than 8 hours	5.0%
8 hours	47.0%
10 hours	20.0%
12 hours	?%
More than 12 hours	22.0%

5. A statistics teacher sets up a two-way table of counts for all 2,000 students in her history of teaching, keeping track of gender and whether or not a student received an E. The table shows that 1,200 were female, and that 500 out of all the students received an E. Overall, 300 female students received an E. Complete the table. Suppose a student were male; what is the chance that they did not receive an E?

	E	not E	Total
Female			
Male			
Total			

6. (NCEA 2017) A survey of young adults produced the following results:

	Obese	**Not Obese**	**Total**
Current smoker	103	317	420
Non-smoker	404	1676	2080
Total	507	1993	2500

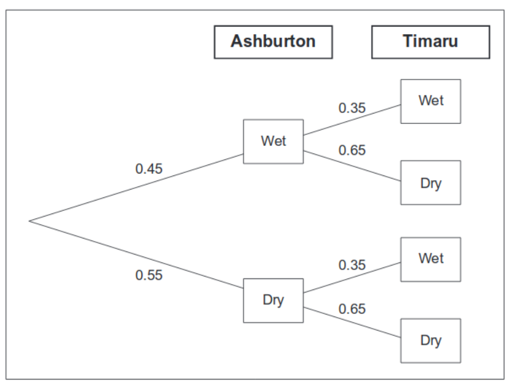
Would it be correct to claim that young adult smokers are more at risk of being obese than young adult non-smokers? (In other words, is the probability that a young adult is obese given that they are a smoker greater than the probability of obesity given that they are not a smoker?)

7. (NCEA 2018)
   (a) Nancy finds some data from NIWA on weather in Ashburton and in Timaru over the past seven years. She analyses the data and finds that:
   - It was wet on 45% of days in Ashburton.
   - If it was wet in Ashburton, the probability that it was wet on the same day in Timaru was 63%.
   - If it was dry in Ashburton, the probability that it was dry on the same day in Timaru was 88%.

   i. Find the probability that it was dry in both Ashburton and Timaru on a randomly chosen day.
   ii. Find the probability that only one of the towns was wet on a given day.
   iii. If it was a dry day in Timaru, what is the probability that it was also dry in Ashburton on the same day?

   (b) Nancy's friend Teri uses the NIWA data for the past seven years to find out that:
   - 45% of days in Ashburton were wet
   - 35% of days in Timaru were wet.

   Teri constructs the tree diagram below.



   Elain why Teri's tree diagram would not give a correct answer to the probability that it is dry in both towns on the same day. (Note: the distance between Ashburton and Timaru is around 75 km. Both towns lie on the Canterbury plains.)

Relative risk

8. A study is conducted of 1500 randomly selected candidates for an international examination to investigate whether Y12 candidates are more successful than Y13 candidates. The findings are summarised in the table below.

	Y12	**Y13**	**Total**
Pass	347	853	1200
Fail	33	267	300
Total	380	1120	1500

(a) What proportion of candidates passed the exam?

(b) What proportion of candidates who failed were in year 12?

(c) There were about 52 500 candidates from year 12 and year 13 who attempted the exam. Based on the results of the study, how many candidates would be expected to be in year 13 and pass the exam?

(d) It is claimed that year 13 candidates are four more times more likely to fail the exam than year 12 candidates. Discuss your agreement or disagreement with this statement.

9. Polygraph (lie-detector) tests are often routinely administered to government employees or prospective employees in sensitive positions. A study performed in the US in 2002 found that lie detector results are "better than chance, but well below perfection". Typically, the test will conclude someone is a spy 80% of the time when he or she actually is a spy; but 16% of the time, the test will conclude someone is a spy when he or she is not.

   (a) Assuming that 10 out of every 10 000 employees are actual spies, compute the following:
      i. The probability that an employee is a spy and is "detected" to be one.
      ii. The probability that an employee is **not** a spy and is "detected" to be one.
      iii. The overall probability that the polygraph "detects" that an employee is a spy.

   (b) On the basis of the given information, and the probabilities that you have calculated, is the use of the polygraph "worth it"?

10. A survey was conducted in 2000, asking respondents about various driving habits. This table classifies 1 086 participants according to type of car driven, and whether or not they were in the habit of making insulting gestures at other drivers.

	Economy	**Family**	**Luxury**	**Sports**	**Truck**	**Utility**	**Van**	**Total**
Gestures	79	65	16	58	42	32	8	300
No Gestures	281	170	45	95	77	79	39	786
Total	360	235	61	153	119	111	47	1086

   (a) Before calculating any conditional probabilities, identify the types of cars whose owners you would suspect to have a tendancy to make insulting gestures at other drivers.

   (b) For each type of car, find the (conditional) probability that surveyed drivers of that type of car make insulting gestures (e.g. given that a person drives a truck, what is the probability that they make insulting gestures).

   (c) Comment on whether your suspicions in part (a) were correct.

   (d) To three decimal places each, find (using the table):
      i. The probability of driving a van ($P(V)$).
      ii. The probability of making insulting gestures ($P(G)$).
      iii. The probability of driving a van *and* making insulting gestures ($P(V \text{ and } G)$).

   (e) Check if $P(V \text{ and } G) = P(V) \times P(G)$ to see if the events V and G are independent. Explain the outcome.

   (f) You found $P(G|V)$ in part (b). Check if $P(V \text{ and } G) = P(V) \times P(G|V)$, and explain the outcome.

   (g) Find the overall probability of driving an economy car, a family car, a luxury car, or a van.

   (h) Find the probability of making insulting gestures, given that someone drives an economy car, a family car, a luxury car, or a van.

   (i) Are drivers of economy cars, family cars, luxury cars, or vans less likely in general to make insulting gestures?