Chapter 3.2 - Probability Set Theory

> "Later mathematicians will regard set theory as a disease from which one has recovered." - Henri Poincaré

We don't really discuss set theory properly in statistics. The majority of the science doesn't rely on it for more than a few concepts. That said those concepts are difficult to introduce *without* discussing sets, and if we can't make sense of them we're doomed to struggle with **every other** concept in this book.

> Imagine you're a quantitative researcher working for a national park. An outbreak of Chronic Wasting Disease (CWD), a prion disease that typically affects deer, has been identified on the park and you've been asked to help the park identify possible causes. The data you're given is small but very comprehensive, but you've decided to focus on the possibility that the disease is being spread by ticks. You've constructed a list of deer harvested from the same general location in the park and whether or not ticks were found on their body post-mortem.

	Ticks	**No Ticks**
CWD +	42	18
CWD −	78	62

One of the goals of statistics is to take small samples like this one and extrapolate them to a larger population. In this case we can think of the sample as a representation of *chance outcomes*, like a set of complex coin flips. Given this we should be able to ask questions such as: What is the *probability* that any given deer is going to end up CWD positive?

As with most of statistics, the answer to this question is found in a model.

Probability model: A mathematical function that assigns a probability to each possible event constructed from the simple events in a particular sample space describing a particular experiment.

For a finite sample space with n simple events (denoted E), i.e. $S = \{E_1, E_2, \ldots, E_n\}$, the **probability model** assigns a number p_i to event E_i where $P(E_i) = p_i$ so that:

$$0 \leq p_i \leq 1 \quad \textbf{and} \quad p_1 + p_2 + \cdots + p_n = 1$$

For an **equally-likely** probability model, the probability of observing E_i is:

$$P(E_i) = p_i = \frac{1}{n}$$

If A is an event in an equally-likely sample space S and contains k outcomes, then:

$$P(A) = \frac{\text{No. of outcomes in } A}{\text{No. of outcomes in } S} = \frac{k}{n}$$

With the CWD data, you run an experiment where you select 1 deer from the group and check if they're disease positive.

	Ticks	**No Ticks**
CWD +	42	18
CWD −	78	62

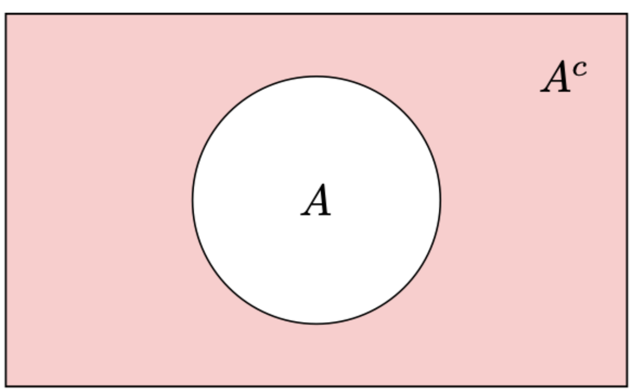
We can refer to the outcome where a deer is positive for CWD as an event, A, and calculate it's probability:

$$P(A) = \frac{60}{200} = 0.3$$

But what about the probability that a deer is *negative* for CWD? If A is an event in S, then the event where A does not occur is called the **complement** of A. We denote the complement of A by A^c – read as "A-complement".

Complements

Consider the sample space, S, to be a square. The event, A, is shown as a circle inside of that square. The complement of A would be everything else in the sample space that **is not** in A.



Since we're discussing probability here, the chance of *anything* occurring in our sample space is 1.00 (or 100%). This is a rule that holds up moving forward: any sample space has a total probability of 1.00.

If the goal is to look at everything in the sample space that **isn't** A, and the sample space has a probability of 1.00, we can define $P(A^c)$ as $P(A^c) = 1 - P(A)$.

This can be flipped around to give a definition for $P(A)$, $P(A) = 1 - P(A^c)$.

Suppose we roll a fair 6-sided die twice, then S contains 36 equally-likely outcomes in the form of 36 ordered pairs, i.e. $(1,1),(1,2),\ldots,(6,5),(6,6)$. Let A be "roll doubles".

$$P(A) = \frac{6}{36} = \frac{1}{6}$$

A^c is the event we "do not roll doubles", and:

$$P(A^c) = 1 - P(A) = 1 - \frac{1}{6} = \frac{5}{6}$$

We could have counted the number of non-doubles in S, but this requires more effort. We're always searching for better ways to be lazy in statistics.

At face value, this rule is useful when $P(A)$ is difficult to calculate but $P(A^c)$ is easy (or vice versa). In later chapters this rule is the foundation for working with several key methods in statistical inference; take the time to understand it.

We'll swap between the dice example and the CWD example to get a good view of how these concepts translate from simple experiments to "real life" scenarios. Returning to the CWD data, what's the probability that a deer will be *negative* for CWD?

	Ticks	**No Ticks**
CWD +	42	18
CWD −	78	62

While we *could* consider a negative status to be it's own event, B, we'll find that $B = A^c$

$$P(B) = \frac{140}{200} = 0.7$$

$$P(A) = \frac{60}{200} = 0.3$$

$$P(A^c) = 1 - 0.3 = 0.7 = P(B)$$

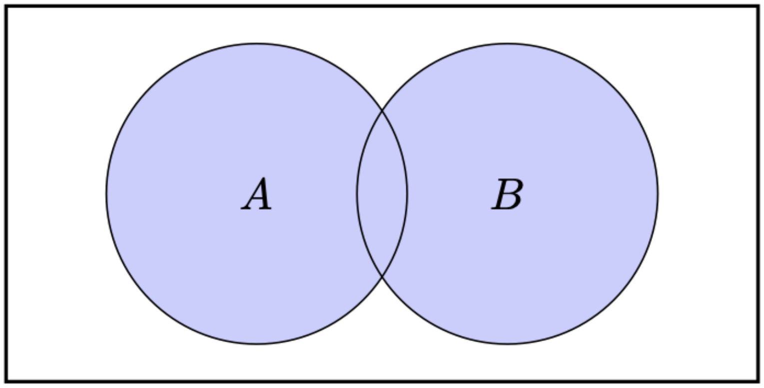
Again, you'll see early on that this method is primarily a convenience but it *will* become vital.

Unions and Intersections

In trying to determine whether or not ticks could be the cause of this outbreak, it might be helpful to look at the deer that had a positive CWD status *or* had ticks. We should be able to recognize at this point that these are two separate events and what we're trying to do is **join them together**. This is referred to as a **union**.

The **union** of two events A and B, denoted $A \cup B$, are all outcomes that belong to A, B, or both. Saying $A \cup B$ is equivalent to saying "**A or B**".

Continuing with our circles example, the events A and B exist as separate event spaces in the total sample space (the box) with a slight overlap. The union of the events is all of the space that each "event circle" (as we'll refer to them now) takes up *including* their overlap.



When we think about *calculating* the union of two events we should think about describing the space in the image above. If we take all of the space in A and add it to the space in B we have an inherent problem of double counting the **overlap** between A and B since they both include that little sliver in their individual areas.

Let A = the deer is CWD + and B = the deer has ticks.

$$P(A) = \frac{60}{200} = 0.3$$

$$P(B) = \frac{120}{200} = 0.6$$

$$P(A) + P(B) = 0.3 + 0.6 = 0.9$$

To say that 90% of the deer were either CWD positive *or* had ticks is ridiculous, and we can prove it with a *complement*. Let's focus on just the CWD negative deer that didn't have ticks and call that event C:

$$P(C) = \frac{62}{200} = 0.31$$

The complement of this event *should* be the union of A and B, since it's the only event that's excluded by the union.

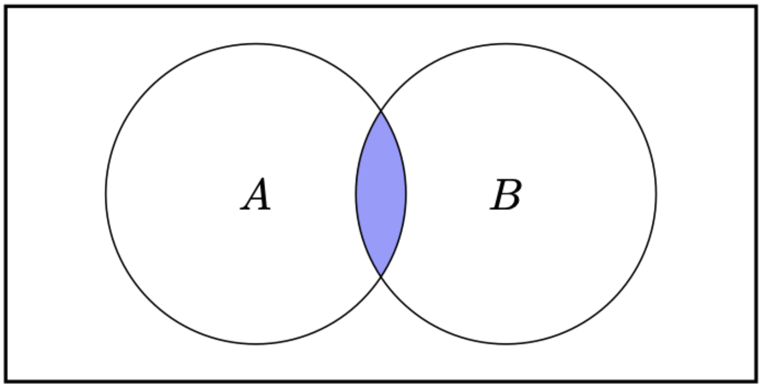
$$P(C^c) = 1 - 0.31 = 0.69 \neq P(A) + P(B)$$

When we look at the difference between this complement and our original calculation and multiply it by the sample size to recover the original event from it's probability, we'll find something familiar:

$$0.9 - 0.69 = 0.21 \times 200 = 42$$

C^c just so happens to be the deer that were CWD positive *and* had ticks. This complement is actually referred to as an **intersection**.

The **intersection** of two events A and B, denoted $A \cap B$, are all outcomes that belong to both A and B. Saying $A \cap B$ is equivalent to saying "**A and B**".



In rolling a 6 sided dice one time, consider events A and B:

- A: Roll an even number: $\{2, 4, 6\}$
- B: Roll a number greater than 4: $\{5, 6\}$

$$A \cup B = A \text{ or } B = \{2, 4, 5, 6\}$$

$$A \cap B = A \text{ and } B = \{6\}$$

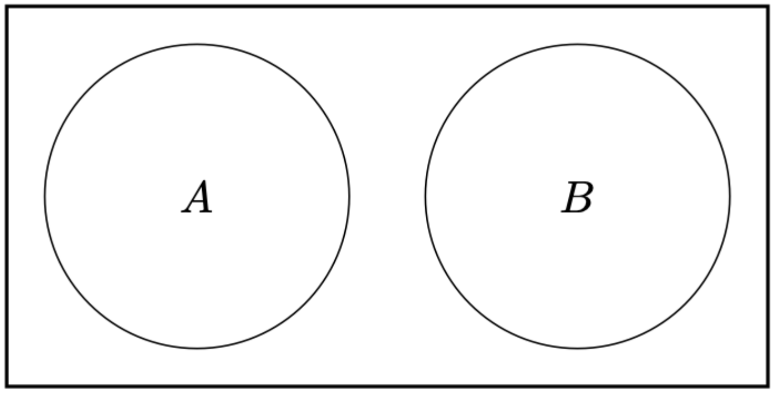
So to solve the problem of double counting when calculating a union we can simply *subtract* the intersection between A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = 0.3 + 0.6 - 0.21 = 0.69$$

Mutual Exclusivity

We can still describe the union of two events even if they lack any overlap. It should be clear though that we can't describe the intersection of two events that lack an overlap. We refer to this idea as **mutual exclusivity**.



Two events A and B are **mutually exclusive** if they do not share any common outcomes.

Roll a die:

- A: Roll a 1 or a 2: $\{1, 2\}$
- B: Roll an even number: $\{2, 4, 6\}$
- C: Roll a 3, 4, or 5: $\{3, 4, 5\}$

Events A and C are mutually exclusive. If a 1 was rolled then none of the events in C could have occurred. Vice versa, if any events in C occur we can confidently say that A did not occur. Conversely, events A and B are not mutually exclusive. They share the common outcome 2 as a possibility.

When we originally calculated the union of two events we had to adjust for double counting of their overlap in probability. A nice property of mutual exclusivity is that we don't have that overlap, so we can skip this step entirely. When we account for two mutually exclusive events, A and B, having an intersection with probability 0 we end up at the **addition rule** for mutually exclusive events:

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

While it may seem ridiculous the example in this case would be calculating the union between CWD positive status and CWD negative status. A deer cannot be both, this is a biologic impossibility. We can denote the event for CWD positive as P and negative as N.

$$P(P \cup N) = P(P) + P(N) + P(P \cap N)$$

$$P(P \cap N) = 0$$

$$P(P \cup N) = \frac{60}{200} + \frac{140}{200} + 0 = \frac{200}{200} = 1.00$$

Despite this being an over-engineered calculation of sample size, we can use this along with our previous calculations to make our table of data much more convenient.

	Ticks	**No Ticks**	**Total**
CWD +	42	18	60
CWD −	78	62	140
Total	120	80	200

This table is now something referred to as a **contingency table** and it allows us to circumvent a lot of the math we've been previously working through:

Event	B	B^c	**Total**
A	$A \cap B$	$A \cap B^c$	A
A^c	$A^c \cap B$	$A^c \cap B^c$	A^c
Total	B	B^c	S

Conditional Probability

If ticks are a relevant explanation for CWD status, we should be able to determine the status of a deer just by observing whether or not is has ticks. We might ask the question "What are the chances that a deer is CWD positive given that we observed that it has ticks?". To do this we would have to confine our view of the table to only those deer that **have** ticks:

	Ticks
CWD +	42
CWD −	78
Total	120

From here we would just have to treat the total number of deer with ticks to be our new sample space and calculate the proportion that were CWD positive.

$$\frac{\text{CWD +}}{\text{Total}} = \frac{42}{120} = 0.35$$

If we think back to our "general" contingency table:

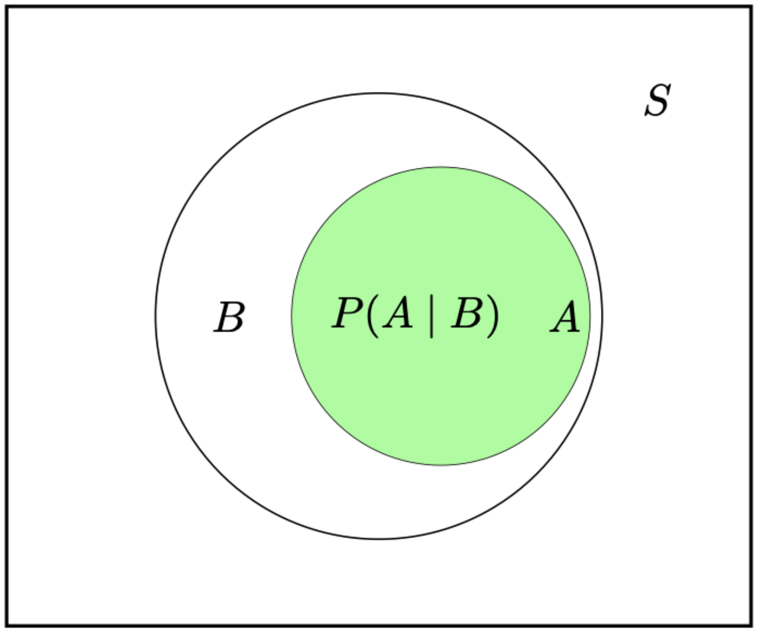
Event	B	B^c	**Total**
A	$A \cap B$	$A \cap B^c$	A
A^c	$A^c \cap B$	$A^c \cap B^c$	A^c
Total	B	B^c	S

The formula for what we've calculated is:

$$\frac{P(A \cap B)}{P(B)}$$

Which is the exact formula for **conditional probability**.

A **conditional probability** of an event is a probability obtained with the **additional information** that some other event has already occurred. In a sense we're *re-scaling* our **sample space** to instead be a specific **event space** and considering all of the possible outcomes of a separate event.



We denote the probability of A conditional on B as $P(A|B)$ which reads as "the probability of A **given** B".

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

As a note: the use of A and B are arbitrary here, the structure of the formula is only based on the order of the events *inside* the conditional.

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

Using this definition of conditional probability, we can apply some basic algebraic knowledge to produce a definition for the intersection of A and B.

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

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

This formula is sometimes referred to as the **multiplication rule** for *intersections*.

Independence

This is one the many instances where independence makes our lives easier. In terms of general probability, events A and B are considered to be independent if the outcome of A does not affect the outcome of B and vice versa.

With regard to conditional probability this means that the following are true:

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

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

We can leverage this to prove an interesting result about independent **intersections**. Given that events A and B are independent:

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

$$P(A) = \frac{P(A \cap B)}{P(B)}$$

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

This is referred to as the **multiplication rule** for *independent events*.