WEEK 2

A quick recap of probability

Felipe Campelo

bristol.ac.uk

Based on Dr. Rihuan Ke's SCEM notes and
Dr. Felipe Campelo's course on Design
and Analysis of Experiments

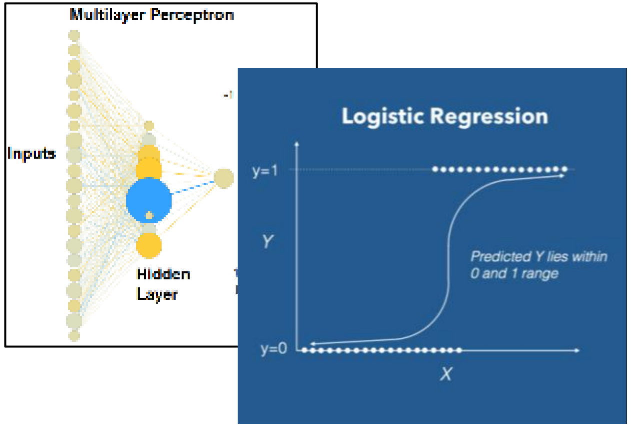
In this lecture...

- We will introduce the fundamental problem of stochastic variability.
- We will introduce some fundamental ideas from probability – event, observation, sample, and statistical population.
- We will discuss the definition of probability, the axioms of probability and some derived properties.
- We will familiarise ourselves with the concepts of independence, conditional probability, and Bayes’ theorem.

The problem of variability

Are Adelie penguins lighter on average than Chinstrap penguins?

Is the effect of a generic drug the same as the reference brand?



For a given problem, does a logistic regression model return a greater expected accuracy than a neural network?

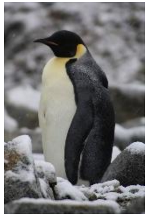
The problem of variability

We attempt to answer such questions by looking at data.

Our data sets are usually ***samples*** from a much larger (statistical) **population**.

A **population** is a large set of objects of a similar nature which are of interest as a whole. It can be an actual set (e.g., all penguins in a specific region) or a hypothetical one (e.g., all possible outcomes of an experiment).

A **sample** is a subset of a population. A sample is chosen to make inferences about the population by examining or measuring the elements in the sample.



An **observation** (or **unit**) is a single element of a given sample, an independent data point. An observation can also be considered as a sample of size one.

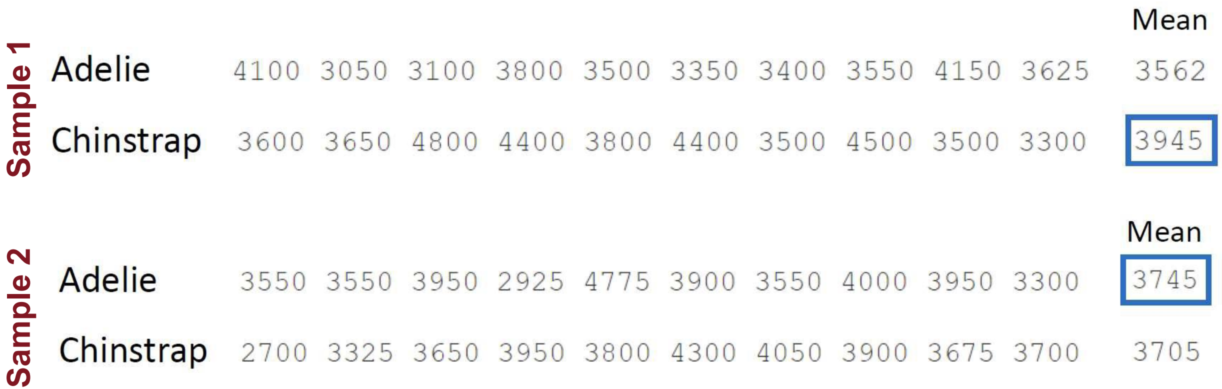
The problem of variability

A common first approach is to compare *sample means* – but these are conditional on the sample and subject to *estimation error.*

If the population has any level of variability in it, then samples will also be inherently variable.

This *stochastic variation* needs to be considered when trying to learn things about a *population* by looking at a *sample*.

For that, we need to introduce ideas from the fields of *probability* and *statistics.*



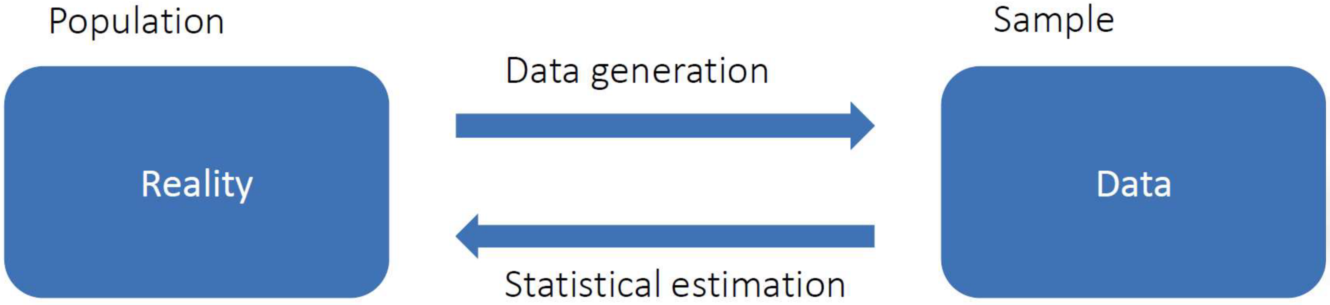
												Mean
Sample 1	Adelie	4100	3050	3100	3800	3500	3350	3400	3550	4150	3625	3562
	Chinstrap	3600	3650	4800	4400	3800	4400	3500	4500	3500	3300	3945

												Mean
Sample 2	Adelie	3550	3550	3950	2925	4775	3900	3550	4000	3950	3300	3745
	Chinstrap	2700	3325	3650	3950	3800	4300	4050	3900	3675	3700	3705

Probability and statistics

Even if we used a larger sample, the problem of variability would remain - we (usually) cannot measure every unit in a population - e.g., weight every penguin in an entire species or test a new medication on all patients, current and future).

We must consider how a *finite sample* reflects a larger (potentially infinite) population of interest. This is the problem of *statistical modelling and inference.*



Population

Sample

Data generation

Reality

Data

Statistical estimation

Sometimes it is also useful to model the data-generating process. This is commonly a problem of *probabilistic modelling*.

Probability and statistics

This course is mostly focused on the usual statistical problems:

1. **Statistical estimation**: How can we design an effective statistic $\hat{\theta}$ to estimate a parameter of interest θ?
2. **Uncertainty quantification**: How can we quantify our uncertainty about the estimated value of θ?
3. **Hypothesis testing**: Hoe can we use estimates $\hat{\theta}$ to test hypotheses about θ?
4. **Prediction**: How can we use our estimates to make effective predictions about new data?

We need some concepts of probability theory to adequately tackle these problems. We'll focus on those for the remainder of this lecture.

Some basic definitions

In basic probability theory, a ***random experiment*** is a procedure (real or imagined) which:

1. Has a well-defined set of possible outcomes;
2. Could (at least in principle) be repeated arbitrarily many times

An **event** is a set (i.e., a collection) of possible outcomes of a given random experiment.

A **sample space** is the set of all possible outcomes of interest for a random experiment

What is probability

We often make statements about the **probability**, **likelihood**, or **chance** of different events.

“*Given how cloudy it is, there’s a high* ***likelihood*** *it will rain.*”

“*There is a good* ***chance*** *that the level of inflation will fall due to the rise in interest rates*”

“Bristol City Football Club will ***probably*** *not win the Football Association Challenge cup this year*”

We need some basic probability theory to make such statements precise so we can reason about them quantitatively.

How to define probability

A formal definition of probability can be built on a few simple rules.

Toy example / intuition **(Experiment: roll of a 6-sided die)**	**Formalisation**
Events (e.g., {1,2}, {4}, …)	Events $E_1, E_2, ...$
Sample space {1,2,3,4,5,6}	Sample space Ω
Probability of {1,2} is 1/3 ≥ 0	**Rule 1**: P(*E*) ≥ 0 for any event *E*
Probability of {1,2,3,4,5,6} is 1	**Rule 2**: P(Ω) = 1 for a sample space Ω
Probability of {1,2,3} = Probability of {1} + Probability of {2,3}	**Rule 3**: For two mutually exclusive events, P($E_1 \cup E_2$) = P(E_1) + P(E_2)

How to define probability

Rules 1, 2 and 3 are know as the **axioms of probability** or the **Kolmogorov axioms.**

Definition of probability

Given a sample space Ω and an event space *F*, a probability *P* is a function which assigns a number *P(E)* to each event *E* ∈ *F* and satisfies the following rules:

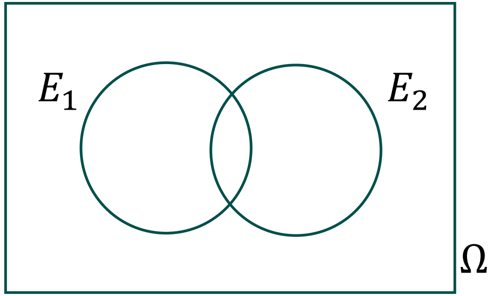
Rule 1: $P(E) \in \mathbb{R}, P(E) \geq 0 \; \forall E \in F$
Rule 2: $P(\Omega) = 1$
Rule 3: For disjoint (i.e., mutually exclusive) events $E_1, E_2, \ldots,$ we have

$$P\left(\bigcup_{\forall i} E_i\right) = \sum_{\forall i} P(E_i)$$

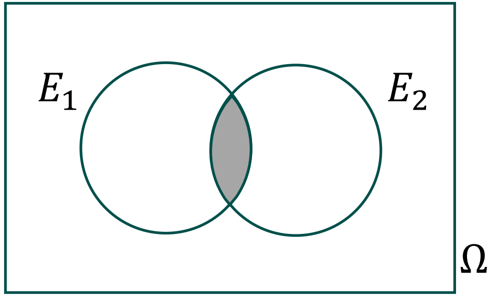
Venn diagrams

We can use Venn diagrams to to represent a sample space and events in it.

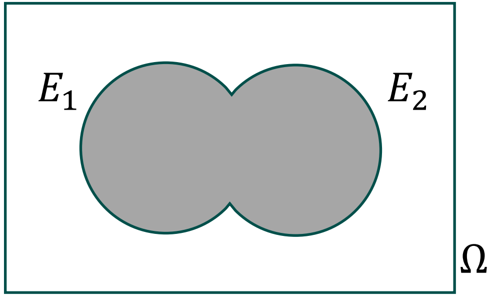
Two events E_1 and E_2



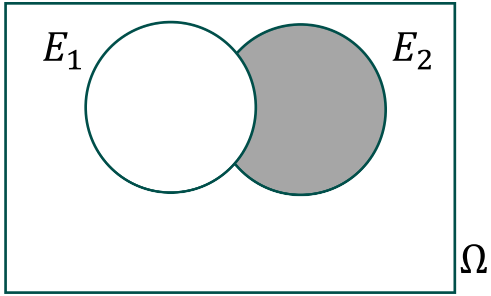
$E_1 \cap E_2$



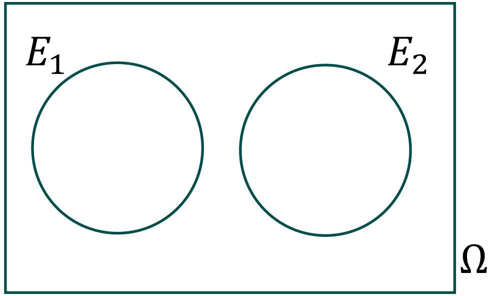
$E_1 \cup E_2$



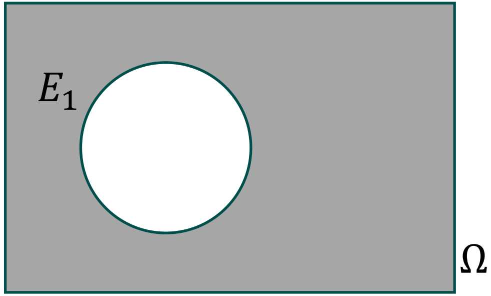
$E_2 \backslash E_1$



Disjoint events $E_1$and E_2



Complement of event E_1



Other properties derived from the 3 rules

Rule 1: $P(E) \geq 0$
Rule 2: $P(\Omega) = 1$
Rule 3: For mutually exclusive events, $P(\bigcup_{\forall i} E_i) = \sum_{\forall i} P(E_i)$

The 3 axioms of probability result in a few additional properties that are of interest:

Probability of the empty set: $P(\emptyset) = 0$

Proof:
$\emptyset \cup \emptyset = \emptyset$

Hence, $P(\emptyset \cup \emptyset) = P(\emptyset)$

$P(\emptyset \cup \emptyset) = P(\emptyset) + P(\emptyset)$ *(rule 3)*

$P(\emptyset) + P(\emptyset) = P(\emptyset)$

Therefore, $P(\emptyset) = 0$

Note: the empty set $\emptyset$ can be thought of as an impossible event in a given sample space, e.g., "rolling a value of 8 in a regular 6-sided die".

Other properties derived from the 3 rules

Rule 1: $P(E) \geq 0$
Rule 2: $P(\Omega) = 1$
Rule 3: For mutually exclusive events, $P(\cup_{\forall i} E_i) = \sum_{\forall i} P(E_i)$

The 3 axioms of probability result in a few additional properties that are of interest:

Monotonicity: If $E_1 \subseteq E_2$, then $P(E_1) \leq P(E_2)$

Proof:

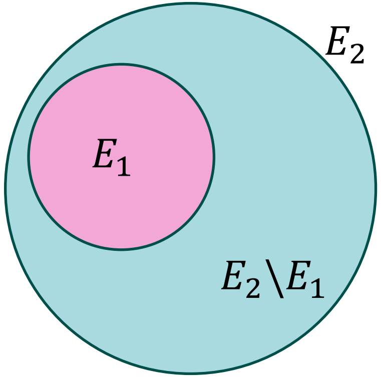
Let $E_1, E_2 \in F$ and $E_1 \subseteq E_2$.

Clearly, E_1 and $E_2 \backslash E_1$ are disjoint. Then:

$$P(E_2) = P(E_1 \cup E_2 \backslash E_1)$$

$$P(E_2) = P(E_1) + P(E_2 \backslash E_1) \quad \textit{(rule 3)}$$

Since $P(E_2 \backslash E_1) \geq 0$, then
$P(E_2) \geq P(E_1) \rightarrow P(E_1) \leq P(E_2)$.



Example: E1 could be the event “student in this SCEM class is 24 years old”, and E2 the event “student in this class is between 20-25 years old.

Other properties derived from the 3 rules

Rule 1: $P(E) \geq 0$
Rule 2: $P(\Omega) = 1$
Rule 3: For mutually exclusive events, $P(\bigcup_{\forall i} E_i) = \sum_{\forall i} P(E_i)$

The 3 axioms of probability result in a few additional properties that are of interest:

Numeric bound: For any event $E \in F$, $0 \leq P(E) \leq 1$

Proof:

$P(E) \geq 0$ *(rule 1)*

Since $E \subseteq \Omega$, then $P(E) \leq P(\Omega)$, *(previous result)*

therefore, $P(E) \leq 1$

Other properties derived from the 3 rules

Rule 1: $P(E) \geq 0$
Rule 2: $P(\Omega) = 1$
Rule 3: For mutually exclusive events, $P(\bigcup_{\forall i} E_i) = \sum_{\forall i} P(E_i)$

The 3 axioms of probability result in a few additional properties that are of interest:

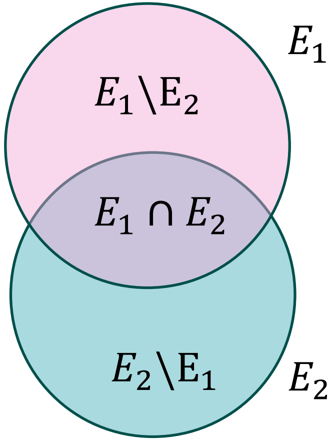
Union bound (Boole's inequality):
For any series of events $E_1, E_2 \ldots \in F$, $P(\bigcup_{\forall i} E_i) \leq \sum_{\forall i} P(E_i)$

Proof: For 2 events E_1 and E_2

$P(E_1 \cup E_2) = P(E_1) + P(E_2 \backslash (E_1 \cap E_2))$ *(rule 3)*
$P(E_2) = P(E_2 \backslash (E_1 \cap E_2)) + P(E_1 \cap E_2)$
$P(E_2 \backslash (E_1 \cap E_2)) = P(E_2) - P(E_1 \cap E_2)$

So,
$P(E_1 \cup E_2) = P(E_1) + P(E_2 \backslash (E_1 \cap E_2)) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

Since $P(E_1 \cap E_2) \geq 0$, then $P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$.
(rule 1)



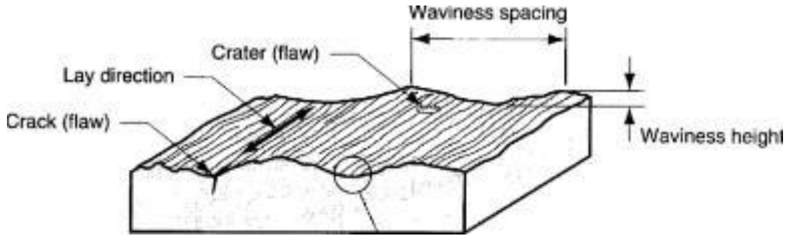
(easy to generalise for n events by recursion)

Interlude

The properties described in the previous slides are quite useful as they will help us understand concepts such as probability mass / density functions, joint probabilities, and Bayes' Theorem – which are important both for statistical modelling and inference and for machine learning more broadly.

Before we continue, let's catch our breath for a few seconds.

Conditional probability



Suppose the following scenario:

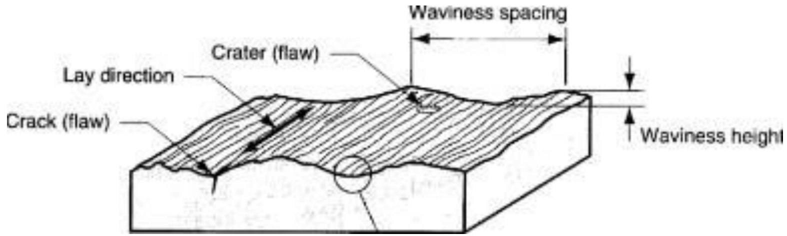
In a manufacturing process, 10% of the parts contain visible surface flaws and 25% of the parts with surface flaws are functionally defective parts. However, only 5% of parts without surface flaws are defective.

The probability of a defective part depends on our knowledge of the presence or absence of a surface flaw.

This example illustrates that probabilities (sometimes) need to be reevaluated as additional information becomes available.

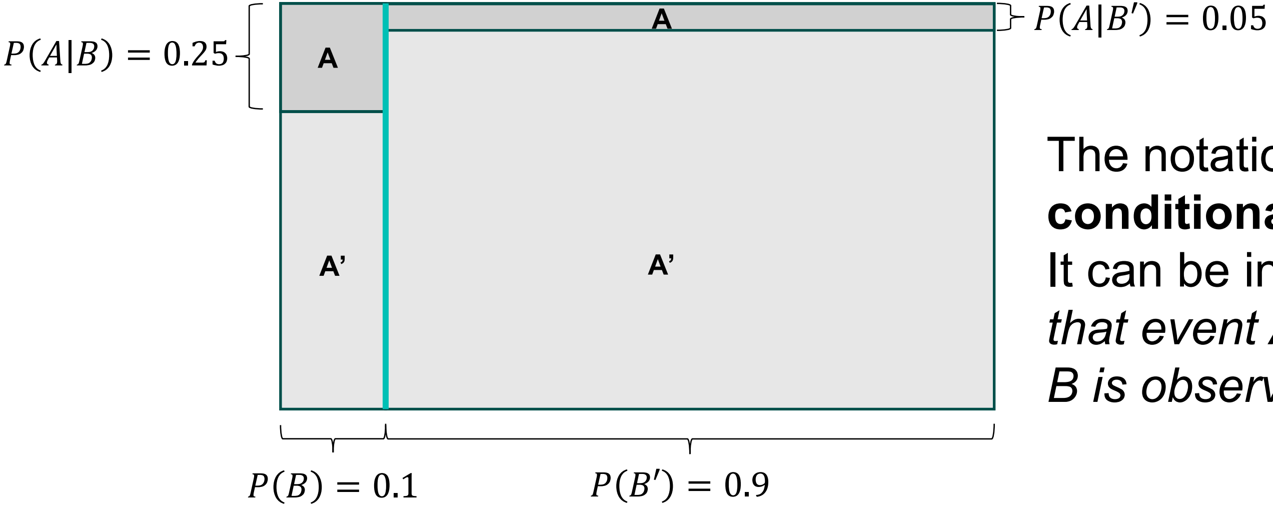
Image: https://tinyurl.com/5ptkcuac

Conditional probability



Let *A* denote the event that a part is defective, and let *B* denote the event that a part has a surface flaw.

Then, we write the probability of *A* given (or assuming) that a part has a surface flaw as $P(A|B)$.



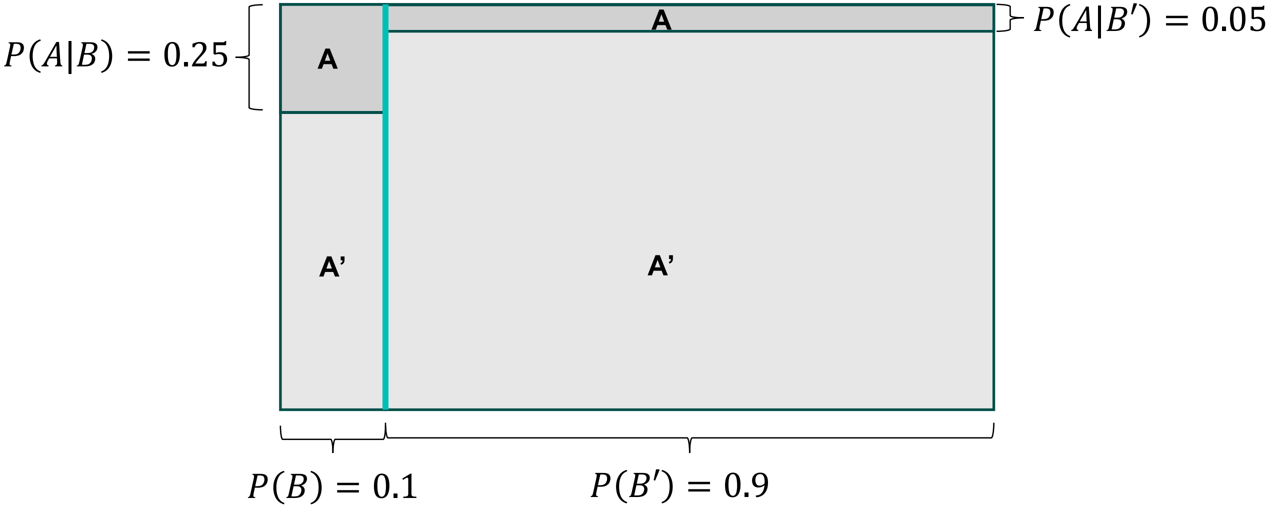
The notation $P(A|B)$ is read as the **conditional probability** of *A* given *B.* It can be interpreted as the *probability that event A happens, given that event B is observed*.

Image: https://tinyurl.com/5ptkcuac

Conditional probability

If we want to define conditional probabilities using the the set notation, we can write that the **conditional probability** of an event *A* given an event *B*, denoted as $P(A|B)$, can be calculated as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ for } P(B) > 0$$



(Note: remember that union and intersection are commutative, i.e., $A \cap B = B \cap A$ and $A \cup B = B \cup A$)

Multiplication rule

From the definition of conditional probabilities, we have that:

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

If we isolate the $P(A \cap B)$ in each equation, we can write:

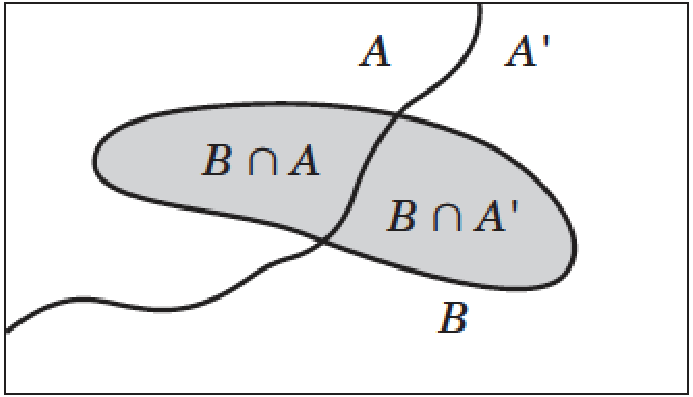
$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

This is known as the *multiplication rule.*

Total probability rule

The multiplication rule is useful for determining the probability of an event that depends on other events.

We can use the multiplication rule and the Rule 3 of probability (union of mutually exclusive events) to derive the *total probability rule*:



A and *A′* are mutually exclusive.

$A \cap B$ and $A' \cap B$ are mutually exclusive

$B \ = \ (A \cap B) \cup \ (A' \cap B)$

So, for two events A and B, we can write:

$$\begin{aligned} P(B) &= P(A \cap B) + P(A' \cap B) \\ &= P(B|A)P(A) + P(B|A')P(A') \end{aligned}$$

Image: 2019 John Wiley & Sons

Independence

In some cases, the conditional probability $P(A|B)$ might be equal to $P(A)$. This happens when knowledge of the occurrence of *A* does not affect the probability that the outcome is *B*. If this is the case, we say that events A and B are ***independent***.

Notice that if $P(A|B) = P(A)$, then:

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

That is, the *joint probability* of two independent events A and B is given as the product of their *marginal probabilities* $P(A)$ and $P(B)$.

(Note: the same is true for multiple independent events A, B, C…)

Bayes' Theorem

One of the most important results in probability theory (and in statistical thinking) is known as Bayes' theorem. It can be derived directly from the total probability rule:

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

It follows that:

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

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

This is a useful result that enables us to solve for $P(A|B)$ from $P(B|A)$.

Bayes' Theorem - example

Suppose that medical study screens everyone in a city for a certain disease.

Assume that:

- The test has a *sensitivity* of 0.99: $P(test = + \mid disease = TRUE) = 0.99$
- The test has a *specificity* of 0.95: $P(test = - \mid disease = FALSE) = 0.95$
- The incidence of the disease in the general population is 0.0001: $\mathrm{P}(disease = TRUE) = 0.0001$

  If someone gets a positive test, what is the probability that they have the illness? That is, what is $P(disease = TRUE \mid test = +)$?

Note: Total probability rule:

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

Bayes’ Theorem - example

From Bayes’ Theorem, we have:

$$P(disease = TRUE \mid test = +) = \frac{P(test = + \mid disease = TRUE)P(disease = TRUE)}{P(test = \ +)}$$

$$= \frac{0.99 \times 0.0001}{P(test = \ +)} = \frac{0.000099}{P(test = \ +)}$$

We can calculate $P(test = \ +)$ from the total probability rule:

$$P(test = \ +) = P(test = + \mid disease = TRUE)P(disease = TRUE) + \cdots$$
$$\ldots + \ P(test = + \mid disease = FALSE)P(disease = FALSE)$$

Bayes' Theorem - example

Since the test only has 2 possible results (+ or -) the probability of a *false positive* is

$$P(test = + \mid disease = FALSE) = 1 - P(test = - \mid disease = FALSE) = 0.05$$

Similarly, for the *baseline probability* of the person not having the disease,

$$P(disease = FALSE) = 1 - \mathrm{P}(disease = TRUE) = 0.9999$$

Which leads to:

$$P(test = +) = 0.99 \times 0.0001 + 0.05 \times 0.9999 = 0.050094$$

and finally:

$$P(disease = TRUE \mid test = +) = \frac{0.000099}{0.050094} = 0.00198$$

Bayes' Theorem - example

$$P(disease = TRUE) = 0.0001$$

$$P(disease = TRUE\ |test = \ +) = 0.00198$$

Some interesting observations:

- Note that even receiving a positive test, the probability of the person really having the disease remains quite low (about 0.2%) - although the probability was **updated**: it is almost **20 times greater** than in the general population.

- This is a common problem when you have *low-incidence events.* It will show up again when we discuss the learning problem of *classification.*

As a side note, this is also one of the reasons why doctors usually don't do indiscriminate testing, but tend to prescribe testing in the context of other evidence (symptoms, age, family history etc.)

In this lecture we discussed…

- The fundamental problem of stochastic variability.
- Some fundamental ideas from probability – event, observation, sample, and statistical population.
- The definition of probability, the axioms of probability and some derived properties.
- The concepts of independence, conditional probability, and Bayes' theorem.

Further reading

DC Montgomery, GC Runger, *Applied statistics and probability for engineers, 5th ed.*
[Chapters 2.1 – 2.7]

- *Read the chapter and try to solve the worked examples by yourself*.