3.1 The Meaning of Probability

Case study (genetics): Tay-Sachs (T-S) disease is what medical researchers call a **storage disease** in humans (see, e.g., `www.ninds.nih.gov/disorders/taysachs/taysachs.htm` for more details).

The **symptoms** of the disease result from the **abnormal accumulation** of a **fatty substance** called **ganglioside** G_{M2} in many cells of the body, especially the **central nervous system**.

Lack of a particular **enzyme**, **beta-hexosaminidase A** (**Hex A** for short), which normally **breaks down ganglioside** G_{M2}, is the **cause** of the disorder.

An **infant** with the disease will have **virtually none** of the enzyme, and at present this condition is typically **fatal** (usually by the age of 4), with no cure in sight.

A **carrier**, perfectly healthy in every way, will produce the enzyme in **about half the usual amount**.

The production of **Hex A** in the body is determined by a particular pair of **genes**.

Carriers have within their body cells **one gene** (H) that operates **normally** and **one gene** (h) that does **not**.

Adults can be **classified** as **carriers** or **noncarriers** by a **blood test** to see how much Hex A is present in their cells — if your Hex A level is **100%** of normal you're a **noncarrier**; if it's about **50%** of normal you're a **carrier**.

A **man** and a **woman** thinking of getting **married** have come into the **family health clinic** where you work, and the **blood tests** have shown them **both** to be **carriers**.

They're planning a family of **five children**, and they need your advice on the possibility of having **one or more T-S babies**.

Frequentist and Bayesian

Set up a simple **genetic model** for this situation and use it to work out the **possibilities** for any one of their children – is it **possible** for their **second child**, say, to be **normal**? a **carrier**? a **T-S baby**?

What are the **chances** of each of these happening for any given child?

What is the **probability** that, if they do have **five children**, they will have **one or more T-S babies**?

The meaning of probability. Two main ways to think about the **meaning of probability** have been developed:

- the **frequentist** (or **relative frequency**) approach, in which attention is restricted to phenomena that are **repeatable** under **identical conditions** (with each repetition logically **independent** of the others) and the **probability** $P(A)$ of an **event** A is regarded as the **long-run relative frequency** with which A would occur in the repetitions; and

- the **Bayesian** approach, in which A can be **any (true/false) proposition** you want (in other words, in this approach **attention need not be restricted to repeatable phenomena**) and $P(A)$ is a **numerical measure** of the **weight of evidence** in favor of the **statement** that A is **true**.

Evidently the **Bayesian** approach is **more general** (it **includes** the **frequentist approach** as a **special case**), but it turns out that the **math** is a lot **harder** in the **Bayesian** world, so in this introductory course we'll concentrate on the **frequentist** story and you can hear more about the **Bayesian** story later (if you have time and interest for **more study** in **probability** and **statistics**).

The Genetic Story

Let's **define**

$A = \{$1 or more T-S babies in a family of 5 children of 2 parents, both of whom are carriers $(Hh)\}$;

we **want** $P(A)$ (the **frequentist interpretation** of $P(A)$ involves **imagining many families** of **5 children**, each with **2 parents** who are **both carriers**, and asking what's the **relative frequency** of **1 or more T-S babies** among these families).

First let's work out the **possibilities** for **each** of their **children** one by one — given that we know **each parent** has the **genetic makeup** (Hh), the **standard way** to do this in **genetics** is with what's called a **Punnett square**, in which **one parent** forms the **rows** and the **other** the **columns** of a 2×2 **table**:

		Father's Genes	
		H	h
Mother's	H	(H, H)	(H, h)
Genes	h	(H, h)	(h, h)

The **simplest way** we can make sense of the **evidence** about the **level** of **Hex A** in the **blood** is to **theorize** that

- if you have the **genetic makeup** (H, H) you'll have **100%** of the **normal level** of Hex A (i.e., you're **normal**);
- if you have the **genetic makeup** (H, h) you'll have **50%** of the **normal level** of Hex A (i.e., you're a **carrier**); and
- if you have the **genetic makeup** (h, h) you'll have **0%** of the **normal level** of Hex A (i.e., you're a **T-S baby**).

Equally Likely Model

(**Genetics** note: if we define **phenotype** at the level of **presence** or **absence** of the **disease** (**2 phenotypes**), this is a **dominant-recessive genetic model** with H **dominant** and h **recessive**; if instead we define **phenotype** by the **amount** of **Hex A** in the **blood** (**3 phenotypes: 100%, 50%, 0%** of normal), this is an **additive genetic model**.)

This **answers some** of the **questions** above: yes, it's **possible** for any one of their children to be **normal**, or a **carrier**, or a **T-S baby**; but what about the **chances of these outcomes**?

When **conception** takes place, our current **best understanding** is that **all 4 of the possibilities** in the **4 cells** of the **Punnett square** are **equally likely** — this means that we can **apply** what must certainly be the **simplest useful probability model** for understanding the real world, the **equally likely model**:

Equally likely model (ELM): If you can **enumerate** {**all** the ways the **repeatable phenomenon** you're thinking about can come out} in such a way that **all** of these **possible outcomes** are **equally likely**, then for any event A

$$P(A) = \frac{\text{number of outcomes } \textbf{favorable} \text{ to } A}{\textbf{total number} \text{ of possible outcomes}}.$$

Example: If I make **one draw** Y **at random** from the little fake **population data set** $(1, 2, 9)$ I've discussed before, by **definition** of the phrase **"at random"** the **ELM** applies, and immediately $P(Y = 9) = \frac{1}{3} \doteq 33\%$ and $P(Y \text{ is odd}) = \frac{2}{3} \doteq 67\%$.

Applying the **ELM** to the **T-S case study**, evidently for **each** of this couple's children

$P(\textbf{normal}) = \frac{1}{4} = \textbf{25\%}$, $P(\textbf{carrier}) = \frac{2}{4} = \frac{1}{2} = \textbf{50\%}$, and $P(\textbf{T-S baby}) = \frac{1}{4} = \textbf{25\%}$.

Logical Equivalents

Turning now to the **main question** of interest in the **case study**, evidently {1 or more T-S babies in a family of 5 children} is **logically equivalent** to

({exactly 1 T-S baby} **or** {exactly 2 T-S babies} **or** {exactly 3 T-S babies} **or** {exactly 4 T-S babies} **or** {exactly 5 T-S babies}),

so it looks like one **strategy** for working out $P(A)$ is to **break it down** into a bunch of **simpler possibilities** linked together by a **logical connective** like **or** and work out how **or** behaves — in other words, that makes me wonder, for any events A and B, how $P(A \text{ or } B)$ **relates** to the two **simpler** ingredients $\{P(A), P(B)\}$.

Notice also that there's **only one other possibility** — if these people are **not** going to have {1 or more T-S babies} then they would **have** to have {exactly 0 T-S babies}, so

A = {1 or more T-S babies} = **not** {exactly 0 T-S babies};

this in turn **makes me wonder** how $P(A)$ and $P(\text{not } A)$ are **related**.

And finally if these people **were indeed to have** {exactly 0 T-S babies}, this would be **logically equivalent** to

({not a T-S baby on child 1} **and** {not a T-S baby on child 2} **and** {not a T-S baby on child 3} **and** {not a T-S baby on child 4} **and** {not a T-S baby on child 5}),

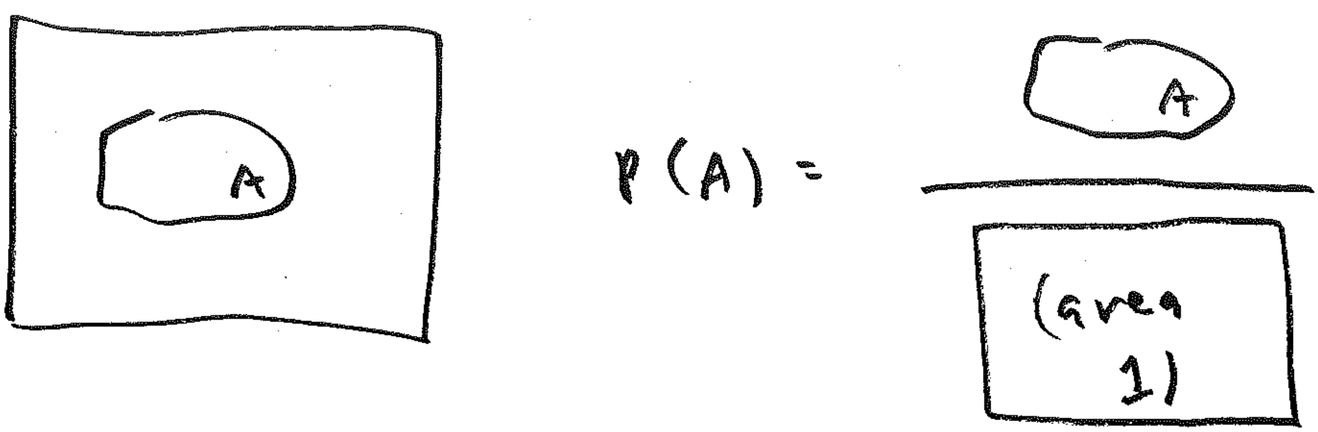
so I'm also left **wondering**, for any events A and B, how $P(A \text{ and } B)$ **relates** to the two **simpler** ingredients $\{P(A), P(B)\}$.

Venn Diagrams

In **intuitively** working out how **and**, **or**, and **not** **behave**, it helps (as Triola and Triola note in Section 3–3) to make use of what are called **Venn diagrams**.

The idea is to draw a **rectangular box** to stand for **all the different ways the repeatable experiment you're interested in could come out** and put one or more **blobs** $A, B, \ldots$ inside the box to stand for **all the ways** $A, B, \ldots$ could turn out to be **true**; then I imagine **shooting** at the box in such a way that (a) the shot **must fall somewhere inside** and (b) **every point inside the box** has the **same chance** of being where the shot falls.

From this, **graphically** $P(A)$ must equal the **ratio** of the **area** of the **blob** for A to the **total area of the box**:



Using the **relative frequency** intuitive idea of probability, the next basic thing to notice is that for any event A the **relative frequency** with which it could happen, in **(imaginary) repetitions** of the basic thing you're imagining repeating, **can't be less than 0** (or 0%) or **more than 1** (or 100%):

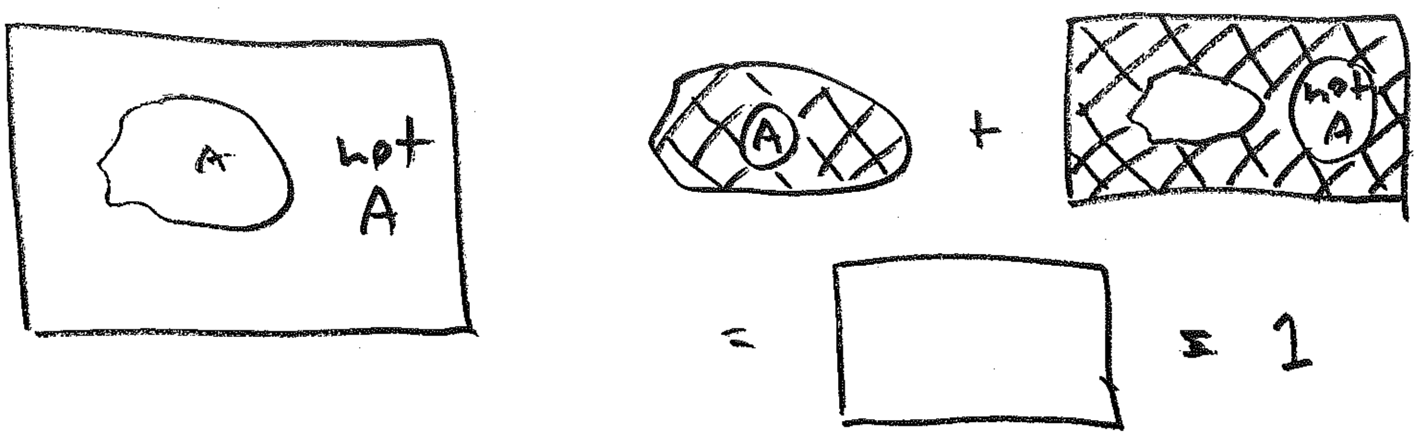
For any event A, $\quad 0\% = 0 \le P(A) \le 1 = 100\%.$

In other words, the **total area** of the **box** is **1** or 100%.

Basic Probability Rules

The next thing to **notice** is that, for any event A, the shot either **has to** fall **inside** A (which is like saying that A is **true**) or it **has** to fall **inside** (not A) (which is like saying that A is **false**), and it can't do **both**.

This means that (A, not A) forms what's called a **partition**: a way of expressing **all the different possible outcomes** so that the **events** making up the **partition** (in this case, A and (not A)) are **mutually exclusive** (if **one** of them is **true** the **other** one **can't be**) and **exhaustive** (**one** of them **has to be true**) — in the **Venn diagram** this just corresponds to the idea that [the **area** for A] + [the **area** for (not A)] has to **equal** the **total area of the box** (which, by the argument above, is **1** or **100%**):



This gives rise to **another basic rule**:

For any event A, $\quad P(A) + P(\text{not } A) = 1 = 100\%.$

This may seem **trivial**, but a **simple rearrangement** of this fact actually turns out to be a **valuable way** to **compute probabilities**:

For any event A, $\quad P(A) = 1 - P(\text{not } A).$

Basic Rules (continued)

In other words, if you're having **trouble** working out $P(A)$ **directly** you can **try** to **compute** $P(\text{not } A)$, which may be **easier**, and **subtract from 1**.

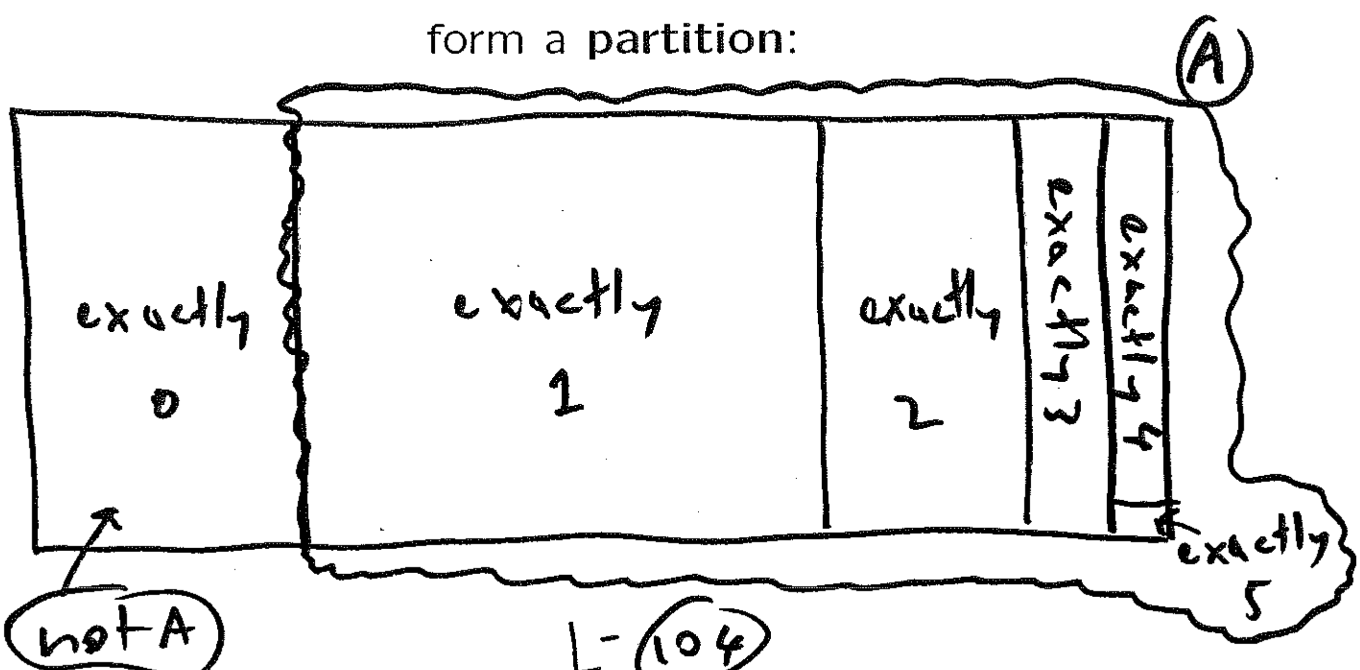
In the **T-S case study**, this observation is **helpful**: as we saw above, if these parents are going to have **1 or more T-S babies** there are **5 different ways** that could occur (**exactly 1**, **exactly 2**, **...**, **exactly 5**), so $P(A) = P(1 \text{ or more T-S babies})$ sounds **difficult** to compute **directly**, but there's **only one way** they can have **(not {1 or more T-S babies})**, namely {**exactly 0 T-S babies**}, so it'll be **easier** to compute $P(A)$ **indirectly** using the rule for **not**:

$$P(\mathbf{1 \text{ or more}} \text{ T-S babies}) = 1 - P(\mathbf{no} \text{ T-S babies}).$$

Another way to put it, using **Venn diagrams**, is that the **events**

({**exactly 0** T-S babies}, {**exactly 1** T-S baby}, {**exactly 2** T-S babies}, {**exactly 3** T-S babies}, {**exactly 4** T-S babies}, {**exactly 5** T-S babies})

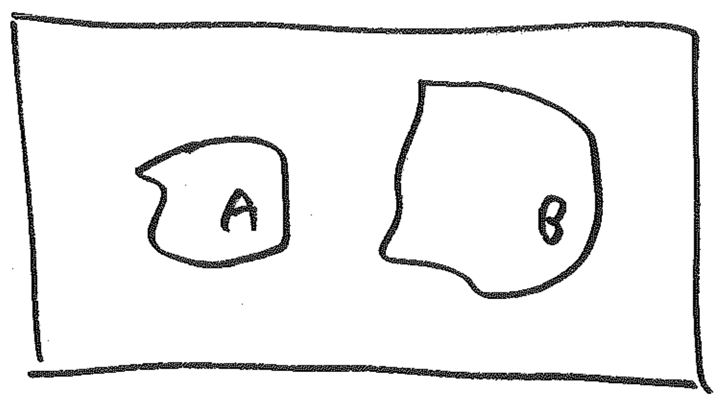
form a **partition**:



Working With Or

OK, we've seen how **not** works; how about **or**? — in other words, how does $P(A \text{ or } B)$ **relate** to the two **simpler** ingredients $\{P(A), P(B)\}$.

It turns out there are **two cases** to consider — suppose the **Venn diagram** looks like this:



In this picture A and B **don't overlap**, which is **equivalent** to saying that they're **mutually exclusive** — in this case, to compute the **chance** the **random shot** lands in (A or B) evidently you can just **add** the **separate chances** $\{P(A), P(B)$ that it lands **either** in A **or** in B:

> For two **mutually exclusive** events A and B,
> $P(A \text{ or } B) = P(A) + P(B)$.

Evidently this rule can **easily be extended** to **three** or more **mutually exclusive** events: if A, B and C have **no (pairwise) overlap**, then

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

and so on.

What about if A and B do **overlap**? — then the **Venn diagram** would look like this: