Probability, Entropy, and Inference / More About Inference

Mário S. Alvim
(msalvim@dcc.ufmg.br)

Information Theory

DCC-UFMG
(2020/01)
Probability, Entropy, and Inference
Definitions and notation

- An **ensemble** $X$ is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$, where:
  - $x$ is the **outcome** of a random variable;
  - $\mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ is the set of possible values for the random variable, called an **alphabet**;
  - $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$ are the **probability** of each value, with
    
    $$P(x = a_i) = p_i,$$
    
    $$p_i \geq 0,$$
    
    and
    
    $$\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1.$$ 

- We may write $P(x = a_i)$ as $P(a_i)$ or $P(x)$ when no confusion is possible.
Definitions and notation

  
  The outcome is a letter that is randomly selected from an English document.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$a_i$</th>
<th>$p_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>0.0575</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>0.0128</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>0.0263</td>
</tr>
<tr>
<td>4</td>
<td>d</td>
<td>0.0285</td>
</tr>
<tr>
<td>5</td>
<td>e</td>
<td>0.0913</td>
</tr>
<tr>
<td>6</td>
<td>f</td>
<td>0.0173</td>
</tr>
<tr>
<td>7</td>
<td>g</td>
<td>0.0133</td>
</tr>
<tr>
<td>8</td>
<td>h</td>
<td>0.0313</td>
</tr>
<tr>
<td>9</td>
<td>i</td>
<td>0.0599</td>
</tr>
<tr>
<td>10</td>
<td>j</td>
<td>0.0006</td>
</tr>
<tr>
<td>11</td>
<td>k</td>
<td>0.0084</td>
</tr>
<tr>
<td>12</td>
<td>l</td>
<td>0.0335</td>
</tr>
<tr>
<td>13</td>
<td>m</td>
<td>0.0235</td>
</tr>
<tr>
<td>14</td>
<td>n</td>
<td>0.0596</td>
</tr>
<tr>
<td>15</td>
<td>o</td>
<td>0.0689</td>
</tr>
<tr>
<td>16</td>
<td>p</td>
<td>0.0192</td>
</tr>
<tr>
<td>17</td>
<td>q</td>
<td>0.0008</td>
</tr>
<tr>
<td>18</td>
<td>r</td>
<td>0.0508</td>
</tr>
<tr>
<td>19</td>
<td>s</td>
<td>0.0567</td>
</tr>
<tr>
<td>20</td>
<td>t</td>
<td>0.0706</td>
</tr>
<tr>
<td>21</td>
<td>u</td>
<td>0.0334</td>
</tr>
<tr>
<td>22</td>
<td>v</td>
<td>0.0069</td>
</tr>
<tr>
<td>23</td>
<td>w</td>
<td>0.0119</td>
</tr>
<tr>
<td>24</td>
<td>x</td>
<td>0.0073</td>
</tr>
<tr>
<td>25</td>
<td>y</td>
<td>0.0164</td>
</tr>
<tr>
<td>26</td>
<td>z</td>
<td>0.0007</td>
</tr>
<tr>
<td>27</td>
<td>-</td>
<td>0.1928</td>
</tr>
</tbody>
</table>
Definitions and notation

- **Probability of a subset.** If $T$ is a subset of $\mathcal{A}_X$ then

  \[ P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i). \]

- **Example 1** (Continued)

  If we define $V$ to be vowels, $V = \{a, e, i, o, u\}$, then

  \[
  P(V) = p(a) + p(e) + p(i) + p(o) + p(u)
  = 0.0575 + 0.0913 + 0.0599 + 0.0689 + 0.0334
  = 0.3110.
  \]
Joint ensemble. $XY$ is an ensemble in which each outcome is an ordered pair $x, y$ with $x \in \mathcal{A}_X = \{a_1, \ldots, a_I\}$ and $y \in \mathcal{A}_Y = \{b_1, \ldots, b_J\}$. We call $P(x, y)$ the joint probability of $x$ and $y$.

Commas are optional when writing ordered pairs, so $xy \Leftrightarrow x, y$.

In a joint ensemble $XY$ the two random variables may or may not be independent.
Definition and notation

Example 1 (Continued)

Frequency of bigrams (pair \(xy\) of letters) in “The FAQ Manual for Linux.” The outcome is a pair of consecutive letters randomly selected from an English document.
We can obtain the **marginal probability** $P(x)$ from the joint probability $P(x, y)$ by summation:

$$P(x = a_i) \equiv \sum_{y \in A_Y} P(x = a_i, y).$$

Similarly, using briefer notation, the marginal probability of $y$ is

$$P(y) \equiv \sum_{x \in A_X} P(x, y).$$

**Conditional probability.** If $P(y = b_j) \neq 0$,

$$P(x = a_i \mid y = b_j) \equiv \frac{P(x = a_i, y = b_j)}{P(y = b_j)}.$$  

If $P(y = b_j) = 0$, then $P(x = a_i \mid y = b_j)$ is undefined.
Definitions and notation

- **Example 1** (Continued)

Conditional probabilities of bigrams in “The FAQ Manual for Linux.”

![Conditional probabilities](image_url)
Definitions and notation

- We will use $\mathcal{H}$ to denote the assumptions on which the probabilities are based.

- **Product rule / Chain rule:**

  $$P(x, y \mid \mathcal{H}) = P(x \mid \mathcal{H})P(y \mid x, \mathcal{H})$$
  $$= P(y \mid \mathcal{H})P(x \mid y, \mathcal{H}).$$

- **Sum rule:**

  $$P(x \mid \mathcal{H}) = \sum_y P(x, y \mid \mathcal{H})$$
  $$= \sum_y P(y \mid \mathcal{H})P(x \mid y, \mathcal{H}).$$

- **Independence.** Two random variables $X$ and $Y$ are independent (sometimes written as $X \perp Y$) if, and only if, for all $x, y$:

  $$P(x, y) = P(x)P(y).$$
What “probability” means

- Probabilities can be interpreted in two main ways:

  1. The **frequentist interpretation** sees probabilities as frequencies of outcomes in random experiments.

     This is the usual approach we have taken so far.

  2. The **Bayesian interpretation** (also known as **degree-of-belief interpretation**, **knowledge interpretation**, or **subjective interpretation**) sees probabilities as the degree to which one believes in the truth value of propositions depending on random variables.

     For instance, juries in trials usually try to access “the probability that Mr. Jones killed his wife, given the evidence”.

     The degree-of-belief interpretation of probabilities is essential for probabilistic reasoning, in particular, Bayesian reasoning.
Cox axioms for reasoning about belief

**Notation.** Let

- “the degree of belief in proposition $x$” be denoted by $B(x)$,
- the negation of $x$ be denoted by $\bar{x}$, and
- the degree of belief in a conditional proposition, “$x$, assuming proposition $y$ to be true”, be denoted by $B(x \mid y)$. 
The Cox axioms establish what is essential to reason about belief:

A1 Degrees of belief can be ordered.
Using $\succeq$ to denote "is greater than":

if $B(x) \succeq B(y)$ and $B(y) \succeq B(z)$ then $B(x) \succeq B(z)$.

(As a consequence, beliefs can be mapped onto real numbers.)

A2 The degree of belief in a proposition $x$ and its negation $\overline{x}$ are related.
There is a function $f$ such that

$$B(x) = f(B(\overline{x})).$$

A3 The degree of belief in a conjunction of propositions $x$, $y$ (i.e., $x \land y$) is related to the degree of belief in the conditional proposition $x \mid y$ and the degree of belief in the proposition $y$.
There is a function $g$ such that

$$B(x, y) = g(B(x \mid y), B(y)).$$
Probabilities as belief

- If a set of beliefs satisfy the Cox axioms, we can map them onto probabilities satisfying

\[ p(\text{FALSE}) = 0, \quad p(\text{TRUE}) = 1, \quad \text{and} \quad 0 \leq p(x) \leq 1, \]

and also satisfying the rules of probability

\[ p(x) = 1 - p(\overline{x}), \]

and

\[ p(x, y) = p(x \mid y)p(y). \]

- This is actually what we have been using all along when applying Bayesian reasoning.

- Probabilities as belief or knowledge is a (very beautiful!) generalization of logic.

I strongly recommend E. T. Jaynes's book *Probability Theory: The Logic of Science* [link].
Using probabilities as degrees of belief, we can make inference of future outcomes and of past parameters of an experiment.

We can talk, then, about forward and inverse probabilities.
Forward and inverse probabilities

- **Forward probability** problems involve a generative model that describes a process that is assumed to give rise to some data.

  The task is to compute the probability distribution or expectation of some quantity that depends on the data.
**Example 2** An urn contains $K$ balls, of which $B$ are black and $W = K - B$ are white. Fred draws a ball at random from the urn and replaces it, $N$ times.

a) What is the probability distribution of the number of times a black ball is drawn, $n_B$?

b) What is the expectation of $n_B$? What is the variance of $n_B$? What is the standard deviation of $n_B$? Give numerical answers for the cases $N = 5$ and $N = 400$, when $B = 2$ and $K = 10$.

**Solution.**

a) Let us define $f_B = B/K$ as the fraction of black balls in the urn.

The number of black balls follow has a binomial distribution.

$$p(n_B \mid f_B, N) = \binom{N}{n_B} f_B^{n_B} (1 - f_B)^{N-n_B}.$$
Forward and inverse probabilities

Example 2  (Continued)

b) Since we have a binomial distribution, its expecatation is

\[ E(n_B) = Nf_B, \]

its variance is

\[ Var(n_B) = Nf_B(1 - f_B), \]

and its standard deviation is

\[ \sigma(n_B) = \sqrt{Nf_B(1 - f_B)}. \]

Hence, when \( B = 2 \) and \( K = 10 \) we have \( f_b = \frac{1}{5} \), and

- for \( N = 5 \) we have

\[ E(n_B) = 1, \ Var(n_B) = 0.8, \text{ and } \sigma(n_B) = 0.89. \]

- for \( N = 400 \) we have

\[ E(n_B) = 80, \ Var(n_B) = 64, \text{ and } \sigma(n_B) = 8. \]
Forward and inverse probabilities

Like forward probability problems, inverse probability problems involve a generative model of a process.

But instead of computing the probability distribution of some quantity produced by the process, we compute the conditional probability of one or more of the unobserved variables in the process, given the observed variables. This invariably requires the use of Bayes’ theorem.
Example 3
Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white.

One of the urns is selected at random an one ball is drawn. The ball is black. What is the probability that the selected urn is A?

Solution.
We have two alternative hypotheses: the selected urn is A or it is B.

We have two possible evidences: the ball drawn is black (b) or it’s white (w).

We can find \( p(\text{urn} = A \mid \text{ball} = \text{black}) \) using Bayes’s theorem:

\[
p(\text{urn} = A \mid \text{ball} = b) = \frac{p(\text{urn} = A, \text{ball} = b)}{p(\text{ball} = b)}
\]

\[
= \frac{p(\text{urn} = A)p(\text{ball} = b \mid \text{urn} = A)}{p(\text{urn} = A)p(\text{ball} = b \mid \text{urn} = A) + p(\text{urn} = B)p(\text{ball} = b \mid \text{urn} = B)}
\]

\[
= \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 1/2 \cdot 2/3}
\]

\[
= \frac{1}{3}
\]
Like we said, using probabilities as degrees of belief, we can make inference of future outcomes and of past parameters of an experiment.

**Notation.** Let

- \( \mathcal{H} \) be the overall hypothesis space;
- \( \theta \) be the parameters of an experiment; and
- \( D \) be the data obtained as the outcome of the experiment.

- \( p(\theta \mid \mathcal{H}) \) is the **prior probability** of the parameter \( \theta \), that is, the probability of the parameter before the experiment is run.
The conditional distribution $p(D \mid \theta, \mathcal{H})$ represents how the experiment evolves, relating the parameter $\theta$ and the data $D$ given some hypothesis $\mathcal{H}$.

Depending on what we are reasoning about, this quantity has different names:

1. If we fix the value of the parameter $\theta$, $p(D \mid \theta, \mathcal{H})$ is the **conditional probability** of observing data $D$ if the parameter assumed value $\theta$ under hypothesis $\mathcal{H}$.

2. If we fix the value of the data $D$, $p(D \mid \theta, \mathcal{H})$ is the **likelihood** of the parameter $\theta$ given we observed data $D$, under hypothesis $\mathcal{H}$.
Forward and inverse probabilities

- $p(D \mid \mathcal{H})$ is the **probability of evidence**, that is, the probability of producing data $D$ under hypothesis $\mathcal{H}$, after the experiment is run.

  Naturally

  $$p(D \mid \mathcal{H}) = \sum_\theta p(\theta, D \mid \mathcal{H})$$

  $$= \sum_\theta p(\theta \mid \mathcal{H}) p(D \mid \theta, \mathcal{H}).$$

- The general equation

  $$p(\theta \mid D, \mathcal{H}) = \frac{p(\theta \mid \mathcal{H}) p(D \mid \theta, \mathcal{H})}{p(D \mid \mathcal{H})}$$

  is written as

  **posterior probability** = \( \frac{\text{prior probability \times likelihood}}{\text{evidence}} \).
Example 4  Consider the following scenario:

1. There are eleven urns labelled by \( u \in \{0, 1, 2, \ldots, 10\} \), each containing ten balls.

2. Urn \( u \) contains \( u \) black balls and \( 10 - u \) white balls.

3. Fred selects an urn \( u \) at random and draws \( N \) times with replacement from that urn, obtaining \( n_B \) blacks and \( N - n_B \) whites.

4. Fred’s friend, Bill, looks on. If after \( N = 10 \) draws and \( n_B = 3 \) blacks have been drawn, what is the probability that the urn Fred is using is urn \( u \), from Bill’s point of view? (Bill doesn’t know the value of \( u \).)

Solution.

The joint probability distribution of the random variables \( u \) and \( n_B \) can be written

\[
P(u, n_B \mid N) = P(u)P(n_B \mid u, N).
\]
From the joint probability of $u$ and $n_B$, we can obtain the condition distribution of $u$ given $n_B$:

\[ P(u \mid n_B, N) = \frac{P(u, n_B \mid N)}{P(n_B \mid N)} \]

\[ = \frac{P(u)P(n_B \mid u, N)}{P(n_B \mid N)} \]

The marginal probability of $u$ is

\[ P(u) = \frac{1}{11} \]

for all $u$.  

Example 4 (Continued)
As we solved in a previous exercise, we know that

\[ P(n_B \mid u, N) = \binom{N}{n_B} f_u^{n_B} (1 - f_u)^{N-n_B} \]

The denominator can be computed as

\[ P(n_B \mid N) = \sum_u P(u, n_B \mid N) \]

\[ = \sum_u P(u)P(n_B \mid u, N) . \]
Example 4  (Continued)

Hence the conditional probability of \( u \) given \( n_B \) is:

\[
P(u \mid n_B, N) = \frac{P(u)P(n_B \mid u, N)}{P(n_B \mid N)} = \frac{1}{P(n_B \mid N)} \cdot \left( \begin{array}{c} N \\ n_B \end{array} \right) f_u^{n_B} (1 - f_u)^{N-n_B}.
\]

Substituting \( n_B = 3 \) and \( N = 10 \) we find the conditional probabilities for all values of \( u \), as in the figure on the side.
Example 5 Consider the same scenario of the previous example, but now:

1. Bill has observed $n_B = 3$ blacks in $N = 10$ draws.
2. Fred is going to draw another ball from the same urn.
3. What is the probability that the next drawn ball is black?

Solution. By the sum rule:

$$P(\text{ball}_{N+1} \text{ is black} | n_B, N) = \sum_u P(u | n_B, N)P(\text{ball}_{N+1} \text{ is black} | u, n_B, N).$$

Since the balls are drawn with replacement from the chosen urn, we have

$$P(\text{ball}_{N+1} \text{ is black} | u, n_B, N) = f_u = u/10,$$

no matter what $n_B$ and $N$ are.

Using the values $P(u | n_B, N)$ from the previous example we obtain

$$P(\text{ball}_{N+1} \text{ is black} | n_B = 3, N = 10) = 0.333.$$
Inverse probability and prediction

In the previous example, we made our prediction using probability theory. More precisely, we:

1. computed the probability of every urn being used, then
2. predicted the probability of the next ball being black averaging over the probability of each urn being the one producing the next ball.

This contrasts with a common practice in statistics of making predictions by:

1. first selecting the most plausible hypothesis (which here would be that the urn is urn \( u = 3 \)), and then
2. making the predictions assuming that hypothesis to be true (which would give a probability of 0.3 that the next ball is black).

The correct prediction is the one that takes into account the uncertainty by marginalizing over the possible values of the hypothesis \( u \).

Marginalization here leads to slightly more moderate, less extreme prediction.
We can say:

- “the probability of the data”, meaning the probability of obtaining data $D$ given parameter $\theta$, under hypothesis $\mathcal{H}$.
- “the likelihood of the parameter”, meaning how likely it is that a given parameter $\theta$ generated the data $D$ we observed, under assumption $\mathcal{H}$.

We should not say “the likelihood of the data”.

Let us consider now a “more complicated” example involving urns and balls.

Example 6. Urn \( A \) contains five balls: one black, two white, one green, and one pink; urn \( B \) contains five hundred balls: two hundred black, one hundred white, 50 yellow, 40 cyan, 30 sienna, 25 olive, 25 silver, 20 gold, and 10 purple. (One fifth of \( A \)'s balls are black; two-fifths of \( B \)'s balls are black.)

One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the selected urn is \( A \)?

Solution.

You can do it, folks! I know you can!
In the previous two examples, does each answer depend on the detailed contents of each urn?

No! All that matters is the probability of the outcome that actually happened (here, that the ball drawn was black) given the different hypotheses.

We need only to know the likelihood, i.e., how the probability of the data that happened varies with the hypothesis.

This simple rule about inference is known as the likelihood principle.

The likelihood principle: given a generative model for data $D$ given parameters $\theta$, $p(D \mid \theta)$, and having observed a particular outcome $d$, all inferences and predictions should depend only on the function $P(d \mid \theta)$.

In other words, the likelihood principle says that once an outcome $d$ is observed, the probabilities of other outcomes do not have an influence on the inferences we make about the parameter $\theta$. 
More About Inference
Bayes Theorem at the heart of inference

- We have seen many examples showing that Bayes’ Theorem is very useful when doing inference.

If we are presented with data $D$ and have a possible explanation $E$ for the data, Bayes’ theorem tells us that

$$p(E \mid D, \mathcal{H}) = \frac{p(E \mid \mathcal{H})p(D \mid E, \mathcal{H})}{p(D \mid \mathcal{H})}.$$  

(Recall that $\mathcal{H}$ denotes the assumptions on which the probabilities are based.)
Bayes Theorem at the heart of inference

- The probability $p(E \mid D, \mathcal{H})$ we will encounter will depend on the assumptions we made, such as:
  - the hypotheses $\mathcal{H}$, and
  - the *a priori* probability $p(E \mid \mathcal{H})$ of the explanation.

- However:

  you can’t do inference –or data compression– without making assumptions, and Bayesians won’t apologize for it.
1. Once assumptions are made, the inferences are:

   - objective,
   - unique, and
   - reproducible with complete agreement by anyone who has the same information and makes the same assumptions.

2. Bayesian inference makes assumptions explicit, so you are forced to think clearly about what hypotheses your model is based on.
   
   Moreover, once assumptions are explicit, they are easier to criticize, and easier to modify.
Why Bayesian inference is good inference

3. When we are not sure which of various alternative assumptions is the most appropriate for a problem, we can treat this question as another inference task.

Given data $D$, we can compare alternative assumptions $\mathcal{H}$ using

$$P(\mathcal{H} \mid D, I) = \frac{P(\mathcal{H} \mid I) P(D \mid \mathcal{H}, I)}{P(D \mid I)},$$

where $I$ denotes the highest assumptions, which we are not questioning.

4. We can take into account our uncertainty about the assumptions themselves when making subsequent predictions.

Rather than choosing one particular assumption $\mathcal{H}^*$, and predicting quantity $t$ using $P(t \mid D, \mathcal{H}^*, I)$, we can do

$$P(t \mid D, I) = \sum_{\mathcal{H}} P(\mathcal{H} \mid D, I) P(t \mid D, \mathcal{H}, I).$$
Assume we have two competing models for a phenomenon, represented by hypothesis $H_1$ and hypothesis $H_2$, and that we run an experiment and observe data $D$.

The best model ($H_1$ or $H_2$) fitting the data ($D$) is indicated by the ratio:

$$\frac{p(H_1 \mid D)}{p(H_2 \mid D)},$$

so that

- If $\frac{p(H_1 \mid D)}{p(H_2 \mid D)} > 1$, hypothesis $H_1$ is a more probable explanation for $D$ than $H_2$.
- If $\frac{p(H_1 \mid D)}{p(H_2 \mid D)} < 1$, hypothesis $H_2$ is a more probable explanation for $D$ than $H_1$.
- If $\frac{p(H_1 \mid D)}{p(H_2 \mid D)} = 1$, hypotheses $H_1$ and $H_2$ are equally good in explaining $D$. 
Note that this ratio can be decomposed into:

$$\frac{p(H_1 \mid D)}{p(H_2 \mid D)} = \frac{p(H_1)p(D \mid H_1)}{p(D)} \cdot \frac{p(D)}{p(H_2)p(D \mid H_2)} = \frac{p(H_1)}{p(H_2)} \cdot \frac{p(D \mid H_1)}{p(D \mid H_2)},$$

where:

- $\frac{p(H_1)}{p(H_2)}$ is the ratio of the prior probabilities of the competing models.

This ratio indicates the relative probability of the hypotheses before data $D$ is collected.

$\frac{p(H_1)}{p(H_2)} = k$ means that $H_1$ is a priori $k$ times more probable than $H_2$. 

- $p(H_1)$ is the prior probability of $H_1$.
- $p(H_2)$ is the prior probability of $H_2$.
- $p(D \mid H_1)$ is the likelihood of the data given $H_1$.
- $p(D \mid H_2)$ is the likelihood of the data given $H_2$. 

- $p(D)$ is the marginal likelihood of the data.
Model comparison

- Note that this ratio can be decomposed into:

\[
\frac{p(\mathcal{H}_1 | D)}{p(\mathcal{H}_2 | D)} = \frac{p(\mathcal{H}_1)p(D | \mathcal{H}_1)}{p(D)} \cdot \frac{p(D)}{p(\mathcal{H}_2)p(D | \mathcal{H}_2)} = \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_2)} \cdot \frac{p(D | \mathcal{H}_1)}{p(D | \mathcal{H}_2)},
\]

where:

- The **likelihood ratio** \( \frac{p(D | \mathcal{H}_1)}{p(D | \mathcal{H}_2)} \) is the ratio of the evidence for each model.

This ratio serves as a “correction factor” for the prior ratio, taking into consideration the weight evidence \( D \) collected has on each hypothesis.

\[
\frac{p(D | \mathcal{H}_1)}{p(D | \mathcal{H}_2)} = c \text{ means that hypothesis } \mathcal{H}_1 \text{ is } c \text{ times more likely an explanation for data } D \text{ than hypothesis } \mathcal{H}_2:
\]

- \( c > 1 \) means that evidence \( D \) favors hypothesis \( \mathcal{H}_1 \) over hypothesis \( \mathcal{H}_2 \).
- \( c < 1 \) means that evidence \( D \) favors hypothesis \( \mathcal{H}_2 \) over hypothesis \( \mathcal{H}_1 \).
- \( c = 1 \) means that evidence \( D \) favors no hypothesis.
Example 7 Two people have left traces of their own blood at the scene of a crime.

A suspect, Oliver, is tested and found to have type “O” blood. The blood groups of the two traces are found to be:

- of type “O” (a common type in the local population, having frequency 60%), and,
- of type “AB” (a rare type, with frequency 1%).

Do these data (type “O” and “AB” blood were found at the scene) give evidence in favor of the proposition that Oliver was one of the two people present at the crime?
Solution.

- A careless lawyer might claim that the fact that the suspect’s blood type was found at the crime scene is evidence for the theory that he was present. This is wrong.

- Let $S$ be the proposition “the suspect and one unknown person were present” and the alternative proposition $\overline{S}$ be “two unknown people from the population were present”.

- Let $D$ be the data collected (i.e., there is one sample of blood type “O” and one of blood type “AB” in the scene), and let $\mathcal{H}$ be the set of underlying hypotheses.

- We want to evaluate the contribution the data $D$ gives as evidence for each supposition ($S$ or $\overline{S}$), i.e., the ratio

$$\frac{p(D \mid S, \mathcal{H})}{p(D \mid \overline{S}, \mathcal{H})}.$$
Let us call $p_{AB}$ and $p_O$ the probabilities of a person in the general population having blood type “AB” and “O”, respectively.

We know that

$$p(D \mid S, \mathcal{H}) = p_{AB},$$

since given $S$, we already know that one trace will be of type “O”.

We also know that

$$p(D \mid \bar{S}, \mathcal{H}) = 2p_O p_{AB}.$$

Hence, the ratio of the evidences is

$$\frac{p(D \mid S, \mathcal{H})}{p(D \mid \bar{S}, \mathcal{H})} = \frac{1}{2p_O} = \frac{1}{2 \cdot 0.6} = 0.83,$$

so the data provides evidence against the supposition that the suspect, Oliver, was present.
Example 8  Assume the same scenario of the example above, but now there is another suspect, Abigail, whose blood type is “AB”.

Does the evidence found (one sample of blood type “O” and one sample of blood type “AB”) favors the proposition that Abigail was one of the two people present at the crime?
Example 8 (Continued)

Solution.

Intuitively, it may seem that the data does provide evidence that Abigail was at the crime scene.

But let’s be careful and do the right calculations.

Let $S'$ be the proposition “the suspect (Abigail) and one unknown person were present” and the alternative proposition $S$ be “two unknown people from the population were present”.

We know that

$$p(D \mid S', \mathcal{H}) = p_O,$$

since given $S'$, we already know that one trace will be of type “AB”.

The ratio of the evidences is

$$\frac{p(D \mid S', \mathcal{H})}{p(D \mid S, \mathcal{H})} = \frac{p_O}{2p_Op_{AB}} = \frac{1}{2p_{AB}} = \frac{1}{2 \cdot 0.01} = 50,$$

so the data provides evidence for the supposition that Abigail was present.
Example 9 Assume the same scenario of the example above, but now consider that 99% of the population is of blood type “O”, and 1% is of blood type “AB” (and that no other blood type exists).

Intuitively the evidence found (one sample of blood type “O” and one sample of blood type “AB”) favors the proposition that Abigail was one of the two people present at the crime.

On the other hand, is it reasonable to assume that the evidence found favors the proposition that Oliver was one of the two people present at the crime?
Example 9 (Continued)

Solution.

Intuitively (since there exist only blood types “AB” and “O”) the evidence cannot increase the probability that both a person of blood type “AB” (such as Abigail) and a person of blood type “O” (such as Oliver) were present.

That would mean that the evidence increases the probability of all suspects, no matter whom, being at the crime scene, which is nonsense.

In other words:

data may be compatible with several theories, but

if some data provides evidence for some theories,

it necessarily provides evidence against some other theories.
Example 10 Assume once again the same scenario of the example above. Consider that were found at the crime scene:

- $n_O$ blood stains of individuals of type “O”, and
- $n_{AB}$ blood stains of individuals of type “AB”

from a total of $N$ blood samples found in total.

Consider that the frequencies of blood type “O” in the general population is $p_O$, and the frequency of blood type “AB” in the general population is $p_{AB}$ (there may be other blood types too).

Find a closed formula for the likelihood ratio of the following two competing hypotheses:

- $S$: “the type “O” suspect (Oliver) and $N-1$ unknown others left $N$ stains”, and
- $\bar{S}$: “$N$ unknowns left $N$ stains”.
Example 10 (Continued)

Solution.

- For hypothesis $\overline{S}$ we have

\[
P(n_O, n_{AB} \mid \overline{S}) = \frac{N!}{n_O! n_{AB}!} p_O^{n_O} p_{AB}^{n_{AB}}.
\]

- For hypothesis $S$ we have

\[
P(n_O, n_{AB} \mid S) = \frac{(N - 1)!}{(n_O - 1)! n_{AB}!} p_O^{n_O - 1} p_{AB}^{n_{AB}},
\]

since because Oliver is fixed, we only need the distribution of the other $N - 1$ individuals.
The likelihood ratio is

\[
\frac{P(n_O, n_{AB} \mid S)}{P(n_O, n_{AB} \mid \overline{S})} = \frac{(N - 1)! p_O^{n_O - 1} p_{AB}^{n_{AB}}}{(n_O - 1)! n_{AB}!} \cdot \frac{n_O! n_{AB}!}{N! p_O^{n_O} p_{AB}^{n_{AB}}} = \frac{n_O/N}{p_O}.
\]

This likelihood ration means that the evidence depends on a comparison on the frequency of Oliver’s blood type in the data \((n_O/N)\), and the frequency of Oliver’s blood type in the general population \((p_O)\):

1. If Oliver’s blood type is more frequent in the data than in the population, the evidence counts for his presence at the crime scene.
2. If Oliver’s blood type is less frequent in the data than in the population, the evidence counts for his absence from the crime scene.
3. If Oliver’s blood type is as frequent in the data than as it is in the population, the evidence is neutral for his presence in the crime scene.
Example 11 A die is selected at random from two twenty-faced dice on which the symbols 1-10 are written with nonuniform frequency as follows.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of faces of die A</strong></td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Number of faces of die B</strong></td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The randomly chosen die is rolled 7 times, with the following outcomes:

5, 3, 9, 3, 8, 4, 7.

What is the probability that the die is die A?
Example 11 (Continued)

The observed data $D$ is the sequence of outcomes 3, 9, 3, 8, 4, 7.
The concurrent hypotheses are whether the die is $A$ or $B$.
We can, then, compute the ratio of posterior probabilities:

$$\frac{p(A | D)}{p(B | D)} = \frac{p(A)}{p(B)} \cdot \frac{p(D | A)}{p(D | B)}$$

$$= \frac{1/2 \cdot 1/20 \cdot 3/20 \cdot 1/20 \cdot 3/20 \cdot 1/20 \cdot 2/20 \cdot 1/20}{1/2 \cdot 2/20 \cdot 2/20 \cdot 1/20 \cdot 2/20 \cdot 2/20 \cdot 2/20 \cdot 2/20}$$

$$= \frac{9}{32}$$

Note that, knowing that $p(A | D) + p(B | D) = 1$, we can go further and infer that

$$p(A | D) = \frac{9}{41}, \quad \text{and} \quad p(B | D) = \frac{32}{41}.$$
Example 12  Consider again the same dice $A$ and $B$ from the previous example, but assume now that there is a third twenty-faced die, die $C$, on which the symbols 1-20 are written once each.

As above, one of the three dice is selected at random and rolled 7 times, giving the outcomes:

$$3, 5, 4, 8, 3, 9, 7.$$ 

What is the probability that the die is

a) die $A$?

b) die $B$?

c) die $C$?
Model comparison - Examples

Example 12 (Continued)

Solution.

We can compute the quantities directly using the formulas

\[ p(A | D) = \frac{p(A)p(D | A)}{p(D)}, \quad p(B | D) = \frac{p(B)p(D | B)}{p(D)}, \quad \text{and} \quad p(C | D) = \frac{p(C)p(D | C)}{p(D)}. \]

First, let us compute

\[ p(D | A) = \frac{3}{20} \cdot \frac{1}{20} \cdot \frac{2}{20} \cdot \frac{1}{20} \cdot \frac{3}{20} \cdot \frac{1}{20} \cdot \frac{1}{20} = \frac{18}{20^7} \]

\[ p(D | B) = \frac{2}{20} \cdot \frac{2}{20} \cdot \frac{2}{20} \cdot \frac{2}{20} \cdot \frac{2}{20} \cdot \frac{1}{20} \cdot \frac{2}{20} = \frac{64}{20^7} \]

\[ p(D | C) = \frac{1}{20} \cdot \frac{1}{20} \cdot \frac{1}{20} \cdot \frac{1}{20} \cdot \frac{1}{20} \cdot \frac{1}{20} \cdot \frac{1}{20} = \frac{1}{20^7} \]
Example 12  (Continued)

Let us assume that all dice are equally likely, i.e.,

\[ p(A) = p(B) = p(C) = \frac{1}{3}. \]

Then, we can compute

\[
p(D) = p(A)p(D \mid A) + p(B)p(D \mid B) + p(C)p(D \mid C) \\
= \frac{1}{3} \cdot \frac{18}{20^7} + \frac{1}{3} \cdot \frac{64}{20^7} + \frac{1}{3} \cdot \frac{1}{20^7} \\
= \frac{83}{3 \cdot 20^7}
\]
Model comparison - Examples

- **Example 12**  (Continued)

And, finally, we can compute:

\[
p(A \mid D) = \frac{p(A) \cdot p(D \mid A)}{p(D)} = \frac{1/3 \cdot 18/20^7}{83/(3 \cdot 20^7)} = \frac{18}{83}
\]

\[
p(B \mid D) = \frac{p(B) \cdot p(D \mid B)}{p(D)} = \frac{1/3 \cdot 64/20^7}{83/(3 \cdot 20^7)} = \frac{64}{83}
\]

\[
p(C \mid D) = \frac{p(C) \cdot p(D \mid C)}{p(D)} = \frac{1/3 \cdot 1/20^7}{83/(3 \cdot 20^7)} = \frac{1}{83}
\]