Notes on probability

Statistics and Linguistic Applications
Hale
February 8, 2010

# Motivation

Probability is the mathematical theory of *chance*. In experiments, financial markets and in dealing with the real world generally, there are aspects of reality that we cannot control or ascertain in advance.
Typical examples are games...

- coin flips $\{H, T\}$
- dice $\{1, 2, 3, 4, 5, 6\}$
- the suite of playing cards drawn from a pack $\{\text{hearts}, \text{spades}, \text{diamonds}, \text{clubs}\}$

...but there are some linguistic examples as well:

1. presence or absence of optional complementizer in English relative clause
2. past-tense "go" realized by child as "goed" or "went"
3. part of speech of a word
4. number of optional postmodifiers of an NP
5. duration of silent pause while holding floor in....uh....discourse

Each of these has a **sample space** of possible outcomes. These outcomes are just labels – names for how things could turn out. And the sample space is a **set**. This set that the outcomes inhabit may be finite or infinite (i.e. under Chomskyan competence/performance there could be an infinite number of optional postmodifiers). The sample space is sometimes indicated by the Greek letter $\Omega$ ("omega").

An **event** is a subset $A$ of the sample space. 'Simple' or 'elementary' events like a coin coming up heads are singletons. But the event defined as drawing one of the clubs incorporates 13 possible elementary events. How many ways are there to roll a 7 with a pair of 6-sided dice?

# Set-theoretic basis

1. $A \cup B$ "either $A$ or $B$ or both"
2. $A \cap B$ "both $A$ and $B$"
3. $A'$ "not $A$" (notated as an overbar as in $\overline{A}$ in M&S)
4. $A - B = A \cap B'$ "$A$ but not $B$"

$A$ and $B$ disjoint i.e. $A \cap B = \emptyset$ is called "mutually exclusive."

## Axioms of probability

As Krenn and Samuelsson succinctly put it in the Linguist's Guide to Statistics, a **probability measure** is a function $P$ from events in the sample space $\Omega$ to real numbers in $[0, 1]$ satisfying these properties

1. $P(A) \geq 0$ for each event $A \subseteq \Omega$
2. $P(\Omega) = 1$
3. $A \cap B = \emptyset \;\Rightarrow\; P(A \cup B) = P(A) + P(B)$

The events themselves must also have the right structure, they must include the impossible event $\emptyset$, the certain event $\Omega$ itself and be closed under union $\cap$ and complementation (the $'$).

Notation: capital letters for events, lowercase letters for outcomes.

### Stability of Relative Frequency

Some subscribe to a **frequentist interpretation**: a probability is an ideal around which relative frequency stabilizes after a large ($n$) number of trials. For instance, after a large enough number of coin tosses the proportion of tails should get closer and closer to $\frac{1}{2}$. That such a limit exists is the idealization of probability theory.

## Conditional probability

If $A$ and $B$ are two events, and $A$ doesn't have zero probability, then it makes sense to talk about the **conditional probability** of $B$ given that $A$ happened.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1}$$

E.g the conditional probability that the next word is "bucket" given that the last two were "kick" and "the." Conditionalizing zooms in, replacing the entire sample space with just the conditioning event $A$. (M&S figure 2.1) What is the probability that a single toss of a 6-sided die will result in a number less than 4, given that the toss resulted in an *odd* number?

Some neat theorems on conditional probability:

### What independence means

If conditioning on some event fails to change the probability of another event, the two are said to be **independent**, i.e. $P(B|A) = P(B)$. This state of affairs holds when

$$P(A \cap B) = P(A)P(B) \tag{2}$$

That is, when the probability of two events both happening equals the product of each event occuring on its own, then those two events are independent of one another. To make an **independence assumption** is to treat $P(B|A)$ as $P(B)$, thus presuming that $A$ and $B$ are independent.

### Chain rule

By the definition of conditional probability the probability of the conjunction of two events can be rewritten as the product of the first event's probability, times the conditional probability of the second given the first.

$$P(A \cap B) = P(A)P(B|A) \tag{3}$$

Equation 3 is sometimes known as the **multiplication rule**. More generally, the probability of the conjunction of events $A_1, A_2, ..., A_n$ can be rewritten using the chain rule

$$P(A_1 \cap \cdots \cap A_n) = P(A_1|A_2...A_n) \cdot P(A_2|A_3...A_n) \cdot P(A_3|A_4...A_n) \cdot ... \cdot P(A_{n-1}|A_n) \cdot P(A_n) \tag{4}$$

The order of choosing the $A$s is irrelevant. For instance, one might write the probability of an 8-letter word actually being "linguist" from left to right using conditional probabilities of individual letters given shorter substrings:

$$\begin{aligned} &P(W_{1,8} = \texttt{linguist}) = \\ &\quad P(W_1 = \texttt{l})P(W_2 = \texttt{i}|\texttt{l})P(W_3 = \texttt{n}|\texttt{li})P(W_4 = \texttt{g}|\texttt{lin})P(W_5 = u|\texttt{ling}) \times \\ &\quad P(W_6 = \texttt{i}|\texttt{lingu})P(W_7 = \texttt{s}|\texttt{lingui})P(W_8 = \texttt{t}|\texttt{linguis}) \end{aligned}$$

The multiplication occurs in opposite order, as compared to 4. Note that $W_{1,8} = W_1 \cap W_2 \cap \cdots \cap W_8$ and conjunction is commutative, so we may select conjuncts in any order.

### Bayes' rule

This one is helpful in turning around the events on either side of the conditional probability vertical bar symbol.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

This is useful for finding the most likely "posterior" $P(A|B)$ even when you can't calculate $P(B)$. For instance, say you want to choose the Hypothesis that maximizes the conditional probability of the Hypothesis given the available Evidence as in equation 6

$$\arg\max_{Hypothesis} P(Hypothesis|Evidence) = \arg\max_{Hypothesis} \frac{P(Evi|Hyp)P(Hyp)}{P(Evi)} \tag{6}$$

Since we're talking about the same Evidence in all cases, $P(Evidence)$ stays the same as we try different Hypotheses. Thus for purposes of maximizing the conditional, we can forget about the constant denominator since it scales all of our conditional probabilities by the same amount.

$$\arg\max_{Hypothesis} P(Evi|Hyp)P(Hyp) \tag{7}$$

Typical application: automatic speech recognition where *Evi* is acoustic/phonetic measurements and *Hyp* is a string of letters as in the $W_{1,8}$ = `linguist` example.

## Now You Try

Table 5.2. *Numbers of monolingual or bilingual adults in two hypothetical populations cross-tabulated by sex*

| | Male | Female | Total |
|---|---|---|---|
| *Population A* | | | |
| Bilingual | 2 080 | 1 920 | 4 000 |
| Monolingual | 3 120 | 2 880 | 6 000 |
| | 5 200 | 4 800 | 10 000 |
| *Population B* | | | |
| Bilingual | 2 500 | 1 500 | 4 000 |
| Monolingual | 2 700 | 3 300 | 6 000 |
| | 5 200 | 4 800 | 10 000 |

Figure 1: Made-up data somewhat similar to Totonac population in Central Mexico

1. Using the table in figure 1, calculate two conditional probabilities for population B. Calculate $P$(male|bilingual) and $P$(female|bilingual).

2. A box contains 3 blue and 2 red marbles, while another box contains 2 blue and 5 red marbles. A marble drawn at random from one of the boxes turns out to be blue. What is the probability that it came from the first box?

3. (M&S p59 exercise 2.4) Are $X$ and $Y$ as defined in the following table independently distributed?

| | | | | |
|---|---|---|---|---|
| x | 0 | 0 | 1 | 1 |
| y | 0 | 1 | 0 | 1 |
| $P(X = x \cap Y = y)$ | 0.32 | 0.08 | 0.48 | 0.12 |