A Concise Introduction to Probability

(Revised 20 February 2004)

1. Outcomes, Sample Spaces, and Events

When an experimenter (physical scientist, social scientist, student doing a class project, or whatever) performs an experiment, there is a set of potential *outcomes*. What this set is depends on the nature of the experiment. The set of all possible outcomes is called the *sample space*. Usually, the experimenter is going to examine the data to see whether the outcome has some characteristic. The set of outcomes having the characteristic of interest is called an *event*.

The sample space is just the set of all possible outcomes. An event is a subset of the sample space.

Example 1. You collar a registered voter coming out of the Democratic Precinct Caucuses in the 46^{th} Legislative District of Washington. You ask the voter whether he or she was elected his/her precinct's delegate to the legislative district caucus.

> The *sample space* is the set of registered voters in the 46^{th} district who attended the Democratic precinct caucus for the voter's precinct.
>
> The *outcomes* are the registered voters in the 46^{th} district who attended the Democratic precinct caucus for the voter's precinct.
>
> The *event* is that the voter was elected a delegate to the 46^{th} district legislative district caucus.

Example 2. A marine chemist takes a 1-deciliter sample of sea water from Puget Sound, and analyzes it to determine whether the concentration of polychlorobiphenols in the sample exceeds 1 part per million.

> The *outcomes* are all possible 1-deciliter samples of Puget Sound sea water.
>
> The *sample space* is the set of all these outcomes.
>
> The *event* is that the concentration of polychlorobiphenols in the sample of sea water exceeds 1 part per million. This is a subset of the sample space.

Note that the sample space in this example is much larger than that in Example 1. It is also more poorly defined (what exactly are the boundaries of Puget Sound, anyway?), but the marine chemist doesn't give a hoot about that.

Example 3. A social scientist interviews families of murder victims. The social scientist asks whether any member of the family knew the perpetrator before the murder.

> The *sample space* consists of all families interviewed.

The *outcomes* are the families interviewed.

The *event* is that one or more family members acknowledges knowing the perpetrator in advance.

Note 1: By now you have noticed that asking what the outcomes are, and asking what the sample space is, are just two different ways of asking the same question. *Sample space* is just a name given to the set of outcomes.

Note 2: There are a lot more murder victims than those whose families were interviewed by our social scientist. The statistics folk have a term for this: the set of all families of murder victims is a *population*. The set of families actually interviewed by our researcher is a *sample* selected from that population. Hence the term "sample space". Statistics is the business of inferring information about a population from data collected from a sample – and estimating the accuracy of that inference. That's all we'll say about statistics here.

Example 4. Choose a number at random from the set {1, 2, 3, …, 100}. Is it even?

The sample space is {1, 2, 3, …, 100}. The outcomes are the integers in the range 1 to 100. The event is that the selected number is even.

Equiprobable Outcomes, and Simple Probability

In Example 4, by saying "Choose a number at random…" we implied that each of the 100 outcomes was equally likely (*equiprobable*). In Examples 1 through 3, that all outcomes were equally likely was subtly implied. If there are N outcomes in the sample space, then the probability that any specific outcome is selected is 1/N. In example 4, if you choose a number at random from {1, 2, 3, …, 100}, the probability that you choose 37 is 1/100.

(Not all real-world experiments result in equiprobable outcomes. Suppose our marine chemist had taken the bulk of the samples a short distance offshore from the outflow of the sewage treatment plant. Then all possible samples from Puget Sound would not be equiprobable – and the experimenter would have introduced *bias* into the results. But in these notes, we'll only consider equiprobable situations.)

Now its easy to assign a probability to an event: since an event is a subset of the sample space, just add up the probabilities of the outcomes that make up the event. In the equiprobable case, all you have to do is count the outcomes in the event and divide by the total number of outcomes in the sample space.

Example 5. In Example 4, what is the probability that a selected number is even? The event is the set of even numbers in the range 1 to 100. There are 50 elements in this set, and there are 100 elements in the sample space. The probability of "even" is therefore 50/100 = ½.

Example 6. In Example 5, what is the probability that the selected number is less than 11? It's 10/100.

Example 7. On February 7, 2004, the Democratic Party held caucuses in 6,497 precincts in Washington State, at which 23,505 delegates to the next level caucuses (legislative district) were elected. Here is an excerpt from the delegate count. King County includes the Seattle area, and the 46th District is in north Seattle. "Uncom." means "Uncommited".
(Source: http://www.wa-democrats.org/tech/caucus_results_04.php?report=countydel .)

Jurisdiction	Total	Clark	Dean	Edwards	Kerry	Kucinich	Sharpton	Uncom.
Washington State	23505	770	7048	1571	11370	1928	19	799
King County	9562	268	3297	617	4247	821	7	305
46th District	949	14	368	54	372	128	0	13

Pick a delegate at random.

a. What is the probability that the delegate supports Kerry? 11370/23505, or 48.37%.
b. What is the probability that the delegate is from King County? 9562/23505 = 40.68%.
c. What is the probability that the delegate supports a candidate who did not withdraw before 20 February 2004? That's Edwards, Kerry, and Sharpton, who have a total of 12960 delegates. Therefore the answer is 12960/23505 = 55.14%.
d. What is the probability that a delegate supports Dean, given that the delegate is from King County? Oops, we just changed the sample space! We are restricting our attention to the space of King County delegates, of which there are 9562, of which 3297, or 34.48%, support Dean. The answer is 0.3448.
e. What is the probability that a delegate resides in the 46th District, given that the delegate supports Dean? Oops, we just changed the sample space again! Now our sample space consists of all the Dean delegates in the state (7048 of them). Of these, 368 are from the 46th District. So the probability is 368/7048 = 0.0522.

Conditional Probability

In Examples 7d and 7e, I snuck in the concept of Conditional Probability. Despite the fancy name, it is nothing more than deciding to inspect a new sample space that is a subset of the original sample space. The "condition" defines our new, smaller sample space that is a subset of the original sample space, so it is an event! In 7d, the "condition" event is "the delegate is from King County". In 7e, the "condition" event is "the delegate supports Dean".

Some Formulas that Describe What We Have Been Doing

Let S be a sample space. Other letters will represent events in S (i.e. subsets of S). If E is an event, we write $P(E)$ for the probability of E. At this point, each of these should be clear to you:

a. $P(E)$ = (number of elements in E)/(number of elements in S)
b. $P(S) = 1$
c. $P(\emptyset) = 0$

d. $P(A^c) = 1 - P(A)$
e. If $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$

We just talked about conditional probability. The conditional probability of event B given event A is denoted by $P(B|A)$. How did we compute it? First, we noted that we are going to take A as the new sample space. Then we looked at the part of B that is relevant to our new sample space; that's just $B \cap A$. We counted the outcomes in $B \cap A$, and divided by the number of outcomes in A:

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

Independent Events

Here are some observations on the data in Example 7:
P(Edwards) = 1571/23505 = .0668
P(Edwards | King County) = 617/9562 = .0645
P(Edwards | 46th District) = 54/949 = .0569

These data show that the probability that a delegate supports Edwards *depends* (although not by very much) on the conditions imposed – whether the delegate was selected from the entire state, the county, or the district. Edwards and King County are *dependent events*. Edwards and 46th District are dependent events.

On the other hand, look at Example 2:
P(even) = ½ (as we saw before)
P(even | x > 50) = ½ too, since 25 of the numbers in {51, …, 100} are even.
"Even" and "Greater than 50" are *independent events*.

That is, events A and B are independent if (and only if, since this is a definition!)

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

Using the above formula for $P(B \mid A)$, we have

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

which can be written more simply as $P(B \cap A) = P(B) \bullet P(A)$.

Exercises:

(1) Make up two simple (but nontrivial) probability questions and two conditional probability questions that can be answered using the Washington State data provided above, and post them as a response to this topic.

(2) Look at another student's posting, and answer the questions therein, and post your answers.
(3) Check the answers to your questions that were posted by other students, and respond, indicating whether you agree.

[To spread the fun around more or less evenly, for "another student" please choose the student next in the alphabetical listing, (or the next, etc, if the next student hasn't posted yet). I suppose I could have set up "Study Groups" for this, but that seemed like overkill.]

(4) Can you generalize formula (e) to the case when If $A \cap B \neq \varnothing$?