3rd Class **6/9/10**

"Events with a-million-to-one odds happen 295 times a day in America."
Michael Shermer
Why People Believe Weird Things

Last time I asked for a 95% CI for the population of South Korea
[specifically, the early-2010 (updated as of 3/19/10) population of 48,636,068,
according to the U.S. Census on-line international database:
http://www.census.gov/ipc/www/idb/ranks.php]
[show link to IDB, including link to list of countries ranked by population size]
[note Shermer is wrong; actually happen 310 times a day in the U.S. now—pop. 310,232,863]

results on South Korea population question from last class: of the 12 people in class, 2 people had too high of a confidence interval, while 5 people had too low of a confidence interval, and 5 people had the correct value of 48,636,068 within the confidence interval—narrowest correct interval was 40 to 60 million; widest correct interview was1 0 to 100 million (note one top bound would be the Earth's population of 6,825,869,677; in previous classes people have given really wide intervals like 0 to 50 billion, or open-ended intervals like 1 to infinity--which would have been a correct guess); so 42 percent of the class got it right (as opposed to ninety-five percent getting it right).
[earth population from http://www.census.gov/ipc/www/popclockworld.html; note this estimate is not without controversy]
[handout on guessing]

--Ch 3
introduces some important definitions and rules and shows how to calculate probabilities

f in the book means the number of times a particular value appears in the sample, or the absolute frequency
--often we are interested in, or only have access to, relative frequencies, i.e., the

proportion of the sample that is a particular value: $\frac{f}{n}$

--if data are grouped, i.e., assigned a common value even though they fall in a range around that data, then some error is introduced into the procedure for calculating means and variances; it is not error intrinsic to using frequency-type data

probability ≡ limiting relative frequency, i.e. Pr ≡ $\lim(\frac{f}{n})$ as n approaches ∞

There are several different ways to visualize probabilities diagrammatically

--tree diagrams

--lists of events/outcomes

an event E is a subset of the outcome set S; E can include multiple outcomes e: $Pr(E) = \Sigma Pr(e)$

S is also called the sample space: $Pr(S) = 1$

--Venn diagrams

consider make-up of a three-child family

when you use symbols, be sure to define them to avoid ambiguity and to distinguish symbols for events from symbols for outcomes

e.g. in problem in book, E stands for the event "at least two girls" and is equivalent to the outcome list $E = \{e_4, e_6, e_7, e_8\}$ as the outcomes are defined

probabilities are given either as fractions or decimals bounded by 0 and 1 inclusive, or as percentages bounded by 0 and 100% inclusive; e.g., $p = \frac{9}{10}$ or .9 or 90%

can always express probabilities in the form of <u>odds</u> instead:

e.g. probability of getting a "1" on a fair dice throw is $p = \frac{1}{6}$ or .167 or 16.7%

odds for getting a "1" are defined as one to five: $d = \frac{1}{5}$ or .20 or 20%

$$\text{probability, } p = \frac{d}{d+1} = \frac{\frac{1}{5}}{\frac{6}{5}} = \frac{1}{6}$$

$$\text{odds, } d = \frac{p}{1-p} = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5}$$

Question: If I were having a baby and I offer you even odds on placing bets with me on whether my baby is going to be a boy or a girl, which should you pick, boy or girl?

answer: actual odds of bearing a boy are 106 to 100, or 1 to .94, or, in probability terms, 51.5%. Since I am offering a return of $1 to every $1 placed on the winning side, you would get more than the fair return by picking boy, and less than the fair return by picking girl, therefore pick boy.

[note. Diet apparently affects gender; high levels of magnesium, potassium, and calcium produce more boys. Nonsmokers tend to produce more boys. But vegetarians are apparently more likely to have daughters (even though they have diets with sufficient levels of those three minerals, and are also less likely to smoke). A 1998 study as reported in the London Times, 8/7/00, found 85 boys to 100 girls for vegetarian women and the standard 106 to 100 for nonvegetarians--the same as the Great Britain average.]

Introduce commonly-used symbols from set theory. If A_1 and A_2 are sets, then

the union of A_1 and $A_2 \equiv A_1 \cup A_2$, includes everything in A_1 or A_2

the intersection of A_1 and $A_2 \equiv A_1 \cap A_2$, includes everything in A_1 and A_2

[show Venn diagrams]

If A_1 and A_2 are mutually exclusive, or disjoint, then

$$Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2)$$

this is a special case of the general rule, or addition law of probabilities, that:

$$Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2) - Pr(A_1 \cap A_2)$$

(if A_1 and A_2 are mutually exclusive, then $Pr(A_1 \cap A_2) = 0$)

note generalizability of the addition law of probabilities to any number of sets:

$$Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2) - Pr(A_1 \cap A_2)$$ [show Venn]

$$Pr(A_1 \cup A_2 \cup A_3) = Pr(A_1) + Pr(A_2) + Pr(A_3) - Pr(A_1 \cap A_2) - Pr(A_2 \cap A_3) - Pr(A_1 \cap A_3) + Pr(A_1 \cap A_2 \cap A_3)$$ [show Venn]

and can keep going by alternating + and - signs by line

It is often easier to calculate the probability of an event E by first calculating the probability of the <u>complement</u> of E: $\overline{E}$ ≡ the set of points that are <u>not</u> in E

since $Pr(E) + Pr(\overline{E}) = 1$,

then $Pr(E) = 1 - Pr(\overline{E})$, which turns out to be a very useful formula

Often we are interested in the <u>conditional probability</u> of something happening. In other words, for those cases where A_1 has occurred, how often will A_2 occur? This is called the conditional probability of A_2, given A_1, and is denoted as $Pr(A_2 | A_1)$. Note that:

$$Pr(A_2 | A_1) = Pr(A_2 \cap A_1) \ / \ Pr(A_1)$$

notice this can be expressed in a couple of other ways:

$$Pr(A_2 \cap A_1) = Pr(A_1)Pr(A_2 | A_1)$$,

which gives us an easy way to calculate the probability of the intersection of A_1 and A_2

also, switching A_1 and A_2 in place in the formula:

$$Pr(A_1 \cap A_2) = Pr(A_2)Pr(A_1 \mid A_2),$$

so:

$$Pr(A_2 \cap A_1) = Pr(A_2)Pr(A_1 \mid A_2) ,$$

since it doesn't matter what order sets A_1 and A_2 are listed in the intersection set

The multiplication law of probabilities is also generalizable to any number of sets:

$$Pr(A_1 \cap A_2) = Pr(A_2 \mid A_1) * Pr(A_1)$$

$$Pr(A_1 \cap A_2 \cap A_3) = Pr(A_3 \mid A_2 \cap A_1) * Pr(A_2 \mid A_1) * Pr(A_1)$$

Statistical independence is defined as follows:

A_2 is called statistically independent of A_1 if $Pr(A_2 \mid A_1) = Pr(A_2)$

We will generally just refer to this as independence from now on rather than as statistical independence. Idea of the boundaries of whether or not we can prove that two occurrences are truly independent of each other in causality.

Note that if A_2 is independent of A_1, then A_1 must be independent of A_2; i.e., symmetry must hold:

remember $Pr(A_1 \cap A_2) = Pr(A_1)Pr(A_2 \mid A_1)$

If A_2 is independent of A_1, then $Pr(A_2 \mid A_1) = Pr(A_2)$ and

$$Pr(A_1 \cap A_2) = Pr(A_1) * Pr(A_2)$$

dividing by $Pr(A_2)$, we get:

$$\frac{Pr(A_1 \cap A_2)}{Pr(A_2)} = Pr(A_1)$$

where the left side can be expressed as the expression for conditional probability:

$$Pr(A_1 \mid A_2) = Pr(A_1)$$

So we can say A_1 and A_2 are independent of each other, or just say that A_1 and A_2 are independent

So when events A_1 and A_2 are independent, the multiplication law reduces to:

$$Pr(A_1 \cap A_2) = Pr(A_1) * Pr(A_2)$$

and, by substituting the above in, the addition law becomes:

$$Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2) - Pr(A_1) * Pr(A_2)$$

To illustrate the use of independence and the usefulness of the idea of the complement, consider the Birthday problem (ignore leap year complication):

group size	prob. of no repeated birthday	prob. of at least one repeated bday
1	1	0
2	$1 * \frac{364}{365} = .997260$	1 - .997260 = .002740
3	$1 * \frac{364}{365} * \frac{363}{365} = .991796$	1 - .991796 = .008204
…		
366	0	1

question: How many people need to be in a group in order for the probability to be more than 50% that at least two of them have the same birthday?

answer: 23 people (50.7%--jumps from 47.6% with 22 people)

question: In our room of 13 people, what is the probability that at least two of us have the same birthday?

answer: 19%

question: How many people need to be in a group in order for the probability to be more than 50% that at least one other person shares my birthday?

answer: consider a group of n people. The probability is $\frac{1}{365}$ that any one person will share my birthday, thus the probability is $1 - \frac{1}{365}$ that any one person will not share my birthday. So the probability is $(1 - \frac{1}{365})^n$ that none of the n people will share my birthday. Therefore the probability is $1 - (1 - \frac{1}{365})^n$ that at least one of the n people will share my birthday.

So need to find n large enough so that $1 - (1 - \frac{1}{365})^n \geq .5$

Can solve this by substitution, but neater to solve it analytically using logarithms:

$$1 - (\frac{364}{365})^n \geq \frac{1}{2}$$

$$\frac{1}{2} \geq (\frac{364}{365})^n$$

$$\log(\frac{1}{2}) \geq n\log(\frac{364}{365})$$

$$-.693 \geq -.0027435n$$

$$.0027435n \geq .693$$

$$n \geq 253$$

question: In our group of 12 people (not counting me), what is the probability that at least one other person shares my birthday?

$$1 - (1 - \frac{1}{365})^{12} = 3\%$$

[note. as with birth probabilities, people are not randomly distributed across birthdates--fewer births occur in the spring]

next class, finish reading Ch. 3 and start Ch. 4-- will go over Bayes' theorem and consider the relevance of probability theory for decisionmaking; will work some more problems in class