MSPP PPOL 501–02 & -06: Fall 2014
Course Notes #5: Basic Probability Rules
Professor Carolyn Hill

** Note: The notes on probability draw on the following textbooks:

Moore, David S., and George P. McCabe. 2003. *Introduction to the Practice of Statistics*, 4th ed. (New York: W.H. Freeman & Company).

Runyon, Richard P., Kay A. Coleman, and David J. Pittenger. 2000. *Fundamentals of Behavioral Statistics*, 9th Ed. (Boston: McGrawHill).

Weiss, Neil A. 2002. *Introductory Statistics*, 6th Ed. (Boston: Addison Wesley)

Wonnacott, Thomas H., and Ronald J. Wonnacott. 1990. *Introductory Statistics*, 5th Ed. (New York: John Wiley & Sons).

I. PROBABILITY BASICS

- Why study probability?

  -- As many authors and observers point out, we all have a basic sense of probability and often make decisions based on this.

  -- Moore & McCabe (*Introduction to the Practice of Statistics* 2003, pp. 284-285) describe different ways that basic probability shows up in our lives:

  > "Probability theory originated in the study of games of chance….It is only a mild simplification to say that probability as a branch of mathematics arose when seventeenth-century French gamblers asked the mathematicians Blaise Pascal and Pierre de Fermat for help….Careful measurements in astronomy and surveying led to further advances in probability in the eighteenth and nineteenth centuries because the results of repeated measurements are random and can be described by distributions much like those arising from random sampling.…Now, we employ the mathematics of probability to describe the flow of traffic through a highway system, the Internet, or a computer processor; the genetic makeup of individuals or populations; the energy states of subatomic particles; the spread of epidemics or rumors; and the rate of return on risky investments. Although we are interested in probability because of its usefulness in statistics, the mathematics of chance is important in many fields of study."

  -- We want to move beyond our subjective sense of probability (because, being human, we aren't always so good at processing things clearly or fairly), and develop more formally the properties of probability that will stand you in good stead.

- How is probability defined and what are its basic properties?

  Probability of an event = P = Pr = $\frac{f}{N} = \frac{Number\ of\ \ ways\ an\ event\ can\ occur}{Total\ number\ of\ \ possible\ outcomes}$

  Define some terms:

Sample Space :	A full listing of all possible results or events
Event:	A defined subset of the sample space

Basic Properties:

1. The probability of an event is always between 0 and 1.
2. The probability that an event cannot occur is 0 (i.e., "impossible").
3. The probability of an event that must occur is 1 (i.e., "certain").
4. The probability that an event will *not* occur is equal to 1 minus the probability that it will occur: *P (not A)* = 1- *P(A)*
   (in other words, *P(not A)* is the **complement** of *P(A)*

Example 1: What's the probability that a youth selected at random in the NLSY97 will be a girl (1=male, 2=female).

```
. tabulate gender97

      youth |
gender 1997 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      4,599       51.19       51.19
          2 |      4,385       48.81      100.00
------------+-----------------------------------
      Total |      8,984      100.00
```

$$\text{P(Girl)} = \frac{f}{N} = \frac{4{,}385}{8{,}984} = 0.4881$$

Example 2: If you roll a fair die what's the probability of rolling a 2?

(a) List all possible outcomes: 1
2
3
4
5
6

(b) List all possible ways the event of interest (roll a 2) could occur: 2

So, there's one way that rolling a 2 could occur, out of 6 possible outcomes:
$\frac{1}{6} = 0.1666667$.

Example 3: What's the probability that a child in a one-child family will be a girl?

Assuming that the birth of a girl or a boy is equally likely, you probably would immediately answer ½, or 0.50. Using the notation above:

(a) List all possible outcomes: **Girl**
Boy

(b) List all possible ways the event of interest (birth of a girl) could occur: **Girl**
So, there's one way that the birth of a girl could occur, out of two possible outcomes: $\frac{1}{2} = 0.50$.

In this simple example, it seems like we're making thing harder than they have to be. Let's make things a little more interesting:

Example 4: What's the probability that both children in a two-child family will be girls?

(a) List all possible outcomes: **GG**
GB
BG
BB

(b) List all possible ways the event of interest (birth of 2 girls) could occur: **GG**
Now, there's one way that the birth of two girls could occur, out of four possible outcomes: $\frac{1}{4} = 0.25$

Example 5: The FBI publishes data on people arrested each year. Records for one year show that 79.6% of the people arrested were male, and 18.3% were under 18 years of age. If a person arrested during that year is selected at random, what's the probability that the person is female?

Pr(Female) = (1 - 0.796) = 0.204.

- Fascinating, huh? We're usually interested in more complicated probabilities, and so we'll rely on some basic probability rules to help get us there.

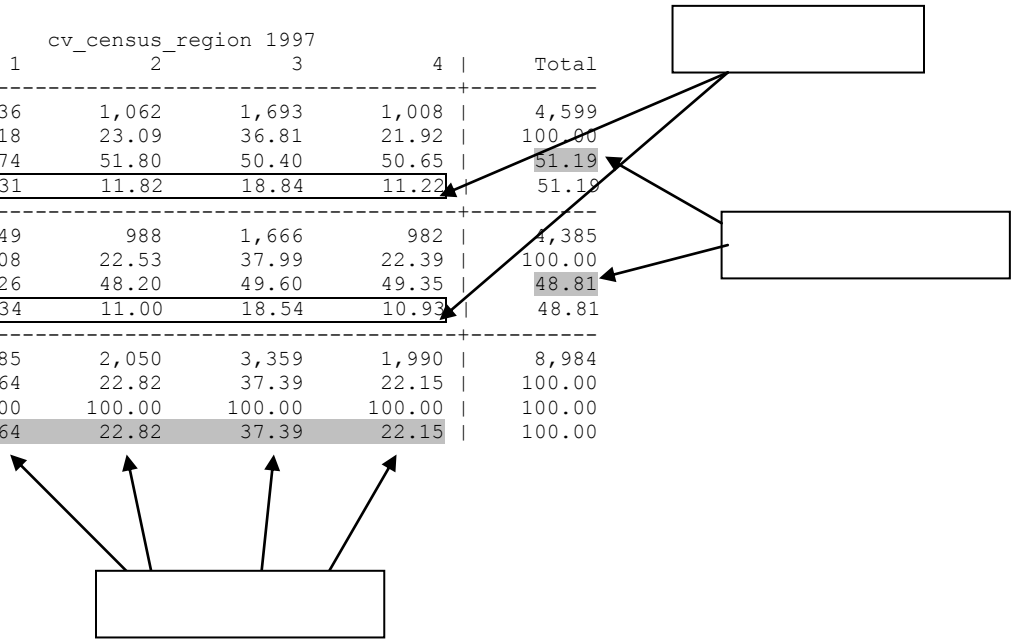
II. MARGINAL, JOINT, AND CONDITIONAL PROBABILITIES

- When we talked about cross-tabs, you already were introduced to these concepts. They're important in the study of probability, so we'll go over them again now, adding in some notation.

  Using the gender by region table: (1=NE, 2=NC, 3=S, 4=W)
  (1=male; 2=female)

```
tabulate gender region97, row col cell
+--------------------+
| Key                |
|--------------------|
|     frequency      |
|  row percentage    |
| column percentage  |
|  cell percentage   |
+--------------------+
```

key!sex (symbol) 1997	cv_census_region 1997 1	2	3	4	Total
1	836	1,062	1,693	1,008	4,599
	18.18	23.09	36.81	21.92	100.00
	52.74	51.80	50.40	50.65	51.19
	9.31	11.82	18.84	11.22	51.19
2	749	988	1,666	982	4,385
	17.08	22.53	37.99	22.39	100.00
	47.26	48.20	49.60	49.35	48.81
	8.34	11.00	18.54	10.93	48.81
Total	1,585	2,050	3,359	1,990	8,984
	17.64	22.82	37.39	22.15	100.00
	100.00	100.00	100.00	100.00	100.00
	17.64	22.82	37.39	22.15	100.00



-- Marginal probabilities. E.g.: *P* (female) = 0.4881
P (from the south) = 0.3739

-- Joint probabilities: E.g.: *P* (female and from the South) = 0.1854
P (male and from the West) = 0.1122

-- Conditional probabilities (definition): the probability that event B occurs, given that event A has occurred:

P(B|A) = "probability of B given A"

Example 6:

```
tabulate gender region97, row col cell
+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+

   key!sex |
  (symbol) |            cv_census_region 1997
      1997 |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |       836      1,062      1,693      1,008 |     4,599
           |     18.18      23.09      36.81      21.92 |    100.00
           |     52.74      51.80      50.40      50.65 |     51.19
           |      9.31      11.82      18.84      11.22 |     51.19
-----------+--------------------------------------------+----------
         2 |       749        988      1,666        982 |     4,385
           |     17.08      22.53      37.99      22.39 |    100.00
           |     47.26      48.20      49.60      49.35 |     48.81
           |      8.34      11.00      18.54      10.93 |     48.81
-----------+--------------------------------------------+----------
     Total |     1,585      2,050      3,359      1,990 |     8,984
           |     17.64      22.82      37.39      22.15 |    100.00
           |    100.00     100.00     100.00     100.00 |    100.00
           |     17.64      22.82      37.39      22.15 |    100.00
```

e.g.: *P(*from south | female) = probability of being from the south, conditional on being female
= 1666/ 4385 = 0.3799

P (from west | male) = 1008 / 4599 = 0.2192

P (female | from northeast) = 749 / 1585 = 0.4726

P (male | from south) = 1693 / 3359 = 0.50

Stated generally: $P(B \mid A) = \dfrac{P(A \& B)}{P(A)}$

III. A LITTLE MORE CONDITIONAL PROBABILITY: *P*(B|A)

- Let's invoke your friend the dice  to provide a little more insight into the conditional probability formula:

Example 7:

You roll the die once. What's the probability that you roll a 5, given that the die comes up with an odd number?: *P*(roll a 5 | odd number).

-- think for a minute: What do you think the answer is without going through the calcs? Two ways to figure this out:

(1) (a) List all possible outcomes: 1
3
5

(b) List all possible ways the event of interest (roll a 5) could occur: 5

There's one way that you could roll a 5, conditional on an odd number being rolled, so: $\frac{f}{N} = \frac{1}{3} = 0.33333$

$$(2)\ P(B \mid A) = \frac{P(A \,\&\, B)}{P(A)} = \frac{P(roll\ a\ 5\ and\ odd\ number)}{P(odd\ number)} = \frac{1/6}{1/2} = \left(\frac{1}{6} * \frac{2}{1}\right) = \frac{2}{6} = \frac{1}{3}$$

- O.K., let's go back to the frequency table and look at this formula just a little bit more.

Example 8:

```
tabulate gender region97, row col cell
   key!sex |
  (symbol) |            cv_census_region 1997
      1997 |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |       836      1,062      1,693      1,008 |     4,599
           |     18.18      23.09      36.81      21.92 |    100.00
           |     52.74      51.80      50.40      50.65 |     51.19
           |      9.31      11.82      18.84      11.22 |     51.19
-----------+--------------------------------------------+----------
         2 |       749        988      1,666        982 |     4,385
           |     17.08      22.53      37.99      22.39 |    100.00
           |     47.26      48.20      49.60      49.35 |     48.81
           |      8.34      11.00      18.54      10.93 |     48.81
-----------+--------------------------------------------+----------
     Total |     1,585      2,050      3,359      1,990 |     8,984
           |     17.64      22.82      37.39      22.15 |    100.00
           |    100.00     100.00     100.00     100.00 |    100.00
           |     17.64      22.82      37.39      22.15 |    100.00
```

- Using the *f*/N formula, we calculated that *P* (female | from northeast) = $\frac{f}{N} = \frac{749}{1{,}585} = 0.4726$
- Using the conditional probability rule, we then calculate the same number using the probabilities that STATA gave us in the table:

$$P(\textit{female} \mid \textit{northeast}) = \frac{P(\textit{female} \,\&\, \textit{northeast})}{P(\textit{northeast})} = \frac{0.0834}{0.1764} = 0.4728$$ (different than above due to rounding)

- This might have flown by quickly, so let's see where it comes from and why it makes sense:

$$P(\textit{female} \mid \textit{northeast}) = \frac{P(\textit{female} \,\&\, \textit{northeast})}{P(\textit{northeast})} = \frac{749 \,/\, 8{,}984}{1{,}585 \,/\, 8{,}984} = \left(\frac{749}{8{,}984} * \frac{8{,}984}{1{,}585}\right) = \frac{749}{1{,}585} = 0.4726$$

- It's important to know and be able to apply the conditional probability rule because you may not always have the information you need to calculate it using $\frac{f}{N}$.

Example 9:

The FBI issues reports on property crimes (among other things). They reports that 4.9% are committed in rural areas and 1.5% are burglaries (a type of property crime) committed in rural areas. Among the property crimes committed in rural areas, what percentage are burglaries?

-- The first step when approaching a problem like this is to write the words into notation that might be helpful to you:

* How can you dichotomize the information given?
  -- Rural Areas (R), and Not Rural Areas (NR)
  -- Buglaries (B), and Not Burglaries (NB)

* What's the probability that a property crime chosen at random would be in a rural area?:

$$P(\text{R}) = 0.049$$

* What's the probability that a property crime is a burglary committed in a rural area (i.e., a burglary & committed in a rural area)?

$$P(\text{burglary \& in a rural area}) = 0.015$$

* What are we trying to find out?: Among the property crimes committed in rural areas, what percentage are burglaries?

  -- Translate this into familiar phrasings: Given that property crimes are committed in rural areas, what's the probability that they are burglaries?:

  $P(\text{Burglary} \mid \text{Rural}) = ?$

$$P(B \mid R) = \frac{P(B \,\&\, R)}{P(R)} == \frac{0.015}{0.049} = 0.3061$$

  -- So, 30.61% of property crimes committed in rural areas are burglaries.

- Question: What percentage of burglaries are committed in rural areas?

TIP: Make a table:

IV. MULTIPLICATION RULE

- O.K., so we were just looking at the conditional probability rule: $P(B \mid A) = \dfrac{P(A \& B)}{P(A)}$

- What do we like to do with equations? MANIPULATE THEM. So:

$$P(B \mid A) = \frac{P(A \& B)}{P(A)} \quad \Leftrightarrow \quad \underbrace{P(A \& B) = P(A) * P(B \mid A)}$$

This formula is known as the **general multiplication rule** for probabilities.

Example 10:

For the 106th U.S. Congress, 18.7% of the members were senators and 45% of the senators were Democrats. What's the probability that a randomly selected member of the 106th Congress was a Democratic senator?

* Breathe.

* What do we know? P (senator) = 0.187
P (Democrat | senator) = 0.45

* What do we want to know? P (Democratic senator), i.e., P (Democrat & senator)

$$P(Democrat \& senator) = P(senator) * P(Democrat \mid senator) = (0.187)(0.45) = 0.08415$$

So, the probability that a randomly selected member of the 106th Congress was a Democratic senator is 0.08415.

TIP: Make a table:

Example 11:

An article in *Science News* (2000, v. 157) reported on research about the effects of regular Internet usage. According to the article, 36% of Americans with Internet access are regular Internet users (log on for 5+ hrs per week). Among regular Internet users, 25% say that the Web has reduced their social contact (e.g., talking with family and friends and going out).

What's the probability that a randomly selected American with Internet access is a regular Internet user who feels that the Web has reduced his or her social contact?

TIP: Make a table:

Example 12:

The U.S. population, broken down by region and attitude to legalization of marijuana, roughly turned out as follows:

Region	In Favor	Opposed
East	0.078	0.222
All except East	0.182	0.518

(a) What kind of probabilities are given in this table?

(b) What's the probability that an individual drawn at random will favor legalization?

(c) What's the probability that a randomly selected person from the East is in favor of legalization?

(d) Are region and attitude toward legalization (measured as above) independent?

(Note: two events are **independent** if knowing that one occurs does not influence the probability that the other occurs)

- **Independent events**: Two events are independent if:

$$P(A \& B) = P(A) * P(B) \text{ or, alternatively, if } P(B \mid A) = P(B)$$

  * If you can show that either of their two formulas is true (and if one is true then the other must also be true), then two events are independent.
    If two events are known to be independent, then you know that these formulas must be true.

  * The second formula provides some insight in words: for two independent events, knowing that one of them happened (A) will not affect the probability that the other will happen (B).

    i.e., knowing something about one doesn't tell you anything about the other

  * Generally, for *k* independent events,
    $P(\text{A\&B\&C\&D\&E}\ldots..) = P(A)*P(B)*P(C)*P(D)*P(E)*.....$

<u>Example 13: Roulette (a game of chance!):</u> 

A roulette wheel has 38 numbers: 18 are red, 18 are black, and 2 are green. When the wheel is spun, the ball is equally likely to land on any of the 38 numbers (i.e., each number has a 1/38 chance of having the ball land on it).

* What's the probability of landing on red?

* What's the probability of landing on black?

* What's the probability of landing on green?

In four plays at the wheel, what's the probability that the ball will land on green the first time and on black the second, third, and fourth times?

$$P(green \& black \& black \& black) = \frac{2}{38} * \frac{18}{38} * \frac{18}{38} * \frac{18}{38} =$$

(we'll look at tree diagrams after looking at the addition rule)….

V. ADDITION RULE: $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$

Note: $P(A \& B)$ is known as the ***joint probability*** of two events occurring.

Example 14: If you roll a fair die once, what's the probability of getting either a 2 or a 4?

Pr(rolling a 2 or a 4) = *Pr*(rolling a 2) + *Pr*(rolling a 4) – *Pr*(rolling a 2 and a 4)
= 1/6 + 1/6 - 0
= 2/6 = 1/3 = 0.333333

Question: why is *Pr*(rolling a 2 and a 4) = 0?

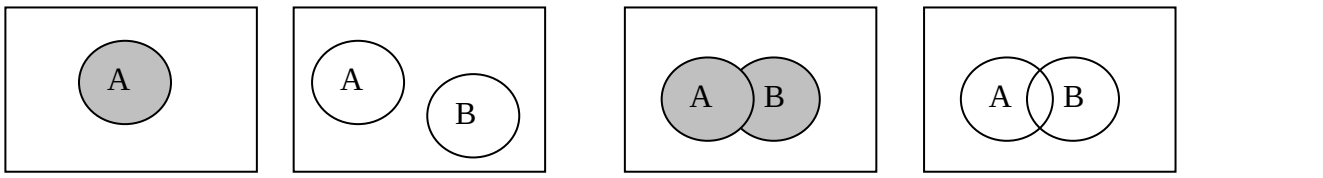
-- **Mutually Exclusive events:** Two (or more) events are mutually exclusive if at most one of them can occur, i.e., no two events have outcomes in common. (also known as "disjoint" events)

Example 15: Expanding Example 5 above: what's the probability that a person randomly selected is either male or under 18? (Extra info: 13.5% of arrestees in that year are males under 18 years of age).

Pr(male or under18) = *Pr*(male) + *Pr*(under18) – *Pr*(male and under18)
= 0.796 + 0.183 - 0.135
= 0.844

The probability that a person obtained is either male or under 18 is 0.844. (i.e., 84.4% of those arrested during the ear were either male, or under 18 years old, or both)

- This example shows that the subtracting off of the joint probability (in this case, both male and under 18) is necessary to avoid double-counting those groups in calculating the probability. Venn diagrams are often used to show this:



Event A | A & B are mut excl | A or B | A & B (intersection filled in)

Example 16: What's the probability that a youth selected at random from the NLSY97 will be a female or from the South?

Pr(female or from the South) = *Pr*(female) + *Pr*(from the South) – *Pr*(female and from the South)

```
tabulate gender region97, row col cell
+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+
   key!sex |
  (symbol) |            cv_census_region 1997
      1997 |         1          2          3          4 |     Total
-----------+----------------------------------------------+----------
         1 |       836      1,062      1,693      1,008 |     4,599
           |     18.18      23.09      36.81      21.92 |    100.00
           |     52.74      51.80      50.40      50.65 |     51.19
           |      9.31      11.82      18.84      11.22 |     51.19
-----------+----------------------------------------------+----------
         2 |       749        988      1,666        982 |     4,385
           |     17.08      22.53      37.99      22.39 |    100.00
           |     47.26      48.20      49.60      49.35 |     48.81
           |      8.34      11.00      18.54      10.93 |     48.81
-----------+----------------------------------------------+----------
     Total |     1,585      2,050      3,359      1,990 |     8,984
           |     17.64      22.82      37.39      22.15 |    100.00
           |    100.00     100.00     100.00     100.00 |    100.00
           |     17.64      22.82      37.39      22.15 |    100.00
```

Pr( ) = 0.1854

Pr( ) = 0.4881

Pr( ) = 0.3739

Pr(female or from the South) = *Pr*(female) + *Pr*(from the South) – *Pr*(female and from the South)
= 0.4881 + 0.3739 - 0.1854
= 0.6766

Example 17: The 2000 census allowed each person to choose from a long list of races. A separate category is available for ethnicity (Hispanic/Latino): Hispanics may be of any race. If we choose a resident of the U.S. at random, the 2000 Census gives these probabilities:

	Ethnicity	
Race	Hispanic	**Not Hispanic**
Asian	0.000	**0.036**
Black	0.003	**0.121**
White	0.060	**0.691**
Other	**0.062**	**0.027**

Q1: How would you verify that this table gives a legitimate assignment of probabilities?

Q2: What's the probability that a randomly chosen person is Hispanic?

Q3: What's the complement of a randomly-selected person being White, and what is the probability of this event occurring? (2 ways)

Q4: What's the probability that a randomly-selected person is a White non-Hispanic?

Q5: What's the probability that a randomly-selected person is White or non-Hispanic?

Q6: Are the events of "Hispanic" and "White" mutually exclusive?

Q7: Are the events of "Hispanic" and "White" independent?

VI: BAYES RULE

-- The conditional probability rule can be stated in this simple way, or (by simple substitution) it can be stated in a more complicated (but sometimes more useful) form.

-- Because P(A) = P(A & B) + P(A & not B), we can substitute this into the denominator.

-- Because the multiplication rule tells us that P(A & B) = P(B)P(A|B), we can substitute this into the numerator

$$P(B \mid A) = \frac{P(A \& B)}{P(A)} = \frac{P(B)P(A \mid B)}{P(A \& B) + P(A \& notB)}$$

-- We can then further apply the multiplication rule to substitute for both terms in the denominator. This creates:

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

-- This expression is called *Bayes' Rule*. It is named after the Reverend Thomas Bayes (1702-1761), who first derived it. It is often used in both game theory and in advanced statistics. But it can also sometimes be a useful formula for dealing with simple problems involving conditional probability.

Example:

Suppose you have a new test for HIV. You know that 0.76 % of the entire population has HIV. You also know that if someone is infected, there is a .99 probability that their test will be positive. In contrast, if they are not infected, there is a .015 probability of a positive test.

So far, this seems like a pretty good test for HIV. But you would also like to know what the probability of being infected is if their test comes back positive. We can use Bayes' rule to calculate that.

If B stands for being infected, and A stands for getting a positive test result, we know that:

P(B) = .0076
P(A|B) = .99
P(A|not B) = .015

Substituting these numbers into Bayes' rule produces:

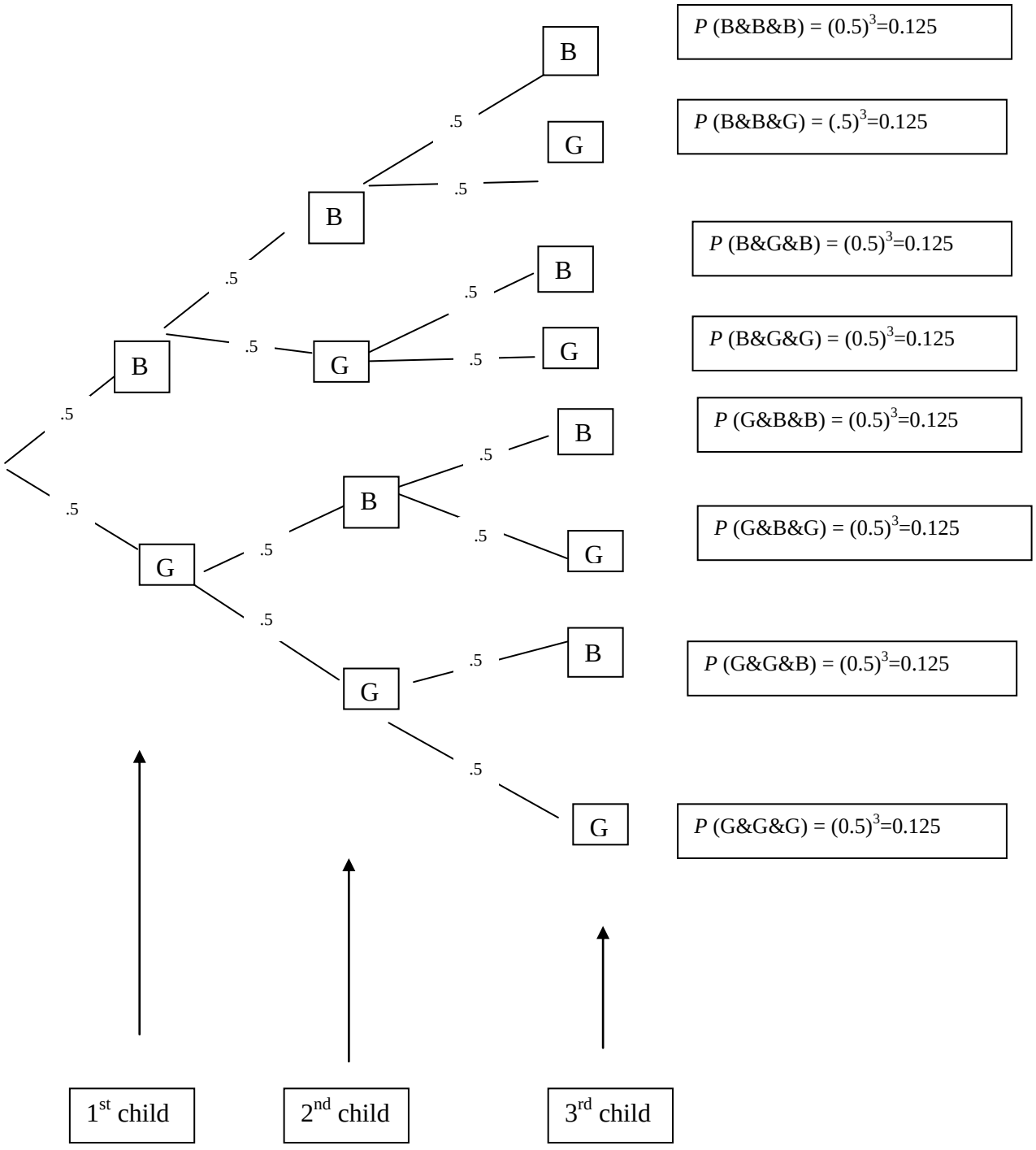
$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(notB)P(A \mid notB)} = \frac{(.0076)(.99)}{(.0076)(.99) + (1 - .0076)(.015)} = .336$$

Given a positive HIV test, one has a .336 probability of being infected. Even for this test which seems very accurate, over 66% of those testing positive will not have the virus.

TIP: Make a table:

VII. TREE DIAGRAMS

- Tree diagrams can also be helpful for calculating probabilities:



Examples of questions you can answer:

Q1: What's the probability that all 3 kids will be boys?

Q2: What's the probability that all 3 kids will be the same gender?

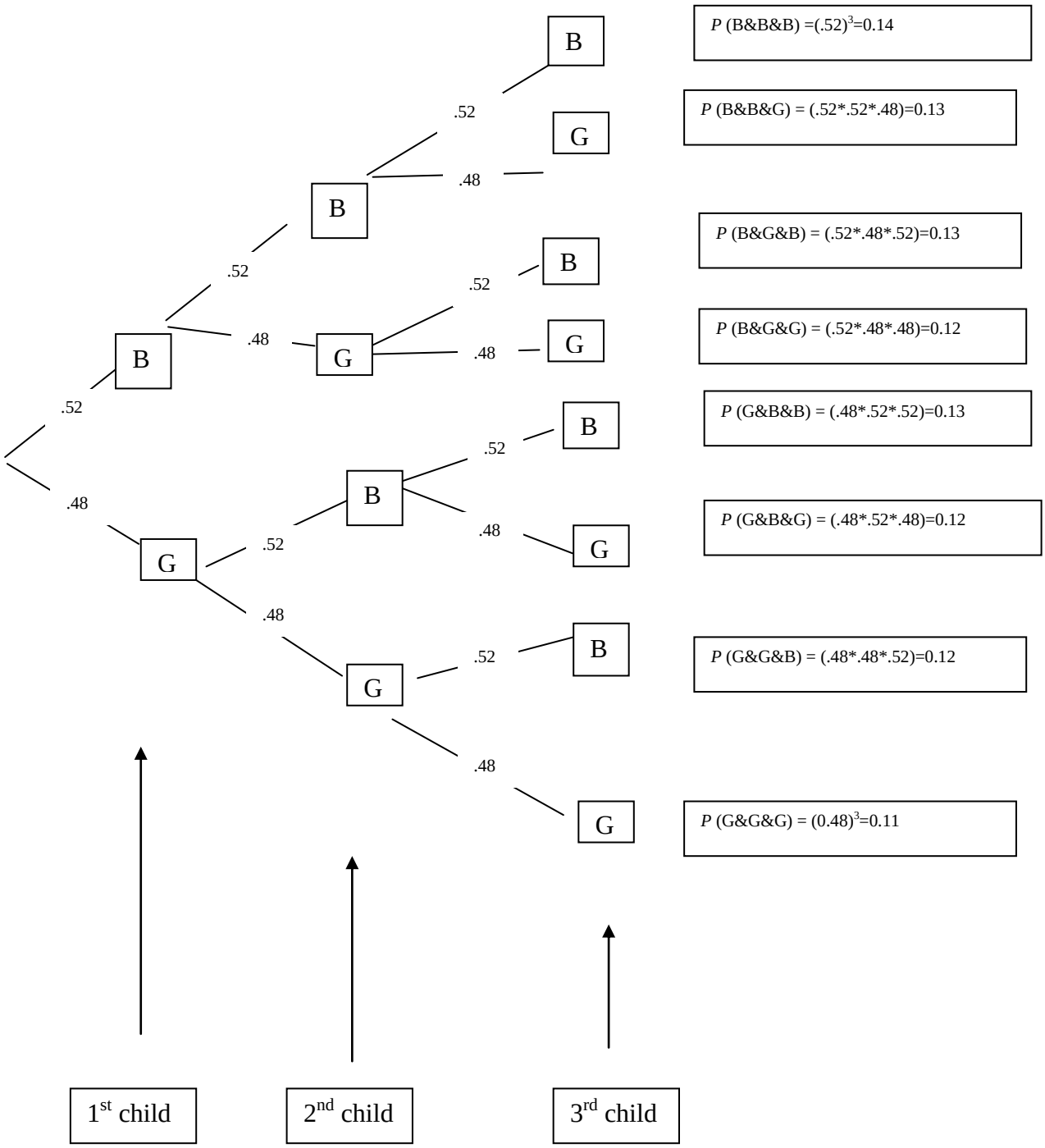
Q3: What's the probability that a family has at least two girls?

Q4: What's the probability that the last two kids are girls?

Q5: What's the probability that family has exactly 1 girl?

Q6: What's the probability that a family has exactly 2 girls?

- Suppose the outcomes aren't equally likely. E.g.:



- Now answer the previous set of questions with this example.