Class 6: Conditional Probability (Text: Sections 4.5)

Ex. Use the fact that 13% of the men and 13% of women are left-handed to fill in the following table. What proportion of people are left-handed?

	Men	Women	Total
Left-handed	780	520	1300
Right-handed	5220	3480	8700
Total	6000	4000	10,000

The probability of a person (either man or woman) being left-handed is 1300/10,000 = 0.13

How this table relates to Independence of handedness and gender:
The probability of a man being left-handed is 780/6000 = 0.13.
Probability of a woman being left-handed is 520/4000 = 0.13.
Probability of a person being left-handed is 0.13.
Since all three probabilities are the same, gender and handedness ***are independent***.

Ex. Use the fact that 7% of men and 0.4% of women are color blind to fill in the following table. What proportion of people are color blind?

	Men	Women	Total
Color blind	420	16	436
Not CB	5580	3984	9564
Total	6000	4000	10,000

The probability of a person (either man or woman) being color blind is 436/10,000 = 0.0436 = 4.36%
(Notice this value is between the 0.4% and 7%)

How this table relates to Independence of colorblindness and gender:
The probability of a man being color blind is 420/6000 = 0.07.
Probability of a woman being color blind is 16/4000 = 0.004.
Probability of a person being color blind is 0.0436 (not the same as in the previous class because the gender breakdown is different)
Since the probabilities are ***not*** the same, gender and colorblindness are ***not independent***.

Conditional Probability
A **conditional probability** tells us the likelihood of one event occurring *given that* another event has occurred.

We write conditional probabilities with a vertical line which is read as "given that" or "conditional on", so

P(Color Blind | Male) = 7%.

This tells us that the probability that someone is color blind, given that the person is male, is 7%. In other words, we restrict our attention to males. The proportion of males who are color blind is 7%.
Similarly, we have

P(Color Blind | Female) = 0.4%.

Here, we restrict our attention to females. This tells us that the probability that a female is color blind is 0.4%.
Note: The ordinary probability $P(\text{Color blind}) = 4.36\%$ can be called an *unconditional probability* if we want to distinguish it from a conditional probability.

To calculate conditional probability:

For ***any*** events A and B (not just independent events), we have

$$P(A|B) = \frac{\text{Number of occurences of } A \text{ and } B}{\text{Number of occurences of } B}$$

This can be written as

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Ex: What is P(Male|Color Blind)? What does it tell us? Explain why it has the value it does.

Now we restrict our attention to people who are color blind and ask what proportion are male. Look at the table:

P(Male|Color Blind) = 420/436 = 0.963 = 96.3%

This number is nearly 100% because many more men than women are color blind *and* there are more men in this population. Hence most color blind people are men.

Ex: Find P(Female|Color Blind). Explain why the number you get has the magnitude it does.

Since P(Male|Color Blind) = 96.3%, we have

P(Female|Color Blind) = 100 – 96.3 = 3.7%.

We can also calculate this directly as

P(Female|Color Blind) = 16/436 = 0.037 = 3.7%.

The proportion is small because there are fewer women in the population and only small proportion of them are color blind.

Ex: Describe in words the difference between P(Color Blind|Male) and P(Male|Color Blind)

P(Color Blind|Male) gives the proportion of males that are color blind, and

P(Male|Color Blind) gives the proportion of color blind people who are male.

Testing for Independence: We can use any one of the following

$P(A \text{ and } B) = P(A) \cdot P(B)$

Or $P(A|B) = P(A)$

Or $P(B|A) = P(B)$

In general, for any events (not just independent):

$P(A \text{ and } B) = P(A|B) \cdot P(B)$ and $P(A \text{ and } B) = P(B|A) \cdot P(A)$

Example: Smoking and Lung Disease

According to the American Lung Association, 7% of the US population has lung disease. Of those with the disease, 90% are smokers; of those without the disease, 25.3% are smokers.[1]

Ex: With no calculation, decide if smoking and lung disease are independent.
Smoking and lung disease are not independent because the proportion of smokers is different in the two groups. If there were independent, the two proportions, $(S|L)$ and $(S|NL)$, would be the same.

Ex: Use the data given to fill in the joint probability distribution table below.

	Lung Disease	No Lung Disease	Total
Smoking	0.063	0.23529	0.29829
Non Smoking	0.007	0.69471	0.70171
Total	0.07	0.93	1.00

Ex: What is the probability that a randomly selected smoker has lung disease?
The probability that a randomly selected smoker has lung disease is

$$P(L|S) = \frac{0.063}{0.29829} = 0.211 = 21.1\%.$$

Ex: What is the probability that a randomly selected non-smoker has lung disease?
The probability that a randomly selected non-smoker has lung disease is

$$P(L|S) = \frac{0.007}{0.70171} = 0.010 = 1.0\%.$$

Ex: Why do the answers to last two questions not add to 1?

The probabilities in parts (c) and (d) do not add to 1 because they are proportions of different subsets—the first of all smokers and the second of all non-smokers.

[1] www.lung.org. Reported in *Intro Statistics* 9th ed, p. 193, by N. Weiss (Pearson, 2012)

Testing for Independence using conditional probabilities

Ex: Using smokers and non-smokers, calculate conditional probabilities to check that lung disease and smoking are not independent.

We calculate the probability of lung disease, given smoking:

$P(L) = 7\%$

$P(L|S) = 21.1\%$

$P(L|NS) = 0.007/0.70171 = 0.010 = 1.0\%$

Thus, people are more likely to have lung disease if they smoke. Smoking and lung disease are ***not independent***.

Testing for Independence using multiplication of probabilities

Ex: Use the probability that someone has lung disease and is a smoker to show that lung disease and smoking are not independent:

From the table: $P(L \text{ and } S) = 0.063 = 6.3\%$

$P(L) * P(S) = 0.07 * 0.29829 = 0.02088 = 2.088\%$.

Thus, we see that

$$P(l \text{ and } S) \neq P(L) * P(S).$$

Example: Medical Testing: HIV Screening

If we have a medical screening test, for example a mammogram or an HIV test, several outcomes are possible:

- A true positive (test is positive and person has the disease)
- A false positive (test is positive, but person does not have the disease)
- A true negative (test is negative, and person does not have the disease)
- A false negative (test is negative, but person does have the disease)

How likely are each type? What does this depend on? How common the disease is (measured by prevalence) and how accurate the test is (measured by sensitivity and specificity).

The **prevalence** of the disease is the rate at which the disease occurs in the population. For the US:

P(HIV) = 0.006

Accuracy of a medical test

The **sensitivity** of the test is the likelihood that test results are positive among patients with the disease. For current HIV tests:

Sensitivity = P(Test+ | HIV) = 99.9%

The **specificity** of the test is the likelihood that the test results are negative among patients without the disease. Currently, for HIV

Specificity = P(Test- | No HIV) = 99.6%

Unfortunately, as the sensitivity increases (the test recognize smaller traces), the specificity tends to decrease.

Ex: In practice we want to know the probability of having HIV if we have a positive test result. What conditional probability do we want to know? This is called the **positive predictive value**.

P(HIV|Test+).

Ex: Fill in the table to calculate the probability of having HIV if given a positive test result.

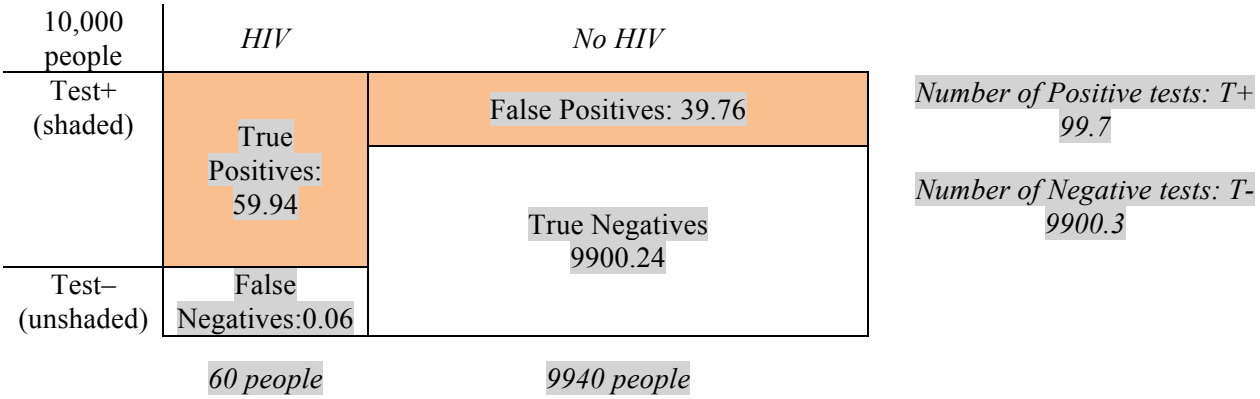
For a Sample of 10,000 People

	HIV	No HIV	Total
Positive test	59.94	39.76	99.7
Negative test	0.06	9900.24	9900.3
Total	60	9940	10,000

Notice that we have to fill in the marginal values first and then the center of the table.
Thus we have

$$P(HIV|\text{Test}+) = \frac{59.94}{99.7} = 0.601 = 60.1\%$$

Ex: Since the test is so accurate, explain using a Venn diagram why the positive predictive value is only 60.1%.



The people with positive tests (99.7 out of 10,000) come from two groups: A large proportion (99.9%) of the people who are HIV positive (59.94 people), and a small proportion of the people who are not HIV positive (39.76 people). Since the proportion in the US who are HIV positive is relatively small (0.6%), we are taking a large percentage of a small number and a small percentage of a large number. The true positives therefore turn out to be only about 60% of the positive test results; the false positives are about 40%.