



1

CONDITIONAL PROBABILITY AND INDEPENDENCE

Unit 3A: Probability



Now we will discuss independent events and conditional probability. These topics, although very important on their own, will also give us the background needed for our two rules for finding $P(A \text{ and } B)$ when we cannot easily use logic and counting.

We will begin with a logical definition of independent events.

The first thing to note is to throw away any definition you may have of the word “independent” or “independence” in the English language. You must base what you understand about our definition of independent events only on what we have to say about it.

In English, we often think of independent as being separate, not overlapping. However, we already have a definition for that, which is disjoint events. In probability DISJOINT events are events that do not overlap, that cannot occur at the same time.

Independent Events

- Two events are independent if knowing one event occurs does not change the probability of the other
- This is not the same as “disjoint” events which are separate in that they cannot occur together
- These are two different concepts entirely
- **Independence** is a statement about the equality of the **probability** of one event whether or not the other event occurs (or is occurring, or has occurred)



The probability definition of independent events is a statement about the equality of probabilities. We will introduce the definition here and discuss some simple examples. Later we will formalize the definition in probability notation.

Two events are independent if knowing one event occurs does not change the probability of the other event.

For example. If I roll a die twice and record the result for each toss. The fact that I rolled a 1 on the first toss does not change the probability of rolling a 1 on the second toss. The events

- A = getting a 1 on the first toss and
- B = getting a 1 on the second toss

Are independent. Knowing that A happens, does not change the probability of B.

Another example, more appropriate for us, is what if we draw two people from the entire US population at random and ask them whether or not they have ever had a heart attack. If the first person says yes, this will not change the probability that the second person will say yes. These two events are independent.

Before illustrating further, we point out that it isn't necessary to fully understand all of the mathematical details for the differences between disjoint and independent events but it is very important to not confuse these two concepts when applying rules as this type of confusion can easily lead to picking the wrong rule.

Independent Events

ASSUME INDEPENDENT:

- Repeated Sampling from any population where individuals are “replaced”
- Repeated Sampling from very large populations where individuals are NOT “replaced”

TEST FOR:

- Results from unrelated questions asked of individuals in the population will result in independent events

We will be able to assume that two events are independent when:

We take repeated samples from any population where the individuals are replaced.

- Draw a card, put it back and reshuffle, then draw another card
- Pick a coin from your pocket, put it back, then draw another coin

We take repeated samples from a very large population where the individuals are not replaced.

- Randomly select two U.S. Adults and ask if they have ever had a heart attack

We will also be interested in testing whether events are independent in our population. For example, are the events having diabetes and having heart disease independent in the population.

In this situation we will need to learn ways to test to answer this question.

If the questions asked of individuals in the population are not associated, they are unrelated, then the resulting events will be independent, otherwise they will be dependent.

Dependent Events

- Are not independent :-)

ASSUME DEPENDENT:

- Repeated Sampling from small populations where individuals are NOT “replaced”

TEST FOR:

- Results from RELATED questions asked of individuals in the population will result in dependent events

Dependent events are events which are NOT independent.

In repeated sampling, events will be dependent if individuals are not replaced.

If we have a classroom of 25 students and we randomly select two students. The probabilities will change.

If we draw a card from a deck and then draw another card, the probabilities are changing.

Technically, this change will occur even for large populations, but we will see that for large populations, the change is negligible and thus the events will be taken to be “effectively” independent.

We will look at these calculations soon.

We will be interested in testing to see if two events are associated in the population. The concept of association between variables in data and the idea of dependent events are closely tied.

Multiplication Rule (Independent Events)

- IF A and B are INDEPENDENT, THEN (and ONLY THEN)
 - $P(A \text{ and } B) = P(A) \cdot P(B)$ {Rule #6}
- Can be used to test for independence by calculating
 - $P(A \text{ and } B)$ using logic
 - $P(A)$ using logic
 - $P(B)$ using logic
 - Compare $P(A) \cdot P(B)$ to that found for $P(A \text{ and } B)$
 - If equal then A and B are INDEPENDENT otherwise DEPENDENT



Probability rule #6 states that IF two events A and B are independent, then (and only then) we have

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

We can multiply the two probabilities $P(A)$ and $P(B)$ to find $P(A \text{ and } B)$. But ONLY when A and B are independent events.

This is a very nice property of independent events but as many events are dependent, we cannot use this rule in just any situation.

If we are not sure whether two events are independent or dependent, we can use this equation as a test.

We find the three probabilities $P(A)$, $P(B)$, and $P(A \text{ and } B)$ using logic and then multiply $P(A)$ times $P(B)$. If the result is equal to the value we found for $P(A \text{ and } B)$ then the two events are independent. Otherwise the events are dependent.

Let's do a few calculations in a scenario where we can assume independence.

Suppose 35% of the US population over age 20 are obese and we select two people completely at random from the entire US population over age 20.

Use Rule #6 to Calculate P(A and B)

- Suppose 35% of the U.S. population over age 20 are obese.
- If we select two people at random from the U.S. Population over age 20:
 - $P(\text{Both will be obese}) = (0.35)(0.35) = 0.1225$
 - $P(\text{Neither will be obese}) = (0.65)(0.65) = 0.4225$
 - $P(\text{First will be obese but the 2nd will not}) = (0.35)(0.65) = 0.2275$
 - $P(\text{First will not be obese but the 2nd will}) = (0.65)(0.35) = 0.2275$



What is the probability that both people selected will be obese? Since this is such a large population and the individuals are chosen completely randomly, the first person's status will not have any impact on the 2nd person's status and so we can assume these two events {the first person is obese} and {the second person is obese} are independent. Therefore we can use the multiplication rule for independent events.

To find $P(\text{Both are obese})$ we note that this is an "AND" situation. We are trying to find the probability that the first person is obese AND the second person is obese. Which we can calculate using the multiplication rule for independent events by

$$P(1^{\text{st}} \text{ obese AND } 2^{\text{nd}} \text{ obese}) = P(1^{\text{st}} \text{ obese}) P(2^{\text{nd}} \text{ obese}) = (0.35)(0.35) = 0.1225.$$

$$\begin{aligned} \text{Similarly we can find } P(\text{neither are obese}) &= P(1^{\text{st}} \text{ NOT obese AND } 2^{\text{nd}} \text{ NOT obese}) \\ &= P(1^{\text{st}} \text{ NOT obese}) P(2^{\text{nd}} \text{ NOT obese}) = (0.65)(0.65) = 0.4225 \end{aligned}$$

There are also two other possibilities that result from getting one of each.

$$\text{We can find } P(1^{\text{st}} \text{ obese AND } 2^{\text{nd}} \text{ NOT obese}) = P(1^{\text{st}} \text{ obese}) P(2^{\text{nd}} \text{ NOT obese}) = (0.35)(0.65) = 0.2275$$

You might wonder if the order matters but you can easily see that so far we have only accounted for 0.7725 by adding the probabilities so far. Thus, we must have more to go! In fact, subtracting from 1, we get $1 - 0.7725 = 0.2275$!

The last probability is

$$P(1^{\text{st}} \text{ NOT obese AND } 2^{\text{nd}} \text{ obese}) = P(1^{\text{st}} \text{ NOT obese}) P(2^{\text{nd}} \text{ obese}) = (0.65)(0.35) = 0.2275.$$

This is every possible outcome for picking two people at random from the US population and

recording whether or not the individual is obese.

Rule #6 as a Test for Independence

- $P(65+ \text{ and YES}) \approx 0.0928$
- $P(65+) \approx 0.4608$
- $P(\text{YES}) \approx 0.1486$

Age:	Yes	No	Total
35-44	30	762	792
45-54	117	992	1109
55-64	220	1428	1648
65+	611	2422	3033
Total	978	5604	6582

- $P(65+) \cdot P(\text{YES}) = (0.4608)(0.1486) = 0.0685$
- \neq
- $P(65+ \text{ and YES}) = 0.0928$
- \Rightarrow DEPENDENT

As we mentioned, we can use the multiplication rule for independent events as a test to determine whether or not two events are independent.

Let's use our example from age groups and need for special equipment.

We found, using only logic and counting, that

- $P(65+ \text{ and YES})$ was 0.0928
- $P(65+)$ was 0.4608 and
- $P(\text{YES})$ was 0.1486

When we multiple $P(65+)$ times $P(\text{YES})$ we get (0.4608) times (0.1486) equals 0.0685.

This is NOT EQUAL to the TRUE value of $P(65+ \text{ and YES})$ found by logic of 0.0928.

And so these two events are dependent. We will come back and talk more about what this means later.

We will see this is only one of a number of ways we can test for independent events.

Conditional Probability

- So far, we have divided by the TOTAL
- Sometimes, however, we have additional CONDITIONS that cause us to alter the denominator (bottom) of our probability calculation
- **Given** the individual is Age 65+, what is the probability the individual requires special equipment?
- “Conditional” refers to the fact that we have these additional conditions, restrictions, or other information



Before we can formalize the definition of independent events, we need to introduce a new type of probability called conditional probability. It is useful to point out that this is the same as the row and column percents we learned to calculate in the section on exploratory data analysis in Case C-C. It would be good to review that material as it will reinforce what we are doing now as well as where we will be heading in the future.

Notice that every probability we have calculated so far: $P(A)$, $P(A \text{ and } B)$, $P(A \text{ or } B)$, has divided by the overall total. However, if we have additional conditions, this can change the total used in our denominator (bottom) of our probability calculation.

Suppose, when choosing one person from the CDC data on the need for special equipment vs. age group that we ask:

Given the individual is Age 65+, what is the probability the individual requires special equipment?

Now, we are restricting our attention to only the 65+ age group.

Before, in our discussion on Case C-C in Unit 1, we would have said:

Among individuals Age 65+, what percent require special equipment?

The question is exactly the same and the only difference in our answer is that we will use decimal notation as opposed to percentages (unless percentages are specifically requested).

Let's answer this question.

Example: Conditional Probability

- **Given** the individual is Age 65+, what is the probability the individual requires special equipment?

- $P(\text{YES} \mid 65+)$

- $= \frac{611}{3033}$

- ≈ 0.20145

Age:	Yes	No	Total
35-44	30	762	792
45-54	117	992	1109
55-64	220	1428	1648
65+	611	2422	3033
Total	978	5604	6582

The question again is:

Given the individual is Age 65+, what is the probability the individual requires special equipment?

Given the individual is Age 65+ tells us we can restrict our attention to only that row in our table. All other values have been covered here.

The new total is the total number of individuals who are 65+ which is 3033 and of those, 611 require special equipment. These are the values we need to calculate this probability.

The notation we use for conditional probability is the vertical line.

For this probability, the “given” is 65+ and we want to find the probability of YES.

The notation may seem backwards but we have

$P(\text{YES} \mid 65+)$ which reads the probability of YES given 65+. The event AFTER the vertical line is the GIVEN and indicates our new DENOMINATOR.

Notice that we just calculated this probability using only logic. We will now introduce a rule for calculating conditional probabilities.

Rule #7: Conditional Probability:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \quad \{\text{Rule \#7}\}$$

- **Given** the individual is Age 65+, what is the probability the individual requires special equipment?

- $P(\text{YES} | 65+)$

- $= \frac{P(\text{YES and } 65+)}{P(65+)}$

- $= \frac{611/6582}{3033/6582}$

- ≈ 0.20145

Age:	Yes	No	Total
35-44	30	762	792
45-54	117	992	1109
55-64	220	1428	1648
65+	611	2422	3033
Total	978	5604	6582

By using logic we simply took a short-cut.

The equation for $P(B | A)$ is $P(A \text{ and } B)$ divided by $P(A)$. Whatever is GIVEN is still the new denominator but now we are working with the probabilities instead of the frequencies.

Last time, we found the total number who were 65+ which was our new denominator of 3033 and then we found how many were 65+ and answered YES for our numerator. This is the same as using the equation except that we are converting each of those values into probabilities before dividing.

So, when we have data, using the formula is more work. However, if we already knew the needed probabilities and we did not have the frequencies, the rule allows us to still find the answer.

Here to find $P(\text{YES} | 65+)$ we need $P(\text{YES and } 65+)$ which is $611/6582$ and $P(65+)$ which is $3033/6582$.

When we divide $P(\text{YES and } 65+)$ by $P(65+)$ we get the same result as before.

You might also notice that mathematically, what happens is that the 6582 in both probabilities cancel to leave us right back at our previous answer of $611/3033$.

One common mistake is dividing by the wrong total. If we are GIVEN 65+, this must be our new total or denominator. Sometimes students divide by the total of the other event, YES in this case, possibly due to confusion of the notation or wording. Be careful to divide by whichever event is GIVEN or KNOWN, not the event for which you are asked to find the probability.

Conditional Probability

- **Given** the individual is Age 65+, what is the probability the individual requires special equipment?
- **IF** the individual selected is Age 65+, what is the probability the individual requires special equipment?
- What is the probability that an individual requires special equipment AMONG individuals Age 65+.
- READ CAREFULLY & UNDERSTAND NOTATION



We can also reword this question as:

What is the probability the individual requires special equipment given the individual is Age 65+.

We can also use IF or occasionally AMONG to represent the “given” such as

IF the individual selected is Age 65+, what is the probability the individual requires special equipment?

What is the probability that an individual requires special equipment AMONG individuals Age 65+.

So you must read carefully to determine the exact question being asked.

It is also very helpful to understand the notation of $P(A \mid B)$ so that you can simplify what you need to write in order to solve these types of problems while still understanding what you are doing in the process.

More About Independent Events

- **Tests for Independent Events:** Two events A and B are independent if any one of the following hold:
- $P(B | A) = P(B)$
- $P(A | B) = P(A)$
- $P(B | A) = P(B | \text{not } A)$
- $P(A \text{ and } B) = P(A) * P(B)$



Now that we have defined conditional probability we can more formally define the concept of dependent events. Our verbal definition of independent events was: Knowing one event occurs (translation: GIVEN one event occurs), the probability of the other event stays the same. Thus we can say, in notation, that A and B are independent events if:

- $P(B | A) = P(B)$. In words this says, given A happens, the probability of B has not changed - it is the same as $P(B)$ overall.
- Since we could approach it from the other direction we can also say $P(A | B) = P(A)$. In words this says, given B happens, the probability of A has not changed - it is the same as $P(A)$ overall.
- Another way to think about it is to write: $P(B | A) = P(B | \text{not } A)$. In words, given A happens, the probability of B is the same as if A did NOT happen. And in fact both of these are then equal to $P(B)$ overall.

Thus we have the following tests for Independent Events: Two events A and B are independent if any one of the following hold:

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

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

$$P(B | A) = P(B | \text{not } A)$$

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

It is up to you and which is easier for you in a given problem to determine which of these tests to use. In each case you must find the probabilities needed on each side of the equal sign by logic or a general equation and be careful not to mistakenly make an assumptions of independence in these calculations that would invalidate your test.

Then you compare the left hand side to the right hand side of the test equations. If they are equal then the events are independent, otherwise the events are dependent.

Tests for Independence

- $P(65+) \approx 0.4608$
- $P(\text{YES}) \approx 0.1486$
- $P(65+ \text{ and YES}) \approx 0.0928$
- $P(\text{YES} \mid 65+) \approx 0.2015$
- $P(65+ \mid \text{YES}) = 611/978 \approx 0.6247$
- $P(65+ \mid \text{NO}) = 2422/5604 \approx 0.4322$

Age:	Yes	No	Total
35-44	30	762	792
45-54	117	992	1109
55-64	220	1428	1648
65+	611	2422	3033
Total	978	5604	6582

Let's return to our example.

We have already calculated the first four probabilities listed here. And, we have used the first three as our first test for independence. We found that $P(65+) \times P(\text{YES})$ was 0.0685 which did not equal $P(65+ \text{ and YES})$ which is 0.0928, found by logic and counting.

Now we can look at the tests based upon conditional probability.

We calculated $P(\text{YES} \mid 65+)$ earlier it was about 0.2015. Now we see that we can compare this to $P(\text{YES})$ which we found earlier to be 0.1486. Since these two values are not equal, we can say the events 65+ and YES are dependent in our sample.

To find $P(65+ \mid \text{YES})$ we note that there are 978 total who said YES. This will be our denominator. Of those 978 who said YES, 611 are 65+. Thus $P(65+ \mid \text{YES}) = 611/978$ or about 0.6247.

We can compare this to $P(65+)$ which is 0.4608 to see that these values are different and thus, again, we confirm that these two events are dependent.

We could also compare $P(65+ \mid \text{YES})$ to $P(65+ \mid \text{NO})$. We can see that there are 5604 total individuals who answered NO and of those, 2422 are 65+. This gives $P(65+ \mid \text{NO}) = 2422/5604 = 0.4322$. Comparing this to $P(65+ \mid \text{YES})$ which was 0.6247 we see that they are not equal and so we arrive at the same conclusion.

You only need to pick one of these tests. The tests will always give the same answer when conducted correctly. Choose whichever test seems easiest to you for the given question.

Conditional Probability in Practice

- $P(\text{YES}) = 978/6582 \approx 0.15$
- $P(\text{YES} \mid 35-44) = 30/792 \approx 0.04$
- $P(\text{YES} \mid 45-54) = 117/1109 \approx 0.11$
- $P(\text{YES} \mid 55-64) = 220/1658 \approx 0.13$
- $P(\text{YES} \mid 65+) = 611/3033 \approx 0.20$

Age:	Yes	No	Total
35-44	30	762	792
45-54	117	992	1109
55-64	220	1428	1648
65+	611	2422	3033
Total	978	5604	6582

Before we discuss the last rule in this unit, let's talk a little about the application of conditional probabilities and how this is the same as what we did in Unit 1 relating to row and column percentages.

What we have here is the prevalence of the need for special equipment

- Overall
- And in each age group.

We can see that overall, around 15% of individuals in our sample required special equipment. You may notice in the data that the sample size is increasing with age group so we must be extra careful.

We can see, not surprisingly, that among younger individuals we have a lower prevalence of the need for special equipment. As we increase in age groups, the probability that individuals need special equipment increases.

For the youngest age group, only 4% require special equipment, this increases to 11% among those 45-54, and 13% among those 55-64. In the 65+ age group, we see that around 20% require special equipment.

We can see that the overall percentage [of 0.15] falls somewhere between that for the 55-64 and 65+ age groups due to the fact that a vast majority of our sample falls in this [age] range.

Breaking down variables in this way and looking at conditional percentages (now called probabilities) allows us to investigate the relationship between these two variables and to quantify the trends that we see.

General Multiplication Rule

- For any events A and B

$$P(A \text{ and } B) = P(A) \cdot P(B | A) = P(B) \cdot P(A | B) \quad \{\text{Rule \#8}\}$$

Probability rule #7 is a simple rearrangement of the definition of conditional probability which gives

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

Since A and B are interchangeable we can also write: $P(A \text{ and } B) = P(B) \cdot P(A | B)$

Remember that the given event should also appear as the single event: P(A) with P(B | A).

The letters in the “front” should never agree P(A) doesn’t go with P(A | B) in this rule.

This rule is useful for calculating the probabilities in repeated sampling for dependent events. Otherwise, in this course, you can always use logic easier than trying to apply this rule. It will always work but it is often much more effort than necessary to solve a problem using this rule.

Let’s look at a useful example of applying this rule.

Use Rule #8 to Calculate P(A and B)

- Suppose 5 out of 8 nurses in a particular unit have a certain certification
- If we select two nurses at random from this unit, what is

$P(\text{Both have certification})$

$$= (5/8)(4/7) \approx 0.357$$

$$= P(1^{\text{st}} \text{ has}) \cdot P(2^{\text{nd}} \text{ has} \mid 1^{\text{st}} \text{ has})$$

Suppose that 5 out of 8 nurses in a particular unit have a certain certification. We need two nurses to handle a situation which will require at least one of the nurses to have this certification. If we randomly selected the two nurses, what is the probability that we will get, both, none, at least one nurse with certification.

We can answer these questions using the general multiplication rule but we will see that we don't necessarily need to write out the rule, only consider the situation logically. Let's start with $P(\text{Both have certification})$. There are 8 total nurses and of them 5 have the certification.

The first time we randomly select a nurse there is a $5/8$ probability of selecting a nurse with the certification. In order to end up with both nurses having certification, it must be that I pick a nurse with the certification on the first pick and so there are 4 nurses with certification left out of the 7 total nurses left. When I randomly select the next nurse there is a $4/7$ probability that I pick a nurse with certification.

Since this is an "AND" problem. We must have the first with the certification AND the 2nd with the certification we are using a multiplication rule and clearly the probability is changing so the events are dependent but notice that we didn't really need to write down the rule to find the answer.

We simply multiply $(5/8)$ times $(4/7)$ to get 0.357. However, for this first question, we did write out the rule in case it helps you to solve these problems.

We have $P(\text{first has the certification})$ which was $5/8$, times $P(2^{\text{nd}} \text{ has the certification} \mid \text{the first had the certification})$. It is the given information that lets me say that since the first selected had the certification, there are only 4 left to choose from who have the certification in my remaining 7 nurses.

Use Rule #8 to Calculate P(A and B)

- Suppose 5 out of 8 nurses have a certain certification
- If we select two nurses at random from this unit, what is
 - $P(\text{Both have certification}) = (5/8)(4/7) \approx 0.357$
 - $P(\text{Neither will have certification}) = (3/8)(2/7) \approx 0.107$
 - $P(\text{First will but the 2nd will not}) = (5/8)(3/7) \approx 0.268$
 - $P(\text{First will not but the 2nd will}) = (3/8)(5/7) \approx 0.268$
- $P(\text{At least 1 certified}) = 1 - P(\text{Neither certified})$
 $= 1 - 0.107$
 ≈ 0.893



To find $P(\text{Neither will have the certification})$ we see that for my first random selection there are 3 nurses without the certification out of the 8 total. After having selected a nurse without certification, there are only 2 nurses without certification left of the 7 remaining nurses. So we get $(3/8) * (2/7) = 0.107$.

Finally we can calculate the combinations where we have one of each.

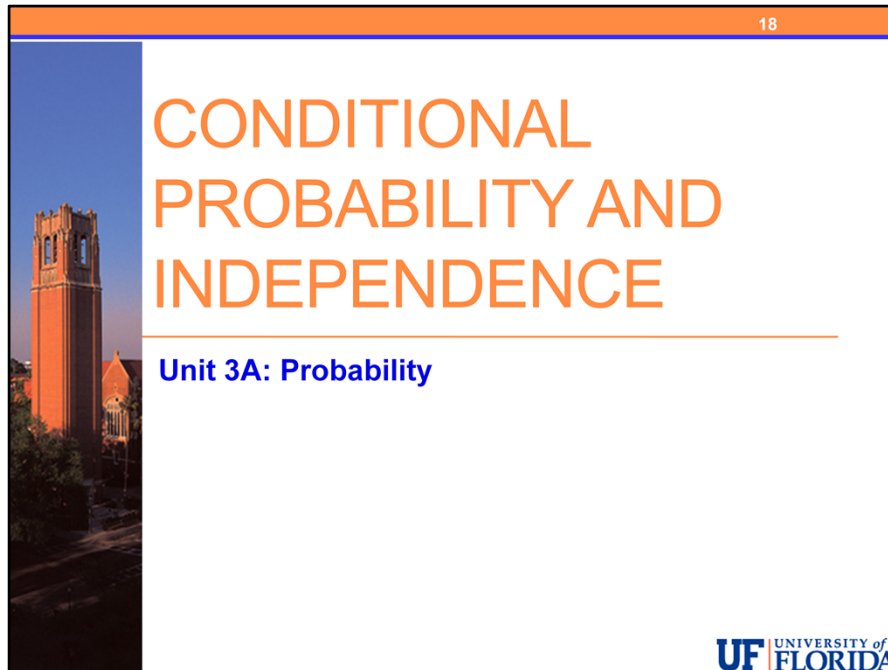
If the first does and the second does not. The first time we randomly select a nurse there is a $5/8$ chance to have the certification. Having picked a certified nurse, there are still 3 non-certified nurses remaining of the 7 so that when we randomly select the 2nd nurse there is a $3/7$ chance of selecting one which is not certified.

The last probability – where the first is not and the second is certified – is the same answer but the order of the numerators is reversed.

Thus if we needed to know what is the probability that we would get at least one certified nurse if we randomly selected two nurses. We can either

- Add the three appropriate probabilities:
 $P(\text{Both have certification}) + P(\text{First does and 2nd doesn't}) + P(\text{First doesn't and 2nd does})$
 $= 0.357 + 0.268 + 0.268 = 0.893$
- OR, we could recognize that
 $P(\text{At least 1 certified}) = 1 - P(\text{Neither certified}) = 1 - 0.107 = 0.893$.

We close with a few comments.



Conditional probability is very important to both our study of probability and statistics. In practical situations where we need to compare percentages based upon another categorical variable we are applying conditional probability, even if we may prefer to consider them to be row and column percentages.

Independence is also an extremely important concept both in our study of probability and for our future study of statistics. Testing for independence is an important skill in this Unit.

Notice that all of the problems where we calculated probabilities using the multiplication rules were situations of repeated sampling. Determining whether such a situation involves independent or dependent events is the first step. If you aren't sure then use the general multiplication rule as it will work in either situation.

Even in this section, with more complex probability rules, you have seen that we present the solutions logically. The logical solutions simply represent the logic that exists in the rules themselves, written in probability notation.

Remember that if you can solve the problem using logic, this is often the best approach! Especially in this course where our goal is to give you a basic understanding of probability and how it might be applicable in practice.

There is much more to learn about probability than we are able to cover. Our goal is to get back to statistics as quickly as possible and see how probability ties in with inferential statistics.