

## Independence

---

### 1000: The Idea Informally

---

Independence is an idea that involves information, or the lack of it. Informally, if two properties are independent, then the information that someone (or something) has one of the properties provides no information about whether that someone (or something) has the other property.

For example, suppose that you know that 25% of Americans are smokers. If I ask you the chances that a particular person, e.g., John Doe, is a smoker, and you know only that he is American, then you would respond "25%." Now suppose I first tell you that John Doe has heavy nicotine stains on his fingers. Since that piece of information raises the chances that John Doe is a smoker (to some number greater than 25%), we can say that the properties "Smoker" and "Has Nicotine Stained Fingers" are **dependent** or **associated**. Suppose instead of telling you about John Doe's fingers, I tell you that he is right-handed. Assuming that knowing he is right-handed tells you nothing about whether he is a smoker, i.e., the chances of John Doe being a smoker given he is right-handed are still 25%, then the properties "Smoker" and "Right-handed" are **independent**.

In general, if the frequency of one property A is the same as the frequency of A conditional on another property B, then properties A and B are **independent**.

Properties A and B are **independent** if and only if:

$$Fr_5(A) = Fr_5(A | B)$$

FIGURE 1000-1

The intuitive idea is simple -- if knowing that an individual has property B (i.e., conditioning on B) tells me nothing about whether they have A, then A and B are independent. Two properties that are not independent are **dependent**, or **associated**.

For an example of dependence, consider the properties being female and being pregnant. The frequency of being female in America is approximately 0.5. The frequency of being female given you are pregnant is 1.0, so these two properties are highly dependent.

For an example of (at least approximate) independence, consider the properties being female and living East of the Mississippi. The frequency of being female in America is 0.5, and the frequency of being female given you live East of the Mississippi is also 0.5, so these two properties are independent.

Independence and association are connected to causation, but are not to be confused with it. For example, consider two properties that are dependent: being bald (property A), and having a brother who is bald (property B). Clearly A and B are associated. The frequency of being bald is not the same as the frequency of being bald given you have a brother who is bald. Yet being bald doesn't cause you to have a brother who is bald, or vice versa. In fact the two are related by a common cause -- having a mother who carries the gene for baldness.

In general, independence and dependence are symmetric, while causation is not. If one property A is independent of another B, then B is also independent of A. The same is true for dependence. For example, as we said above, the frequency of being female is associated with being pregnant, and thus it is also true that being pregnant is associated with being female. Causation is not symmetric, however. Knowing that A is a cause of B does not allow us to infer that B is a cause of A. For example, knowing that breaking a bone is a cause of great pain, we cannot infer that great pain is a cause of breaking a bone.

<b>ASSOCIATION IS SYMMETRIC</b>	property A is <b>ASSOCIATED</b> with property B	→	property B is <b>ASSOCIATED</b> with property A
<b>INDEPENDENCE IS SYMMETRIC</b>	property A is <b>INDEPENDENT</b> of property B	→	property B is <b>INDEPENDENT</b> of property A
<b>CAUSATION IS NOT SYMMETRIC</b>	property A is a <b>CAUSE</b> of property B	→ <del>X</del>	property B is a <b>CAUSE</b> of property A

FIGURE 1000-2

< A link to exercises in the interactive version of this module. >

---

## 2000: Independence Among Properties

---

### 2100: Representing Independence Among Properties

---

Two properties A and B are independent if the frequency of A is the same as the frequency of A conditional on B. Using frequency notation, they are independent when  $Fr(A) = Fr(A | B)$ . So if we want to use pie charts and histograms to graphically represent independence, we need at least two charts: one for the frequency of A, and another for the frequency of A **conditional** on B.

For example, suppose I wanted to use histograms to show that picking an Ace and picking a Diamond are independent properties. Then I would need to display a histogram for the frequency of picking Aces, and another for picking Aces **conditional** on having picked a Diamond.

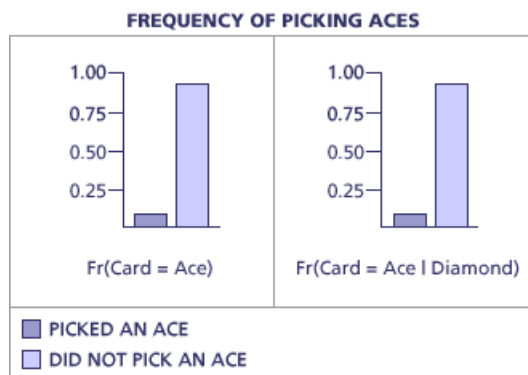


FIGURE 2100-1

< A link to exercises in the interactive version of this module. >

---

## 2200: Determining Independence Among Properties

---

Sometimes, instead of histograms, we are given a data table or a contingency table. To determine whether two properties are independent given this kind of information, we need to apply the definition of independence. Two properties A and B are independent if and only if  $\text{Fr}(A) = \text{Fr}(A | B)$ . So, we need to compute these two frequencies from the data table or contingency table, and then we can find out if the properties are independent. Consider the following data table:

TABLE 2200-1: DATA TABLE

Individual	Class Year	Taken Calculus
1	Freshman	Yes
2	Junior	Yes
3	Senior	No
4	Senior	Yes
5	Junior	Yes
6	Freshman	No
7	Senior	No
8	Sophomore	No
9	Junior	No
10	Junior	Yes

Is being a junior independent of taking calculus? To answer this question, we need to ask whether:

$$\text{Fr}_S(\text{Class} = \text{Junior}) = \text{Fr}_S(\text{Class} = \text{Junior} | \text{Taken calculus?} = \text{Yes}).$$

Looking at the data table, we can see that:

- +  $\text{Fr}_S(\text{Class} = \text{Junior}) = 4/10 = 2/5$ ; and
- +  $\text{Fr}_S(\text{Class} = \text{Junior} | \text{Taken calculus?} = \text{Yes}) = 3/5$ .

Since these two frequencies are different, the properties are associated (or dependent). In other words, the chance of picking a junior at random is  $2/5$ . But if we only look at the people who have taken calculus, the chance of picking a junior is  $3/5$ . Since the chances change when we get information about the person having taken calculus, the properties are associated.

Now let's see how to determine whether two properties are independent from a contingency table. Once again, we just need to determine which two frequencies to compare, and then compute the frequencies from the contingency table. For example, consider the following contingency table for Transportation [Drives car, Takes bus] and UCSD student [Yes, No]:

TABLE 2200-2: CONTINGENCY TABLE

Mode of Transport	Student = Yes	Student = No	Total
Drives Car	50	400	450
Takes Bus	25	200	225
Total	75	600	675

Is driving a car independent of being a UCSD student? To determine whether these properties are independent, we need to know whether:

$$\text{Fr}_S(\text{Transportation} = \text{Drives car}) = \text{Fr}_S(\text{Transportation} = \text{Drives car} | \text{UCSD student} = \text{Yes}).$$

According to the contingency table:

- +  $\text{Fr}_S(\text{Transportation} = \text{Drives car}) = 450/675 = 2/3$ ; and
- +  $\text{Fr}_S(\text{Transportation} = \text{Drives car} | \text{UCSD student} = \text{Yes}) = 50/75 = 2/3$ .

Since these two frequencies are equal, the properties are independent. Learning that someone is a UCSD student doesn't alter the chances that he or she drives a car.

< [A link to exercises in the interactive version of this module.](#) >

### 3000: Interactive Exploration

The exercises that follow all use the Set Builder applet. If you aren't familiar with Setbuilder, take five minutes to look at the Set Builder Manual.

The Setbuilder applet allows you to create samples of any size from a given set of "atoms" with certain properties. For example, in the following instance of Setbuilder there are eight "atoms":

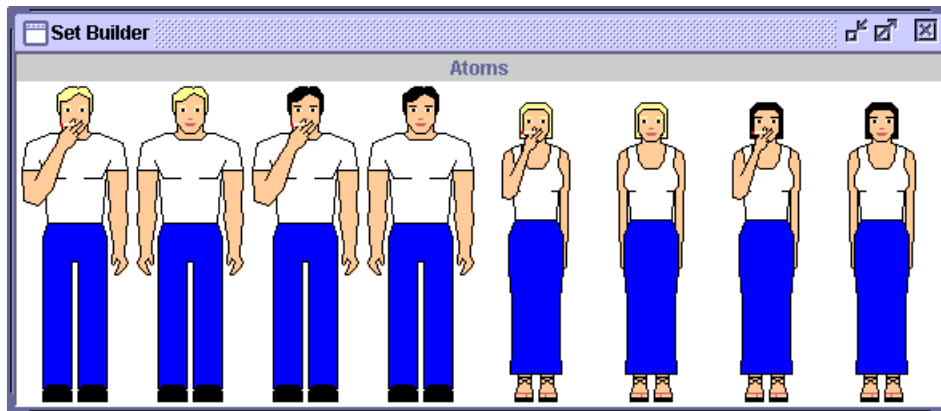


FIGURE 3000-1

- + Male, Blond, Smoker
- + Male, Blond, Non-smoker
- + Male, Dark-haired, Smoker
- + Male, Dark-haired, Non-smoker
- + Female, Blond, Smoker
- + Female, Blond, Non-smoker
- + Female, Dark-haired, Smoker
- + Female, Dark-haired, Non-smoker

This set of atoms involve every combination of three binary properties:

- + SEX: (Male or Female)
- + HAIR-COLOR: (Blond or Dark)
- + SMOKER: (Smoker or Non-smoker)

After the modules "Relative Frequency" and "Conditional Relative Frequency", you should be reasonably comfortable creating sets in the SetBuilder. Creating sets in which certain independencies hold is much harder, though. We suggest that, before you start a problem, you write down the frequencies that must be equal for each pair of properties that are independent. Then, if frequency values are provided for you, see if those values determine what value the equal frequencies must be. Otherwise, you can assign whatever value you want to each pair of frequencies, as long as the equalities still hold.

< [A link to exercises in the interactive version of this module.](#) >

### 4000: The Idea Formally

---

## 4100: Table of Notations

---

Let  $S$  be a sample or population, i.e., any non-empty collection of objects. Objects in the collection may have various properties. If  $A$  is a property, the set of objects that have  $A$  in  $S$  will be signified by " $A$ " itself, and the set of objects that do not have  $A$  in  $S$  -- the complement of  $A$  in  $S$  -- will be signified by " $\sim A$ ". The set of objects that have both properties  $A$  and  $B$  is denoted by " $A \& B$ "; the set of objects that have either  $A$  or  $B$  or both is denoted by " $A \vee B$ ".

<b>A</b>	The subset of the sample that has property <b>A</b>
<b><math>\sim A</math></b>	The subset of the sample that does not have property <b>A</b>
<b>A &amp; B</b>	The subset of the sample that has both property <b>A</b> and property <b>B</b>
<b>A <math>\vee</math> B</b>	The subset of the sample that has either property <b>A</b> , or property <b>B</b> , or both

FIGURE 4100-1

The **cardinality** of a set is the number of elements in the set. If  $S$  is a set, we write the cardinality of  $S$  as:  $|S|$ . For example, if I use the letter  $P$  to represent the set of major planets in our solar system, then  $|P|$  will be 9. If  $A$  represents the set of individuals in a study who were HIV positive and 5% of the 1000 people studied were HIV positive, then  $|A|$  would be 50.

Two properties  $A$  and  $B$  are said to be **exclusive** if no one in the sample has both  $A$  and  $B$ , i.e., if the subset of the sample that has both property  $A$  and property  $B$  is the empty set ( $A \& B = \emptyset$ ). For example the properties male and female are exclusive, but the properties male and being a smoker are not.

Two properties are said to be **exhaustive** if everyone in the sample  $S$  has at least one of them ( $A \vee B = S$ ). For example, the properties male and female are exhaustive, but the properties male and smoker are not.

---

## 4200: Definitions

---

### 4210: Relative Frequency

---

If  $S$  is a finite sample and  $A$  is a property in the sample, the relative frequency of  $A$  in  $S$  is defined to be:

$$Fr_S(A) = \frac{|A|}{|S|}$$

FIGURE 4210-1

---

### 4220: Conditional Relative Frequency

---

If S is a sample, and A and B are properties in the sample, the frequency of A conditional on B is defined to be:

$$\text{Fr}_S(\mathbf{A} \mid \mathbf{B}) = \frac{\text{Fr}_S(\mathbf{A} \ \& \ \mathbf{B})}{\text{Fr}_S(\mathbf{B})}$$

FIGURE 4220-1

Applying the definition of frequency to both the numerator and denominator of the right hand side, we get:

$$\text{Fr}_S(\mathbf{A} \mid \mathbf{B}) = \frac{\frac{|\mathbf{A} \ \& \ \mathbf{B}|}{|S|}}{\frac{|\mathbf{B}|}{|S|}}$$

FIGURE 4220-2

And since the # in S cancels, we end up with:

$$\text{Fr}_S(\mathbf{A} \mid \mathbf{B}) = \frac{|\mathbf{A} \ \& \ \mathbf{B}|}{|\mathbf{B}|}$$

FIGURE 4220-3

---

## 4230: Independence and Association

---

Properties A and B are independent in sample S (which we write as  $\mathbf{A} \perp\!\!\!\perp_S \mathbf{B}$ ) if and only if the frequency of A in S equals the frequency of A conditional on B in S. Formally:

$$\mathbf{A} \perp\!\!\!\perp_S \mathbf{B}$$

if and only if

$$\text{Fr}_S(\mathbf{A}) = \text{Fr}_S(\mathbf{A} \mid \mathbf{B})$$

FIGURE 4230-1

Properties A and B are associated (dependent) in sample S (written as  $\mathbf{A} \not\perp\!\!\!\perp_S \mathbf{B}$ ) if and only if they are not independent in S.

An equivalent definition of independence is: Properties A and B are independent in a sample S if and only if the frequency of both A and B equals the frequency of A multiplied by the frequency of B. Formally:

$$\mathbf{A} \perp\!\!\!\perp_S \mathbf{B}$$

if and only if

$$\text{Fr}_S(\mathbf{A} \ \& \ \mathbf{B}) = \text{Fr}_S(\mathbf{A}) * \text{Fr}_S(\mathbf{B})$$

FIGURE 4230-2

This equivalent definition follows straightforwardly from the first definition of independence, and the definition of conditional relative frequency:  $\mathbf{A} \perp\!\!\!\perp_S \mathbf{B}$  if and only if  $\text{Fr}(\mathbf{A}) = \text{Fr}(\mathbf{A} \mid \mathbf{B})$ , and  $\text{Fr}(\mathbf{A} \ \& \ \mathbf{B}) = \text{Fr}(\mathbf{A} \ \& \ \mathbf{B}) / \text{Fr}(\mathbf{B})$  by definition.

---

### 4300: Facts about Independence

---

### 4310: Independence is Symmetric

---

Unlike causation, association and independence are entirely symmetric. Whenever one property A is independent of another property B, then B is also independent of A:

$$A \perp\!\!\!\perp B$$

if and only if

$$B \perp\!\!\!\perp A$$

FIGURE 4310-1

So, looking back at our definition of independence, we now know that  $A \perp\!\!\!\perp B$  if and only if  $\text{Fr}_S(B) = \text{Fr}_S(B | A)$ .

[< A link to exercises in the interactive version of this module. >](#)

---

### 4320: Independence and Complex Properties

---

If A and B are independent properties, then we can use our second definition of independence to find the frequency of the complex property  $A \& B$ . Since  $A \perp\!\!\!\perp B$  implies that  $\text{Fr}(A \& B) = \text{Fr}(A) * \text{Fr}(B)$ , we can multiply the frequencies of the simple properties to get the frequency of the complex property. For example, suppose I tell you that the frequency of drawing a King is  $1/13$ , and that the frequency of drawing a red card is  $1/2$ . Since drawing a King and red card are independent, the frequency of drawing a red King is  $1/13 * 1/2 = 1/26$ .

[< A link to exercises in the interactive version of this module. >](#)

---

### 4330: Independence of Properties and Their Complements

---

Recall that if A is a property, then  $\sim A$  refers to the property of not having A (so  $\sim A$  is the complement of A). If we have two independent properties, then each property (and its complement) is also independent of the complement of the other property:

$$A \perp\!\!\!\perp B$$

if and only if

$$\sim A \perp\!\!\!\perp B$$

if and only if

$$A \perp\!\!\!\perp \sim B$$

if and only if

$$\sim A \perp\!\!\!\perp \sim B$$

FIGURE 4330-1

For example, consider the properties Male and Smoker. If we know that Smoker is independent of Male, then we can infer that:

- + Smoker is independent of Female (the complement of Male);
- + Non-smoker (the complement of Smoker) is independent of Male; and
- + Non-smoker (the complement of Smoker) is independent of Female (the complement of Male).

## 5000: Independence for Variables

## 5100: Definition of Independence for Variables

The value of a variable is the same as a property. For instance, the variable **SEX** takes on values Male and Female, and a person can have the property of being male or being female. The property of having a disease and the property of not having the disease can be thought of as values for a variable, **DISEASE**.

In these examples the variables have only two values, but others have more values. For example, we might include several values (as the U.S. Census does) for the variable **RACE**. Or we might pick cut-offs and divide the variable **INCOME** into Low, Middle, and High. Some variables are naturally thought of as having an infinity of values: **HEIGHT**, **WEIGHT**, **LENGTH**, and so on.

There are some rules about variables. Here are two of them:

Values of variables (i.e., properties) are mutually exclusive

No individual has two or more properties that are alternative values of one and the same variable. No one, for example, is supposed to be both male and female, or have both high income and low income. Sometimes this assumption is quite unrealistic (consider the values of **RACE**).

Values of variables (i.e., properties) are exhaustive.

For every variable to be studied, every individual has one of its values. It isn't required that the values always be known, or measured, only that they exist. For example, that we can't consider the variable **IQ** for a population that includes, say, a desk. A desk doesn't have any IQ at all. Sometimes this problem is sidestepped by assuming a special value, for example arbitrarily assigning the property of having 0 IQ to things for which IQ scores make no sense.

### Definition of Independence among Variables

Two variables are independent in a sample or population if each value of one variable is independent of each value of the other variable.

So, if we have the variables [and values]: **SEX** [Male, Female], and **HAIR COLOR** [Blond, Red, Dark], then **SEX** and **HAIR COLOR** are independent if and only if:

- + **SEX** = Male \_||\_ **HAIR COLOR** = Blond
- + **SEX** = Male \_||\_ **HAIR COLOR** = Red
- + **SEX** = Male \_||\_ **HAIR COLOR** = Dark
- + **SEX** = Female \_||\_ **HAIR COLOR** = Blond
- + **SEX** = Female \_||\_ **HAIR COLOR** = Red
- + **SEX** = Female \_||\_ **HAIR COLOR** = Dark

If any of these six independencies fails to hold (i.e., if any of the values are associated), then the two variables are associated. The variables are independent only if every value of one variable is independent of every value of the other variable.

< [A link to exercises in the interactive version of this module.](#) >

## 5200: Independence for Binary Variables

A binary variable has only two values. Since the values of variables are just properties, this means that the values of a binary variable are a property and its complement. We earlier saw that if a property is independent of another property, then its complement is also independent of that property. So, if one value of a binary variable is independent of a value of another variable, then the other value of the binary variable must also be independent of the value of the other variable. And this means that, if one value of a binary variable is independent of a value (property) or variable, then the binary variable itself is independent of the value (property) or variable.

For example, consider the variables **SEX** [Male, Female] and the property Right-handed. If Male  $\perp$  Right-handed, then Female  $\perp$  Right-handed, which means that **SEX**  $\perp$  Right-handed (since every value of **SEX** is independent of Right-handed).

This fact about binary variables is especially useful when we are considering two binary variables. Consider the variables: **SEX** [Male, Female], and **SMOKES** [Yes, No]. In section 4330, we saw that Male  $\perp$  Smoker implies three other independencies. But this one independence among properties also implies the variable independence: **SEX**  $\perp$  **SMOKES**, since every value of **SEX** is independent of every value of **SMOKES**.

## 6000: Case Studies

### 6100: Sex and Race

In a Pittsburgh study of Pneumonia patients, the Sex (Male, Female) and Race (White, Non\_white) of approximately two thousand Pneumonia patients was recorded. Are the variables **SEX** and **RACE** independent among pneumonia patients? Remember that they are independent if the frequency of a patient being male or female is the same regardless of the race of the patient. Formally, **SEX** and **RACE** are independent when:

$$\begin{aligned} \text{Fr}(\mathbf{SEX} = \text{Male}) &= \text{Fr}(\mathbf{SEX} = \text{Male} \mid \mathbf{RACE} = \text{White}) \\ &= \text{Fr}(\mathbf{SEX} = \text{Male} \mid \mathbf{RACE} = \text{Non-white}) \end{aligned}$$

and

$$\begin{aligned} \text{Fr}(\mathbf{SEX} = \text{Female}) &= \text{Fr}(\mathbf{SEX} = \text{Female} \mid \mathbf{RACE} = \text{White}) \\ &= \text{Fr}(\mathbf{SEX} = \text{Female} \mid \mathbf{RACE} = \text{Non-white}) \end{aligned}$$

FIGURE 6100-1

Here is a table that gives the breakdown of Sex and Race among 2,287 Pneumonia patients in the study:

TABLE 6100-1: SEX AND RACE AMONG PNEUMONIA PATIENTS

Sex	White	Non-white	Total
Male	. 972	. 172	. 1144
Female	. 977	. 166	. 1143
Both	. 1949	. 338	. 2287

[< A link to exercises in the interactive version of this module. >](#)

---

## 7000: Summary

---

Two properties are independent when learning about whether an individual has one property does not change the chances that the individual has the other property. In terms of relative frequencies, this means that two properties are independent when the frequency of one is the same whether we look in the whole population, or just in the sub-population with the other property. Formally, A is independent of B when  $\text{Fr}(A) = \text{Fr}(A | B)$ . Alternately,  $A \perp\!\!\!\perp B$  if and only if  $\text{Fr}(A \& B) = \text{Fr}(A) * \text{Fr}(B)$ . So, when we need to know whether two properties are independent, we either check to see whether these two frequencies are equal (by looking at histograms, or computing frequencies from a data table or contingency table), or else we check whether the frequency of the complex property is just the product of the frequencies of the simple properties. Two properties are associated if they are not independent.

Unlike causation, independence (and so also association) is symmetric; if A is independent of (associated with) B, then B is independent of (associated with) A. Also, if two properties are independent, then their complements are also independent (of the other property and the other complement). So, if we learn that A and B are independent (or associated), then we automatically learn three other independencies (associations).

It is harder to determine if two variables are independent than it is to determine if two properties are independent, because every value of one variable must be independent of every value of the other variable. So, we typically have quite a few independencies that must be true. Values of binary variables, though, are just properties and their complements. So, learning that one value of a binary variable is independent of a property or another variable implies that the binary variable itself is independent of the other property or variable.

Finally, we should finish by remembering that independence and association are evidence for causation, but they are not equivalent to it.

---