Conditional Relative Frequency

1000: The Idea Informally

The relative frequency of one property A conditional on another B is just the relative frequency of A within the sub-group of individuals that have property B. Recall the definition of the relative frequency of a property A in a group S: the number of individuals in the group who have property A divided by the total number of individuals in S:

$$\mathrm{Fr}(\mathbf{A}) = \frac{\text{\# OF INDIVIDUALS WITH } \mathbf{A} \text{ IN S}}{\text{\# OF INDIVIDUALS IN S}}$$

FIGURE 1000-1

Pictorially, we can represent the relative frequency of A with a box in which the large box is the whole group and the light purple box labeled "A" represents those with property A. The size of the A box in relation to the size of the whole box is the frequency of A.

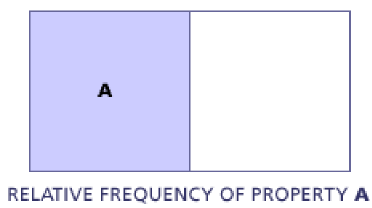


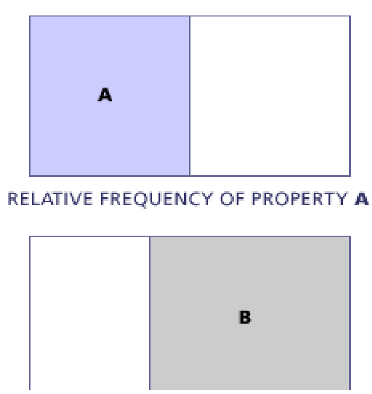
FIGURE 1000-2

The relative frequency of property A **conditional** on property B is simply the relative frequency of A in the sub-group with property B:

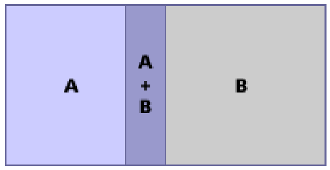
$$\mathrm{Fr}(\mathbf{A} \mid \mathbf{B}) = \frac{\text{\# OF INDIVIDUALS WITH } \mathbf{A} \text{ AND } \mathbf{B} \text{ IN S}}{\text{\# OF INDIVIDUALS WITH } \mathbf{B} \text{ IN S}}$$

FIGURE 1000-3

Intuitively, conditioning on B is the same as restricting the "group" to only those individuals with property B. In the picture below, we depict the relative frequency of property A, the relative frequency of property B, and the relative frequency of A conditional on B. The relative frequency of A conditional on B is the size of the dark purple region (individuals who have property A among those with B) relative to the size of the gray box B (individuals who have B).



RELATIVE FREQUENCY OF PROPERTY **B**



RELATIVE FREQUENCY OF PROPERTY **A**
CONDITIONAL ON PROPERTY **B**

FIGURE 1000-4

The relative frequency of a property A can be dramatically different from its conditional relative frequency on another property B, and its relative frequency conditional on B can be very different from its relative frequency conditional on another property C, as you can see in the examples above. In Figure 1000-3, the relative frequency of A is approximately 1/2. In Figure 1000-4, however, the frequency of A conditional on B (the size of the dark purple region relative to the size of B) is closer to 1/8 than to 1/2.

For a concrete example, consider people who live in the United States, and the property of having a beard (B). Suppose the relative frequency of having a beard among people in the U.S. is 0.05. That is, Fr(B) = 0.05. Now consider two properties that we might condition on:

+ Property F: Being Female
+ Property EM: Living east of the Mississippi River

Now consider the conditional relative frequencies:

+ Fr(B | F): The frequency of beards among females
+ Fr(B | EM): The frequency of beards among people who live east of the Mississippi

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

If the relative frequency (also called the unconditional relative frequency) of a property A is different than the relative frequency of A conditional on B, then we say that properties A and B are **associated**. So the properties of having a beard and being female are associated, as are the properties of having a beard and being male. If the relative frequency of a property A is the same as the relative frequency of A conditional on B, then we way that properties A and B are **independent**. Conditional relative frequencies are typically used to indicate when two properties are associated, which in turn is often used to suggest a causal relation. (Most of this course is about how to investigate whether such a suggestion is or is not a good one.) Conditional relative frequencies are also used to indicate when two properties are not associated -- in other words, that the properties are independent -- which in turn is used to rule out a causal relation.

It is important to be able to distinguish between relative frequencies and conditional relative frequencies in everyday speech and writing. The conditional relative frequency of A on B is the relative frequency of A in the sub-population of individuals with B. This sub-population is often indicated by the phrases "conditional on B," "among those with B," "given B," or "within the population with B." As an example, the following statements are equivalent to Fr(A | B) = 0.20:

+ The relative frequency of A conditional on B is 0.20.
+ The frequency of A among those with B is 0.20.
+ The frequency of A given B is 0.20.
+ The frequency of A within the population with B is 0.20.

Note that none of these statements refer to the frequency of A and B. This is because the frequency of A and B is the number of individuals that have both A and B relative to the number of individuals in the whole population. This is very different from the frequency of A conditional on B, which is the number of individuals with A relative to the sub-population of individuals that have B.

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

2000: Representing Conditional Relative Frequencies

Histograms and pie charts are used to represent relative frequencies, and the same graphical devices are used to represent conditional frequencies as well. For example, suppose we have the following set of 16 individuals created with Set Builder:

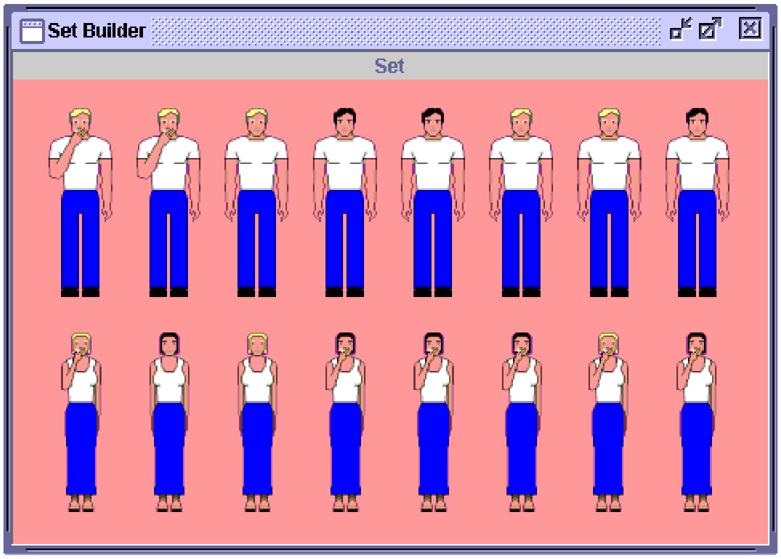


FIGURE 2000-1

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

As you can see, although 1/2 of the sample are males, 1/2 are females, 1/2 are smokers, and 1/2 non-smokers, the frequency of smokers changes dramatically when we condition on the sex of the individual.

The two **conditional** histograms below capture this immediately. The left histogram displays the relative frequency among females of smokers vs. non-smokers, and the histogram just to the right of it displays the relative frequency of smoking among males.

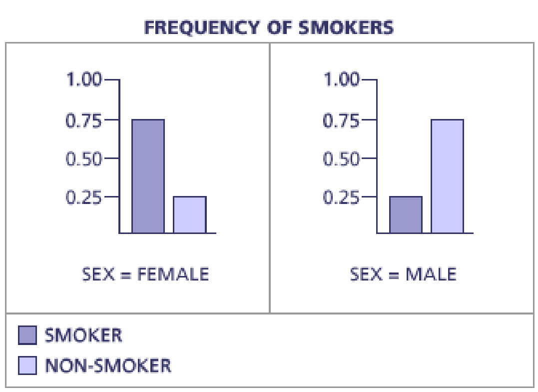


FIGURE 2000-2

Conditional frequencies are sometimes displayed in charts that combine two or more conditional histograms in the same plot:

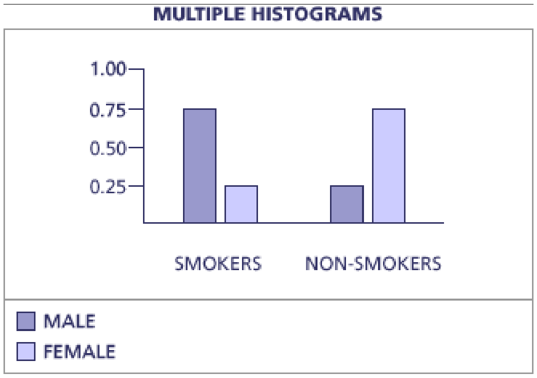


FIGURE 2000-3

In this case, we are plotting four conditional frequencies simultaneously:

- Fr(Smokes = Smoker | Sex = Male)
- Fr(Smokes = Smoker | Sex = Female)
- Fr(Smokes = Non-smoker | Sex = Male)
- Fr(Smokes = Non-smoker | Sex = Female)

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

3000: Calculating Conditional Relative Frequencies

3100: Conditional Relative Frequencies and Data Tables

Scientists collect statistical data in tables in which each row corresponds to an individual, and each column to a variable. The cells are the values of the variable in the column applied to the individual in the row. Suppose you are given the following data table:

TABLE 3100-1: DATA TABLE

Individual	Age	Political Party
1	Old	Democrat
2	Old	Republican
3	Young	Republican
4	Old	Republican
5	Young	Republican
6	Young	Democrat
7	Old	Republican
8	Young	Democrat

To calculate the conditional frequency of Republicans among Old people, Fr(Party = Republican | Age = Old), we need to calculate:

$$\text{Fr}(\textbf{Republican} \mid \textbf{Old}) = \frac{\text{\# OF INDIVIDUALS THAT ARE } \textbf{REPUBLICAN} \text{ AND } \textbf{OLD}}{\text{\# OF INDIVIDUALS THAT ARE } \textbf{OLD}}$$

FIGURE 3100-1

To do this, we only need to focus on those rows in which the individual is Old:

TABLE 3100-2: DATA TABLE EMPHASIZING ROWS IN WHICH THE INDIVIDUAL IS OLD

Individual	Age	Political Party
1	**Old**	**Democrat**
2	**Old**	**Republican**
3	Young	Republican
4	**Old**	**Republican**
5	Young	Republican
6	Young	Democrat
7	**Old**	**Republican**
8	Young	Democrat

Since there are 3 out of 4 of such rows in which the individual is a Republican, then Fr(Party=Republican | Age = Old) = 0.75.

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

3200: Conditional Relative Frequencies and Contingency Tables

---

Consider the following table of comparing plans for graduate school among males and females (From p. 15 of "Introductory Statistics" by J.K. Lindsey, Clarendon Press, 1995.):

TABLE 3200-1: CONTINGENCY TABLE

Sex	PhD Plans = Yes	PhD Plans = No	Total
Male	5	12	17
Female	6	5	11
Male and Female	11	17	28

Suppose we wanted to compare the frequency of Ph.D planners conditional upon being male with the frequency of Ph.D planners conditional upon being female. That is, we want to compare:

- Fr(Ph.D. Plans = Yes | Sex = Male)
- Fr(Ph.D. Plans = Yes | Sex = Female)

To calculate the first conditional relative frequency, we need to calculate the number of individuals who are both Male and Ph.D Planners, and divide that amount by the number of individuals who are Male:

$$\text{Fr}(\textbf{PhD Planner} \mid \textbf{Male}) = \frac{\text{\# OF INDIVIDUALS THAT ARE } \textbf{PHD PLANNERS} \text{ AND } \textbf{MALE}}{\text{\# OF INDIVIDUALS THAT ARE } \textbf{MALE}}$$

FIGURE 3200-1

To do this, we only need to focus on the two columns in the Male row that state:

- the number Ph.D. planners
- the total number of males

TABLE 3200-2: CONTINGENCY TABLE EMPHASIZING COLUMNS IN THE MALE ROW

Sex	PhD Plans = Yes	PhD Plans = No	Total
Male	**5**	12	**17**
Female	6	5	11
Male and Female	11	17	28

Now we can calculate the frequency of Ph.D planners conditional upon being male:

$$\text{Fr}(\textbf{PhD Planner} \mid \textbf{Male}) = \frac{5}{17} = 0.2941$$

FIGURE 3200-2

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

4000: 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":

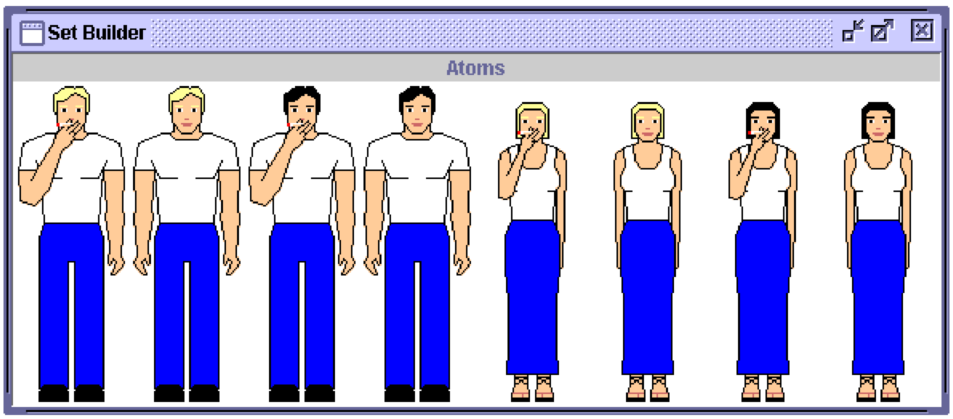


FIGURE 4000-1

- 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)

We suggest you take the following steps to do SetBuilder exercises:

- Read the instructions fully, and decide which graphical displays would help you with the exercise. For example, if the instructions ask you to construct a sample that has Fr(SEX = Male) = 0.5, then display a pie chart of the variable SEX
- Figure out how many individuals total you will need in the set. For example, if the problem asks you to create a set in which Fr(SEX = Male) = 0.4, then you can't do this if you have only 3 individuals in the set. Five would do, or ten, but not three or four, etc. If the problem asks you for a set in which Fr(SEX = Male) = 0.5 and Fr(HAIR-COLOR = Blond) = 0.33, then you need a set such that the total number of individuals can be broken evenly into halves and into thirds. For example, if your set had 10 individuals, then you might make five of them male so that Fr(SEX = Male) = 0.5, but there would be no possible way to assign hair-color such that Fr(HAIR-COLOR = Blond) = 1/3 = 0.33, because 10 doesn't divide into thirds. So the set must contain a number that can be divided into halves and into thirds, like 12.
- Do one property at a time. For example, if you are asked to create a set in which Fr(SEX = Male) = 0.5 and Fr(HAIR-COLOR = Blond) = 0.33, start with SEX. Create a set in which Fr(SEX = Male) = 0.5, and then proceed to HAIR-COLOR. When you adjust the frequency of HAIR_COLOR in the set, swap individuals that are identical except for HAIR-COLOR. For example, if you have 6 dark-haired male smokers, and 6 dark-haired female smokers, remove 1 dark-haired male smoker from the set and add 1 blond-haired male smoker. By doing this you will keep the total number of individuals the same -- and you won't disturb the frequency of SEX or of SMOKING. You will, however, change the frequency of BLOND, which is what you want and only what you want.
- Repeat this until your set satisfies the specified frequencies (which you should verify with Histograms or Pie Charts), and then click SUBMIT.

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

5000: The Idea Formally

5100: 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 "~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 v B".

A	The subset of the sample that has property **A**
~A	The subset of the sample that does not have property **A**
A & B	The subset of the sample that has both property **A** and property **B**
A v B	The subset of the sample that has either property **A**, or property **B**, or both

FIGURE 5100-1

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 v B = S). For example, the properties male and female are exhaustive, but the properties male and smoker are not.

5200: Definitions

5210: 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:

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

FIGURE 5210-1

5220: 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:

$$Fr_S(\mathbf{A}\mid\mathbf{B}) = \frac{Fr_S(\mathbf{A}\ \&\ \mathbf{B})}{Fr_S(\mathbf{B})}$$

FIGURE 5220-1

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

$$Fr_S(\mathbf{A}\mid\mathbf{B}) = \frac{\dfrac{|\,\mathbf{A}\ \&\ \mathbf{B}\,|}{|\,\mathbf{S}\,|}}{\dfrac{|\,\mathbf{B}\,|}{|\,\mathbf{S}\,|}}$$

FIGURE 5220-2

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

$$Fr_S(\mathbf{A}\mid\mathbf{B}) = \frac{|\,\mathbf{A}\ \&\ \mathbf{B}\,|}{|\,\mathbf{B}\,|}$$

FIGURE 5220-3

5300: Properties of Conditional Relative Frequency

5310: Complementary Properties Among a Sub-Population

It is always the case that the sum of the relative frequency of a property "A" and the relative frequency of the property "not A" is equal to one.

$$Fr_S(\mathbf{A}) + Fr_S(\sim\mathbf{A}) = 1$$

FIGURE 5310-1

Since the conditional relative frequency FrS(A | B) is just the relative frequency of A within the sub-group of individuals with B, the sum of the conditional relative frequency of the property "A" on the property "B" and the conditional relative frequency of the property "not A" on the property "B" is equal to one.

$$Fr_S(\mathbf{A} \mid \mathbf{B}) + Fr_S(\sim\mathbf{A} \mid \mathbf{B}) = 1$$

FIGURE 5310-2

5320: Properties Among Complementary Sub-Populations

In general, however, the sum of the relative frequency of property "A" within the sub-population with property "B" and the relative frequency of the property "A" within the sub-population with the property "not B" is not equal to one.

It is **not** always the case that

$$Fr_S(\mathbf{A} \mid \mathbf{B}) + Fr_S(\mathbf{A} \mid \sim\mathbf{B}) = 1$$

FIGURE 5320-1

The reason that this is not always true is that, in general, the relative frequency of a property within one sub-population has no relation to the relative frequency of that property within a different sub-population.

5330: The Asymmetry of Conditional Relative Frequency

The relative frequency of property "A" within the sub-population with property "B" is not the same as the relative frequency of property "B" within the sub-population with property "A."

It is **not** always the case that

$$Fr_S(\mathbf{A} \mid \mathbf{B}) = Fr_S(\mathbf{B} \mid \mathbf{A})$$

$$Fr_S(\mathbf{A} \mid \mathbf{B}) = Fr_S(\mathbf{B} \mid \mathbf{A})$$

only in the special circumstance in which

$$Fr_S(\mathbf{A}) = Fr_S(\mathbf{B})$$

FIGURE 5330-1

Thus, the frequency of A conditional on B is the same as the frequency of B conditional on A if and only if the frequency of A is the same as the frequency of B.

CLAIM:

$Fr_S(\mathbf{A} \mid \mathbf{B}) = Fr_S(\mathbf{B} \mid \mathbf{A}) \Longleftrightarrow Fr_S(\mathbf{A}) = Fr_S(\mathbf{B})$

PROOF:

1	$Fr_S(\mathbf{A}\ \&\ \mathbf{B}) = Fr_S(\mathbf{B} \mid \mathbf{A})$	Assumption
2	$Fr_S(\mathbf{A}\ \&\ \mathbf{B}) / Fr_S(\mathbf{B}) = Fr_S(\mathbf{B}\ \&\ \mathbf{A}) / Fr_S(\mathbf{A})$	Apply Definition of Cond. Relative Frequency to both sides of Eq. 1
3	$Fr_S(\mathbf{A}\ \&\ \mathbf{B}) / Fr_S(\mathbf{B}\ \&\ \mathbf{A}) = Fr_S(\mathbf{B}) / Fr_S(\mathbf{A})$	Algebra on Line 2
4	$1 = Fr_S(\mathbf{B}) / Fr_S(\mathbf{A})$	$Fr_S(\mathbf{A}\ \&\ \mathbf{B}) = Fr_S(\mathbf{B}\ \&\ \mathbf{A})$
5	$Fr_S(\mathbf{A}) = Fr_S(\mathbf{B})$	
6	$Fr_S(\mathbf{A} \mid \mathbf{B}) = Fr_S(\mathbf{B} \mid \mathbf{A}) \rightarrow Fr_S(\mathbf{A}) = Fr_S(\mathbf{B})$	Lines 1, 5
7	$Fr_S(\mathbf{A} \mid \mathbf{B}) = Fr_S(\mathbf{B} \mid \mathbf{A}) \leftarrow Fr_S(\mathbf{A}) = Fr_S(\mathbf{B})$	

FIGURE 5330-2

6000: Case Studies

6100: Seat Belts and Fatalities

In 1988, the Department of Highway Safety and Motor Vehicles compiled statistics on the frequency of injuries in car crashes and the frequency of seat belt use. Here is the contingency table they constructed:

TABLE 6100-1: CONTINGENCY TABLE

Seat Belt	Injury = Fatal	Injury = Non-fatal	Total
No	1601	162,527	164,128
Yes	510	412,368	412,878
Total	2111	574,895	577,006

Since the causal question at hand was whether seat belts prevent fatalities, the authors were interested in whether there was a negative association between seat belt use and fatal injuries. That is, their hypothesis was that the frequency of fatal injuries conditional on wearing a seat belt was lower than the frequency of fatal injuries in the total sample.

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

7000: Summary

The relative frequency of property A in a population S is just the percentage of individuals with the property A in the population S, and is given by:

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

FIGURE 7000-1

The conditional relative frequency of property A on a property B is just the percentage of individuals with the property A in the sub-population of individuals with B, and is given by:

$$Fr_S(A \mid B) = \frac{|A \& B|}{|B|}$$

FIGURE 7000-2

Conditioning on the property B, then, is the same as restricting the population we are considering from the total population S to the sub-population of individuals with B. Thus, the conditional relative frequency of A on B is just the relative frequency of A with respect to the sub-population B.

In everyday speech and writing, the conditional relative frequency of A on B is indicated by such phrases as:

- The relative frequency of A conditional on B
- The frequency of A among those with B
- The frequency of A given B
- The frequency of A within the population with B

In scientific practice the concepts of relative frequency and conditional relative frequency are very important for helping investigators decide whether two properties are associated or independent, which in turn bear on whether the two properties are causally related.
