

Relative Frequency

1000: The Idea Informally

In previous modules, we have been building an account of causation through causal graphs, response structures, and ideal interventions. Yet, so far we have not quite seen what happens commonly with researchers: They compare seemingly different units in a set of associations in order to form hypotheses. Finding ways to compare what appear to be entirely different units is at the root of judging causal claims, and is the topic of this module.

For example, consider how you would determine whether a particular city is a safe place. To start, you could look to the number of murders in that city in any given year. The assumption would be that the safest cities have the lowest number of murders. But would the number of murders alone be a good indicator for level of safeness, or would you have to know something more about each city, say its population? That is, how could you compare New York City with Richmond, VA or Atlanta, GA? If you were told that New York had 770 murders in 1997 while Richmond had 139 and Atlanta had 150, would you conclude that New York was the least safe among the three, while Richmond was the safest?

TABLE 1000-1: Consider the following statistics for the year 1997:

| City | . New York | . Richmond | . Atlanta |
|--------------|-------------|------------|-----------|
| # of Murders | . 770 | . 139 | . 150 |
| Population | . 7,320,477 | . 206,692 | . 420,865 |

As you will see, knowing something about a property, in this case the number of murders, is pretty meaningless unless we know something about the group that property came from. In this module you will see that by dividing the occurrence of a property, e.g. Murders, by the number of individuals in a group, e.g. city population, gives a **relative frequency** that allows for the comparison. That is, if you were you divide the number of Murders by the cities population you could calculate the **relative frequency** of murder in each city. For example:

TABLE 1000-2: Relative frequency of murder in select cities for 1997:

| City | . New York | . Richmond | . Atlanta |
|------------------------------|------------|------------|-----------|
| Relative Frequency of murder | . .000105 | . .000672 | . .000356 |

Using the relative frequency of murder in each city, rather than just the number of murders, we can see that you are actually much less likely to be a victim in New York, than in Richmond or Atlanta. And, you are twice as likely to be killed in Richmond than in Atlanta. Calculating and comparing the relative frequency of a property between different groups is the topic of this module.

Like percent, a relative frequency is a special kind of ratio or proportion that expresses how common something is in a certain group. For example, a relative frequency or a percent could be used to express how common red hair color is among residents of Holland, or how common NATO membership is among European countries. So, a relative frequency involves both a property (red hair color, being a member of NATO), and a group (residents of Holland, countries of Europe). A property is the same as a value of a variable, and we will use the two interchangeably. An individual having a property is the same as a variable taking on a value for that individual. For example, suppose we define a variable **HAIR COLOR**, with values: [Red, Brunette, Blond, Other]. If I have the property of having blond hair, that is the same as the variable **HAIR COLOR** taking on the value blond when applied to me.

The relative frequency of property A in a group S is a fraction, or ratio: it is the ratio of the number of individuals in the group S who have property A to the total number of individuals in the group S. When it is clear what group S we are discussing, we may just speak of the relative frequency of A, which we denote $Fr(A)$. The formula for $Fr(A)$ is:

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

FIGURE 1000-1

The percent of group S that have property A is just 100 times the relative frequency of A. If no individual in the group has the property A, then the relative frequency of A is 0, i.e., $\text{Fr(A)} = 0$. If every individual in the group has the property A, then the relative frequency of A is 1, i.e., $\text{Fr(A)} = 1$. Thus, relative frequencies must be numbers between 0 and 1 (assuming, that is, that the number in the group is finite).

For example, consider a group of 4,000 undergraduate students enrolled at Carnegie Mellon University (CMU). What is the relative frequency of smoking (S) among this group of CMU students? In a formula, it is:

$$\text{Fr(S)} = \frac{\text{\# OF STUDENT SMOKERS AT CMU}}{\text{\# OF STUDENTS AT CMU}}$$

FIGURE 1000-2

Some properties are simple, like being male or having blond hair, but some are complicated, like being a blond female smoker, or like being a female of mixed race born between 1970 and 1980 with three siblings. Complicated properties can also be represented as more than one variable taking on a particular combination of values. For example, if we have three variables: **HAIR COLOR** : [Blond, Brunette], **SEX**: [Male, Female], **SMOKER** : [Yes, No], then saying someone is a blond female smoker is the same as saying that for that person, the variable **HAIR COLOR** takes on the value blond, the variable **SEX** takes on the value female, and the variable **SMOKER** takes on the value yes. The idea of the relative frequency of a property is the same, whether the property is simple or complicated.

For example, the relative frequency (X) of being a female of mixed race born between 1970 and 1980 with three siblings in a group S is:

$$\text{Fr(X)} = \frac{\text{\# OF FEMALES OF MIXED RACE WITH THREE SIBLINGS BORN BETWEEN 1970 AND 1980 IN S}}{\text{\# OF INDIVIDUALS IN S}}$$

FIGURE 1000-3

2000: Representing Relative Frequencies

Relative Frequencies are often shown with a pie chart or a histogram (bar chart). In a pie chart, the size of the pie piece relative to the whole pie corresponds to the size of the relative frequency of the property. In a histogram, the height of each bar relative to the total height of all the bars is the relative frequency. For example, we might represent that the $\text{Fr}(\text{Smokers at CMU}) = 0.25$ with the following pie chart and histogram:

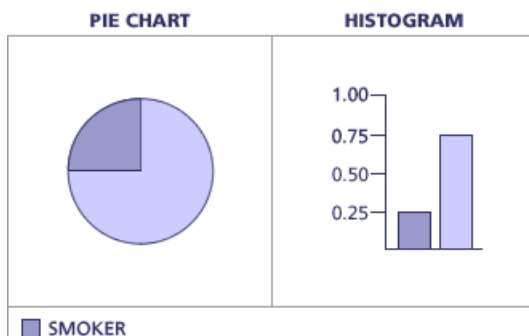




FIGURE 2000-1

When there are more than two categories, the same rules apply. For example, suppose that at University of Amsterdam in Holland, we recorded the smoking habits of 2,400 students, and categorized them into non-smokers, moderate smokers, and heavy smokers:

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

The relative frequencies you calculated in the previous exercise are listed in the table below:

TABLE 2000-1: SMOKING HABITS AT THE UNIVERSITY OF AMSTERDAM

| Group | # of Individuals | Relative Frequency |
|------------------|------------------|--------------------|
| Non-smokers | 1200 | Fr(NS) = 0.500 |
| Moderate-smokers | 800 | Fr(MS) = 0.333 |
| Heavy-smokers | 400 | Fr(HS) = 0.167 |
| Total | 2400 | -- |

To represent these frequencies we might use the following pie chart and histogram:

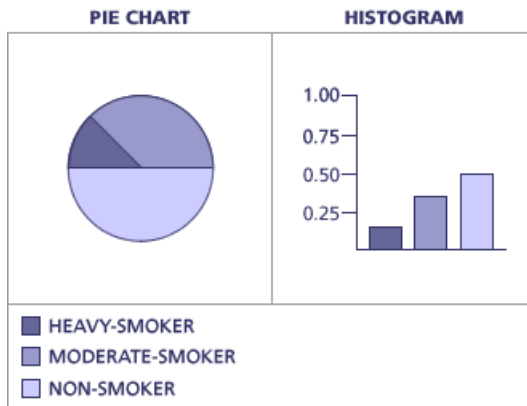


FIGURE 2000-2

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

3000: Calculating Relative Frequencies

3100: Relative Frequencies and Data Tables

Statistical data usually comes in two forms: a data table or a contingency table. In a data table, there are N rows and K columns, where each row corresponds to an individual and each column to a variable. For example, were we to collect data on eight individuals where we measured their **SEX**, **HAIR COLOR**, and whether they were a **SMOKER** or not, we might get the table below.

TABLE 3100-1: DATA TABLE

| Individual | Sex | Hair Color | Smoker |
|------------|--------|------------|--------|
| 1 | Male | Blond | Yes |
| 2 | Female | Dark | Yes |
| 3 | Male | Dark | No |
| 4 | Male | Dark | Yes |
| 5 | Female | Blond | No |
| 6 | Male | Blond | No |
| 7 | Female | Dark | No |
| 8 | Female | Blond | No |

To calculate the relative frequency of a property, A, in this data,...

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

FIGURE 3100-1

Since the number of individuals is always 8, we really only need to calculate:

$$\text{Fr}(\mathbf{A}) = \frac{\text{\# OF INDIVIDUALS WITH A IN S}}{8}$$

FIGURE 3100-2

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

3200: Relative Frequencies and Contingency Tables

Contingency tables are often used to describe the relative frequencies among different combinations of properties. Tables like the one below are typically used to inform us about the relationship between the properties in the rows (Smoking/Non-smoking), and the properties in the columns (Male/Female). For example, consider the following contingency table recording the smoking habits of 400 imaginary students at Carnegie Mellon University (CMU):

TABLE 3200-1: CONTINGENCY TABLE OF SMOKERS VS SEX

| Group | Male | Female |
|-------------|------|--------|
| Smokers | 50 | 60 |
| Non-smokers | 140 | 150 |

The table contains the raw number of people with each of four complicated properties:

TABLE 3200-2: SMOKING HABITS OF A GROUP OF STUDENTS AT CMU

| Group | # in the Group |
|--------------------|----------------|
| Male Smokers | 50 |
| Female Smokers | 60 |
| Male Non-smokers | 140 |
| Female Non-smokers | 150 |

Suppose, however, we wanted to know the relative frequency of smokers in this group:

$$Fr(S) = \frac{\# \text{ OF STUDENT SMOKERS AT CMU}}{\# \text{ OF STUDENTS AT CMU}}$$

FIGURE 3200-1

The information is not available directly from the contingency table, so we must calculate the total number of smokers before we can compute this relative frequency.

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

Having calculated the total number of persons in terms of smoking and sex, it is now possible to calculate the relative frequency:

TABLE 3200-3: CONTINGENCY TABLE OF SMOKERS VS SEX

| Group | Male | Female | Total |
|-------------|------|--------|-------|
| Smokers | 50 | 60 | 110 |
| Non-smokers | 140 | 150 | 290 |
| Total | 190 | 210 | 400 |

< [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":

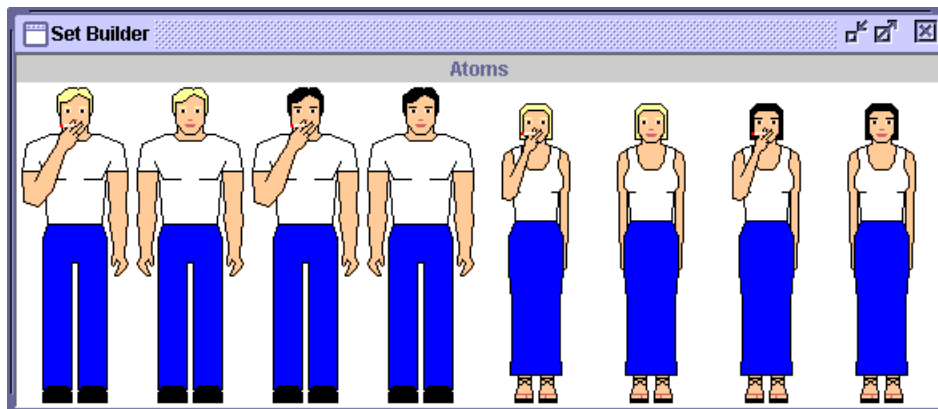


FIGURE 4000-1

- + 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 $\text{Fr}(\text{SEX} = \text{Male}) = 0.5$, 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 $\text{Fr}(\text{SEX} = \text{Male}) = 0.4$, 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 $\text{Fr}(\text{SEX} = \text{Male}) = 0.5$ and $\text{Fr}(\text{HAIR-COLOR} = \text{Blond}) = 0.33$, 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 $\text{Fr}(\text{SEX} = \text{Male}) = 0.5$, but there would be no possible way to assign hair-color such that $\text{Fr}(\text{HAIR-COLOR} = \text{Blond}) = 1/3 = 0.33$, 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 $\text{Fr}(\text{SEX} = \text{Male}) = 0.5$ and $\text{Fr}(\text{HAIR-COLOR} = \text{Blond}) = 0.33$, start with SEX. Create a set in which $\text{Fr}(\text{SEX} = \text{Male}) = 0.5$, 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: Introduction

In this section we take a more formal approach. We do so for several reasons. First, several ideas that we discuss later, e.g., conditional frequency and independence, require that we build on the definitions we give in this module. Second, by being slightly formal we can precisely define concepts with none of the slippage that informal descriptions allow. Third, once you understand the idea informally, the formal expression can serve as a quick reference later.

The section is relatively short, and has only a half dozen or so definitions and concepts. You need not be able to reproduce the proofs we offer, but you should memorize the notation in the next section and especially the definition of a relative frequency. Section 5400 provides some key equations that allow you to compute the frequency of complex properties from the frequencies of simple ones. Thus, you should also memorize the equalities given in these subsections.

5200: 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 " $A \vee B$ ".

| | |
|------------------------------|---|
| A | The subset of the sample that has property A |
| $\sim A$ | 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 \vee B | The subset of the sample that has either property A , or property B , or both |

FIGURE 5200-1

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.

5300: The Definition of 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 5200-1

5400: The Sums of Relative Frequencies

5410: Complementary Properties

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" ($\sim A$) is equal to one.

CLAIM:

$$Fr_S(A) + Fr_S(\sim A) = 1$$

PROOF:

$$\begin{array}{l} 1 \quad Fr_S(A) = \frac{|A|}{|S|} \\ \hline 2 \quad Fr_S(\sim A) = \frac{|\sim A|}{|S|} \\ \hline 3 \quad Fr_S(A) + Fr_S(\sim A) = \frac{|A|}{|S|} + \frac{|\sim A|}{|S|} \\ \hline 4 \quad Fr_S(A) + Fr_S(\sim A) = \frac{|A| + |\sim A|}{|S|} \\ \hline 5 \quad Fr_S(A) + Fr_S(\sim A) = \frac{|S|}{|S|} \end{array}$$

FIGURE 5410-1

The complement of a property A is denoted $\sim A$. For example, if A is the property of being a smoker, then $\sim A$ is the complement of A: not being a smoker (being a non-smoker). If two properties are complementary, e.g., Smoker and Non-smoker, or HIV positive and HIV negative, then every individual in a group must have one property or the other, but not both. Thus the number of individuals in a group that have property A plus the number who have property $\sim A$ must be the number in the group.

5420: One or Both of Two Properties

The relative frequency of having either or both of two overlapping properties is the sum of the individual properties minus their overlap:

$$Fr_S(A \vee B) = Fr_S(A) + Fr_S(B) - Fr_S(A \& B)$$

FIGURE 5420-1

You can see from the following figure that by simply summing the size of A and B, we would double count the region where they overlap, thus we need to subtract it out.

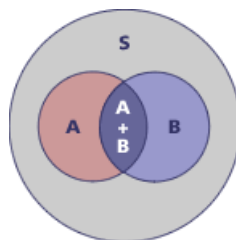


FIGURE 5420-2

Unless we can calculate $Fr(A \& B)$, we cannot in general calculate $Fr(A \vee B)$. If A and B are exclusive, however, then $Fr(A \& B) = 0$, and then we need only the individual frequencies on A and B to calculate $Fr(A \vee B)$.

5430: One or Both of Two Exclusive Properties

Provided that two properties A and B are exclusive (e.g., figure 5430-1), the relative frequency of having one or both is the sum of the individual relative frequencies.

$$Fr_S(A \vee B) = Fr_S(A) + Fr_S(B)$$

FIGURE 5430-1

If A and B are exclusive, then:

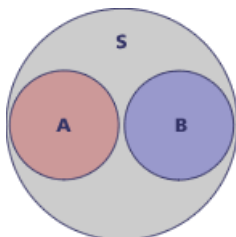


FIGURE 5430-2

6000: Case Studies

6100: The Salk Polio Vaccine

In the 1950s, infantile polio threatened every family with children. Jonas Salk, at the University of Pittsburgh, developed a vaccine which immunized animals against the disease. Salk's initial tests on humans revealed no ill side effects and seemed to increase the immune response, even in those who had previously had polio. In 1954 the Public Health Service of the United States organized two experiments to test the new polio vaccine. In total, more than 400,000 children were inoculated with the vaccine, and more than a million others -- the controls -- were either given a salt water injection (called a placebo), or no inoculation at all.

In the first study, called the Observed Control Study (OCS), dozens of schools were selected, all 2nd graders within a selected school were offered inoculations, and all 1st and 3rd graders were used as "observational controls." In the Observed Control Study table below, the first row records the total number of 2nd graders vaccinated (V) as well as the total number who contracted Polio. The second row records the total number of 2nd graders not vaccinated (~V) and how many of them contracted Polio. The third row records the 1st and 3rd graders who were not vaccinated (~V).

TABLE 6100-1: OBSERVED CONTROL STUDY

| Group | # of Children | # with Polio | Fr(P) |
|------------------|---------------|--------------|---------------------------|
| Grade 2 (V) | 221,998 | 56 | ??? -- see exercise below |
| Grade 2 (~V) | 123,605 | 55 | ??? -- see exercise below |
| Grade 1 & 3 (~V) | 725,173 | 391 | ??? -- see exercise below |
| Total (~V) | 848,778 | 446 | ??? -- see exercise below |
| Total | 1,070,776 | 502 | ??? -- see exercise below |

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

In the second study, called the Placebo Study (PS), a pool of subjects were selected, and, among these subjects, half were randomly selected to be vaccinated and half selected to get a placebo, which in this case was an injection of salt water. Neither the doctors giving the injections nor those diagnosing the patients knew which students had gotten the vaccination and which the placebo. Records were also kept on students in the same classes who got neither a vaccination nor a placebo.

TABLE 6100-2: PLACEBO STUDY

| Group | # of Children | # with Polio | Fr(P) |
|-----------------|---------------|--------------|----------|
| Vaccinated | 200,745 | 57 | 0.000284 |
| Not Vaccinated | 338,778 | 157 | 0.000463 |
| Placebo | 201,229 | 142 | 0.000706 |
| Total (-V & -P) | 540,007 | 299 | 0.000554 |
| Total | 740,752 | 356 | 0.000481 |

7000: Summary

The relative frequency of a property in a group is the proportion of individuals in the group that have that property. If we are considering property A, for example, then: $Fr(A) = \# \text{ of individuals with A} / \# \text{ of individuals in the group}$. The relative frequency of any property is therefore a number between 0 and 1 inclusive.

The relative frequencies of all the properties that correspond to the possible values of a variable are represented graphically with histograms and pie charts.

TABLE 7000-1: SMOKING HABITS AT THE UNIVERSITY OF AMSTERDAM

| Group | # of Individuals | Relative Frequency |
|------------------|------------------|--------------------|
| Non-smokers | 1200 | $Fr(NS) = 0.500$ |
| Moderate-smokers | 800 | $Fr(MS) = 0.333$ |
| Heavy-smokers | 400 | $Fr(HS) = 0.167$ |
| Total | 2400 | -- |

For example, the following pie chart and histogram represent the relative frequencies of the variable **SMOKER**, which can take on values: non-smoker, moderate-smoker, and heavy-smoker.

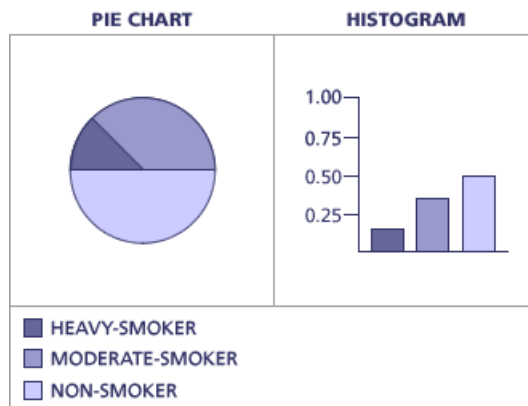


FIGURE 7000-1

The relative frequencies of complimentary properties A and $\sim A$ sum to 1, that is:

$$Fr_S(A) + Fr_S(\sim A) = 1$$

FIGURE 7000-2

If two properties have no overlap, that is, they are exclusive (like having pure blond and pure black hair). The relative frequency of having either of two exclusive properties is the sum of the frequencies for each individual property. That is, if A and B are exclusive, then:

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

FIGURE 7000-3

The relative frequency of having either or both of two properties (whether they overlap or not) is the sum of the individual properties minus their overlap:

$$Fr_S(\mathbf{A} \vee \mathbf{B}) = Fr_S(\mathbf{A}) + Fr_S(\mathbf{B}) - Fr_S(\mathbf{A} \& \mathbf{B})$$

FIGURE 7000-4