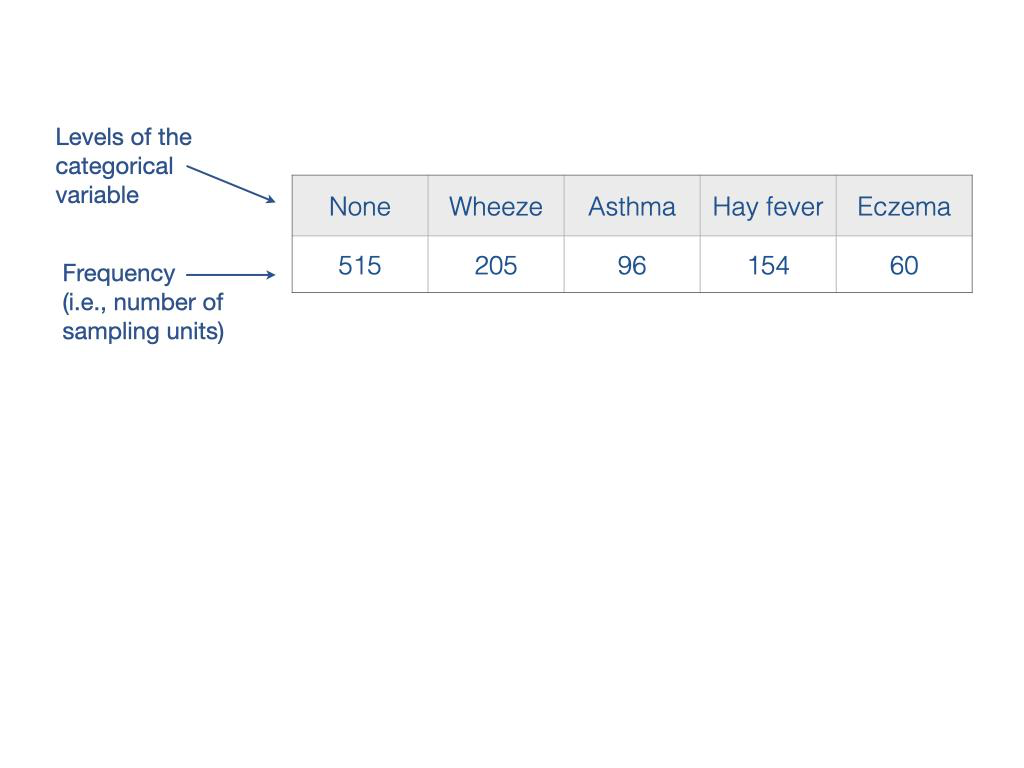
# Visualizing data

### Contingency tables

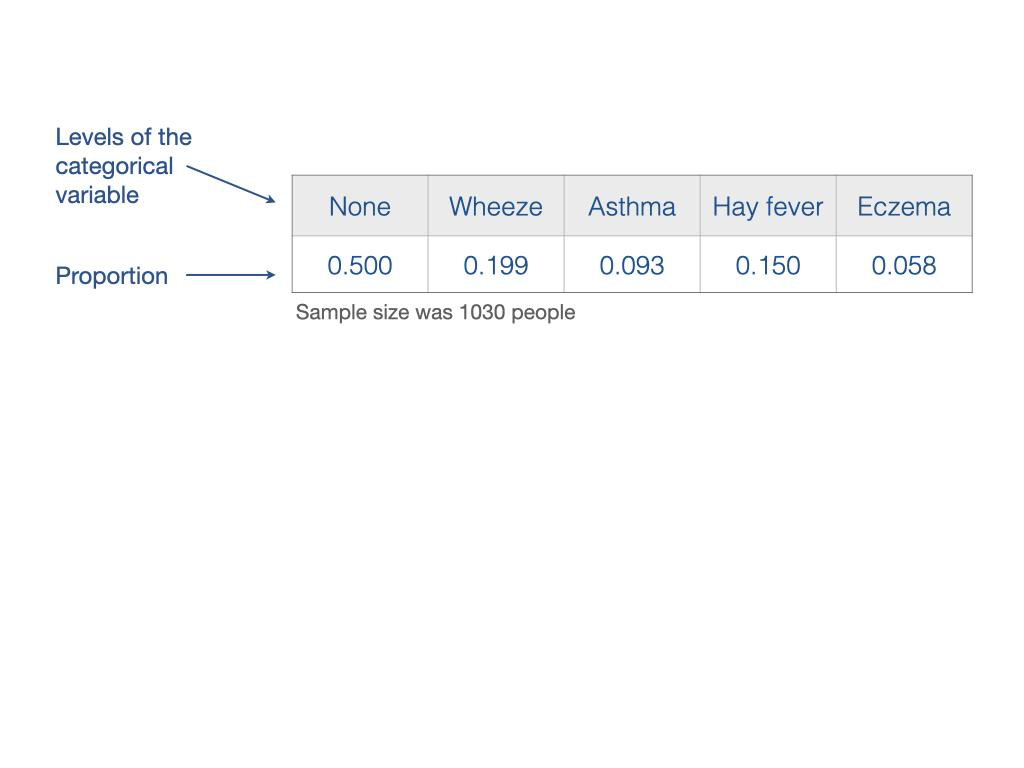
This lesson is about creating tables for categorical variables. By the end, you will be able to:

* Create a contingency table using categorical data
* Identify whether the data should be presented as a one-way or two-way contingency table
* Distinguish between a conditional distributionand marginal distribution

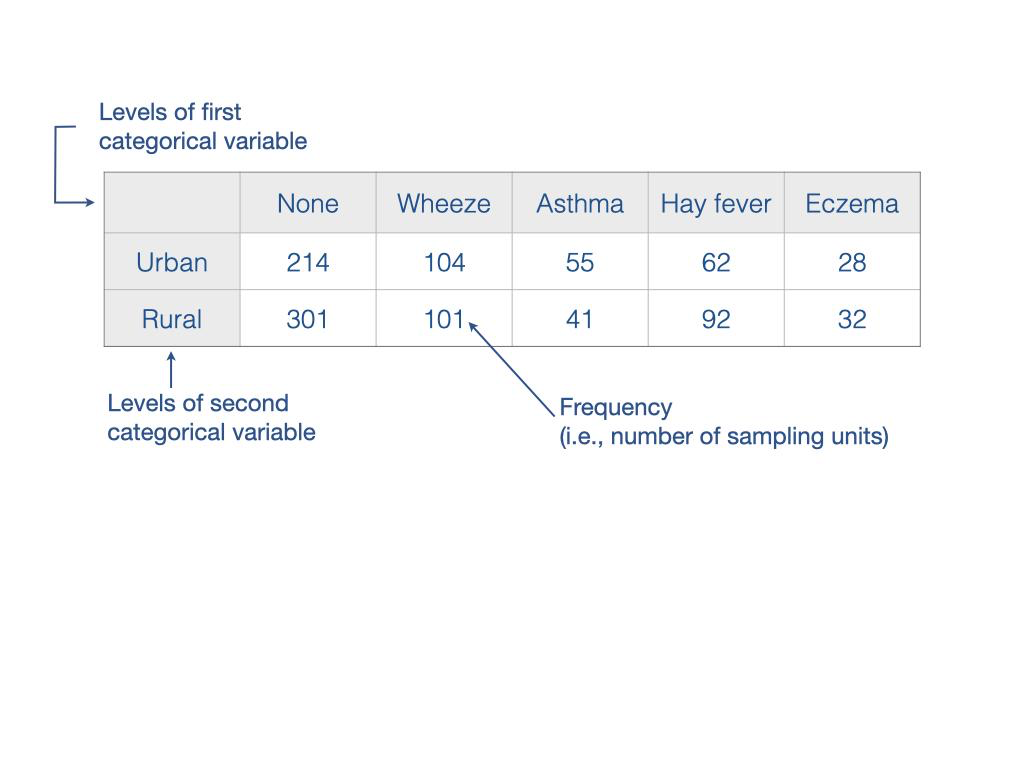
A contingency tableshows the frequency (or proportion) of sampling units in each level of a categorical variable. The frequency is simply the number of sampling units that falls in each level. For example, imagine that we surveyed a group of people to ask about common allergy symptoms and collected the following data:



Here is the same allergy symptom data as proportions.



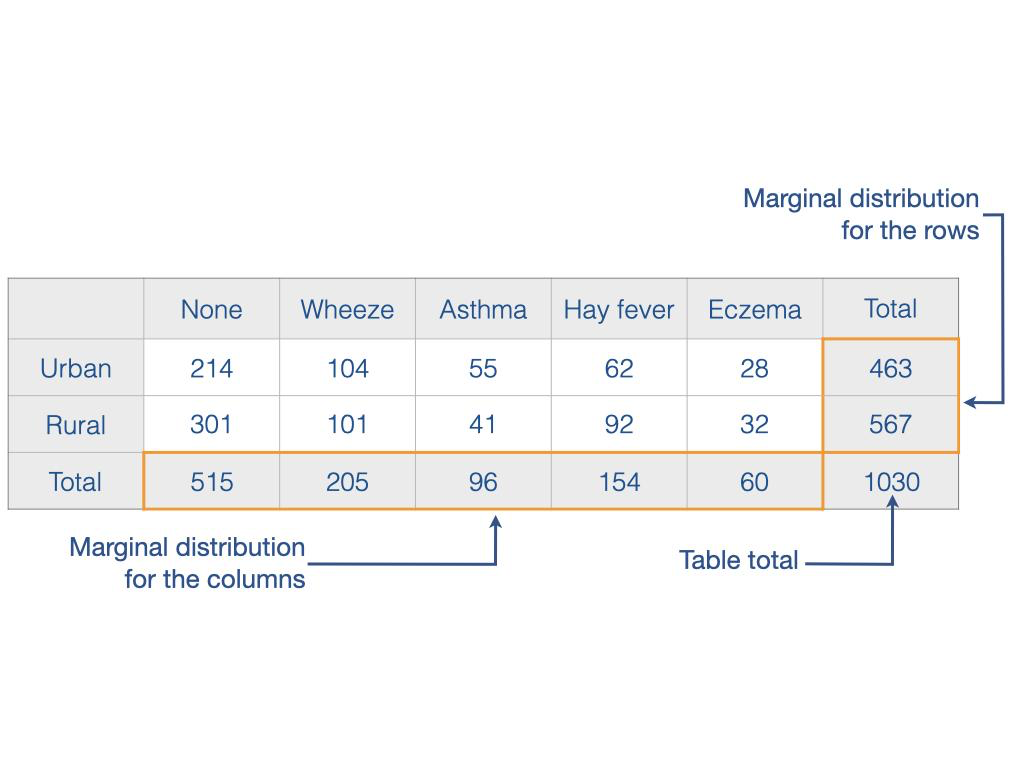
This may seem like odd terminology, but one-way contingency tablesand two-way contingency tablesjust refer to the number of categorical variables you observe for each sampling unit.

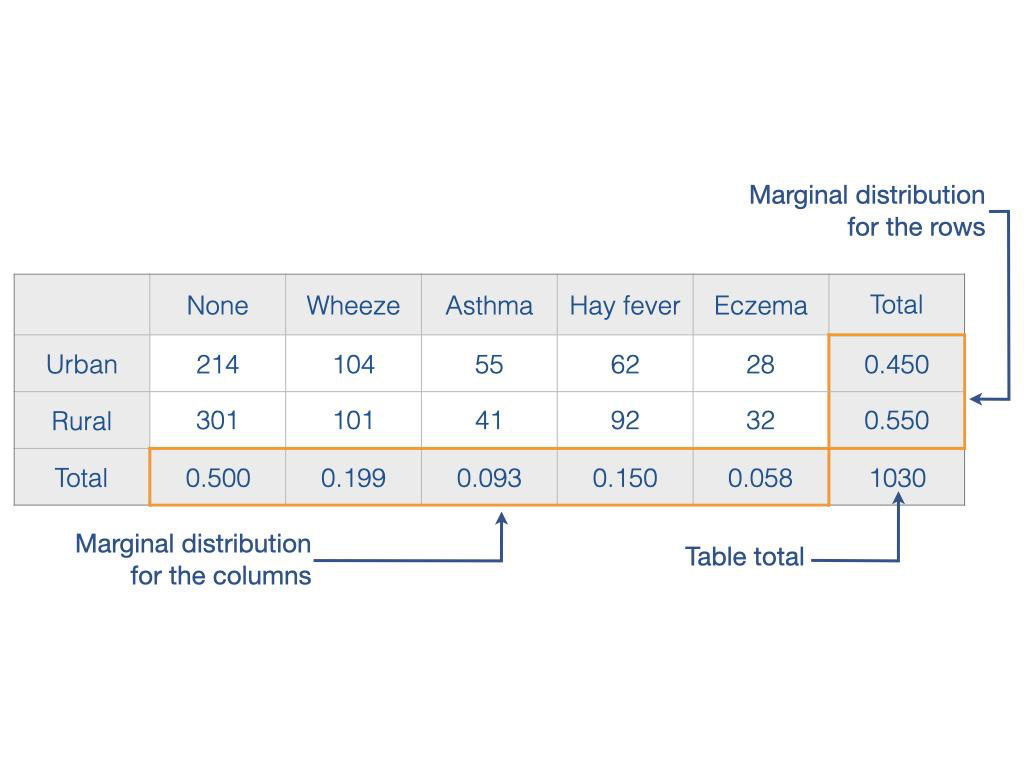


### Marginal distributions

Contingency tablesare an efficient way to summarize and visualize data, but they don't do a good job of highlighting patterns when there are two (or more) categorical variables. One way to see the overall patterns in the data is to calculate row and column frequencies, which are called marginal distributions.

When there are two (or more) categorial variables, it is common to present the total frequencies for each row and column. These are the simple sums across each row and column, and are called marginal distributions*.* The row marginal distributionshows the total counts for each row across all columns, and the column marginal distributionshows the total counts for each column across all rows. Marginal distributions can also be shown as proportions. To calculate marginal distributionsas proportions, the frequencies are divided by the table total.





### Conditional distributions

Contingency tablesare great at showing the raw data from a sample with two (or more) categorical variables, and marginal distributionshighlight the overall pattern of how sampling units fall into each level of the categorical variables. However, neither reveal anything about how the two categorical variables *interact* with each other. For example, if you ask a sample of people about their smoking (1st categorical variable) and alcohol consumption habits (2nd categorical variable), a contingency tablewith marginal distributionstells you the proportion of people that smoke, and the proportion of people that consume alcohol. But it would not tell you whether smokers drink more alcohol than non-smokers, which is a question about the interaction between the categorical variables. The interaction between categorical variables is visualized using conditional distributions.

To create a conditional distributiontable, you first select one of the categorical variables to be the primary variable and the other to be the secondary (conditional) variable. For example, if you want to know whether people who are smokers consume more alcohol compared to non-smokers, then the primary variable would be smoking and the secondary would be alcohol consumption.

