UNIT 6. BIVARIATE FREQUENCY DISTRIBUTIONS AND GRAPHIC PRESENTATIONS

In the previous units we have dealt with the statistical treatment of a single variable. But when we face the study of a population (for example, the socioeconomic situation of families in a city), the usual approach is to use different features of its individuals (family size, expenses and income, level of education, etc.). This not only provides a wider and more rewarding view of reality, but also makes it possible to study whether the different variables influence each other, i.e, "*Does family income affect the structure of family expenditure?*".

We begin by presenting the notation necessary for the joint analysis of two variables.

Bivariate Joint Frequency Distribution

When we jointly observe two variables X and Y from a population of size N, we have a two-dimensional statistical variable represented by (X, Y). Let's suppose that the variable X takes k values x_i (i=1,...,k) and that the variable Y takes values y_j (j =1,...,h). We can define the following frequencies:

- **Absolute joint frequency (n_{ij}):** number of times that the pair (x_i, y_j) has been observed:

$$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = N$$

- **Relative joint frequency (f_{ij}):** percentage of times that the pair (x_i, y_j) has been observed:

$$f_{ij} = \frac{n_{ij}}{N}$$

verifying that $\sum_{i=1}^{k}\sum_{j=1}^{h} f_{ij} = 1$

- **Bivariate joint absolute frequency distribution:** is defined by the three values $\boldsymbol{(x_i, y_j, n_{ij})}$ $\boldsymbol{i = 1 \dots k; j = 1 \dots h}$

The numerical representation of the data of a joint frequency distribution is done by means of a double-entry table called *cross-table*. The values of the variables X and Y are represented in the margins and the frequency of each couple of values is represented in the intersection cell. In addition, the size of the population can be calculated as:

$$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = N$$

X\Y	y_1	y_2	….	y_j	…	y_h
x_1	n_{11}	n_{12}	…	n_{1j}		n_{1h}
x_2	n_{21}	n_{22}	…	n_{2j}		n_{2h}
….	….	….	….	….	….	….
x_i	n_{i1}	n_{i2}	…	n_{ij}		n_{ih}
….	….	….	….	….	….	….
x_k	n_{k1}	n_{k2}	…	n_{kj}		n_{kh}

- **Bivariate joint relative frequency distribution:** is defined by the three values $(x_i, y_j, f_{ij}) \quad i = 1 \dots k; j = 1 \dots h$

  where

$$f_{ij} = \frac{n_{ij}}{N} \sum_{i=1}^{k} \qquad \sum_{j=1}^{h} f_{ij} = 1$$

The values of the variables X and Y are represented in the margins and the relative frequency of each couple of values is represented in the intersection cell.

X\Y	y_1	y_2	….	y_j	…	y_h
x_1	f_{11}	f_{12}	…	f_{1j}		f_{1h}
x_2	f_{21}	f_{22}	…	f_{2j}		f_{2h}
….	….	….	….	….	….	….
x_i	f_{i1}	f_{i2}	…	f_{ij}		f_{ih}
….	….	….	….	….	….	….
x_k	f_{k1}	f_{k2}	…	f_{kj}		f_{kh}

Example:

Given the following variables:

X = Monthly salary of the workers of a company (in euros)

Y = Education level of the workers of a company

The cross table of bivariate joint absolute frequency distribution of the example is represented as:

	Until secondary 1st stage	Secondary 2nd stage	High education
1000 - 1500	18	9	3
1500 - 2000	15	24	6
2000 - 2500	6	19	15
2500 - 3000	1	3	6

$$N = 125 \rightarrow f_{ij} = \frac{n_{ij}}{125}$$

	Until secondary 1st stage	Secondary 2nd stage	High education
1000 - 1500	14.4%	7.2%	2.4%
1500 - 2000	12.0%	19.2%	4.8%
2000 - 2500	4.8%	15.2%	12.0%
2500 - 3000	0.8%	2.4%	4.8%

By using this joint representation, we are going to be able to answer different questions related to both variables. For instance:

- How many workers are there with a higher education level?
- What percentage of workers earn more than €2,500?
- What is approximately the average salary?
- What is the most frequent level of education?

For this, we need to introduce a new concept: marginal distributions

Marginal distributions

From the joint frequency distribution of two variables, univariate distributions can be obtained for each of the variables separately. These distributions are called **marginal distributions** and are obtained by considering the values taken by one of the variables with their respective frequencies, independently of the values of the other variable. Therefore, marginal distributions can be characterised by all the measures of position, dispersion and shape studied in the previous lessons.

Marginal frequencies are obtained by **adding** the **joint frequencies** (absolute or relative) by rows or columns.

X \ Y	y_1	y_2	...	y_j	...	y_h	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1h}	$n_{1.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	$n_{i.}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kh}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.h}$	N

- **Absolute marginal frequency of X ($n_{i.}$)**: number of times the value x_i is observed, regardless of the values of Y

$$n_{i.} = \sum_{j=1}^{h} n_{ij} \qquad \sum_{i=1}^{k} n_{i.} = N$$

In the cross-table, the marginal distribution of X is obtained by adding the values of each row:

X\Y	y_1	y_2	...	y_j	...	y_h	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1h}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2h}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	$n_{i.}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kh}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.h}$	N

- **Absolute marginal frequency of Y ($n_{.j}$):** number of times the value y_j is observed, regardless of the values of X

$$n_{.j} = \sum_{i=1}^{h} n_{ij} \qquad \sum_{j=1}^{k} n_{.j} = N$$

In the cross-table, the marginal distribution of Y is obtained by adding the values of each column:

X\Y	y_1	y_2	...	y_j	...	y_h	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1h}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2h}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	$n_{i.}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kh}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.h}$	N

From the previous *example*:

	Until secondary 1st stage	Secondary 2nd stage	High education	
1000 - 1500	14.4%	7.2%	2.4%	24.0%
1500 - 2000	12.0%	19.2%	4.8%	36.0%
2000 - 2500	4.8%	15.2%	12.0%	32.0%
2500 - 3000	0.8%	2.4%	4.8%	8.0%
	32.0%	44.0%	24.0%	100.0%

We can also express the marginal distributions in relative terms:

X\Y	y_1	y_2	...	y_j	...	y_h	$f_{i.}$
x_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1h}	$f_{1.}$
x_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2h}	$f_{2.}$
...	...	...	...	...	...	...	...
x_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{ih}	$f_{i.}$
...	...	...	...	...	...	...	...
x_k	f_{k1}	f_{k2}	...	f_{kj}	...	f_{kh}	$f_{k.}$
$f_{.j}$	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.h}$	1

- Relative marginal frequency of X ($f_{i.}$):

$$f_{i.} = \frac{n_{i.}}{N} \qquad f_{i.} = \sum_{j=1}^{h} f_{ij} \qquad \sum_{i=1}^{k} f_{i.} = 1$$

- Relative marginal frequency of Y ($f_{.j}$):

$$f_{.j} = \frac{n_{.j}}{N} \qquad f_{.j} = \sum_{i=1}^{k} f_{ij} \qquad \sum_{i=1}^{h} f_{.j} = 1$$

Conditional distributions

Conditional distributions can be defined as the distribution of one of the two variables when a particular value/modality has been observed in the other variable. Like the marginal distributions, they are univariate distributions to which the statistical techniques studied in the previous topics can be applied. We can differentiate between:

- Distributions of X given Y
    - When **a value of Y is fixed** and the **distribution of** the variable **X** is analysed.
    - *Example*: analyse the salary (X) for workers with a certain level of education (a given value of Y).
- Distributions of Y given X

- When **a value of X is fixed** and the **distribution of** the variable **Y** is analysed.
- *Example*: analyse the level of education (Y) for workers with a given amount of salary (a given value of X).

Distributions of X given that Y=y_j

For a bivariate distribution (x_i, y_j, n_{ij}) $i = 1 \dots k; j = 1 \dots h$, the distribution of X given y_j is given by:

$$(x_i, n_{ij}) \quad i = 1 \dots k \text{ or } \left(x_i, f_{X=x_i|Y=y_j}\right) \quad i = 1 \dots k$$

X\Y	y_1	y_2	...	y_j	...	y_h	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1h}	
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2h}	
...	...	...	...	...	...	...	
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	
...	...	...	...	...	...	...	
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kh}	
$n_{.j}$				$n_{.j}$			

X\Y	y_j
x_1	n_{1j}
x_2	n_{2j}
...	...
x_i	n_{ij}
...	...
x_k	n_{kj}
$n_{.j}$	$n_{.j}$

$$f_{X=x_i|Y=y_j} = \frac{n_{ij}}{n_{.j}} \quad \forall\, i = 1, \dots, k \qquad \sum_{i=1}^{k} f_{i/j} = 1$$

X\Y	y_1	y_2	...	y_j	...	y_h	Marginal X
x_1	$f_{X=x1\|Y=y1}$	$f_{X=x1\|Y=y2}$	...	$f_{X=x1\|Y=yj}$	...	$f_{X=x1\|Y=yh}$	$f_{1.}$
x_2	$f_{X=x2\|Y=y1}$	$f_{X=x2\|Y=y2}$	...	$f_{X=x2\|Y=yj}$	...	$f_{X=x2\|Y=yh}$	$f_{2.}$
...	...	...	...	...	...	...	...
x_i	$f_{X=xi\|Y=y1}$	$f_{X=xi\|Y=y2}$	...	$f_{X=xi\|Y=yj}$	...	$f_{X=xi\|Y=yh}$	$f_{i.}$
...	...	...	...	...	...	...	...
x_k	$f_{X=xk\|Y=y1}$	$f_{X=xk\|Y=y2}$	...	$f_{X=xk\|Y=yj}$	...	$f_{X=xk\|Y=yh}$	$f_{k.}$
Total	**1**	**1**	...	**1**	...	**1**	**1**

The set of distributions of X, using relative frequencies, for each value/modality of the variable Y (Y=y_j j =1,...,h), is known as the **column profiles** (the total sum of each column equals 1).

Example of column profiles:

Distributions of salaries given the level of education.

	Until secondary 1st stage	Secondary 2nd stage	High education	$f_{i.}$
1000-1500	45.0%	16.4%	10.0%	24.0%
1500-2000	37.5%	43.6%	20.0%	36.0%
2000-2500	15.0%	34.5%	50.0%	32.0%
2500-3000	2.5%	5.5%	20.0%	8.0%
	100.0%	100.0%	100.0%	100.0%

Distributions of Y given that X=x_i

For a bivariate distribution(x_i, y_j, n_{ij}) $i = 1 \dots k; j = 1 \dots h$, the distribution of Y given x_i is given by:

$$(y_j, n_{ij})\ j = 1, \dots, h \quad \text{or} \quad \left(y_j, f_{Y=y_j|X=x_i}\right) \quad j = 1 \dots h$$

X\Y	y_1	y_2	...	y_j	...	y_h	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1h}	
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2h}	
...	...	...	...	...	...	...	
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	$n_{i.}$
...	...	...	...	...	...	...	
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kh}	
$n_{.j}$							

X\Y	y_1	y_2	...	y_j	...	y_h	$n_{i.}$
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	$n_{i.}$

$$f_{j|i} = f_{Y=y_j|X=x_i} = \frac{n_{ij}}{n_{i.}} \quad \forall\, j = 1, \dots, h \qquad \sum_{j=1}^{h} f_{j/i} = 1$$

X\Y	y_1	y_2	...	y_j	...	y_h	Total
x_1	$f_{Y=y1\|X=x1}$	$f_{Y=y2\|X=x1}$	...	$f_{Y=yj\|X=x1}$	...	$f_{Y=yh\|X=x1}$	1
x_2	$f_{Y=y1\|X=x2}$	$f_{Y=y2\|X=x2}$	...	$f_{Y=yj\|X=x2}$	...	$f_{Y=yh\|X=x2}$	1
...	...	...	...	...	...	...	...
x_i	$f_{Y=y1\|X=xi}$	$f_{Y=y2\|X=xi}$	...	$f_{Y=yj\|X=xi}$	...	$f_{Y=yh\|X=xi}$	1
...	...	...	...	...	...	...	...
x_k	$f_{Y=y1\|X=xk}$	$f_{Y=y2\|X=xk}$	...	$f_{Y=yj\|X=xk}$	...	$f_{Y=yh\|X=xk}$	1
Marginal Y	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.h}$	1

The set of distributions of X, using relative frequencies, for each value/modality of the variable X ($X=x_i$ i=1,…,k), is known as the **row profiles** (the total sum of each row equals 1).

Example of row profiles:

Distributions of the level of education for each range of salaries.

	Until secondary 1st stage	Secondary 2nd stage	High education	
1000-1500	60,0%	30.0%	10.0%	100.0%
1500-2000	33.3%	53.3%	13.3%	100.0%
2000-2500	15.0%	47.5%	37.5%	100.0%
2500-3000	10.0%	30.0%	60.0%	100.0%
$f_{.j}$	32.0%	44.0%	24.0%	100.0%

Statistical Dependence and Independence

Two variables X and Y are said to be **independent** if the values of one of them are not affected by the values of the other

X and Y are statistically independent if one of the following conditions is fulfilled:

1. $f_{i|j} = f_{i.} \quad \forall\ j = 1 \dots h$
2. $f_{j|i} = f_{.j} \quad \forall\ i = 1 \dots k$

} Row and column profiles are **equal** to the corresponding marginal distributions

3. $f_{ij} = f_{i.} \times f_{.j} \quad \forall i = 1 \dots k,\ \forall\ j = 1 \dots h$

All the three conditions are equivalent. If one of them is verified, the other two are verified. These are their meanings:

1. All the conditional distributions of X are the same as the marginal distribution of X
2. All the conditional distributions of Y are the same as the marginal distribution of Y
3. The relative joint frequencies coincide with the product of the corresponding relative marginal frequencies

Types of dependence

When variables are not independent, there are various possible degrees of dependence between two variables X and Y:

- **Functional dependence**: one of the variables can be expressed as a function of the other: **Y = f(X)**
- **Statistical dependence:** there exists a certain relationship between the variables (intermediate situations)

Example:

Distributions of X given Y (column profiles)

	Until secondary 1st stage	Secondary 2nd stage	High education	$f_{i.}$
1000-1500	45.0%	16.4%	10.0%	24.0%
1500-2000	37.5%	43.6%	20.0%	36.0%
2000-2500	15.0%	34.5%	50.0%	32.0%
2500-3000	2.5%	5.5%	20.0%	8.0%
	100.0%	100.0%	100.0%	100.0%

Column profiles ≠ Marginal distribution of X → Variables X and Y are NOT statistically independent

Distributions of Y given X (row profiles)

	Until secondary 1st stage	Secondary 2nd stage	High education	
1000-1500	60,0%	30.0%	10.0%	100.0%
1500-2000	33.3%	53.3%	13.3%	100.0%
2000-2500	15.0%	47.5%	37.5%	100.0%
2500-3000	10.0%	30.0%	60.0%	100.0%
$f_{.j}$	32.0%	44.0%	24.0%	100.0%

Row profiles ≠ Marginal distribution of Y → Variables X and Y are NOT statistically independent

Joint distribution of salary and level of education

	Until secondary 1st stage	Secondary 2nd stage	High education	
1000 - 1500	14.4%	7.2%	2.4%	24.0%
1500 - 2000	12.0%	19.2%	4.8%	36.0%
2000 - 2500	4.8%	15.2%	12.0%	32.0%
2500 - 3000	0.8%	2.4%	4.8%	8.0%
	32.0%	44.0%	24.0%	100.0%

It is necessary to check if $f_{ij} = f_{i.} \times f_{.j} \quad \forall i = 1,\ldots,k \quad j = 1,\ldots,h$

$i = 1, j = 1 \quad f_{11} = 14.4\,\% \quad f_{1.} \times f_{.1} = 32.0\,\% \times 24.0\,\% = 7.68\,\%$

$i = 1, j = 2 \quad f_{12} = 7.2\,\% \quad f_{1.} \times f_{.2} = 44.0\,\% \times 24.0\,\% = 10.56\,\%$

$i = 1, j = 3 \quad f_{12} = 2.4\,\% \quad f_{1.} \times f_{.2} = 24.0\,\% \times 24.0\,\% = 5.76\,\%$

$i = 2, j = 1 \quad f_{21} = 12.0\,\% \quad f_{2.} \times f_{.1} = 32.0\,\% \times 36.0\,\% = 11.52\,\%$

$i = 2, j = 2$ $f_{22} = 19.2\ \%$ $f_{2.} \times f_{.2} = 44.0\ \% \times 36.0\ \% = 15.84\ \%$

$i = 2, j = 3$ $f_{12} = 4.8\ \%$ $f_{1.} \times f_{.2} = 24.0\ \% \times 36.0\ \% = 8.64\ \%$

$i = 3, j = 1$ $f_{21} = 4.8\ \%$ $f_{2.} \times f_{.1} = 32.0\ \% \times 32.0\ \% = 10.24$

$i = 3, j = 2$ $f_{22} = 15.2\ \%$ $f_{2.} \times f_{.2} = 44.0\ \% \times 32.0\ \% = 14.08\ \%$

$i = 3, j = 3$ $f_{12} = 12.0\ \%$ $f_{1.} \times f_{.2} = 24.0\ \% \times 32.0\ \% = 7.68\ \%$

$i = 4, j = 1$ $f_{21} = 0.8\ \%$ $f_{2.} \times f_{.1} = 32.0\ \% \times 8.0\ \% = 2.56\ \%$

$i = 4, j = 2$ $f_{22} = 2.4\ \%$ $f_{2.} \times f_{.2} = 44.0\ \% \times 8.0\ \% = 3.52\ \%$

$i = 4, j = 3$ $f_{12} = 4.8\ \%$ $f_{1.} \times f_{.2} = 24.0\ \% \times 8.0\ \% = 1.92\ \%$

As can be seen, it is proven that the variables X and Y are NOT statistically independent.