4. 2-Dimensional Empirical Distributions

4.1. Contingency Tables

Suppose you want to determine whether working besides studying has an impact upon study progress and you are given the following data:

Study Progress	working	not working	Sum
slower than average	16	11	27
average	25	75	100
faster than average	3	14	17
Sum	44	100	144

This is an exampel of a contingency table (ristiintaulukko/ frekvenssitaulukko/ kontingenssitaulukko).

Generally, suppose you have two statistical variables, x and y say, the first one of which is classified into J categories $E_1, E_2, \ldots, E_J$, and the second one of which is classified into I categegories $G_1, G_2, \ldots, G_I$. Then a contingency table looks like this:

x	E_1	E_2	$\ldots$	E_J	
y					Sum
G_1	f_{11}	f_{12}	$\cdots$	f_{IJ}	$f_{1\bullet}$
G_2	f_{21}	f_{22}	$\cdots$	f_{2J}	$f_{2\bullet}$
$\vdots$			f_{ij}		$\vdots$
G_I	f_{I1}	f_{I2}	$\cdots$	f_{IJ}	$f_{I\bullet}$
Sum	$f_{\bullet 1}$	$f_{\bullet 2}$	$\ldots$	$f_{\bullet J}$	$f_{\bullet\bullet} = n$

where:

x measures the suspected cause of interdependence.
y measures the suspected effect of interdependence.
f_{ij} measures the count of statistical units, whose x-value belongs to class E_j and whose y-value belongs to class G_i. It is called the observed frequency (havaittu frekvenssi) of cell (G_i, E_j).
$f_{i\bullet} = f_{i1} + \ldots + f_{iJ}$ is row i's row total (rivisumma).
$f_{\bullet j} = f_{1j} + \ldots + f_{Ij}$ is the column total (sarakesumma) of column j.
$f_{\bullet\bullet} = n$, called the grand total (kokonaissumma), is the sum of all statistical units.

The collection of all observed frequencies f_{ij} make up the joint distribution (yhteisjakauma) of x and y. The collection of all row totals $f_{i\bullet}$ make up the marginal distribution (reunajakauma) of y, and the collection of all column totals $f_{\bullet j}$ make up the marginal distribution of x. Clearly, both the sum of all row totals $f_{i\bullet}$, and of all column totals $f_{\bullet j}$ equal the grand total $f_{\bullet\bullet} = n$.

We obtain relative frequencies (suhteellinen yhteisjakauma) by dividing each cell frequency f_{ij} by the total number of observations n. Multiplying relative frequencies by hundred yields procentual frequencies (prosentuaalinen yhteisjakauma).

Example. The procentual frequencies

Study Progress	working	not working	Sum
slower than average	11%	8%	19%
average	17%	52%	69%
faster than average	2%	10%	12%
Sum	30%	70%	100%

are obtained from the original table

Study Progress	working	not working	Sum
slower than average	16	11	27
average	25	75	100
faster than average	3	14	17
Sum	44	100	144

by dividing all numbers by 144 and multiplying with 100%.

Consider again our contingency table:

Study Progress	working	not working	Sum
slower than average	16	11	27
average	25	75	100
faster than average	3	14	17
Sum	44	100	144

Each row and column of the joint distribution of x (working or not) and y (study progress) defines a conditional distribution (ehdollinen jakauma) in the following sense:

Each column tells us the value of y (study progress) conditional upon that we know the value of x (working or not). Similarly, each row tells us the value of x (working or not) conditional upon knowing the value y (study progress).

We may obtain relative or procentual conditional distributions (suhteellinen eli prosentuaalinen frekvenssijakauma) by dividing by the relevant row or sum totals (and multiplying with 100%).

Example: (study progress continued.)
Consider again our contingency table:

Study Progress	working	not working	Sum
slower than average	16	11	27
average	25	75	100
faster than average	3	14	17
Sum	44	100	144

We obtain the procentual conditional distributions of y (study progress) conditional upon x (working or not) by dividing by the relevant column total and multiplying with 100%:

Study Progress	working	not working	Sum
slower than average	36%	11%	19%
average	57%	75%	69%
faster than average	7%	14%	12%
Sum	100%	100%	100%

x and y are called statistically independent (tillastollisesti riippumaton) if the relative or procentual conditional distributions do not differ between columns (if conditioned on x) or rows (if conditioned on y). Otherwise they are called statistically dependent (tillastollisesti riippuva).

For each contingency table with given marginal distributions it is possible to calculate exactly one set of observations which would correspond to statistical independence of the row and column variables. These are given by the so called expected frequencies (odotetut eli teoreettiset frekvenssit) e_{ij} determined from the column and row totals as

$$e_{ij} = \frac{f_{i\bullet} \cdot f_{\bullet j}}{n} \quad \text{for all } i, j.$$

In order to see that such frequencies correspond indeed to identical conditional distributions in each row and column, note that division of e_{ij} by $f_{i\bullet}$ yields

$$\frac{e_{ij}}{f_{i\bullet}} = \frac{f_{\bullet j}}{n} \quad \text{for all rows } i,$$

and that division of e_{ij} by $f_{\bullet j}$ yields

$$\frac{e_{ij}}{f_{\bullet j}} = \frac{f_{i\bullet}}{n} \quad \text{for all columns } j.$$

That is, the conditional relative distributions are the same for every row and column.

Example.

The expected frequencies read in our case:

Study Progress	working	not working	Sum
slower than average	8	19	27
average	31	69	100
faster than average	5	12	17
Sum	44	100	144

The conditional distributions of y (study progress) given x (working or not) and the corresponding marginal distribution of the expected frequencies are:

Study Progress	working	not working	Sum
slower than average	19%	19%	19%
average	69%	69%	69%
faster than average	12%	12%	12%
Sum	100%	100%	100%

The conditional distributions of x (working or not) given y (study progress) and the corresponding marginal distribution of the expected frequencies are:

Study Progress	working	not working	Sum
slower than average	31%	69%	100%
average	31%	69%	100%
faster than average	31%	69%	100%
Sum	31%	69%	100%

The difference between expected and observed frequencies displays the amount of statistical dependence between x and y. Such is measured with Pearson's χ^2 statistics defined as:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}.$$

The value of χ^2 is zero if all observed frequencies f_{ij} equal exactly their expected counterpart, that is, when x and y are statistically independent. The stronger the statistical dependence is, the higher the value of χ^2 becomes. However, it can be shown that its value cannot exceed

$$\chi^2_{\text{max}} = (k-1)n,$$

where k is the smaller number of I and J, and n is the number of observations.

The dependence of χ^2_{max} upon I, J and n implies that χ^2 may not be used to compare tables of different size or even just different numbers of observations among each other. It also implies that the seriousness of statistical dependence for nonzero values of χ^2 are hard to judge based upon this measure alone.

The dependence upon the number of observations n may be removed by calculating the so called contingency coefficient (kontingenssikerroin) C defined as

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \text{ with maximum } C_{\text{max}} = \sqrt{\frac{k-1}{k}},$$

where k is again the smaller number of I and J. The value of C is zero if x and y are statistically independent and rises with statistical dependence between x and y. Since C_{max} still depends upon k, C should not be used to compare tables with different numbers of rows and/or columns, unless C is stated as a fraction of C_{max}.

A measure of statistical dependence which is independent of both k and n and furthermore easy to interpret, is given by Cramer's V:

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\text{max}}}} = \sqrt{\frac{\chi^2}{n(k-1)}}.$$

It ranges from 0 to 1 (no/perfect dependence). As a rule of thumb, there is no substantial dependence if $V < 0.1$.

Example.

The expected frequencies in our case were:

Study Progress	working	not working	Sum
slower than average	8	19	27
average	31	69	100
faster than average	5	12	17
Sum	44	100	144

to be compared with the observed frequencies:

Study Progress	working	not working	Sum
slower than average	16	11	27
average	25	75	100
faster than average	3	14	17
Sum	44	100	144

$$\chi^2 = \frac{(16-8)^2}{8} + \frac{(11-19)^2}{19} + \ldots + \frac{(14-12)^2}{12} \approx 14.2,$$

$$\chi^2_{\text{max}} = (k-1)\cdot n = (2-1)\cdot 144 = 144,$$

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\text{max}}}} = \sqrt{\frac{14.2}{144}} \approx 0.31 > 0.1,$$

so there is substantial dependence between x and y.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{14.2}{14.2+144}} \approx 0.3,$$

$$C_{\text{max}} = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{2-1}{2}} \approx 0.7.$$

Two-Way Tables (nelikentäjä)

Contingency tables with only two rows and two columns are called Two-Way Tables (nelikentäjä). For those, the calculation of the χ^2 statistics simplifies to:

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1\bullet}f_{2\bullet}f_{\bullet 1}f_{\bullet 2}}.$$

Example. The variabel x attains the values "husband" or "wife" depending upon who in the couple takes usually care of the talking, and the variabel y attains the values "husband" or "wife" depending upon which partner usually decides common matters. A random sample of 34 married couples yields the following contingency table:

x y	husband	wife	 Sum
husband	13	6	19
wife	5	10	15
Sum	18	16	34

such that:

$$\chi^2 = \frac{34(13 \cdot 10 - 6 \cdot 5)^2}{19 \cdot 15 \cdot 18 \cdot 16} \approx 4.1,$$

$$\chi^2_{\text{max}} = (2-1) \cdot 34 = 34,$$

$$V = \sqrt{\frac{4.1}{34}} \approx 0.35 > 0.1,$$

$$C = \sqrt{\frac{4.1}{4.1 + 34}} \approx 0.33,$$

$$C_{\text{max}} = \sqrt{\frac{2-1}{2}} \approx 0.71.$$

So there is statistical dependence between x and y in the sense that the more talkative spouse tends to decide common matters.