The Null Hypothesis and Some Others

In a class of 30 students, the relation between the grades of the X and Y subjects was calculated. The results are illustrated by the following 5x5 chart.

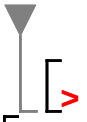
Két tantárgy osztályzatainak kapcsolata		Y tantárgy jegyei					Összeg
		1	2	3	4	5	
X tantárgy jegyei	1	2	6	0	0	0	8
	2	0	6	5	0	0	11
	3	0	1	4	2	0	7
	4	0	0	1	0	1	2
	5	0	0	0	1	1	2
Összeg		2	13	10	3	2	30

the relation between the grades of the 2 subjects, the grades of the Y subject, sum, the grades of the X subject

The number located in the intersection of the ith row and kth column of the chart shows that how many students of the class have grade i in the X subject and grade k in the Y subject. Naturally, the values of i and k can be 1,2,3,4 and 5 independently of each other. For example, grade 2 in the intersection of the 3rd row and 4th column of the chart means that there are two students out of the 30 who have 3 in the X and 4 in the Y subject.

Decide with the help of the [képlet] test whether the grades of the X subject are independent of the grades of the Y subject on a 5% significance level.

Calculate the value of the correlation coefficient between the 5.5=25 data observed and the calculated values based on independency. Plot the values of the two data series in the same coordinate system. Is there a close relation between the two value series?



For the sake of the more flexible operation of the programme, we query the data from the “grades.txt” data base. This file always contains 25 data no matter how many students there are in the class. The chart above was entered continuously into the file. Thus the numbers [2,6,0,0,0] appear divided by spaces in the first row. The entering, editing and modification of the data can be done by the Notepad program located in the directory of the Windows accessories.

We can have the data read from the file with the help of the readdata procedure. The first parameter of the readdata procedure is a file name opened for reading. The types of the data, which is the integer, located in the file were given in its second parameter so every data is an integer. Its third parameter is a

positive integer which shows how much data there is in a row. In this case it is 5. The reading goes row by row, which makes the whole file be put into a list of lists data type the name of which is “data”.

```
> restart:
  f := fopen("osztalyzatok.txt",READ):
  adatok := readdata(f,integer,5);
  fclose(f);
```

$$adatok := [[2, 6, 0, 0, 0], [0, 6, 5, 0, 0], [0, 1, 4, 2, 0], [0, 0, 1, 0, 1], [0, 0, 0, 1, 1]] \quad (1)$$

The file can be opened by the fopen procedure. Its first parameter is the name of the file in the form of a string and its second parameter is the READ key word. This shows that we have opened the “grades. txt” for reading. We tag the result of the opening with an f to which we can refer from now on. The readdata procedure must be given this f name as its first parameter. In the instruction above, we did not give the whole directory path of the file (elérési útvonal). In this case the system looks for the file to be opened in its own working directory. If it does not find it we will get an error alert. And do not forget to close the data files with the fclose procedure if you do not use them.

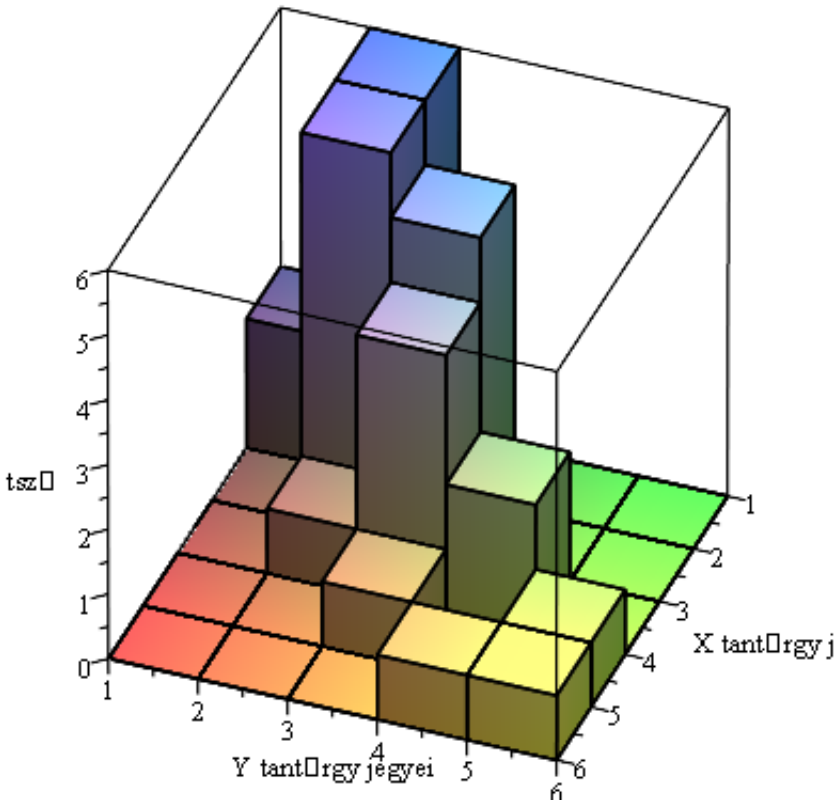
First, let’s convert the data read into Matrix data type because the ChiSquareIndependenceTest procedure will require this data type later in the task.

```
> jegyek := convert(adatok, Matrix)
```

$$jegyek := \begin{bmatrix} 2 & 6 & 0 & 0 & 0 \\ 0 & 6 & 5 & 0 & 0 \\ 0 & 1 & 4 & 2 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad (2)$$

So far we know that the matrix has lots of zeros, it is not symmetrical and the non-zero elements are located one step right and left next to the main diagonal. Before determining its independence, display the distribution of the headcount located in the matrix with the matrixplot graphic procedure.

```
> plots[matrixplot](jegyek,heights=histogram,orientation=[21,54],
  axes=boxed,
                    labels=["X tantárgy jegyei","Y tantárgy
  jegyei","Létszám"]);
```



According to the heights of the columns, there are a lot of (2,2) and (1,2) grade pairs while the number of (5,5) and (4,5) grade pairs is rather low. Unfortunately, the low grades are dominant in this class.

The first part of the task can be easily solved because of the built-in chi square method, called ChiSquareIndependenceTest, related to data independence test in the Statistics package. For the sake of the total display of the replies of the procedure, we set the infolevel variable of the Statistics package to 1 thus each of the results calculated by the procedure is displayed. Let's check the instruction in the case of the infolevelStatistics:=0 setting.

```
> with(Statistics):
> infolevel[Statistics] := 1;
```

$$infolevel_{Statistics} := 1 \qquad (3)$$

```
> statisztika := ChiSquareIndependenceTest(jegyek, level = 0.05);
Chi-Square Test for Independence
--------------------------------
Null Hypothesis:
Two attributes within a population are independent of one
```

```
another
Alt. Hypothesis:
Two attributes within a population are not independent of one
another

Dimensions:              5
Total Elements:          30
Distribution:            ChiSquare(16)
Computed statistic:      36.6563
Computed pvalue:         0.00234355
Critical value:          criticalvalue

Result: [Rejected]
There exists statistical evidence against the null hypothesis
```

$$statisztika := hypothesis = false,\ criticalvalue = 26.29622762,\ distribution = ChiSquare(16),\ pvalue = 0.0023435472,\ statistic = 36.65634366 \quad \textbf{(4)}$$

The first parameter of the ChiSquareIndependenceTest procedure is the joint distribution of the grades given in the form of a 5x5 matrix. This is also called a contingency chart. The second parameter of the procedure is the significance level the default value of which is 5%, that is, level=0.05. This value may not be given. We will get back to the significance level later.

All the data of the independence statistics can be seen on the result. Let's start its interpretation with the login row which shows that an independence test is to come with the help of the chi square method.

Below the horizontal dividing line comes the so-called null hypothesis. It determines that two items, in this case the grades of the X and the Y subjects, are independent of each other. This independence means that there is no relation between them, or if we consider the slogan of the Greek philosophers that "everything is connected" then this means that their connection is not so strong.

Don't misunderstand it: we have not found the answer to the independence. We have only come up with a hypothesis. This H0 hypothesis has to be decided with the chi square method concerning the data. The answer to the hypothesis will either be true or false.

In the next row there is the negation of the H0 hypothesis which can be called an alternative hypothesis. Denote it with H1 according to which the two properties of the population depend on each other. In this case the population is the class and the two properties mean the performance of the students concerning the two subjects.

Then come the dimension (=5) and the number of the elements (=30) rows which show that the X and Y size of the chart is 5 and there are 30 students in the class.

The chi square distribution has only one parameter which is called the degree of freedom. The degree of freedom of the chi square distribution used for the calculation is (n-1) (m-1)=16. Let's explore the topic of the chi square distribution with 16 degrees of freedom. Let's give its density function and its graph.

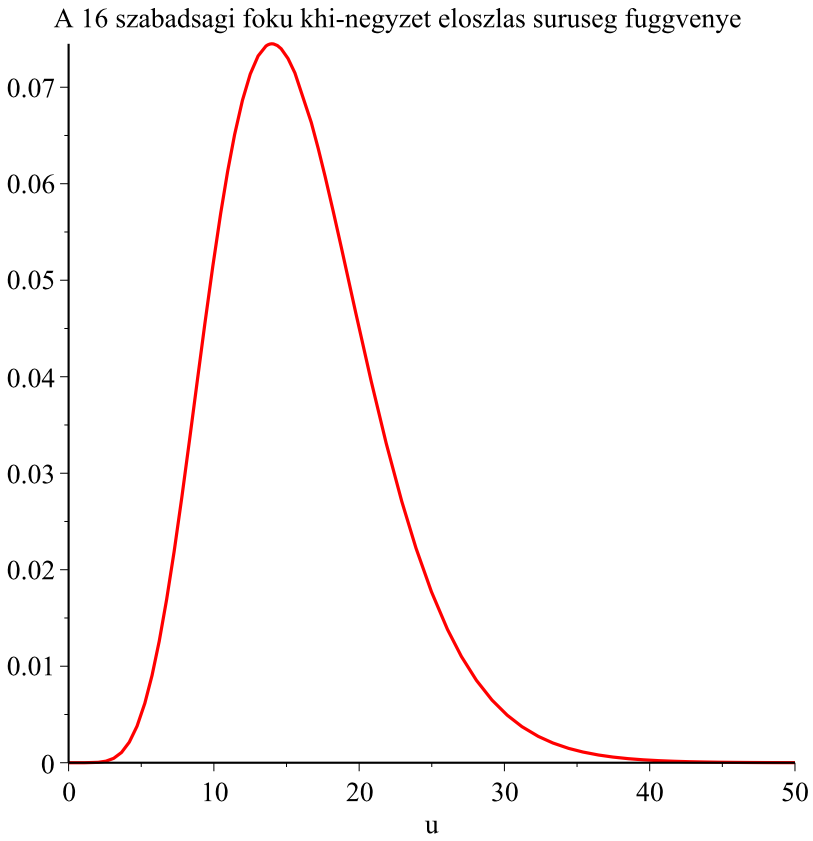
```
> Khi16 := RandomVariable(ChiSquare(16)) :
> suruseg:=PDF(Khi16, u);
```

(5)

$$suruseg := \begin{cases} 0 & u < 0 \\ \frac{1}{1290240}\, u^7\, e^{-\frac{1}{2}u} & otherwise \end{cases} \quad (5)$$

> *rajz16* := *plot*(*suruseg*, *u* = 0 ..50, *title* = "A 16 szabadsagi foku khi-negyzet eloszlas suruseg fuggvenye") :

> *rajz16*



With the help of the RandomVariable procedure, we have entered a random variable called chi16 with 16 degrees of freedom and the type of which is ChiSquare(16). Then we have created its density function with the PDF (Probability Density Function) procedure. The density function is the exponent function of the u independent variable, which we got plotted in the [0,50] interval. Since it is a density function, the area below the non negative and the whole curve is 1.

> *Int*(*suruseg*, *u* = 0..*infinity*) = *int*(*suruseg*, *u* = 0..*infinity*)

(6)

$$\int_0^{\infty} \left(\begin{cases} 0 & u < 0 \\ \frac{1}{1290240}\, u^7\, e^{-\frac{1}{2}u} & otherwise \end{cases} \right) du = 1 \qquad (6)$$

Prove that the area below the curve calculated from the criticalvalue=26.29622762, that is, from the critical value returned by the ChiSquareIndependenceTest to the infinity is exactly 0.05. This 0.05 is the significance level given. We only have to integrate the density function in the [26.29622762, infinity) domain.

```
> statisztika[2];
```

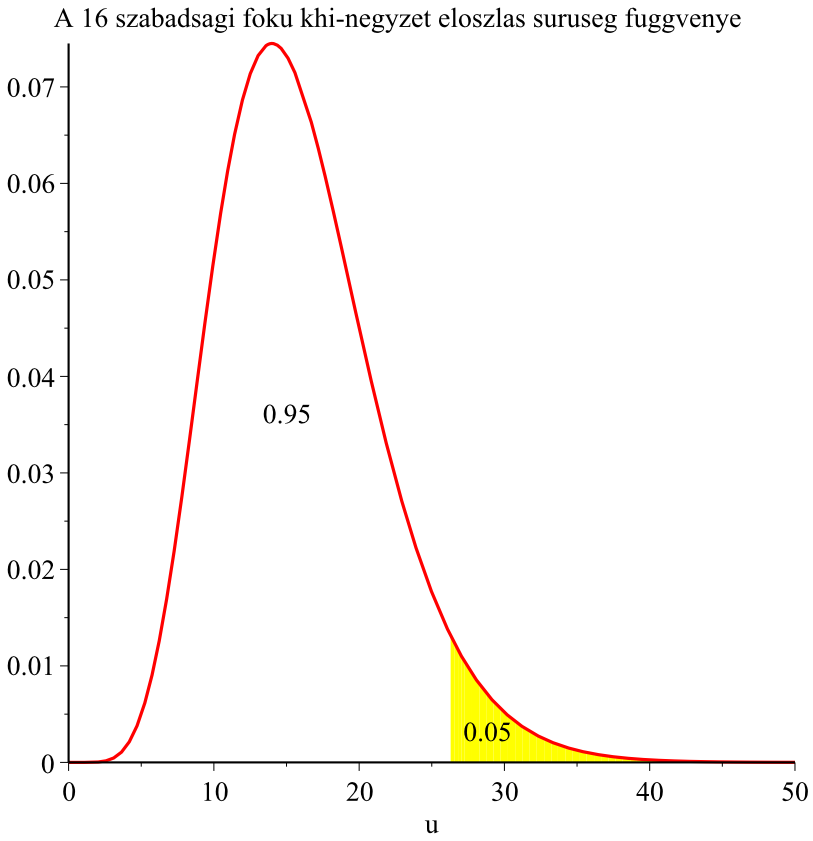
```
> Int(suruseg, u = rhs(statisztika[2])..infinity) = int(suruseg, u = rhs(statisztika[2])..infinity);
```

$$criticalvalue = 26.29622762$$

$$\int_{26.29622762}^{\infty} \left(\begin{cases} 0 & u < 0 \\ \frac{1}{1290240}\, u^7\, e^{-\frac{1}{2}u} & otherwise \end{cases} \right) du = 0.04999999980 \qquad (7)$$

How did we get the critical value? This is the second element of the series returned by the ChiSquareIndependenceTest procedure. We have kept the reply in the “statistics” variable. To illustrate what has been said so far, plot the density function and colour the area the size of which is 0.05 below the curve of the chi square density function to yellow.

```
> satir16:=plot(suruseg,u=rhs(statisztika[2])..50,filled=true,
  color=yellow):
> s16:=plots[textplot]([[28.8,0.003,`0.05`],[15,0.036,`0.95`]]):
> abra16:=plots[display]([rajz16,satir16,s16]):abra16;
```



The fact that we accept or reject the H0 hypothesis and the other fact that the H0 hypothesis is really true or false are two independent statements. According to this, 2x2=4 matches are possible which you can see in the chart below.

Accept\H0 is satisfied	The H0 hypothesis is true	**The H0 hypothesis is false**
We accept the H0 hypothesis	The method works well	type II error
We reject the H0 hypothesis	type I error	The method works well

Our method operates well in case

1.the H0 hypothesis is accepted or

2.the false H0 hypothesis is rejected.

These cases are shown in the main diagonal of the chart. But there are two wrong cases which are considered errors.

1.When we reject the true H0 hypothesis, we make a type I error.

2.If we accept the false H0 hypothesis we make a type I error.

In case we accept an H0 hypothesis with the help of the chi square test on a 5% significance level, then it means that the type I error of this decision will be smaller than 0.05. So we reject a true H0 hypothesis in the cases of less than 5%. Practically, if we do 20 independent experiments on the same true hypothesis then the number of the cases rejected by the chi square test may not be more than 1 because 1/2=0.05.

We have **to** *admit that with the help of the* chi *square test we can only examine the type I errors but* **not** *the type II errors*.
In practise, the 0.001, 0.01 **and** 0.05 *values are used* **for** *the significance level. We can say that*

- *the value between* 5 *%* **and** 1 *% is almost significant*
- *the value between* 1 *%* **and** 0.1 *% is significant*
- *the value below* 0.1 *% is highly significant*

After this, the decision of the statistical hypothesis is done **in** *a way that the procedure calculates a so - called statistical value* **from** *the data of the matrix, which is called computed statistics . We can get this value at two places* : **in** *the text* **and** *as the last, fifth element of the values returned.*

> *statisztika*[5]

$$statistic = 36.65634366 \quad (8)$$

In our case the computed statistics is 36.65634366. We compare this with the former 26.29622762 critical value and we accept the H0 hypothesis if the computed statistics is lower than the critical value calculated from the chi square density. In any other cases we reject the H0 hypothesis.

> (*statisztika*[2] < *statisztika*[5])

> *statisztika*[1];

$$(criticalvalue = 26.29622762) < (statistic = 36.65634366)$$
$$hypothesis = false \quad (9)$$

According to this we have to reject the H0 hypothesis. We can see this decision in this row

Result: [Rejected]

returned by the procedure. The next sentence explains the reason for this.

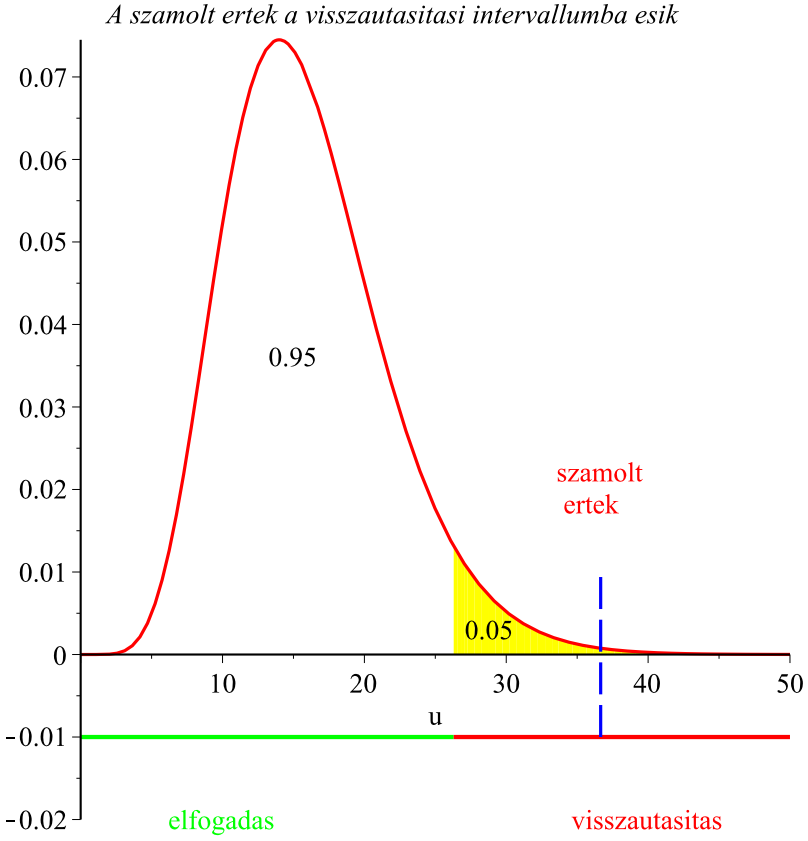
There exists statistical evidence against the null hypothesis.

In our case it means that the H0 independence hypothesis, that is, the truth of the H1 hypothesis written for the independence of the given X and Y subjects has come into force.

So as we have expected, the grades of the X and Y subjects are not independent of each other. We do not know yet how much they depend on each other but it is sure that they affect each other. So if someone has low grades in one subject then it is highly possible that he has the same grades in the other subject. The same is true for the good grades. We mentioned this at the beginning but now we can see them calculated.

We can complete our graph with a horizontal line which illustrates the rejection interval in red and the accept interval in green. Furthermore, we have plotted the statistical value calculated into the graph, which is in the red domain.

```
> zold:=plot([[0,-0.01],[rhs(statisztika[2]),-0.01]],color=green,
  thickness=2):piros:=plot([[rhs(statisztika[2]),-0.01],[50,-0.01]
  ],color=red,thickness=2):igaz:=plots[textplot]([[10,-0.02,
  `elfogadas`]],color=green):
  hamis:=plots[textplot]([[40,-0.02,`visszautasitas`]],color=red):
  stat16:=plot([[rhs(statisztika[5]),-0.01],[rhs(statisztika[5]),
  0.01]],color=blue,linestyle = 3):
  mutat:=plots[textplot]([[rhs(statisztika[5]),0.02,`szamolt\n
  ertek`]],color=red):
  plots[display]([abra16,zold,piros,igaz,hamis,stat16,mutat],title=
  `A szamolt ertek a visszautasitasi intervallumba esik`);
```



Let's start the search for the correlation coefficient by completing the chart. Expand the matrix with a column containing sums of rows and a row containing sums of columns. As a first step, create the XS column vector and the YS row vector.

```
> XS,YS := convert([seq(add(jegyek[i, k], k = 1 .. 5), i = 1 .. 5)
  ], Vector),
```

```
> convert([seq(add(jegyek[i, k], i = 1 .. 5), k = 1 .. 5)], Vector
  [row]);
```

$$XS,\ YS := \begin{bmatrix} 8 \\ 11 \\ 7 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 & 13 & 10 & 3 & 2 \end{bmatrix} \tag{10}$$

For the addition, we have used the add procedure instead of the sum because it is quicker at adding numbers. It is reassuring that the sum of the elements of the X and Y vectors coincide with the headcount, that is, 30.

```
> letszam := add(XS[i], i = 1 ..5); Sum('YS'[i], i = 1 ..5) = add(YS[i], i = 1 ..5);
```

$$letszam := 30$$

$$\sum_{i=1}^{5} YS_i = 30 \tag{11}$$

The expansion of the matrix can be done by the Concatenate procedure of the ArrayTools package. First, we concatenate the XS column vector subsequent to the last column of the grades matrix then we put the YS sum after the last row of the matrix received. Finally, we put the headcount into the lower right corner.

```
> with(ArrayTools) : Concatenate(2, jegyek, XS);
> Observed := Concatenate(1,%, Concatenate(2, YS, < letszam >));
```

$$\begin{bmatrix} 2 & 6 & 0 & 0 & 0 & 8 \\ 0 & 6 & 5 & 0 & 0 & 11 \\ 0 & 1 & 4 & 2 & 0 & 7 \\ 0 & 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{bmatrix}$$

$$Observed := \begin{bmatrix} 2 & 6 & 0 & 0 & 0 & 8 \\ 0 & 6 & 5 & 0 & 0 & 11 \\ 0 & 1 & 4 & 2 & 0 & 7 \\ 0 & 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 2 \\ 2 & 13 & 10 & 3 & 2 & 30 \end{bmatrix} \tag{12}$$

Maybe you can recall from your probability theory studies that we consider an Ai event independent of the Bk event if the

$$P(A_i B_k) = P(A_i)\, P(B_k)$$

equality is fulfilled, that is, the probability of the product event is equal to the product of the probabilities of the factors.

The Ai should denote the event concerning the X subject that the grade is i. Bk should mean that the grade of the Y subject is k. We can approximate the P(Ai) and P(Bk) probabilities with relative probabilities, that is,

$$P(A_i) = \frac{XS_i}{letszam} \quad \text{és} \quad P(B_k) = \frac{YS_k}{letszam}.$$

So if we divide the row and column sums by the headcount then we get the probability of occurrence of the grades of each subject. Thus if we assume the independence of Ai and Bk events then

$$P(A_i B_k) = \frac{XS_i \cdot YS_k}{letszam^2}.$$

If we denote the frequency of the AiBk product event with Ci,k then in the case of the assumption of the independence the

$$\frac{C_{i,k}}{letszam} = \frac{XS_i \cdot YS_k}{letszam \cdot letszam}$$

equality must become fulfilled. We can express the Ci,k frequencies based on this.

$$C_{i,k} = \frac{XS_i\, YS_k}{letszam} \qquad (i, k = 1, 2, 3, 4, 5).$$

Let's create the following 5x5 matrix based on independence. Divide the product of the ith element of the row sums (XS vector) and the kth element of the column sums (YS vector) by the headcount. The values returned should be put into the kth element of the ith row of the matrix to be created.

```
> vartak := convert( [ seq( [ seq( XS[i]·YS[k] / letszam, k = 1..5 ) ], i = 1..5 ) ], Matrix )
```

$$vartak := \begin{bmatrix} \frac{8}{15} & \frac{52}{15} & \frac{8}{3} & \frac{4}{5} & \frac{8}{15} \\ \frac{11}{15} & \frac{143}{30} & \frac{11}{3} & \frac{11}{10} & \frac{11}{15} \\ \frac{7}{15} & \frac{91}{30} & \frac{7}{3} & \frac{7}{10} & \frac{7}{15} \\ \frac{2}{15} & \frac{13}{15} & \frac{2}{3} & \frac{1}{5} & \frac{2}{15} \\ \frac{2}{15} & \frac{13}{15} & \frac{2}{3} & \frac{1}{5} & \frac{2}{15} \end{bmatrix} \qquad \textbf{(13)}$$

This matrix can be called the frequency matrix of the values expected based on independence. It will be your task to check if the row and column sums of the matrix expected coincide with the vectors of the row and column sums of the originally measured data. Let's continue with completing the expected matrix with the column and row sums. We have named the matrix completed Expected.

```
> Concatenate(2, vartak, XS) :
> Expected := Concatenate(1,%, Concatenate(2, YS, < letszam >));
```

$$Expected := \begin{bmatrix} \frac{8}{15} & \frac{52}{15} & \frac{8}{3} & \frac{4}{5} & \frac{8}{15} & 8 \\ \frac{11}{15} & \frac{143}{30} & \frac{11}{3} & \frac{11}{10} & \frac{11}{15} & 11 \\ \frac{7}{15} & \frac{91}{30} & \frac{7}{3} & \frac{7}{10} & \frac{7}{15} & 7 \\ \frac{2}{15} & \frac{13}{15} & \frac{2}{3} & \frac{1}{5} & \frac{2}{15} & 2 \\ \frac{2}{15} & \frac{13}{15} & \frac{2}{3} & \frac{1}{5} & \frac{2}{15} & 2 \\ 2 & 13 & 10 & 3 & 2 & 30 \end{bmatrix} \quad \textbf{(14)}$$

So that the values of the two matrixes should be easier to compare, we have put the two matrixes next to each other. Notice that there is no null element in the latter matrix.

The matrix of the values observed and the sums	The matrix of the values based on independence and the sums
$\begin{bmatrix} 2 & 6 & 0 & 0 & 0 & 8 \\ 0 & 6 & 5 & 0 & 0 & 11 \\ 0 & 1 & 4 & 2 & 0 & 7 \\ 0 & 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 2 \\ 2 & 13 & 10 & 3 & 2 & 30 \end{bmatrix}$	$\begin{bmatrix} 0.5333 & 3.467 & 2.667 & 0.8000 & 0.5333 & 8. \\ 0.7333 & 4.767 & 3.667 & 1.100 & 0.7333 & 11. \\ 0.4667 & 3.033 & 2.333 & 0.7000 & 0.4667 & 7. \\ 0.1333 & 0.8667 & 0.6667 & 0.2000 & 0.1333 & 2. \\ 0.1333 & 0.8667 & 0.6667 & 0.2000 & 0.1333 & 2. \\ 2. & 13. & 10. & 3. & 2. & 30. \end{bmatrix}$

Our aim is to illustrate the numerical data located in the grades and expected matrixes. Thus a kind of technical trick is needed here. Put the 25 values observed and the 25 values based on independence into lists. Put both of them continuously.

```
> megfigyeltek:=map(op,convert(jegyek, listlist));
```

```
> fuggetlenek := map(op, convert(vartak, listlist));
```

$$megfigyeltek := [2, 6, 0, 0, 0, 0, 6, 5, 0, 0, 0, 1, 4, 2, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1]$$

$$fuggetlenek := \left[\frac{8}{15}, \frac{52}{15}, \frac{8}{3}, \frac{4}{5}, \frac{8}{15}, \frac{11}{15}, \frac{143}{30}, \frac{11}{3}, \frac{11}{10}, \frac{11}{15}, \frac{7}{15}, \frac{91}{30}, \frac{7}{3}, \frac{7}{10}, \frac{7}{15}, \frac{2}{15}, \frac{13}{15}, \frac{2}{3}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15}, \frac{13}{15}, \frac{2}{3}, \frac{1}{5}, \frac{2}{15}\right] \quad \textbf{(15)}$$

For the sake of the plotting of the data, match the values in the lists with natural numbers from 1 to 25. Use the zip procedure which matches the two rows of values.

```
> adatsor1 := zip( (x, y) → [x, y], [seq(i, i = 1 ..25) ], megfigyeltek)
```

```
> adatsor2 := zip( (x, y) → [x, y], [seq(i, i = 1 ..25) ], fuggetlenek)
```

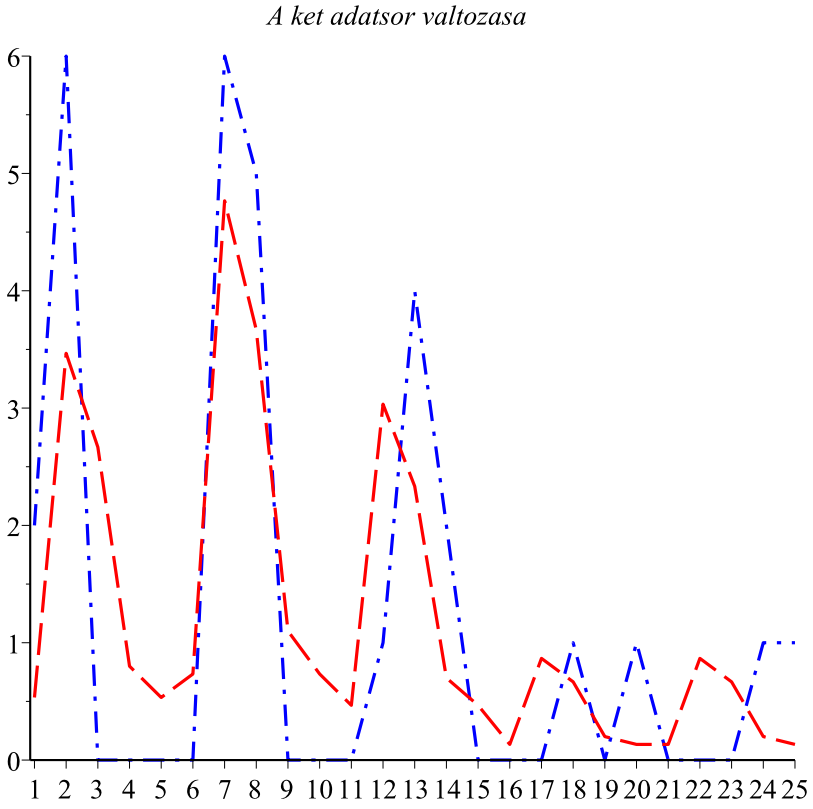
$$adatsor1 := [[1, 2], [2, 6], [3, 0], [4, 0], [5, 0], [6, 0], [7, 6], [8, 5], [9, 0], [10, 0], [11,$$

0], [12, 1], [13, 4], [14, 2], [15, 0], [16, 0], [17, 0], [18, 1], [19, 0], [20, 1], [21, 0], [22, 0], [23, 0], [24, 1], [25, 1]]

$$adatsor2 := \left[\left[1, \frac{8}{15}\right], \left[2, \frac{52}{15}\right], \left[3, \frac{8}{3}\right], \left[4, \frac{4}{5}\right], \left[5, \frac{8}{15}\right], \left[6, \frac{11}{15}\right], \left[7, \frac{143}{30}\right], \left[8, \frac{11}{3}\right], \left[9, \frac{11}{10}\right], \left[10, \frac{11}{15}\right], \left[11, \frac{7}{15}\right], \left[12, \frac{91}{30}\right], \left[13, \frac{7}{3}\right], \left[14, \frac{7}{10}\right], \left[15, \frac{7}{15}\right], \left[16, \frac{2}{15}\right], \left[17, \frac{13}{15}\right], \left[18, \frac{2}{3}\right], \left[19, \frac{1}{5}\right], \left[20, \frac{2}{15}\right], \left[21, \frac{2}{15}\right], \left[22, \frac{13}{15}\right], \left[23, \frac{2}{3}\right], \left[24, \frac{1}{5}\right], \left[25, \frac{2}{15}\right]\right] \quad \textbf{(16)}$$

After such a long struggle the plot procedure is able to plot the two rows of data.

```
> plot([adatsor1, adatsor2], linestyle = [4, 3], color = [blue,
  red], xtickmarks = 26, title = `A ket adatsor valtozasa`);
```



The two graphs can remind us of the BUX index on the stock exchange. Notice that the alternations of the two rows of data follow each other. Where one has a high local maximum the other also has a peak. And where one has a bottom the other also tends to decrease its value.

It is obvious that the two rows of value are related.

After this, we calculate the correlation coefficient of the independence row observed with the Correlation procedure of the Statistics package, which deals with the two 25-length vectors. The correlation coefficient measures the closeness of the relation between the two rows of value. Its value is always between -1 and 1. If it is near 1 and -1 then the relation is close between the two rows of value. The values around zero mean a not so strong connection. Since previously we saw that the X and Y grades depended on each other, we are curiously looking forward to the degree of the dependence.

> `Korrelációs együttható( mért,számított) ` = Correlation(megfigyeltek, fuggetlenek)

Korrelációs együttható(mért,számított) = 0.7861830430 **(17)**

The value received is 0.7861830430 which is bigger than 0,75 thus the data highly depend on each other. If you think that we will get the same value when calculating the correlation of the sums of the XS and YS rows and columns, then you are labouring under a delusion.

> `Korrelációs együttható( X összeg , Y összeg) ` = Correlation(XS, YS)

Korrelációs együttható(X összeg , Y összeg) = 0.7277852244 **(18)**

It is a bit smaller than the previous correlation coefficient but still significant.

We are going to finish our statistical examinations by showing how the 36.65634366 statistical value calculated by the ChiSquareIndependenceTest is created with the help of the values expected, observed and based on the independence.

> $\text{`Számolt statisztikai érték`} = \text{evalf}\left(\sum_{k=1}^{25} \frac{(\text{megfigyeltek}_k - \text{fuggetlenek}_k)^2}{\text{fuggetlenek}_k}\right);$

Számolt statisztikai érték = 36.65634366 **(19)**

What Have You Learnt About Maple?

- The text files can be read by using the fopen, readdata and fclose procedures together. The task of the fopen is to open the file for buffered writing and reading. Its syntax is fopen(filename, method) in which case the file name is the name of the file to be opened given by the whole directory path. The method is one of the READ, WRITE or APPEND key words. The output of the fopen is a so-called file descriptor with which we can refer to the file opened during further operations.

- The readdata procedure reads numerical data from the text file. Its call is readdata(file descriptor, format, n) in which case the file descriptor is an output of a former fopen procedure. The format is one of the integer or float key words and the n is a positive integer which determines the number of the columns to be read.

- The ChiSquareIndependenceTest procedure of the Statistics package executes an independence test between two properties of a population based on the chi square method. The number of the individuals having the properties given has to be entered into the M Matrix. In this case the call sequence of the instruction is ChiSquareIndependenceTest(M, options) in which case we can give the maximum of the type I error allowed concerning the independence hypothesis in the options, which is

called significance level. Its syntax can be level=significance level. The default value of the significance level is level=0.05 thus it need not be given.

- If we want to see all the output lists of the procedures of the Statistics package then the infolevelStatistics:=1 instruction should be used. In the infolevelStatistics:=0 default case only the most important results are returned by the procedures.

- A random probability variable can be created by the RandomVariable procedure of the Statistics package. The

X:=RandomVariable(name of the built in probability distribution) instruction makes the X become a probability value with a certain distribution. The list of the built-in probability variables can be found in the ? Statistics help site.

- The density function of a probability variable with X continuous distribution can be created by the ProbabilityDensityFunction (X,t) or shortly PDF (X,t) instruction. Maple uses the t variable given in the formula of the density function.

- The matrixes can be expanded by the Concatenate procedure of the ArrayTools package. Its call is Concatenate(dimension, A,B) in which case the name of the dimension can be 1 or 2 in the case of matrixes. It denotes that the A and B matrixes have to be concatenated according to the rows or columns. Depending on this, the B matrix is concatenated subsequent to or below the A. Make sure that the dimensions coincide with each other.

- The add(X[k],k=i..j) instruction adds the elements of the X Vector from the ith to the jth element.

- The zip((x,y) > [x,y], X, Y) instruction makes certain elements of the lists or the X and Y same-sized vectors be assigned to each other and it creates the list of [[X1,Y1], [X2,Y2], ..., [Xn, Yn]]] lists.

- The correlation coefficient of two data rows can be calculated by the Correlation procedure of the Statistics package. The correlation coefficient is a number between -1 and 1 which measures the degree of the relation between the two data rows. If the value is near -1 or 1 then the nonlinear relation is close while the values near 0 shows an independent relation. The call of the procedure is Correlation (X,Y) where X and Y are vectors of the same size.

□

Exercises

1. The X and Y subjects can be considered discrete probability variables. Their possible value can be the grades 1,2,3,4 or 5. We can get the fXY joint probability distribution of the X and Y variable if we divide the grades matrix by the headcount. Create the probability distribution of the fXY.

2. We can get the fX marginal distribution of the X subject if we divide the vector of the sums of the XS rows by the headcount. Determine the fX and its Ex expected value. Use the Mean procedure of the Statistics package.

3. Similarly to task 2, determine the fY marginal distribution of the Y subject and determine its Ey expected value.

4. The conditional distribution of the Y variable concerning the X variable can be defined as follows.

Create the

$$f(y \mid x_k) = P(Y=y \mid X=x_k) = \frac{P(\{Y=y\} \cdot \{X=x_k\})}{P(\{X=x_k\})}$$

function (where y is an independent variable), which is called the conditional distribution of the Y probability variable concerning the X=xk fixed value. Determine the vector of the conditional probabilities

$$[\ P(Y=1 \mid X=1),\quad P(Y=2 \mid X=1),\ P(Y=3 \mid X=1),\quad P(Y=4 \mid X=1),\ P(Y=5 \mid X=1)\ ]$$

For this, divide the first row of the fXY matrix by the fX1 sum of row.

Feel free to use the SubMatrix procedure of the LinearAlgebra package.

5. Determine the conditional distribution of the Y probability variable concerning the X=2, X=3, X=4 and X=5 fixed values by using the method of task 4.

6. If the [képlet] joint and marginal distributions of the X and Y discrete probability variables are known, then determine the conditional value expected of the Y probability variable by the X=xk condition with the

$$\mathrm{E}(\mathrm{Y} \mid \mathrm{X}=x_k) = \sum_i y_i f(y_i \mid x_k)$$

weighted average. In this case the

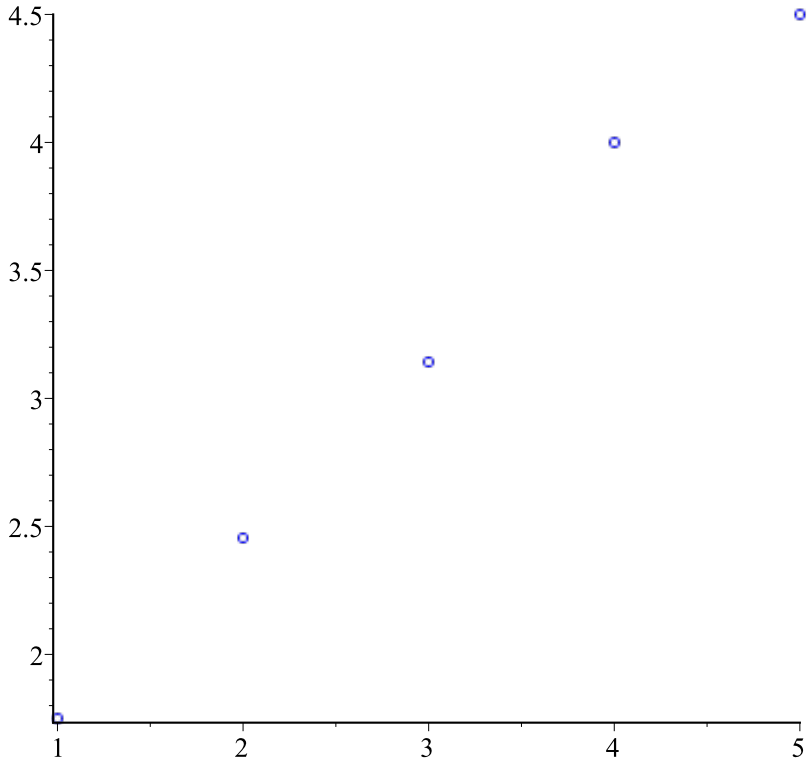
$$f(y \mid x_k) = P(Y=y \mid X=x_k) = \frac{f_{X,Y}(x_k, y)}{f_X(x_k)}$$

Y probability variable is the conditional probability distribution by the X=xk condition. (see tasks 4 and 5)

Calculate the E(Y | X=1), E(Y | X=2), E(Y | X=3), E(Y | X=4) és E(Y | X=5) values.

7. 3. Plot the E(Y | X=1), E(Y | X=2), E(Y | X=3), E(Y | X=4) és E(Y | X=5) values as series of points. Keep trying until you get a graph like this.

Y felteteles varhato ertekei X-re vonatkozoan



>