Data Analysis, Lecture 3 : Mining categorial bivariate data and Introduction to multivariate data analysis

Jérémie Sublime

▶ **To cite this version:**

Jérémie Sublime. Data Analysis, Lecture 3 : Mining categorial bivariate data and Introduction to multivariate data analysis. Engineering school. France. 2022. hal-03845284

HAL Id: hal-03845284
https://hal.science/hal-03845284

Submitted on 9 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Data Analysis - Lecture 3

Mining categorial bivariate data and Introduction to multivariate data analysis

Dr. Jérémie Sublime

LISITE Laboratory - DaSSIP Team - ISEP
LIPN - CNRS UMR 7030

jeremie.sublime@isep.fr

Outline

Outline

Problem presentation

In the last course, we have been interested in finding correlations between quantitative ideally continuous variables.

- What about categorical variables ?
- How to find whether there is a correlation between two categorical variables: hair color and eye color, neighborhood and type of job, etc.

The measures that we have seen so far (correlation, determination, covariance, etc.) cannot be applied to these types of data.

Categorical variables : Chi-squared test

Chi-squared test of independence: χ^2

- The Chi-squared is based on the **contingency table** (cross table) of the possibles values of two variables.
- It is computed based on the difference between the **expected frequencies** and the **observed frequencies** of one or more categories of the contingency table.
- A zero Chi-squared means that the two variables are completely independent.
- A non-zero Chi-squared is more difficult to interpret:
  - Requires to evaluated the likelihood of the resulting Chi-squared based on its known distribution. (Requires tables or a calculator)
  - Can be evaluated using the Chuprov contingency coefficient or Cramer's V.

Categorical variables : Chi-squared test

- Let us consider two variables i and j so that $i \in [1..r]$ and $j \in [1..c]$.
- Let $o_{i,j}$ be the number of observed data so that the first feature's value is i and the second feature's value is j.
- Let n_i be the total number of elements having i as a value for there first feature.
- Let N be the total number of observations.

Then, the expected contingency $e_{i,j}$ computes as follows:

$$e_{i,j} = \frac{n_i \cdot n_j}{N}$$

Computing the χ^2

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Cross Table: Example (1/2)

Total per Category	Cat A1 (30)	Cat A2 (30)	Cat A3 (40)
Cat B1 (40)			
Cat B2 (30)			
Cat B3 (30)			

What are the expected values $e_{i,j}$ in this table, if we suppose that there is no links between the categories in A and B ?

Cross Table: Example (2/2)

Total per Category	Cat A1 (30)	Cat A2 (30)	Cat A3 (40)
Cat B1 (40)	12	12	16
Cat B2 (30)	9	9	12
Cat B3 (30)	9	9	12

If there is no link, the repartition in the cross table should be almost exactly proportional with the size of the categories.

- If the observed values $o_{i,j}$ are far from this supposed proportional repartition, there is probably a link between some of the categories.

Cross Table: Example (3/3)

	Dark hair	Light hair
Brown eyes	32	12
Blue eyes	14	22
Green eyes	6	9

Notations examples

$i \in$ (Dark hair, light hair) $\quad j \in$ (Brown eyes, Blue eyes, Green eyes)

$o_{dark,brown} = 32 \quad e_{dark,brown} = \frac{44 \times 52}{95} = 24.08$

Chi-square interpretation: Hypothesis test

- Unless its value is zero (which is rare), the Chi-squared cannot directly be interpreted.
- In statistics, **p-values** are criteria often linked to what is called a **hypothesis test**.

Hypothesis Test

- State your **null hypothesis** H_0 and alternative hypotheses.
- Choose a value α (usually 0.1, 0.05 or 0.01)
- Compute the p-value for your proposed model:
  - if p-value $< \alpha$: reject H_0.
  - if p-value $\geq \alpha$: you cannot reject H_0.

The Chi-squared is a **test of independence**. Therefore, the hypothesis that you are trying to reject is H_0: "The two variables are independent".

Chi-square interpretation: Hypothesis test

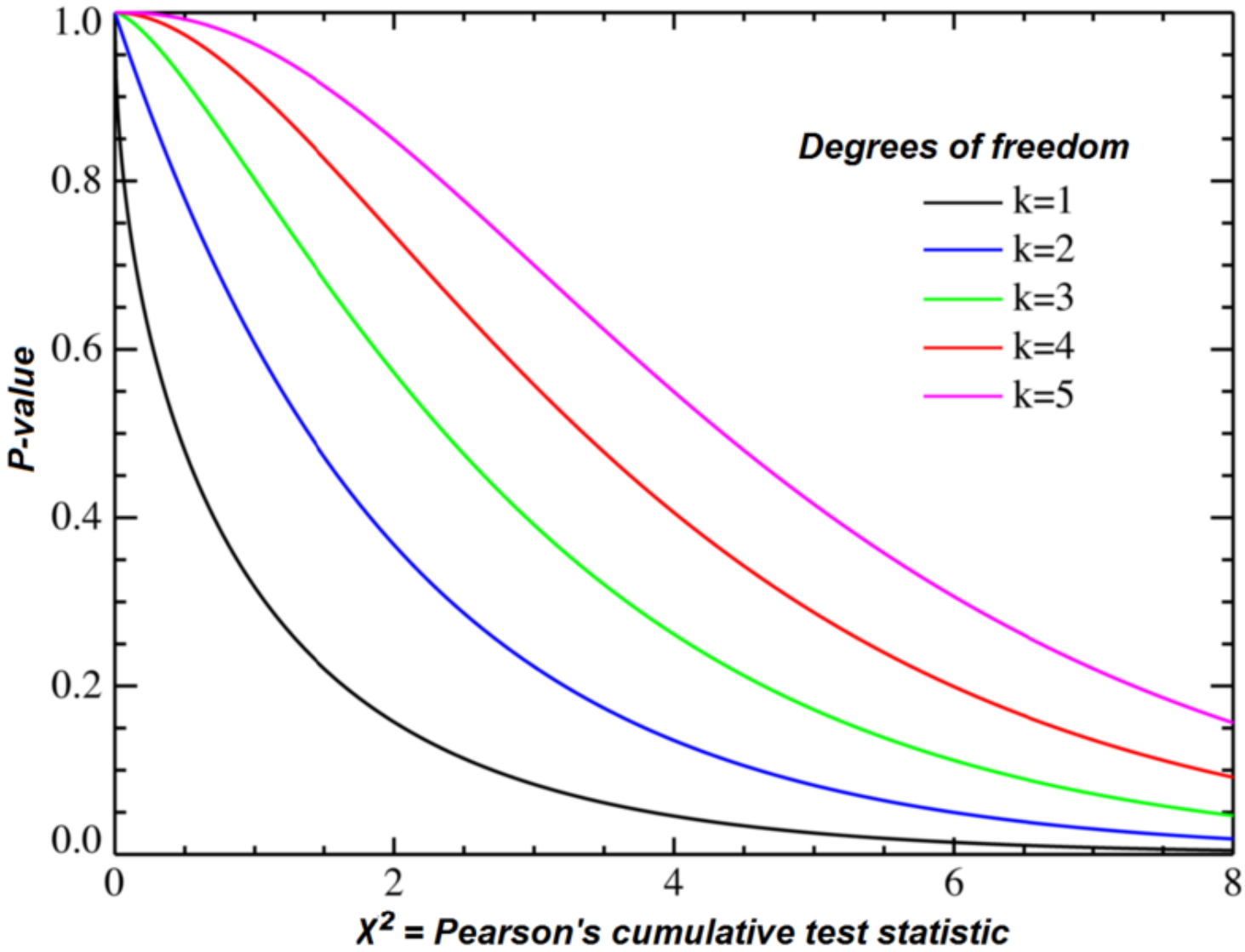
Computing the p-value

The p-value is computed using the known distribution of the Chi-squared function (using tables, a calculator, or a graph) considering degrees of freedom of the problem.

$$\text{Degrees of freedom} = (r-1)(c-1)$$

- A p-value close to zero rejects the hypothesis that the two variables are independent.
- We can say that there is "1−p-value % chance" that the link between the two variables is not random luck.
- Computing further indexes will be necessary to assess the degree of correlation.

Chi-square interpretation: p-value graph



Contribution to the chi2

- One way to start explaining a chi2 result is to look at **which associations have a strong influence** on the chi2.
- This can be done by analyzing the normalized residuals:

$$r_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

- And their contribution to the chi2, which is very useful for large cross-tables:

$$c_{ij} = 100 \times \frac{r_{ij}^2}{\chi^2}$$

Chi-squared interpretation: Chuprov coefficient

The **Chuprov contingency coefficient** (sometimes spelled Tschuprow) can also be used to interpret the result of a Chi-Squared:

- It is denoted $T \in [0, 1]$.
- It measures the amount of dependency between two categorical variables.

Chuprov coefficient

$$T = \sqrt{\frac{\chi^2}{N\sqrt{(c-1)(r-1)}}}$$

- If T is close to 0, we can't predict one from the other.
- If T is close to 1, the link is strong and we can predict one from the other.

Chi-squared interpretation: Cramer's V

Cramer's V is another is another index measuring the amount of dependency between two variables.

Cramer's V

$$V = \sqrt{\frac{\chi^2}{N \cdot min(c-1, r-1)}}$$

- If V is close to 0, we can't predict one from the other.
- If V is close to 1, the link is strong and we can predict one from the other.

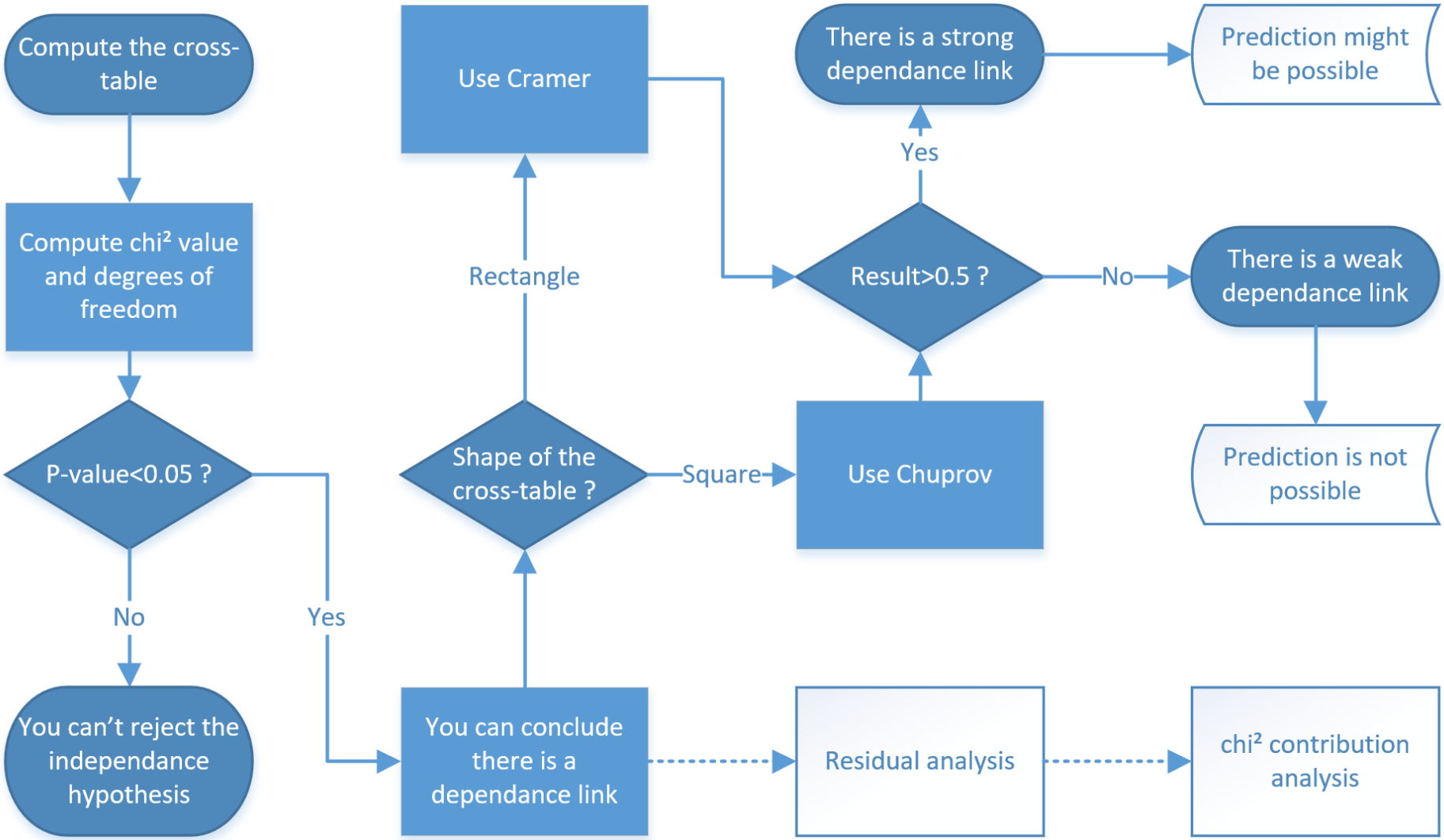
Chi-squared test: Remarks

- While the p-value evaluates whether a result based on the chi-squared has good chances of being significant, **its value is not proportional to the amount of correlation**.
- Chuprov coefficient is best reliable with square contingency tables, while Cramer's V works best with rectangular ones.
- Both Chuprov coefficient and Cramer's V can be very biased: unevenly distributed observations, one variable with much more possible values than the other, etc.
- Both Chuprov coefficients and Cramer's V can't be used if the result of the p-value shows that the chi-squared result is not significant.

Chi-squared test: Remarks

- The Chi-squared test is not recommended with very small data sets (most expected values bellow 10), and can be replaced by the similar **Fisher test** in such cases.
- A significant dependency between two variables using the chi squared test does not imply causality.
- The chi squared result alone with the p-value are not informative enough:
  - Chuprov or Cramer's V are needed to assess strength of the link and determine the reliability of predicting one variable based on the other.
  - Computing the residuals and contribution to the chi2 explains how the different pairing of categories affect the test and might thus be remarkable or related.

Chi-squared test: Summary



Example: Dataset

	Dark hair	Light hair
Brown eyes	32	12
Blue eyes	14	22
Green eyes	6	9

Given this contingency table, is there a correlation between hair color and eye color ?

Example: Summing the columns

	Dark hair	Light hair	n_j
Brown eyes	32	12	44
Blue eyes	14	22	36
Green eyes	6	9	15
n_i	52	43	$N = 95$

Example: expected values

	Dark hair	Light hair	n_j
Brown eyes	$o_{1,1} = 32$ $e_{1,1} = \frac{44 \times 52}{95}$	$o_{1,2} = 12$ $e_{1,2} = ?$	44
Blue eyes	$o_{2,1} = 14$ $e_{2,1} = ?$	$o_{2,2} = 22$ $e_{2,2} = ?$	36
Green eyes	$o_{3,1} = 6$ $e_{3,1} = ?$	$o_{3,2} = 9$ $e_{3,2} = ?$	15
n_i	52	43	$N = 95$

Example: chi squared (1/2)

	Dark hair	Light hair	n_j
Brown eyes	$o_{1,1} = 32$ $e_{1,1} = 24.1$	$o_{1,2} = 12$ $e_{1,2} = 19.9$	44
Blue eyes	$o_{2,1} = 14$ $e_{2,1} = 19.7$	$o_{2,2} = 22$ $e_{2,2} = 16.3$	36
Green eyes	$o_{3,1} = 6$ $e_{3,1} = 8.2$	$o_{3,2} = 9$ $e_{3,2} = 6.8$	15
n_i	52	43	$N = 95$

$$\chi^2 = \frac{(32 - 24.1)^2}{24.1} + \cdots + \frac{(9 - 6.8)^2}{6.8}$$

Example: chi squared (2/2)

	Dark hair	Light hair	n_j
Brown eyes	$o_{1,1} = 32$ $e_{1,1} = 24.1$	$o_{1,2} = 12$ $e_{1,2} = 19.9$	44
Blue eyes	$o_{2,1} = 14$ $e_{2,1} = 19.7$	$o_{2,2} = 22$ $e_{2,2} = 16.3$	36
Green eyes	$o_{3,1} = 6$ $e_{3,1} = 8.2$	$o_{3,2} = 9$ $e_{3,2} = 6.8$	15
n_i	52	43	$N = 95$

$$\chi^2 = 10.67$$

$k = (3-1)(2-1) = 2$ degrees of freedom

Example: chi squared table

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

With $\chi^2 = 10.67$ and $d_f = 2$, we have a p-value bellow 0.01: There is a significant correlation between hair color and eye color.

Example: Interpretation

$\chi^2 = 10.67$ p-value< 0.01

$$T = \sqrt{\frac{\chi^2}{N\sqrt{(c-1)(r-1)}}} = \sqrt{\frac{10.67}{95\sqrt{2}}} = 0.28$$

$$V = \sqrt{\frac{\chi^2}{N \cdot min(c-1, r-1)}} = \sqrt{\frac{10.67}{95 \cdot min(1,2)}} = 0.34$$

Interpretation

- From the p-value, we deduce that the two variables show a significant dependency.
- On a rectangle contingency matrix Cramer's V indicates around 34% of correlation between the two variables (so does the Chuprov coefficient with 28%).

Outline

What about mixed variables ?

- We have seen that for numerical variables we can use the covariance, the correlation (Spearman or Pearson), and the determination coefficient.
- For categorial variables, we have the Chi square (or Fischer) test coupled with a Cramer's V or Chuprov coefficient.
- What about mixed data ?

What about mixed variables ?

- We have seen that for numerical variables we can use the covariance, the correlation (Spearman or Pearson), and the determination coefficient.
- For categorial variables, we have the Chi square (or Fischer) test coupled with a Cramer's V or Chuprov coefficient.
- What about mixed data ?

What are mixed data ?

- Mixed data are datasets that contains both numerical and categorial data (and sometimes other things).
- Numerical bivariate analysis won't work on them (because of the categorial variables)
- Neither should the chi square test (because of the continuous variables)

Example of mixed variables

- Weather forecast : Temperature, humidity, and weather (rainy, sunny, overcast, etc.)

Example of mixed variables

- Weather forecast : Temperature, humidity, and weather (rainy, sunny, overcast, etc.)
- Medical data combining physiological variables as well as personnal and clinical information : blood type, sex, oxygen level, risk group, age, glycemia level, etc.

Example of mixed variables

- Weather forecast : Temperature, humidity, and weather (rainy, sunny, overcast, etc.)
- Medical data combining physiological variables as well as personnal and clinical information : blood type, sex, oxygen level, risk group, age, glycemia level, etc.

Mixed data are actually the norm ...

Pretty much anything can be mixed data as most data nowadays are described by all sorts of variables, some of which are even neither numerical nor categorical (but we will talk about them another time).

From numerical to categorial data

Dealing with mixed data ?

- Numerical bivariate analysis won't work on them (because of the categorial variables)
- Neither should the chi square test (because of the continuous variables)

From numerical to categorial data

Dealing with mixed data ?

- Numerical bivariate analysis won't work on them (because of the categorial variables)
- Neither should the chi square test (because of the continuous variables)

So, how do we deal with them ?!

From numerical to categorial data

Dealing with mixed data ?

- Numerical bivariate analysis won't work on them (because of the categorial variables)
- Neither should the chi square test (because of the continuous variables)

So, how do we deal with them ?!

Discretizing the numerical variables

- Numerical variables can be categorized by building artificial groups between ranges of interest
- The main difficulty is to find the right size/number of groups/categories

Example: Dataset

	A-H	I	J	L-Q	R1a	Other
175	41	13	7	26	0	40
176	64	14	5	17	0	60
177	64	15	0	21	1	61
178	64	13	11	23	8	58
179	69	7	10	22	7	70
180	53	15	5	25	11	54
181	61	0	6	17	3	60
182	62	20	3	18	8	63
183	52	17	2	13	11	50
184	38	21	2	11	9	39
185	53	20	3	10	18	52
186	47	11	0	6	12	46
187	38	17	2	11	0	39
188	33	0	0	0	3	3
189	0	7	0	10	9	20
190	0	7	1	0	7	10
193	28	0	1	5	0	0
195	20	8	0	4	5	13
196	7	9	0	4	0	11
197	6	0	1	0	1	0
200	0	3	0	0	0	4
202	3	0	0	2	0	1
203	0	0	0	0	1	1
207	0	1	0	0	1	0
212	0	1	0	0	0	0

Given this contingency table, is there a correlation between the Y haplogroup and the average height ?

Example: Groups of heights

	A-H	I	J	L-Q	R1a	Other
175-179	302	62	33	109	16	289
180-184	266	73	18	84	42	266
185-189	171	55	5	37	42	160
190-194	28	7	2	5	7	10
195-199	33	17	1	8	6	24
200+	3	5	0	2	2	6

Example: Summing

	A-H	I	J	L-Q	R1a	Other	n_j
175-179	302	62	33	109	16	289	811
180-184	266	73	18	84	42	266	749
185-189	171	55	5	37	42	160	470
190-194	28	7	2	5	7	10	59
195-199	33	17	1	8	6	24	89
200+	3	5	0	2	2	6	18
n_i	803	219	59	245	115	755	2196

Example: Chi 2

$e_{i,j}$	A-H	I	J	L-Q	R1a	Other
175-179	296.55	80.88	21.79	90.48	42.47	278.83
180-184	273.88	74.70	20.12	83.56	39.22	257.51
185-189	171.86	46.87	12.63	52.44	24.61	161.59
190-194	21.57	5.88	1.59	6.58	3.09	20.28
195-199	32.54	8.88	2.39	9.93	4.66	30.60
200+	6.58	1.80	0.48	2.01	0.94	6.19

Example: Chi 2

$e_{i,j}$	A-H	I	J	L-Q	R1a	Other
175-179	296.55	80.88	21.79	90.48	42.47	278.83
180-184	273.88	74.70	20.12	83.56	39.22	257.51
185-189	171.86	46.87	12.63	52.44	24.61	161.59
190-194	21.57	5.88	1.59	6.58	3.09	20.28
195-199	32.54	8.88	2.39	9.93	4.66	30.60
200+	6.58	1.80	0.48	2.01	0.94	6.19

$$\chi^2 = 87.33, \qquad d_f = (6-1) \times (6-1) = 25$$

Example: Chi 2 interpretation

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928

With $\chi^2 = 87.83$ and $d_f = 25$, we have a p-value bellow 0.01: There is a significant link between the Y haplogroup and the height of an individual !

Example: Chi 2 interpretation

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928

With $\chi^2 = 87.83$ and $d_f = 25$, we have a p-value bellow 0.01: There is a significant link between the Y haplogroup and the height of an individual !

We also have $T = V = 0.089$, so the link appears to be weak.

Example: residual analysis

To better explain this result, we can take a look at the normalised residuals:

r_{ij}	A-H	I	J	L-Q	R1a	Other
175-179	0.32	**-2.10**	**2.40**	1.95	**-4.06**	0.61
180-184	-0.48	-0.20	-0.47	0.05	0.44	0.53
185-189	-0.07	1.19	**-2.15**	**-2.13**	**3.50**	-0.13
190-194	1.38	0.46	0.33	-0.62	**2.22**	**-2.28**
195-199	0.08	**2.73**	-0.90	-0.61	0.62	-1.19
200+	-1.40	**2.39**	-0.70	-0.01	1.09	-0.08

We see that a few of the observed values that are too high or too low compared with the expected values.

Example: Contribution to the chi 2

Let us take now a look at the contribution to the chi2:

c_{ij}	A-H	I	J	L-Q	R1a	Other
175-179	0.11	**5.05**	**6.61**	4.34	**18.89**	0.42
180-184	0.26	0.04	0.26	0.00	0.23	0.32
185-189	0.00	1.61	**5.28**	**5.20**	**14.07**	0.02
190-194	2.19	0.24	0.12	0.44	**5.67**	**5.97**
195-199	0.01	**8.52**	0.93	0.43	0.44	1.63
200+	2.23	**6.55**	0.55	0.00	1.36	0.01

Example: Contribution to the chi 2

Let us take now a look at the contribution to the chi2:

c_{ij}	A-H	I	J	L-Q	R1a	Other
175-179	0.11	**5.05**	**6.61**	4.34	**18.89**	0.42
180-184	0.26	0.04	0.26	0.00	0.23	0.32
185-189	0.00	1.61	**5.28**	**5.20**	**14.07**	0.02
190-194	2.19	0.24	0.12	0.44	**5.67**	**5.97**
195-199	0.01	**8.52**	0.93	0.43	0.44	1.63
200+	2.23	**6.55**	0.55	0.00	1.36	0.01

In addition to the previous informations, we clearly see that haplogroups I and R1a have the most influence on the chi2, which means that these 2 groups have an influence on the size:

- It explains why the chi2 is significant
- But it also explain why the link remains weak since only 2 haplogroups are remarkable, which far from enough to predict the size from the haplogroup, or the opposite.