Data Mining - Lecture 2
Mining bivariate data and Introduction to multivariate data

Dr. Issam Falih

issam.falih@uca.fr

Outline

Outline

Introduction

Bivariate data can be stored in a table with two columns:

	X	Y
Obs. 1	2	1
Obs. 2	4	4
Obs. 3	3	1
Obs. 4	7	5
Obs. 5	5	6
Obs. 6	2	1
Obs. 7	7	5
Obs. 8	9	6
Obs. 9	3	2
Obs. 10	7	4

Some examples

- Height (X) and weight (Y) are measured for each individual in a sample.
- Stock market valuation (X) and quarterly corporate earnings (Y) are recorded for each company in a sample.
- A cell culture is treated with varying concentrations of a drug, and the growth rate (X) and drug concentration (Y) are recorded for each trial.
- Temperature (X) and precipitation (Y) are measured on a given day at a set of weather stations.

Remark

- To be clear about the difference between **bivariate** data and **two sample** data: n two sample data, the X and Y values are not paired, and there aren't necessarily the same number of X and Y values.

Remark

- To be clear about the difference between **bivariate** data and **two sample** data: n two sample data, the X and Y values are not paired, and there aren't necessarily the same number of X and Y values.

Two-sample data:

Sample 1: 3,2,5,1,3,4,2,3
Sample 2: 4,4,3,6,5

Objectives of bivariate data analysis

Analyzing bivariate data

When the data are described by two random variables (e.g. X and Y), we are interested in knowing the possible statistical link between these two variables.

- Does the value of X depends on the value of Y (and the other way around) ?
- What is the strength of the link between these two variables ?
  - How precisely can I deduce one variable from the other ?
  - What is their correlation ?
- What is the numerical relation between X and Y ?
  - What is the function linking them ? (e.g. regression)

Examples

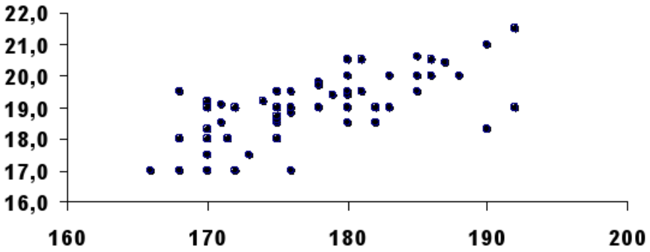


Figure: Egg size depending on the size of the bird

- There is a clear link between the two variables
- It is probably possible to make a regression to estimate the approximate size of the egg depending on the size of the bird, or the opposite (Cf Lecture 3).

Examples

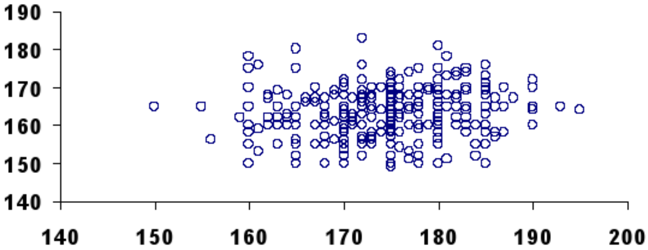


Figure: Respective weights of men and women in a family (in pounds)

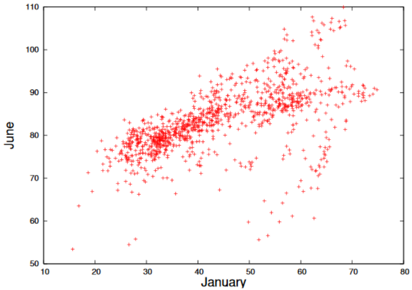
- There is no obvious link.
- No regression seem to be possible here.

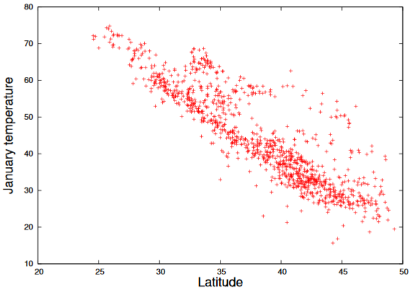
Interest

- The study of pairwise relationships between data is often very useful to the understanding of a phenomenon:
  - Understanding the relationship between different aspects of a phenomenon.
  - Discover redundancies in the description of a phenomenon.
- In this section we will study linear relationships between variables.
  - They are simple to analyze.
  - In most cases variables can be transformed to fall back in the linear case.

Example of visual bivariate analysis

The most important graphical summary of bivariate data is the scatterplot. It is simply a plot of the points (X_i, Y_i) in the plane. The following figures show scatterplots of June maximum temperatures against January maximum temperatures, and of January maximum temperatures against latitude. All temperatures are in degree Fahrenheit.





Example of a visual bivariate analysis

A key feature to look for in a scatterplot is the association, or trend between X and Y.

- Higher January temperatures tend to be paired with higher June temperatures, so these two values have a **positive association**.
- Higher latitudes tend to be paired with lower January temperature decreases, so these values have a **negative association**.

Example of a visual bivariate analysis

A key feature to look for in a scatterplot is the association, or trend between X and Y.

- Higher January temperatures tend to be paired with higher June temperatures, so these two values have a **positive association**.
- Higher latitudes tend to be paired with lower January temperature decreases, so these values have a **negative association**.

Remark: If higher X values are paired with low or with high Y values equally often, there is no association.

Causality in bivariate analysis

It is important to remember that it is ill advised to draw causal implications from statements about associations, unless your data come from a randomized experiment.
Example:

- Just because January and June temperatures increase together does not mean that January temperature cause June temperature to increase (and vice versa).

The only certain way to sort out causality is to move beyond statistical analysis and talk about **mechanisms**: This often requires priori knowledge of the field related to the data, as well as reasoning.

Causality in bivariate analysis

In general, if X and Y have an association, then:

Causality in bivariate analysis

In general, if X and Y have an association, then:

- X could cause Y to change

Causality in bivariate analysis

In general, if X and Y have an association, then:

- X could cause Y to change
- Y could cause X to change

Causality in bivariate analysis

In general, if X and Y have an association, then:

- X could cause Y to change
- Y could cause X to change
- An external variable Z (perhaps unknown) could cause both X and Y to change.

Causality in bivariate analysis

In general, if X and Y have an association, then:

- X could cause Y to change
- Y could cause X to change
- An external variable Z (perhaps unknown) could cause both X and Y to change.

Unless your data come from a randomized experiment, statistical analysis alone is not capable of answering questions about causality.

Causality in bivariate analysis

Back to our example, for the association between January and July temperatures, we can try to propose some simple mechanisms:

Causality in bivariate analysis

Back to our example, for the association between January and July temperatures, we can try to propose some simple mechanisms:

1. Warmer or cooler air masses in January persist in the atmosphere until July, causing similar effects on the July temperature.

Causality in bivariate analysis

Back to our example, for the association between January and July temperatures, we can try to propose some simple mechanisms:

1. Warmer or cooler air masses in January persist in the atmosphere until July, causing similar effects on the July temperature.
2. None, it is impossible for one event to cause another event that preceded it in time.

Causality in bivariate analysis

Back to our example, for the association between January and July temperatures, we can try to propose some simple mechanisms:

1. Warmer or cooler air masses in January persist in the atmosphere until July, causing similar effects on the July temperature.
2. None, it is impossible for one event to cause another event that preceded it in time.
3. If Z is latitude, then latitude influences temperature for both months because it determines the amount of atmosphere that solar energy must traverse to reach a particular point on the Earth's surface.

Causality in bivariate analysis

Back to our example, for the association between January and July temperatures, we can try to propose some simple mechanisms:

1. Warmer or cooler air masses in January persist in the atmosphere until July, causing similar effects on the July temperature.
2. None, it is impossible for one event to cause another event that preceded it in time.
3. If Z is latitude, then latitude influences temperature for both months because it determines the amount of atmosphere that solar energy must traverse to reach a particular point on the Earth's surface.

Case (iii) is the correct one. Yet, just looking at the scatterplot does not give the strength of the relation between latitude and temperature.

Covariance

The **covariance** *between two random variables (or series of* **equal size***) assesses the joint difference between their respective means values.*

- The covariance is denoted $Cov(\mathrm{X}, \mathrm{Y})$ or sometimes $\sigma_{\mathrm{X},\mathrm{Y}}$.

Regular Covariance

$$Cov(\mathrm{X}, \mathrm{Y}) = \frac{1}{n} \sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y)$$

Covariance of a sample

$$Cov(\mathrm{X}, \mathrm{Y}) = \frac{1}{n-1} \sum_{i=1}^{N} (x_i - m_X)(y_i - m_Y)$$

Covariance

Properties

$$Cov(\mathrm{X}, \mathrm{Y}) = Cov(\mathrm{Y}, \mathrm{X})$$
$$Cov(\mathrm{X}, \mathrm{X}) = VAR(X) = \sigma_X^2$$
$$Cov(a \times \mathrm{X}, b \times \mathrm{Y}) = a \times b \times Cov(\mathrm{X}, \mathrm{Y})$$
$$Cov(a + \mathrm{X}, b + \mathrm{Y}) = Cov(\mathrm{X}, \mathrm{Y})$$

Covariance

- When two variables are fully independent, their covariance is *null*. **However, the opposite is not always true !**
- If both greater and lower values from both variables tend to be similar, then the two variables are similar and the covariance is positive.
- When the two variables show opposite behavior, the covariance is negative.
- **The covariance is very sensitive to the unit and scale of the observed variables !**
- The covariance is often difficult to interpret.

Covariance

- When two variables are fully independent, their covariance is *null*. **However, the opposite is not always true !**
- If both greater and lower values from both variables tend to be similar, then the two variables are similar and the covariance is positive.
- When the two variables show opposite behavior, the covariance is negative.
- **The covariance is very sensitive <u>to the unit and scale</u> of the observed variables !**
- The covariance is often difficult to interpret.

We need to find more accurate and easier to use criteria

Correlation coefficient

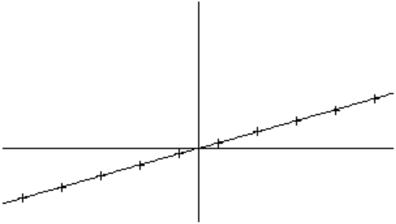
The **Pearson Product-moment** **correlation coefficient** *is a measure of the linear correlation between two random variables.*

- The correlation coefficient takes values between -1 and $+1$ inclusive.
- $+1$ denotes a complete correlation.
- 0 denotes no correlation between the two variables.
- -1 means a total negative correlation.

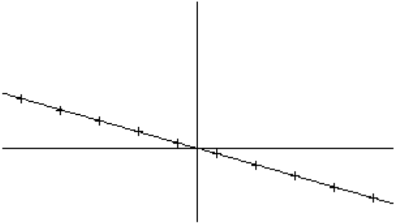
Correlation coefficient between two random variables X and Y

$$r = Cor(\mathrm{X}, \mathrm{Y}) = \frac{Cov(\mathrm{X}, \mathrm{Y})}{s_x s_y}$$

Correlation coefficient: examples

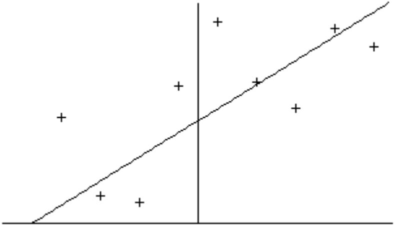


(a) $r = 1$

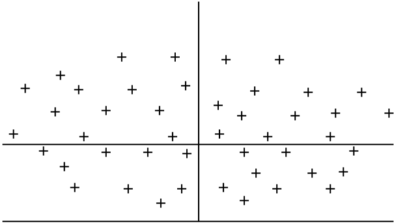


(b) $r = -1$

Correlation coefficient: examples



(c) $r = 0.77$



(d) $r \approx 0$

Correlation coefficient: weaknesses

Warning

- Correlation does not imply causality ! **Example**: The number of sunburn as a function of the number of person.
- The interpretation of the correlation coefficient is sometimes counter-intuitive. **Example**: is $r = -0.6$ is strong correlation ?

Luckily there is a better coefficient: The **coefficient of determination**.

Coefficient of determination

The **coefficient of determination** *is the proportion of the variance of* Y*, which disappears if* X *is fixed (or the other way around).*

- It is the square value of the correlation coefficient.
- r^2 is a proportion between 0 and 1, and is very easy to interpret.

Coefficient of determination between two random variables X and Y

$$r^2 = \left(\frac{Cov(\mathrm{X}, \mathrm{Y})}{s_x s_y} \right)^2$$

Coefficient of determination: Interpretation

The coefficient of determination is independent of the units and scales chosen for X and Y.

- When r^2 is close to zero, the link between the two variables is very weak: we know precisely that both variables provide almost no information on the other.
- When r^2 is close to one, the relationship between the two variables is very strong: We know that X greatly reduces the variability of Y (and the other way around). Therefore we can predict one from the other with a very high probability.

Coefficient of determination: Interpretation

Example

Let X and Y be two random variables with a correlation coefficient $r = -0.6$.

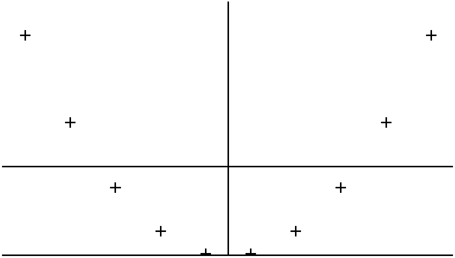
- The correlation is negative.
- Is it strong ?
- $r^2 = 0.36$: it means that 36% of Y information is contained in X.
- It is not bad, but it also means that 64% of Y information cannot be deduced from X.
- The deduction of Y from X is unreliable.

Coefficient of determination: Interpretation

Remarks

1. The correlation coefficient gives us information on the existence of a **linear relationship** between the two considered variables. A correlation coefficient of zero does not mean the absence of any relationship between the two variables. There may be a nonlinear relationship between them.
2. Do not confuse correlation and causality: A strong correlation between two variables can reveal a causal relationship between them, but not necessarily.

Coefficient of determination: Limits



In this example, we have $r = 0$. Yet there is an obvious link between the two variables, a non-linear one.

Coefficient of determination: Limits

A single outlying observation can have a substantial effect on the correlation coefficient. The following scatterplot shows a bivariate data set in which a single point produces a correlation of around -0.75. The correlation would be around 0.01 if the point were removed.

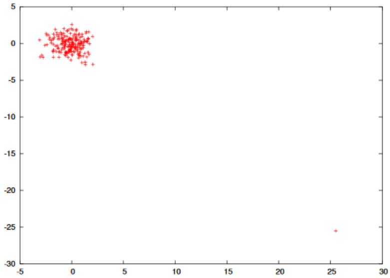


Figure: A correlation of ≈ -0.75 produced by a single outlier

Confidence intervals for the correlation

IC_{95} of r and r^2

- If X (resp. Y) is fixed, and the distribution of Y (resp. X) follows a normal distribution (often difficult to assess):
  - Then $Z = \frac{\ln(1+r)-\ln(1-r)}{2}$ follows a normal distribution
  - With $s_Z = \sqrt{1/(n-3)}$,
  - Then $Z_{inf} = Z - 1.96 s_Z$, $Z_{sup} = Z + 1.96 s_Z$,
  - And $IC_{95}(r) = \left[\frac{e^{2Z_{inf}}-1}{e^{2Z_{inf}}+1}, \frac{e^{2Z_{sup}}-1}{e^{2Z_{sup}}+1}\right]$
- **Otherwise: The bootstrap method still works !**
- For the coefficient of determination: $IC_{95}(r^2) = (IC_{95}(r))^2$

Confidence intervals for the correlation

Interpretation

The interpretation of the confidence interval is the same that what we saw for the mean and standard deviation:

- If the range is too large, then we can't say whether the two variables are correlated or not.
- If the confidence interval of r or r^2 is around 0, then we can't say that there is a link between the two variables.

Once again, the existence of a correlation between two variables doesn't mean that there is a causality link, nor that this correlation is interesting at all.

Outline

Categorical variables : Chi-squared test

So far 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 correlation measures that we have seen so far 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 (40)	Cat A3 (40)
Cat B1 (40)			
Cat B2 (30)			
Cat B3 (40)			

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) $j \in$ (Brown eyes, Blue eyes, Green eyes)

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

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

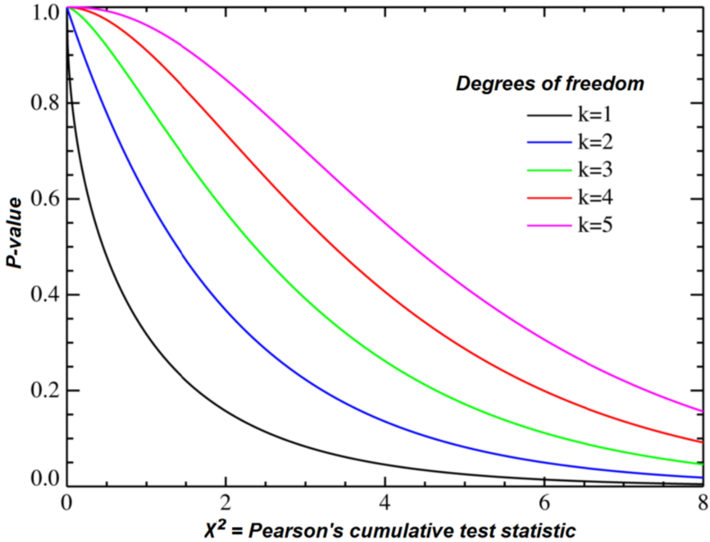
Computing the p-value

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 apparent dependence 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



Chi-squared interpretation: Chuprov coefficient

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

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

Chuprov coefficient

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

- If ρ is close to 0, the two variables are independent.
- If ρ is close to 1, they are dependent.

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, the two variables are independent.
- If V is close to 1, they are dependent.

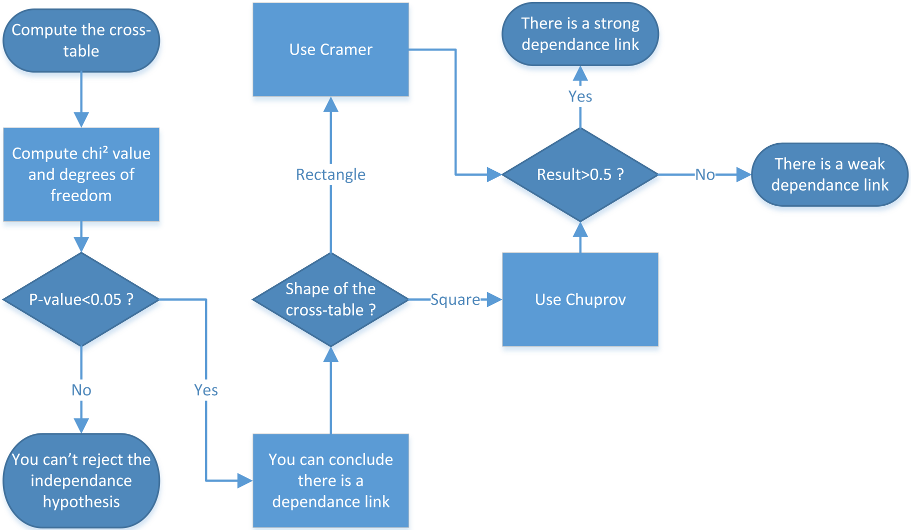
Chi-squared test: Remarks

- 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 can also be used to assess the correlation between categorical and quantitative data: the quantitative data need to be regrouped (examples with grades: grades from 0 to 5, 6 to 10, 11 to 15 and 16 to 20).
- The Chi-squared test is not recommanded with very small data sets (most expected values bellow 10), and can be replaced by the similar **Fisher test** in such cases.

Chi-squared test: Summary



Example

	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

	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

	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

	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

	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

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

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

Example

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

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

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

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

Mining multivariate data

When each data is described by multiple values (i.e. several random variables), the intuitive analysis and visualization of the data becomes difficult or impossible.

Issues with multivariate datasets

- Some features may be redundant.
- The variables may be different in nature (numerical, categorical, binary, etc.)
- There may be missing values.

Preparing the data: missing values

Missing values can be a problem because they make the data unreliable and prevent some calculations to be done. There are several ways to deal with them:

- Ignoring them when it is possible.
- Removing the data that have missing values.
- Trying to fill in the missing values.

Preparing the data: missing values

Ignoring missing values

Most univariate and bivariate statistics can still be done while ignoring missing values.

- Most data analysis softwares have a parameter to ignore missing values.
- **Pairwise deletion** in bivariate statistics can cause impossible mathematical situations where not all measures have the same N.

Removing data with missing values

Removing any data with missing values is a solution of last resort.

- It is the most commonly used solution.
- If the data are not missing randomly, removing them can cause a significant bias.

Preparing the data: missing values

- In statistics, **imputation** is the process of replacing missing data with substituted values.
- It is a complex field and we will only give some basic ideas.

Single Imputation: Examples

- Finding the most similar complete data, and using its value to fill in the missing ones.
  - Weakness: Artificially increases any correlation coefficients.
- Replacing the missing values with the mean of the considered variable.
  - Weakness: Reduces dispersion criteria.
- Using regression techniques (Cf Lecture 3) to imput the missing variable from the values of the others.
  - Weakness: Artificially increases any correlation coefficients due to overfitting problems.

Preparing the data: missing values

Single imputation can be biased and is sometimes replace with multiple imputation techniques when more reliability is needed.

Multiple Imputation: Examples

- Using multiple stochastic regressions to cause the least changes on the position and dispersion criteria of the missing variables.
- Using a generative model and guessing the missing values a posteriori using maximum likelihood expectation.
- Training a supervised classifier (Cf. Lecture 5) to imput the missing values.

Preparing the data: dimensionality reduction

It is sometimes possible to **reduce the dimensionality** of a multivariate data set: If only 2 or 3 variable remain, it is possible to visualize the data with minimal loss of information.

Dimensionality reduction

If several variables are highly correlated, the information they contain is redundant:

- Keep only one of them.
- We keep the variable best correlated with eliminated variables, but less correlated with the remaining variables.

We will see in the next course, that **Principal component analysis** is one of the most effective method for dimensionality reduction.

Preparing the data: Normalizing the data

- Most techniques that we have introduced for bivariate analysis can be applied to the the analysis of multiple variables by pairs of two.
- However, one issue with multivariate data sets is that the different variables have an even higher likelihood to come with very different scales and units.

Normalizing the data becomes a mandatory first step before the data can be analyzed using any technique (bivariate or multivariate).

Normalizing the data: unit normalization

A first method to normalize numerical continuous variables is to scale them between 0 and 1.

- Let us consider a data set X containing N data, each having D variables.

Min-Max normalization

$$\tilde{X} = \begin{pmatrix} \frac{x_{1,1} - min(X_1)}{max(X_1) - min(X_1)} & \cdots & \frac{x_{1,D} - min(X_D)}{max(X_D) - min(X_D)} \\ \vdots & \ddots & \vdots \\ \frac{x_{N,1} - min(X_1)}{max(X_1) - min(X_1)} & \cdots & \frac{x_{N,D} - min(X_D)}{max(X_D) - min(X_D)} \end{pmatrix}$$

Outliers may be an issue with this normalization.

Normalizing the data: logarithmic normalization

- It reduces the effect of outliers.
- It is also useful when the attribute have non-linear correlations.

logarithmic normalization

$$Y = \begin{pmatrix} \log_a(1 + b\frac{x_{1,1} - min(X_1)}{max(X_1) - min(X_1)}) & \cdots & \log_a(1 + b\frac{x_{1,D} - min(X_D)}{max(X_D) - min(X_D)}) \\ \vdots & \ddots & \vdots \\ \log_a(1 + b\frac{x_{N,1} - min(X_1)}{max(X_1) - min(X_1)}) & \cdots & \log_a(1 + b\frac{x_{N,D} - min(X_D)}{max(X_D) - min(X_D)}) \end{pmatrix}$$

- Different values a and b can be used depending on the intended effect.

Normalizing the data: centered normalization

Centering the data around zero based on their mean is another possible normalization.

Centered reduced variables

$$\bar{M} = \begin{pmatrix} x_{1,1} - \mu_1 & \cdots & x_{1,D} - \mu_D \\ \vdots & \ddots & \vdots \\ x_{N,1} - \mu_1 & \cdots & x_{N,D} - \mu_D \end{pmatrix}$$

Standardizing the data: Centering and reducing

The dispersion of the data is a problem left unaddressed when only centering the data. Prior to several operations such as a **Principal Component Analysis**, centering the variables around their means and scaling them may be necessary.

Centered Reduced variables: Standardization

$$\tilde{M} = \begin{pmatrix} \frac{x_{1,1}-\mu_1}{\sigma_1} & \dots & \frac{x_{1,D}-\mu_D}{\sigma_D} \\ \vdots & \ddots & \vdots \\ \frac{x_{N,1}-\mu_1}{\sigma_1} & \dots & \frac{x_{N,D}-\mu_D}{\sigma_D} \end{pmatrix}$$

This way, all variables have a mean of zero and a unit variance.

Variance-Covariance Matrix

Once the data have been normalized (one way or another), regular univariate and bivariate statistics measures can be used to describe the relations between the different variables.

- The Variance-Covariance matrix denoted C or Σ is a common correlation measure for multivariate data.

Variance-Covariance

$$C = \begin{pmatrix} VAR(X_1) & \cdots & Cov(X_D, X_1) \\ \vdots & \ddots & \vdots \\ Cov(X_1, X_D) & \cdots & VAR(X_D) \end{pmatrix}$$

Correlation Matrix

When the data are reduced, then the covariance is equal to the correlation and we obtain a correlation matrix:

Correlation Matrix

$$\tilde{C} = \begin{pmatrix} 1 & \cdots & Cor(X_D, X_1) \\ \vdots & \ddots & \vdots \\ Cor(X_1, X_D) & \cdots & 1 \end{pmatrix}$$

This matrix is easier to read than the regular variance-covariance matrix. However, the individual variance of each variable is lost.

Interpretation

The variance-covariance matrix and the correlation matrix give a comprehensible overview of the relations between the data:

- As long as there are not too many variables, its small size $D \times D$ can be easily visualized.
- Visualizing the matrix makes it possible to easily find correlated variables.

This matrix is also used in many processing techniques: model based clustering, principal component analysis, etc.

Processing Multivariate data

Since it is not possible to directly visualize and easily process multivariate data (at least not as easily as univariate or bivariate data), other possibilities are available:

- **Dimension reduction techniques** (Lecture 3 & 7): They reduce the number of variables to make the problem easier and sometimes reduce it to a bivariate problem.
- **Clustering techniques** (Lecture 4): It is an unsupervised technique that aims at finding groups of similar data on D-dimensional data sets.

Outline

Bibliography

- Ad Feelders, Advanced Data Mining 2011
- Srinivasan Parthasarathy, Introduction to Data Mining
- Min Song, Data Mining