MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet as association measures
- Chi-squared as the contribution to data scatter
- Homework 4 (with a solution)

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- **Nominal feature; probabilistic interpretation**
- Contingency table and bivariate distribution
- Conditional probability and independence; Bayes theorem and total probability rule
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared as a criterion of independence
- Chi-squared and average Quetelet
- Homework 4

Illustrative Data table

Company data; first three companies making product A, next three making product B, and the last two product C.

Company name	Income, $mln	MShare,%	NSup	AA	Sector
Aversiona	19.0	43.7	2	No	Utility
Antyops	29.4	36.0	3	No	Utility
Astonite	23.9	38.0	3	No	Manufacture
Bayermart	18.4	27.9	2	Yes	Utility
Breaktops	25.7	22.3	3	Yes	Manufacture
Bumchista	12.1	16.9	2	Yes	Manufacture
Civiok	23.9	30.2	4	Yes	Retail
Cyberdam	27.2	58.0	5	Yes	Retail

Company Dataset

Metadata: **Object names, Features and Domain knowledge**

1) Income, $ Mln;
2) MShare - Market share, per cent;
3) NSup - Number of principal suppliers;
4) Affirmative Action (AA) - Yes or No;
5) Sector - (a) Retail, (b) Utility, and (c) Manufacture.

Feature: Maps entities to feature values (unlike "random variable": a density function)

Quantitative scale: [1) Income, 2) MShare, 3) Nsup], **Binary scale: [**AA], **Nominal scale** [Sector]

Nominal feature as partition

# Company	Income	MShare,%	NSup	AA	Sector
1 Aversiona	19.0	43.7	2	No	Utility
2 Antyops	29.4	36.0	3	No	Utility
3 Astonite	23.9	38.0	3	No	Manufacture
4 Bayermart	18.4	27.9	2	Yes	Utility
5 Breaktops	25.7	22.3	3	Yes	Manufacture
6 Bumchista	12.1	16.9	2	Yes	Manufacture
7 Civiok	23.9	30.2	4	Yes	Retail
8 Cyberdam	27.2	58.0	5	Yes	Retail

Feature S:

S(1)=S(2)=S(4)=v1 ('Utility')

S(3)=S(5)=S(6)=v2 ('Manuf.')

S(7)=S(8) =v3 ('Retail')

Partition with S^{-1}: $R = \{S^{-1}(v)\}$

R= {{1,2,4}, {3,5,6}, {7,8}}

These two representations are equivalent in Data Science

Nominal feature: probabilistic view

Feature S:

S(1)=S(2)=S(4)=v1 ('Utility')
S(3)=S(5)=S(6)=v2 ('Manuf.')
S(7)=S(8) =v3 ('Retail')

⟺

Partition with S^{-1}: R= {S^{-1}(v)}

R= {{1,2,4}, {3,5,6}, {7,8}}

The two are equivalent in Data Science

Probability Theory PT:

Probability of each out of N dataset entities is 1/N

Probability of a subset is the proportion of its cardinality:

***P(R)*=**{3/8, 3/8, 2/8**}** (totaling to 1)

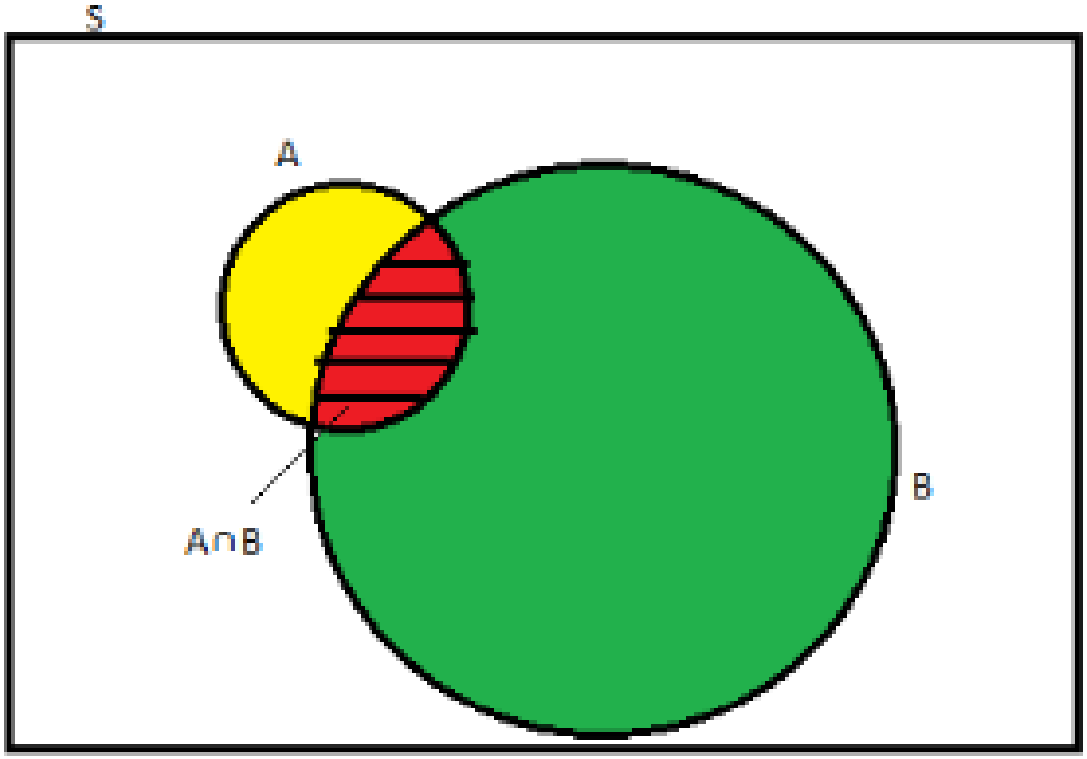
A subset is referred to as an event in PT. A partition in PT is referred to as a total system of events.

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- **Conditional probability and independence; Bayes theorem and total probability rule**
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet
- Homework 4

Conditional probability; Bayes theorem, I



Given events A, B, take A∩B and probabilities p(A), p(B), and p(A∩B).
Conditional probability def:

$p(A|B)=p(A\cap B)/p(B)$
$p(B|A)=p(A\cap B)/p(A)$

Bayes theorem:
$p(B|A)=p(A|B)p(B)/p(A)$

Conditional probability; Bayes theorem, 2

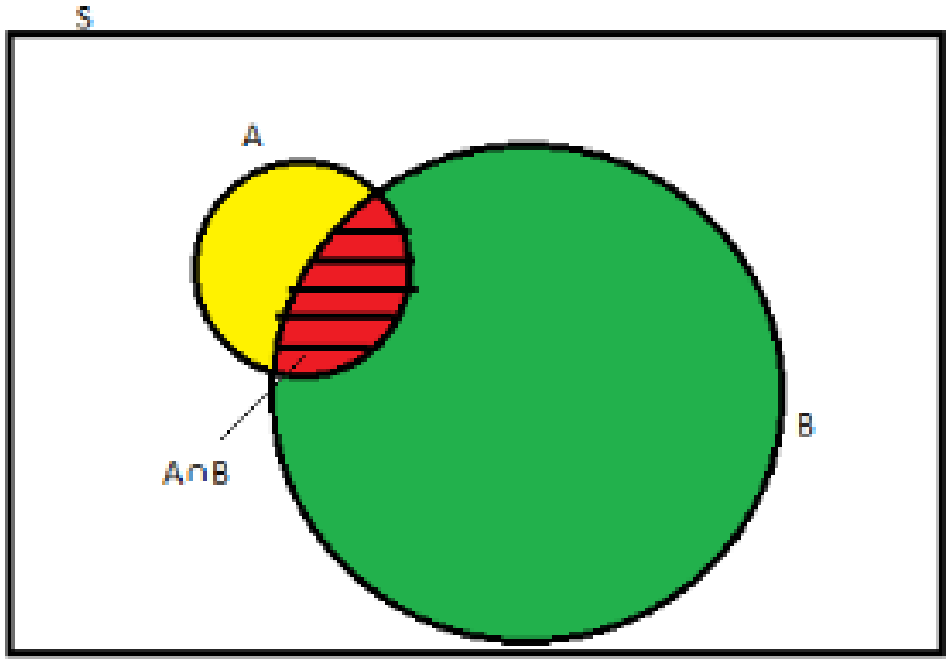


Figure:

p(A)=1/8, p(B)=1/2,

p(A∩B)=1/16.

Then:

p(A|B)=(1/16)/(1/2)=1/8

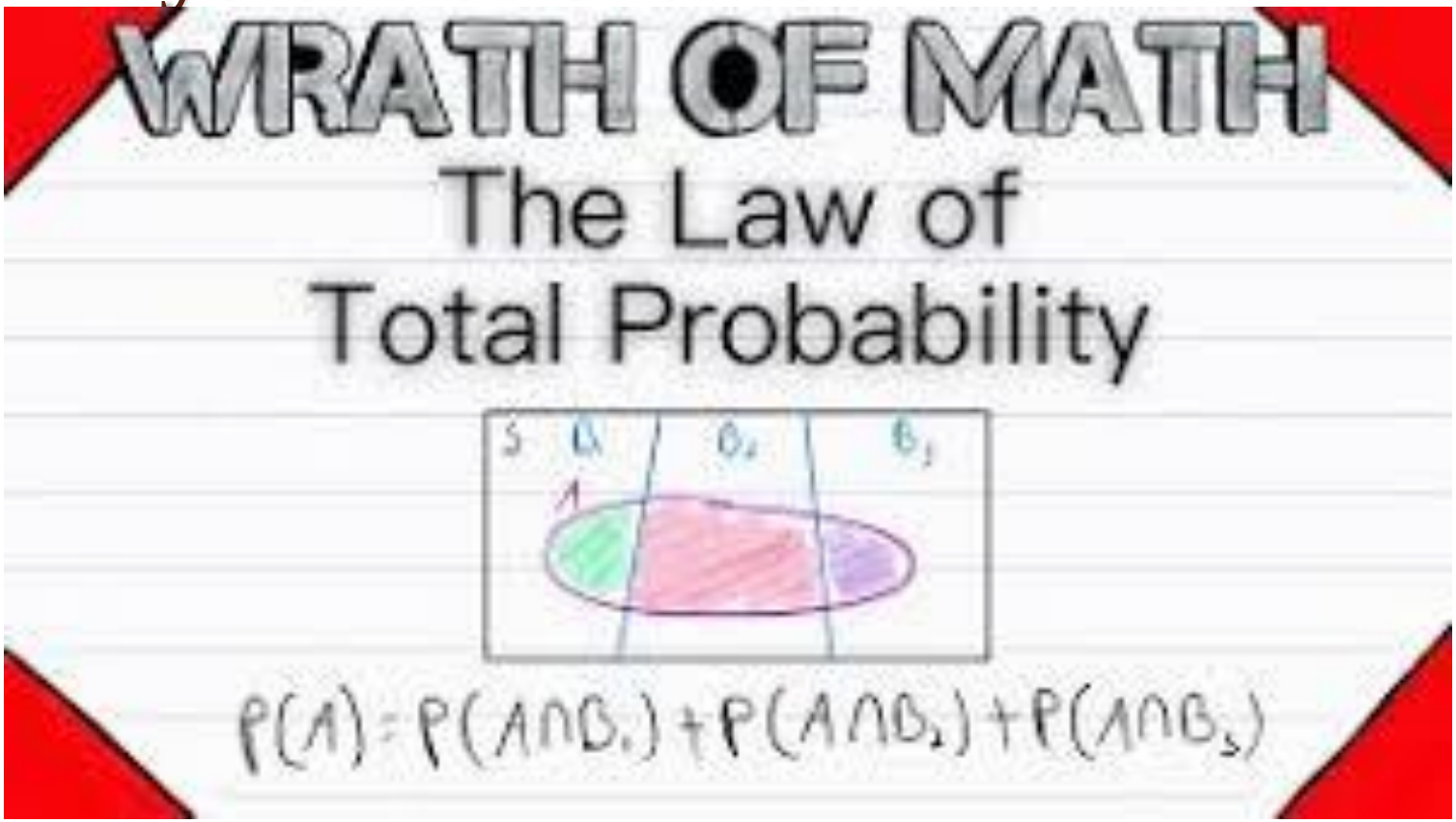
p(B|A)=(1/16)/(1/8)=1/2

Bayes theorem:

p(B|A)=p(A|B)p(B)/p(A)

½ = (1/8*1/2)/(1/8)

Total Probability Rule with total system of events



$$P(A)=p(A|B_1)*p(B_1)+p(A|B_2)*p(B_2)+p(A|B_3)*p(B_3)$$

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet
- Homework 4

Contingency table, I

- Two sets of categories on the same entity set, **k=1,…K**, and **v=1,…,L**
- **Cross-classify them by putting k in rows and v in columns: the frequency of (k,v) cases, N_{kv}, the relative frequency p_{kv}**
- Conditional frequency (probability)

$$p(v/k) = p_{kv} / p_k$$

Contingency table, 2

Iris data: 150×4. Nominal features? What is about 3 of them?

A feature Taxon is available, 3 categories: T1, T2, T3 (50 specimens each) – a partition

Let us develop two nominal features out of Sepal measures:

Sepal Length and Sepal Width

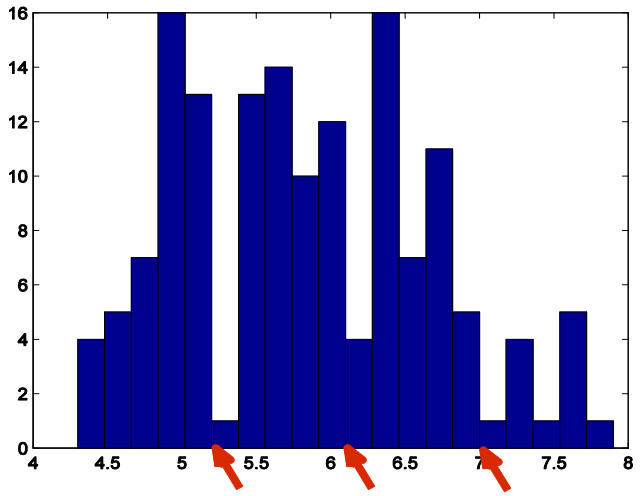
Categorize a quantitative feature, 1

Iris data: 150×4. Nominal features?

Let us categorise

Sepal Length by using

its Histogram S.L.



Minima define boundaries of categories (in red)

Categorize a quantitative feature, 2

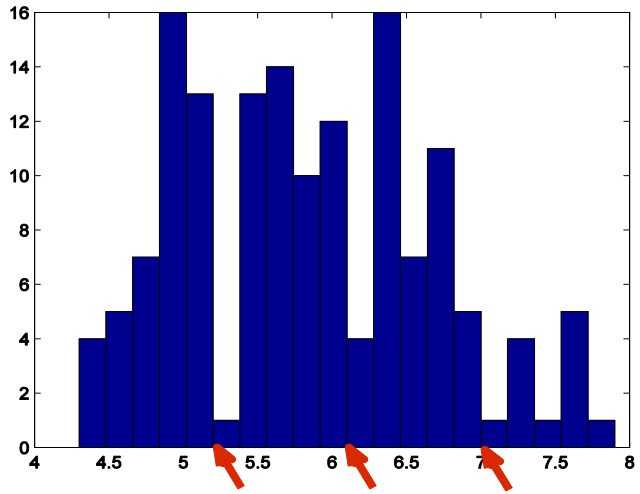
Iris data: 150×4 table. Nominal f

Define categories of

Sepal Length,

separated by “minima” in its

hystogram in 20 bins



Boundaries between categories

>>a=[4 5.2 6.1 7.0 8]

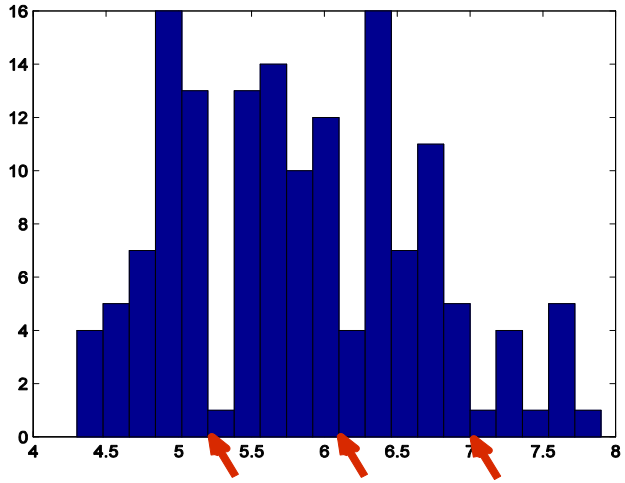
5 items, here **4** and **8** – are endpoints of the range interval.

Categorize a quantitative feature, 3

Define 4 categories of Sepal Length

over dividers in (a):

>>**a=[4 5.2 6.1 7.0 8]**



MatLab:

>>for k=1:4;f=find(sl>=a(k) & sl<a(k+1));g(f)=k; end;

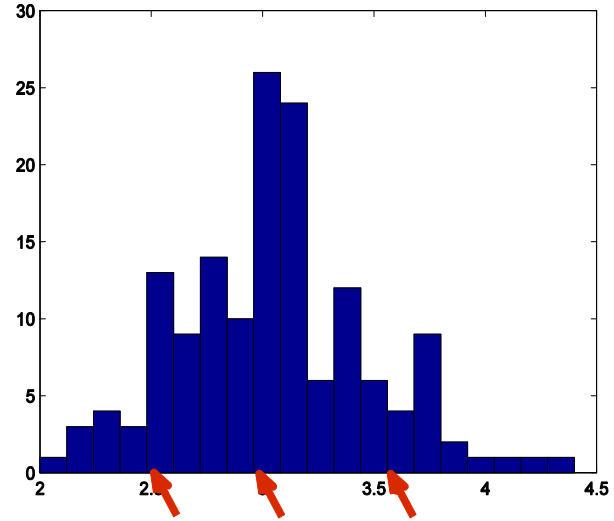
Here **sl** – is a 150x1 array with Sepal Length feature;

find - operation, for finding indices, for which the predicate holds

sl>=a(k) & sl<a(k+1)); "selects indices of the objects falling between boundaries in k-th category, with the exception of a(k+1)"; **g** –the nominal feature being defined

Categorize a quantitative feature, 4

Iris data: 150×4. Define yet one more, this time out of Sepal Width, convenient red points are chosen.



>>b=[2 2.5 3.0 3.6 4.5]

2 and 4.5 – are boundary

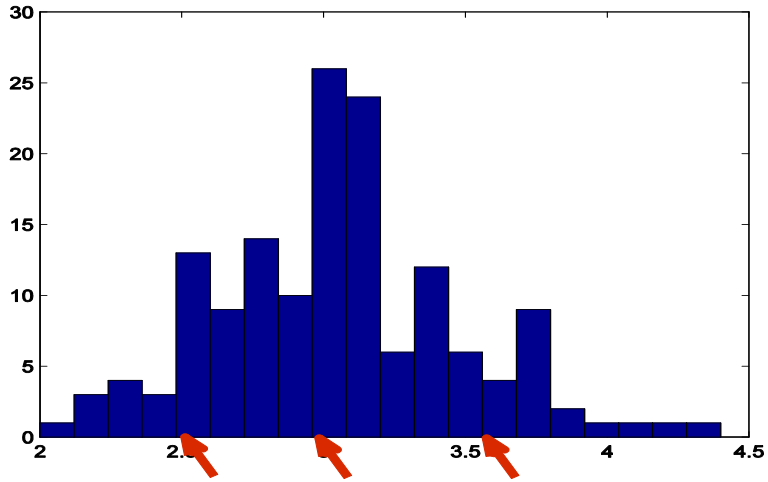
Categorize a quantitative feature, 6

Define categories of

Sepal Width, by using

>>b=[2 2.5 3.0 3.6 4.5]

Nominal ***h***:



>>for k=1:4;f=find(sw>=b(k) & sw<b(k+1));h(f)=k; end;

Here **sw** – is a 150×1 array for Sepal Width feature

find – an operation for choosing indices at which sw satisfies the predicate

sw>=b(k) & sw<b(k+1)); - "between boundaries of k-й category" and,

h, assigns a label, **k**, to all objects of this set.

Nominal Taxon feature

Define Taxon feature, **t** : Beginning 50 indices form T1, next 50, T2, and the last 50, T3:

```
>>for k=1:3;f1=(k-1)*50+1; f2=k*50;
t([f1:f2])=k; end;
```

Here **f1:f2** is an interval of integers between 1 and 50 at k=1, between 51 and 100 at k=2, and between 101 and 150 at k=3.

Contingency table, 3

Taxon (*Tk*) and Categorized_Sepal_Length (*Gl*)

Taxon	Sepal_Length_Category				Total
	G1	G2	**G3**	G4	
T1	36	14	0	0	50
T2	4	26	**19**	1	50
T3	1	8	29	12	50
Total	41	48	48	13	150

19 in cell (T2,G3) – the number of objects in the intersection of T2 and G3

Contingency table, 4

Taxon (*Tk*) and Categorized_Sepal_Length (*Gl*)

Taxon	Sepal_Length_Category				Total
	G1	G2	G3	G4	
T1	36	14	0	0	50
T2	4	26	19	1	50
T3	1	8	29	12	50
Total	41	48	48	13	150

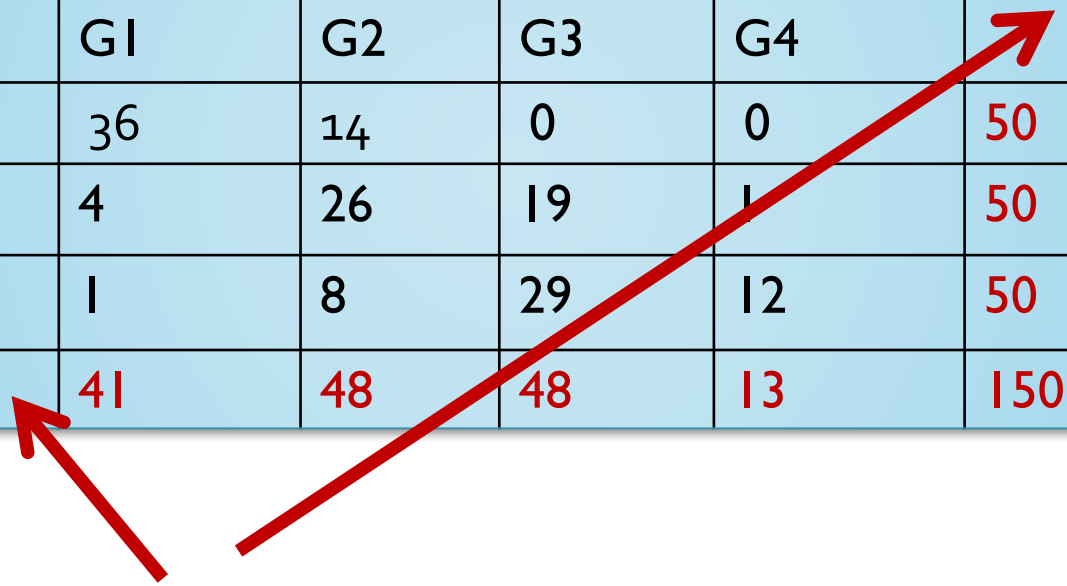
MatLab:

```
>> for k=1:3;for l=1:4; nl(k,l)=length(find(g==l & t==k));end;end
```

Contingency table, 5

Taxon (*Tk*) and Categorized_Sepal_Length (*Gl*)

Taxon	Sepal_Length_Category				Total
	G1	G2	G3	G4	
T1	36	14	0	0	50
T2	4	26	19	1	50
T3	1	8	29	12	50
Total	41	48	48	13	150



Marginal frequencies: totals of rows and columns – frequencies of Tk and Gl

Contingency table, 6

Taxon (*Tk*) and Categorized_Sepal_Length (*Gl*)

Relative frequencies

Taxon	Sepal_Length_Category				Total
	G1	G2	G3	G4	
T1	0.240	0.093	0	0	0.333
T2	0.027	0.173	0.127	0.007	0.333
T3	0.007	0.053	0.193	0.080	0.333
Total	0.273	0.320	0.320	0.087	1

Proportions of the total (found by dividing frequencies by N=150)

Conditional probability, 1

Taxon (*Tk*) and Categorized_Sepal_Length (*Gl*)

Conditional probability *p(Tk|Gl)* of taxon *Tk* at SL-category *Gl*:

Taxon	Sepal_Length_Category				Total
	G1	**G2**	**G3**	**G4**	
T1	36/41	14/48	0/48	0/13	50
T2	4/41	26/48	19/48	1/13	50
T3	1/41	8/48	29/48	12/13	50
Total	41	48	48	13	150

Conditional probability, 2

Taxon *Tk* and Sepal_Length_Category *Gl*

Conditional probability *p(Tk|Gl)* of taxon *Tk* given SL-category *Gl:*

Taxon	Sepal_Length_Category				Total
	G1	**G2**	**G3**	**G4**	
T1	**0.878**	0.292	0	0	50
T2	0.098	**0.542**	0.396	0.077	50
T3	0.024	0.167	**0.604**	**0.923**	50
Total	41	48	48	13	150

Those highlighted by bold are indicative, especially T1 given G1 and T3 given G4: almost conceptual rules G1⇒T1 and G4 ⇒T3.

Conditional probability, 3

Taxon (*Tk*) as a function of Categorized_Sepal_Length (*Gl*)

Taxon	Sepal_Length_Category				Total
	G1	**G2**	**G3**	**G4**	
T1	**0.878**	0.292	0	0	50
T2	0.098	**0.542**	0.396	0.077	50
T3	0.024	0.167	**0.604**	**0.923**	50
Total	41	48	48	13	150

Some values (very high or very low) are useful

Next will be shown a dull table, at which the conditional probability does not help.

A dull contingency table: **no contrasts** (categories of CSW as functions of CSL)

SW-Cat	Cond. Probability SW|SL				Total
	G1	**G2**	**G3**	**G4**	
H1	0.098	0.104	0.042	0	11
H2	0.073	**0.417**	0.396	0.308	46
H3	**0.658**	0.292	**0.562**	**0.462**	74
H4	0.171	0.188	0	0.231	19
Total	41	48	48	13	150

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- **Chi-squared as a criterion of independence**
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet
- Homework 4

Statistical independence, 1

Relative frequencies:

SW-Cat	Sepal_Length_Category				Total
	G1	**G2**	**G3**	**G4**	
H1	0.027	0.033	0.013	0	0.073
H2	0.020	0.133	0.127	0.027	0.307
H3	0.180	0.093	0.180	0.040	0.493
H4	0.047	0.060	0	0.020	0.127
Total	0.273	0.320	0.320	0.087	1

Karl Pearson (1867-1936): Statistical independence can be tested

How?

Karl Pearson (27 March 1857 – 27 April 1936, UK), a founding father of data science and mathematical statistics

Statistical independence, 2

Two features are statistically independent iffor all *k,l*:

SW-Cat	Sepal_Length_Category				Total
	G1	G2	G3	G4	
H1	0.027	0.033	0.013	0	0.073
H2	0.020	0.133	0.127	0.027	0.307
H3	0.180	0.093	0.180	0.040	0.493
H4	0.047	0.060	0	0.020	0.127
Total	0.273	0.320	0.320	0.087	1

$$p(Hk \cap Gl) = p(Hk)P(Gl)$$

***p(Hk), p(Gl)* are marginal probabilities**

Let us check whether $p(H3 \cap G1) = p(H3)P(G1)$???

0.180 0.493*0.273= 0.135

NO!!!

Difference 0.180 - 0.135=0.045, is rather high; H3 & G1 occurs more frequently than at the independence:positive association

Statistical independence, 3

0.027	0.033	0.013	0	**0.073**
0.020	0.133	0.127	0.027	**0.307**
0.180	**0.093**	0.180	0.040	**0.493**
0.047	0.060	0	0.020	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

Observed relative frequencies

$$p(Hk \cap Gl)$$

0.020	0.024	0.024	0.006	**0.073**
0.084	0.098	0.098	0.027	**0.307**
0.135	**0.158**	0.158	0.043	**0.493**
0.035	0.040	0.040	0.011	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

Expected at the independence:

$$p(Hk)P(Gl)$$

A weak correlation!

Only 3 entries out of 16 differ by more than 0.04 (in bold)

Pearson chi-squared, 1

$p(Hk \cap Gl)$

0.027	0.033	0.013	0	**0.073**
0.020	0.133	0.127	0.027	**0.307**
0.180	**0.093**	0.180	0.040	**0.493**
0.047	0.060	0	0.020	**0.127**
0.273	**0.320**	**0.320**	**0.087**	1

$p(Hk)P(Gl)$

0.020	0.024	0.024	0.006	**0.073**
0.084	0.098	0.098	0.027	**0.307**
0.135	**0.158**	0.158	0.043	**0.493**
0.035	0.040	0.040	0.011	**0.127**
0.273	**0.320**	**0.320**	**0.087**	1

Pearson chi-squared sums quadratic differences between the observed bivariate distribution and that would be under the statistical independence:

$$X^2 = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(p(Hk \cap Gl) - p(Hk)p(Gl)\right)^2}{p(Hk)p(Gl)}$$

Pearson chi-squared, 2

$p(Hk \cap Gl)$

0.027	0.033	0.013	0	**0.073**
0.020	0.133	0.127	0.027	**0.307**
0.180	**0.093**	0.180	0.040	**0.493**
0.047	0.060	0	0.020	**0.127**
0.273	**0.320**	**0.320**	**0.087**	1

$p(Hk)P(Gl)$

0.020	0.024	0.024	0.006	**0.073**
0.084	0.098	0.098	0.027	**0.307**
0.135	**0.158**	0.158	0.043	**0.493**
0.035	0.040	0.040	0.011	**0.127**
0.273	**0.320**	**0.320**	**0.087**	1

Pearson: Under the hypothesis that the features are independent in the population, and entity sampling has been done randomly and independently, the density function of random variable NX^2 tends to distribution χ^2 with (K-1)(L-1) degrees of freedom.

Notes: 1. In many textbooks, it is the value of NX^2 which is denoted by X^2, while denoting our X^2 by φ^2.

2. This theorem allows for probabilistic testing of the hypothesis that the features are independent (see next slide).

3. Distribution χ^2 **with n degrees of freedom is the distribution of $x_1^2 + x_2^2 + \ldots + x_n^2$, where every x_i (i=1,2,…n) is a Gaussian N(0,1) random variable**

Pearson chi-squared, 3

p: $p(Hk \cap Gl)$

0.027	0.033	0.013	0	**0.073**
0.020	0.133	0.127	0.027	**0.307**
0.180	**0.093**	0.180	0.040	**0.493**
0.047	0.060	0	0.020	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

pn: $p(Hk)P(Gl)$

0.020	0.024	0.024	0.006	**0.073**
0.084	0.098	0.098	0.027	**0.307**
0.135	**0.158**	0.158	0.043	**0.493**
0.035	0.040	0.040	0.011	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

Pearson: Under the hypothesis that the features are independent in the population, and entity sampling has been done randomly and independently, the density function of random variable NX^2 tends to distribution χ^2 with f=(K-1)(L-1) degrees of freedom.

Apply this theorem to our case.

We have K=4, L=4, therefore f=9. At f=9, there is a 5% chance that the NX^2 value will be greater than 16.92 if the hypothesis of independence is true. In our case X^2 =0.1929, N=150 so that NX^2 =28.93 > 16.92. The hypothesis is to be rejected with 95% confidence. In fact, it is rejected even with 99.9% confidence, because the critical value in this case is 27.88. **Amusingly, if there were only 50 specimen, NX^2 =9.64: the independence could not be rejected.**

Pearson chi-squared, 4

p: $p(Hk \cap Gl)$ are here

0.027	0.033	0.013	0	**0.073**
0.020	0.133	0.127	0.027	**0.307**
0.180	**0.093**	0.180	0.040	**0.493**
0.047	0.060	0	0.020	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

pn: $p(Hk)P(Gl)$ are here

0.020	0.024	0.024	0.006	**0.073**
0.084	0.098	0.098	0.027	**0.307**
0.135	**0.158**	0.158	0.043	**0.493**
0.035	0.040	0.040	0.011	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

Pearson: Under the hypothesis that the features are independent in the population, and entity sampling has been done randomly and independently, the density function of random variable NX^2 tends to distribution χ^2 with f=(K-1)(L-1) degrees of freedom.

In statistics texts, Pearson chi-squared is not recommended as a correlation measure - it is considered only for testing the independence.

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- **Interpretation of clusters using dummy variables**
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet
- Homework 4

Company Dataset: Quantification

Company name	Income, $mln	MShare,%	NSup	AA	Sector
Aversiona	19.0	43.7	2	No	Utility
Antyops	29.4	36.0	3	No	Utility
Astonite	23.9	38.0	3	No	Manufacture
Bayermart	18.4	27.9	2	Yes	Utility
Breaktops	25.7	22.3	3	Yes	Manufacture
Bumchista	12.1	16.9	2	Yes	Manufacture
Civiok	23.9	30.2	4	Yes	Retail
Cyberdam	27.2	58.0	5	Yes	Retail

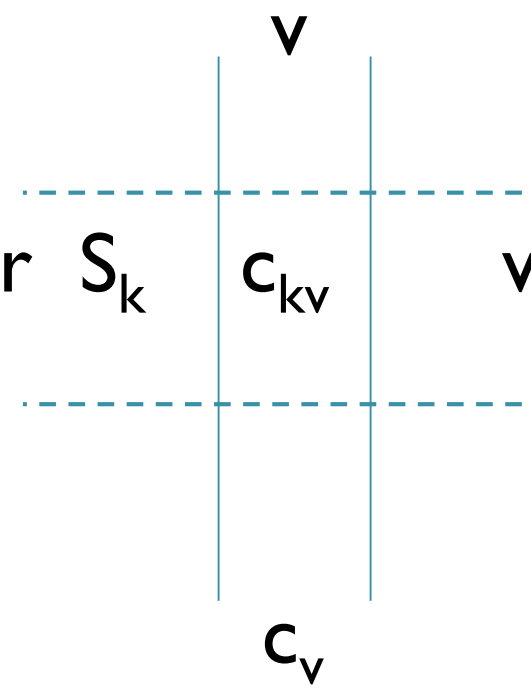
Quantitative coding: Converts categories into 1/0 binary (dummy) features "Does that hold? 1 if Yes, 0 if No."

Entity	Income	MShar	NSup	AA?	Util?	Manu?	Retail?
1	19.0	43.7	2	0	1	0	0
2	29.4	36.0	3	0	1	0	0
3	23.9	38.0	3	0	0	1	0
4	18.4	27.9	2	1	1	0	0
5	25.7	22.3	3	1	0	1	0
6	12.1	16.9	2	1	0	1	0
7	23.9	30.2	4	1	0	0	1
8	27.2	58.0	5	1	0	0	1

Company data 8×5 converted into quantitative format 8×7

Interpretation of nominal feature

- Recall interpretation of a quantitative feature v



Compare $c_{kv} - c_v$ **(absolute),**

$(c_{kv} - c_v)/c_v = c_{kv}/c_v - 1$ **(relative) – interpret !!!**

Cluster interpretation

Center c_k for Interpretation of cluster S_k:

Iris taxon T1: Specimens number 1, 2, …, 50

Taxon T1 Interpretation: SMALL PETAL

	SLength	SWidth	PLength	PWidth
Center	5.006	3.428	1.462	0.246
Grand mean	5.843	3.057	3.758	1.199
Difference	–0.837	0.371	–2.296	–0.953
Difference, %	–14.3	+12.1	**–61.1**	**–79.5**

This is $c_{kv}/c_v - 1$, per cent!

Nominal feature interpretation, I

- Category *v* quantified:

	Category *v*
	1 0
	2 1
N objects	… …
	i 1
	… …
	N 0
Average	p_v

– proportion of v

Nominal feature interpretation, 2

- Category *v* quantified: **Average: p_v - proportion of v**

- **Category *v* within cluster S_k with $|S_k| = N_k$ objects**

	i1	0
	i2	1
N_k objects	…	…
	ij	1
	…	…
	N_k	0

Average: $p(v/k) = N_{vk}/N_k = p_{vk}/p_k$

- proportion of v in S_k = probability [relative frequency] of v under condition S_k

Nominal feature interpretation, 3a

- Category *v* quantified: **Average: p_v - proportion of v**

- **Within-cluster average:**
  p(v/k) - proportion of v in S_k = probability [relative frequency] of v under condition S_k

- **Interpretation, full analogy to the quantitative case: the center of S_k at *v* relates to the mean of *v*:**
- $$I_{kv}=[p(v/k) - p_v]/p_v = p(v/k)/p_v - 1$$

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet
- Chi-squared as the contribution to data scatter
- Homework 4

Nominal feature interpretation, 3b

- Category *v* quantified, at cluster S_k
- **Interpretation, full analogy to the quantitative case: the center of S_k at *v* relates to the mean of *v*:**
- $$q_{kv}=[p(v/k) - p_v]/p_v = p(v/k)/p_v - 1$$
- **Adolphe Quételet (1796–1874) index (1832)**
- $$q_{kv}=p(v/k)/p_v - 1 = \frac{p_{vk}}{p_k p_v} - 1$$
- **Shows the relative change of probability of *v* under condition of S_k (from the average)**

Lambert Adolphe Jacques Quételet (22.02.1796 – 17.02.1874, Belgium), founding father of social statistics

Nominal feature interpretation, 3c

- **Adolphe Quételet (1796–1874) index (1832)**
- $$q_{kv}=p(v/k)/p_v - 1 = \frac{p_{vk}}{p_k p_v} - 1$$
- **Shows the relative change of probability of *v* under condition of S_k (from the average)**
- **Corresponds to "lift" (marketing), "overrepresentation" (biology)**
- **Examples:**
  - 1. p(tbc)=0.001, p(tbc/bad housing)=0.01, q=0.01/0.001-1=900%
  - 2. Company:
    p(Manu)=3/8; p(Manu/B)=2/3
    q= 2/3/3/8 -1=16/9-1=177%

Quetelet index

Quetelet index ***q(Hk/Gl)*** **at SW-categ. *Hk* under condition SL-categ. *Gl*, per cent:**

H4, given G4, is 82% more frequent than on average

SW Cat	Quetelet 100*$q(Hk/Gl)$				Prob. $p(Hk)$
	G1	**G2**	**G3**	**G4**	
H1	33	42	-43	- 100	0.073
H2	-76	36	29	0	0.301
H3	33	-40	14	- 6	0.493
H4	35	48	-100	82	0.127

This could not be seen using conditional probabilities

Расчет индексов Кетле, 2

Индекс Кетле *q(h/g)* категории SW *h* при заданной SL-кат. *G:*

H4 при условии G4, 82% более часта, чем в среднем

По условным вероятностям это не видно:

Напротив:
p(H4|G4) =0.231, минимальное положительное число в столбце

SW-Cat	Cond. Probability SW|SL				Total
	G1	**G2**	**G3**	**G4**	
H1	0.098	0.104	0.042	0	11
H2	0.073	**0.417**	0.396	0.308	46
H3	**0.658**	0.292	**0.562**	**0.462**	74
H4	0.171	0.188	0	0.231	19
Total	41	48	48	13	150

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- **Chi-squared and average Quetelet as association measures**
- Chi-squared as the contribution to data scatter
- Homework 4

Average Quetelet index

Relative contingency table for Sep_L & Sep_W

	G1	G2	G3	G4
H1	0.027	0.033	0.013	0
H2	0.020	0.133	0.127	0.027
H3	0.180	0.093	0.180	0.040
H4	0.047	0.060	0	0.020

Quetelet index table

	G1	G2	G3	G4
H1	0.33	0.42	-0.43	-1
H2	-0.76	**0.**36	0.29	0
H3	0.33	-0.40	0.14	- 0.06
H4	0.35	0.48	-1	**0.82**

Average Quetelet index is the inner product of two matrices:

Observed relative contingency table and the table of Quetelet indices:

$Q(H/G) = 0.193 = 19.3\%$,

Meaning: on average, knowledge of GI categories "adds" 19.3% to the frequency of Hk

Average Quetelet index = Pearson chi-squared

p: $p(Hk \cap Gl)$

0.027	0.033	0.013	0	**0.073**
0.020	0.133	0.127	0.027	**0.307**
0.180	**0.093**	0.180	0.040	**0.493**
0.047	0.060	0	0.020	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

pn: $p(Hk)P(Gl)$

0.020	0.024	0.024	0.006	**0.073**
0.084	0.098	0.098	0.027	**0.307**
0.135	**0.158**	0.158	0.043	**0.493**
0.035	0.040	0.040	0.011	**0.127**
0.273	**0.320**	**0.320**	**0.087**	**1**

MatLab's computation of the chi-squared chi2:

>> d=p-pn; dd=d.*d; chi=dd./pn; chi2=sum(sum(chi))
%=0.1929

MatLab's computation of the summary Quetelet qs:

>> q=p./pn -1; qq=p.*q; qs=sum(sum(qq)%=0.1929

Here **.*, ./** are the operations over corresponding entries, and p, pn are arrays being 4×4 substance fragments of the respective tables above.

Average Quetelet index- some maths, I

Quetelet index is the relative change of probability of Hk from that on average to that given Gl:

$$q(Hk|Gl) = \frac{p(Hk \cap Gl)}{p(Hk)p(Gl)} - 1 = q(Gl|Hk)$$

Average Quetelet index: the sum of Quetelet indices over all k, l weighted by their probabilities $p(Hk \cap Gl)$

in the Contingency Table of relative frequencies:

$$Q = \sum_{k=1}^{K} \sum_{l=1}^{L} p(Hk \cap Gl)\, q(Hk|Gl) = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{p^2(Hk \cap Gl)}{p(Hk)p(Gl)} - 1 \quad (*)$$

Average Quetelet index= Pearson chi-squared, 2

Pearson's Chi-Squared coincides with Q (Mirkin 2019)

$$Q = \sum_{k=1}^{K}\sum_{l=1}^{L} p(Hk \cap Gl)\, q(Hk|Gl) = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{p^2(Hk \cap Gl)}{p(Hk)p(Gl)} - 1 \qquad \text{(i)}$$

$$X^2 = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{\left(p(Hk \cap Gl) - p(Hk)p(Gl)\right)^2}{p(Hk)p(Gl)} \qquad \text{(ii)}$$

(i)=(ii). For a proof see next slide:

X^2 in (ii) can be transformed to the right-hand expression in (i): see next slide.

Pearson's chi-squared is as a measure of correlation as Q is.

The meaning: $Q=X^2$ is the average relative increase in the occurrence of Hk values when Gl become known.

Average Quetelet index= Pearson chi-squared, 3

Proof:

$$Q = \sum_{k=1}^{K}\sum_{l=1}^{L} p(Hk \cap Gl)\, q(Hk|Gl) = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{p^2(Hk \cap Gl)}{p(Hk)p(Gl)} - 1 \qquad \text{(i)}$$

$$X^2 = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{\left(p(Hk \cap Gl) - p(Hk)p(Gl)\right)^2}{p(Hk)p(Gl)} \qquad \text{(ii)}$$

Indeed,

$$\frac{\left(p(Hk \cap Gl) - p(Hk)p(Gl)\right)^2}{p(Hk)p(Gl)} = \frac{p^2(Hk \cap Gl) - 2p(Hk \cap Gl)p(Hk)p(Gl) + p(Hk)^2 p(Gl)^2}{p(Hk)p(Gl)} =$$

$$= \frac{p^2(Hk \cap Gl)}{p(Hk)p(Gl)} - 2p(Hk \cap Gl) + p(Hk)p(Gl).$$

Equations $\sum_{k=1}^{K}\sum_{l=1}^{L} p(Hk \cap Gl) = 1$ **and**

$$\sum_{k=1}^{K}\sum_{l=1}^{L} p(Hk)p(Gl) = 1$$

warrant now that (i)=(ii) indeed.

Chi-squared (and Q) as association measure, 1

Values of Q=X^2 are in interval [0, min(K,L)-1].

Proof:

1) $X^2 \geq 0$, as $\frac{\left(p(Hk \cap Gl) - p(Hk)p(Gl)\right)^2}{p(Hk)p(Gl)} \geq 0$

2) **Min:** $\left(p(Hk \cap Gl) - p(Hk)p(Gl)\right)^2 = 0$ **iff**

$$p(Hk \cap Gl) - p(Hk)p(Gl) = 0 \; -$$

statistical independence

3) Max:

(ii)=

$$\sum_{k=1}^{K} \sum_{l=1}^{L} \frac{p^2(Hk \cap Gl)}{p(Hk)p(Gl)} - 1 = \sum_{k=1}^{K} \sum_{l=1}^{L} p(Hk|Gl)p(Gl|Hk) - 1$$

Chi-squared (and Q) as association measure, 2

Values of Q=X^2 are in interval [0, min(K,L)-1].

Proof:

3) Max: (ii) =

$$\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{p^2(Hk\cap Gl)}{p(Hk)p(Gl)}-1=\sum_{k=1}^{K}\sum_{l=1}^{L}p(Hk|Gl)p(Gl|Hk)-1$$

(ii) $\le \sum_{k=1}^{K} p(Gl|Hk)-1$. Let K ≤ L. Case: at each k, $p(Gl_k|Hk)=1$ .Then (ii)=min(K,L)-1.

	l=1			l=L	
k=1	0	p_1	0	0	p_1
k=2	0	0	p_2	0	p_2
k=K	0	p_K	0	0	p_K

Chi-squared (and Q) as association measure, 3

3) Max Q=X^2 =min(K,L)-1 at

	l=1	l=2	l=3	l=L	
k=1	0	$\mathbf{p_1}$	0	0	$\mathbf{p_1}$
k=2	0	0	$\mathbf{p_2}$	0	$\mathbf{p_2}$
…….	…	…	…	…	…
k=K	0	$\mathbf{p_K}$	0	0	$\mathbf{p_K}$

This is a case of LOGICAL / CONCEPTUAL association (out of statistical association):

k=1 ⇒ l=2

k=2 ⇒ l=3

……

k=K ⇒ l=2

Chi-squared (and Q) as an association measure, 4

Min Q=X^2 = 0

This is a case of STATISTICAL INDEPENDENCE

Max Q=X^2 =min(K,L)-1

This is a case of LOGICAL / CONCEPTUAL association

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet
- **Chi-squared as the contribution to data scatter**
- Homework 4

Chi-squared as contribution, I

- Category *v* quantified: **Average: p_v - proportion of v**
- **Dummy v standardized: subtracting $a_v = p_v$, dividing by some b_v**
- **Within-cluster average: p(v/k) probability of v under condition S_k**

- **After standardization, the center of S_k at *v*:**

$$c_{kv} = [p(v/k) - p_v]/b_v = [p_{vk} - p_v p_k]/[b_v p_k]$$

$p_{vk} = N_{vk}/N$, $p_k = N_k/N$,
N_{vk} is the number of ones in S_k over category v

Chi-squared as contribution, 2

- **Dummy v standardized: subtracting $a_v = p_v$, dividing by b_v**
- **Within-cluster average: p(v/k) probability of v under condition S_k**

- **After standardization, the center of S_k at v:**

$$c_{kv} = [p(v/k) - p_v]/b_v = [p_{vk} - p_v p_k]/[b_v p_k]$$

Here $p_{vk} = N_{vk}/N$, $p_k = N_k/N$,

- **Contribution of category v and cluster S_k to the data scatter:**

$$N_k c_{vk}^2 = N[p_{vk} - p_v p_k]^2/[b_v^2 p_k]$$

Chi-squared as contribution, 3

- **Dummy v standardized: subtracting $a_v = p_v$, dividing by b_v**

- **Contribution of category v and cluster S_k to the data scatter:**

$$N_k c_{vk}^2 = N[p_{vk} - p_v p_k]^2 / [b_v^2 p_k]$$

- **At $b_v = 1$, $N_k c_{vk}^2 = N[p_{vk} - p_v p_k]^2 / p_k$**

- **At $b_v = (p_v)^{1/2}$, $N_k c_{vk}^2 = N[p_{vk} - p_v p_k]^2 / [p_v p_k]$**

Chi-squared as contribution, 4

- Dummy v standardized: subtracting $a_v = p_v$, dividing by b_v
- Contribution of category v and cluster S_k to the data scatter:

  At $b_v = 1$, $N_k c_{vk}^2 = N[p_{vk} - p_v p_k]^2 / p_k$

  At $b_v = (p_v)^{1/2}$, $N_k c_{vk}^2 = N[p_{vk} - p_v p_k]^2 / [p_v p_k]$

- Contribution of nominal feature L and partition $S=\{S_k\}$ to the data scatter:

  At $b_v = 1$, $N \sum_{v \in L} \sum_{k=1}^{K} [p_{vk} - p_v p_k]^2 / p_k$

At $b_v = (p_v)^{1/2}$, $N \sum_{v \in L} \sum_{k=1}^{K} \frac{(p_{vk} - p_v p_k)^2}{p_v p_k}$ -

Pearson's chi-squared (NX^2)

Chi-squared as contribution, 5

- **Contribution of nominal feature L and partition $S=\{S_k\}$ to the data scatter:**

At $b_v = 1$, $N \sum_{v \in L} \sum_{k=1}^{K} [pvk - pvp_k]^2 / p_k$

At $b_v = (p_v)^{1/2}$, $N \sum_{v \in L} \sum_{k=1}^{K} \frac{(p_{vk} - p_v p_k)^2}{p_v p_k}$ **that is Pearson's chi-squared (χ^2)**

[$b_v = (p_v)^{1/2}$ is std at Poisson model]

These are partition-to-partition statistics association indexes conventionally derived within statistics contexts. Also, we see that the difference is due to different normalization options.

Nominal feature interpretation, 9

- **Contribution of nominal feature L and partition $S=\{S_k\}$ to the data scatter:**

At $b_v = (p_v)^{1/2}$, $NX^2 = N \sum_{v \in L} \sum_{k=1}^{K} \frac{(p_{vk} - p_v p_k)^2}{p_v p_k}$

Pearson (1857-1936) chi-squared (X^2):

$b_v = (p_v)^{1/2}$ is std at the Poisson model for binary features

MMDA 2019 Lecture 4: Interpretation of clusters over nominal features and basics of probability and statistics

Contents:

- Nominal feature; probabilistic interpretation
- Conditional probability and independence; Bayes theorem and total probability rule
- Contingency table and bivariate distribution
- Chi-squared as a criterion of independence
- Interpretation of clusters using dummy variables
- Category-to-category association by Quetelet
- Chi-squared and average Quetelet as association measures
- Chi-squared as the contribution to data scatter
- Homework 4

Homework 4: Contingency Table

1. Consider three nominal features (**one** of them, not more, may be taken from nominal features in your data)
2. Build two contingency tables over them: present a conditional frequency table and Quetelet relative index tables. Make comments on relations between categories of the common (to both tables) feature and two others.
3. Compute and visualize the chi-square-summary_Quetelet_index over both tables. Comment on the meaning of the values in the data analysis context.
4. Tell what numbers of observations would suffice to see the features as associated at 95% confidence level; 99% confidence level.

Homework 4: Contingency Table for IRIS

1. Consider three nominal features (**one** of them, not more, may be taken from nominal features in your data):

T is the given taxonomy:

```
>>for k=1:3;f1=(k-1)*50+1; f2=k*50; t([f1:f2])=k; end;
```

SL categorized: >>**a=[4 5.2 6.1 7.0 8];**

```
>>for k=1:4;f=find(sl>=a(k) & sl<a(k+1));g(f)=k; end;
```

SW categorized: >>**b=[2 2.5 3.0 3.6 4.5];**

```
>>for k=1:4;f=find(sw>=b(k) & sw<b(k+1));h(f)=k;
end;
```

Three categorical features, t (categories 1, 2, 3), g (categories 1, 2, 3, 4), h (categories 1, 2, 3, 4), are created.

Homework 4: Contingency Table for IRIS, 2

2. Build two contingency tables over them: present a conditional frequency table and Quetelet relative index tables. Make comments on relations between categories of the common (to both tables) feature and two others.

Table (G→T)

```
>> for k=1:3;for l=1:4;
>>    ng(k,l)=length(find(g==l & t==k));
>>    end;end
```

Table (H→T)

```
>> for k=1:3;for l=1:4;
>>    nh(k,l)=length(find(h==l & t==k));
>>    end;end
```

Homework 4: Contingency Table for IRIS, 3

2. Build two contingency tables

Table (G→T) ng

36	14	0	0	50
4	26	19	1	50
1	8	29	12	50
41	48	48	13	150

Table (H→T) nh

1	1	32	16	50
9	25	16	0	50
1	20	26	3	50
11	46	74	19	150

Homework 4: Contingency Table for IRIS, 4

2. Build two contingency tables

Relative Frequency Table (G→T) (divided by n=150) **ngr**

0.2400	0.0933	0	0	0.3333
0.0267	0.1733	0.1267	0.0067	0.3333
0.0067	0.0533	0.1933	0.0800	0.3333
0.2733	0.3200	0.3200	0.0867	1.0000

Relative Frequency Table (H→T) (divided by n=150) **nhr**

0.0067	0.0067	0.2133	0.1067	0.3333
0.0600	0.1667	0.1067	0	0.3333
0.0067	0.1333	0.1733	0.0200	0.3333
0.0733	0.3067	**0.4933**	0.1267	1.0000

Homework 4: Contingency Table for IRIS, 5

2. Build two contingency tables

Conditional Frequency Table (G→T): >>cng=ng(1:3,:)./repmat(ng(4,:),3,1)

T1	**0.8780**	0.2917	0	0
T2	0.0976	0.5417	0.3958	0.0769
T3	0.0244	0.1667	0.6042	**0.9231**
	G1	G2	G3	G4

The highlighted conditional probabilities show associations:

G1 →T1 (probability .88) – short sepal implies taxon T1 and

G4 →T3 (probability .92) – long sepal implies taxon T3.

Conditional Frequency Table (H→T) >>cnh=nh(1:3,:)./repmat(nh(4,:),3,1)

T1	0.0909	0.0217	0.4324	**0.8421**
T2	**0.8182**	0.5435	0.2162	0
T3	0.0909	0.4348	0.3514	0.1579

The highlighted conditional probabilities show unexpected associations:

H1 →T2 (probability .82) – very narrow sepal implies taxon T2

H4 →T1 (probability .92) – very wide sepal implies taxon T1

H2 →no T1 (probability .98) – rather narrow sepal implies anything but not taxon T1.

Homework 4: Contingency Table for IRIS, 6 2.

Quetelet Index Table (G→T): first, derive the relative frequency under independence:

```
>> ngn=ngr(:,5)*ngr(4,:)
```

T1	0.0911	0.1067	0.1067	0.0289	0.3333
T2	0.0911	0.1067	0.1067	0.0289	0.3333
T3	0.0911	0.1067	0.1067	0.0289	0.3333
Marg	0.2733	0.3200	0.3200	0.0867	1.0000
	G1	G2	G3	G4	Margin

Then define Quetelet index matrix

```
>> qg=ngr./ngn -1
>> qg=qg(1:3,1:4)
```

T1	**1.6341**	-0.1250	-1.0000	-1.0000
T2	-0.7073	0.6250	0.1875	-0.7692
T3	-0.9268	-0.5000	**0.8125**	**1.7692**
	G1	G2	G3	G4

Homework 4: Contingency Table for IRIS, 7

2. Two ways for computing chi-squared

2.1. Classic: matrix pg=((ngr-ngn).*(ngr-ngn))./ngn

T1	**0.2433**	0.0017	**0.1067**	0.0289
T2	0.0456	0.0417	0.0038	0.0171
T3	0.0783	0.0267	0.0704	0.0904
	G1	G2	G3	G4

The sum of this is X2=0.7544. This is about 38% of the maximum value of X2=2. Highlighted are items greater than 0.1: most deviated from the independence

2.2. The average Quetelet index: matrix qpg=ngr.*qg

T1	**0.3922**	-0.0117	0	0
T2	-0.0189	**0.1083**	0.0238	-0.0051
T3	-0.0062	-0.0267	**0.1571**	**0.1415**
	G1	G2	G3	G4

The sum of this is Q=0.7544, again; meaning that, on average, knowledge of Gl-group adds to the probability of taxon, 75%. The highlighted entries are those most contributing to the change: G1→T1 (39%), G2 →T2 (11%), G3 →T3 (16%), G4 →T4 (14%).

Homework 4: Contingency Table for IRIS,8

3. Tell, what numbers of observations would suffice to see the features as associated at 95% or 99% confidence level:

The number of degrees of freedom is 9. According to Google tables, at the 95% probability that chi-squared is less than t value, t=16.919, and t= 21.666 for the 99% probability. At X2=0.7544, what N makes N*X2>t?

Obviously, any N>16.919/0.7544=22.4, that is, at any N>23 the hypothesis of statistical independence should be rejected at 95% confidence level. The number N should be raised to N>21.666/0.7544=28.7, that is, 29 or more, to reject the hypothesis at 99% confidence level.

At N smaller than 22 (or 28), the hypothesis should be accepted at 95% (or 99%) confidence level.

Homework 4: Contingency Table for IRIS,9

The computations related to chi-squared for H and T association are similar; still they should be done. Also a conclusion should be drawn which of the two source partitions is better associated with the target one.