Bus 274:
Further Statistics For Business

Harald Schmidbauer

About These Slides

- The present slides are not self-contained; they need to be explained and discussed. This will be done in the lectures.

- Even though being a "work in progress" and subject to revision, the slides constitute copyrighted material.
  If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer	**harald** at **hs-stat** dot **com**
Angi Rösch	**angi** at **angi-stat** dot **com**

- The slides were produced using LaTeX (www.latex-project.org) and R (www.R-project.org) on a GNU/Linux system, all of which are free and open source software (FOSS).

- R files used for this course are available upon request.

Chapter 13:

Contingency Tables and Independence

13.1 Contingency Tables

Contingency tables.

- A contingency table is the result of classifying a population or sample according to two categorical variables.

- Important questions:
  - What is the degree of dependence between the variables?
  - Could the variables be independent?

    (This question refers to the underlying stochastic model!)

13.1 Contingency Tables

Example 1: Who buys eggs from our supermarket?

- In a random sample of 508 supermarket customers,
  - 366 were female, 142 were male;
  - 63 bought eggs, 445 did not buy eggs.
- Are the variables
  - X = sex (values: f / m) and
  - Y = bought eggs? (values: yes / no)

  interrelated? Or are X and Y independent?
- What kind of data are needed to find out?

13.1 Contingency Tables

Example 1: Who buys eggs from our supermarket?

- Are X and Y independent?
- This question cannot be answered using the univariate distributions of X and Y.
- We need observations from the bivariate variable (X, Y):

  (m,no), (f,no), (f,no), (f,no), (f,yes), . . . , (m,no)

  There are 508 data pairs.

13.1 Contingency Tables

Example 1: Who buys eggs from our supermarket?

- Are X and Y independent?
- The joint empirical distribution of X and Y can be represented in a contingency table:

		Y		
		yes	no	
X	f	49	317	366
	m	14	128	142
		63	445	508

13.1 Contingency Tables

Example 2: Illiteracy in Turkey.

- The population of Turkey (1990), age 6+, by gender and literacy (in million persons):

	l	i	
m	22.7	2.9	25.6
f	18.0	7.0	25.0
	40.7	9.9	50.6

- Are *gender* and *literacy* independent?

13.1 Contingency Tables

Example 3: Browser type and monitor resolution.

- A sample of 163 visitors of a website, taken in early 2007, by browser type and screen resolution:

	1024×768	1280×1024	
MS IE 6	35	14	49
MS IE 7	13	14	27
Mozilla Firefox 6	47	40	87
	95	68	163

- Are *browser type* and *monitor resolution* independent?

13.1 Contingency Tables

A model for 2×2-tables.

- 2×2-table: X, Y can each take on only two values.
- Model behind our observed 2×2-table:

		Y		
		1	0	
X	1	π_{11}	π_{10}	$\pi_{1\bullet}$
	0	π_{01}	π_{00}	$\pi_{0\bullet}$
		$\pi_{\bullet 1}$	$\pi_{\bullet 0}$	1

$$\begin{aligned} \pi_{ij} &= P(X = i \text{ and } Y = j) \\ \pi_{i\bullet} &= P(X = i) = \pi_{i1} + \pi_{i0} \\ \pi_{\bullet j} &= P(Y = j) = \pi_{1j} + \pi_{0j} \end{aligned}$$

- In case X and Y are independent: $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$.

13.1 Contingency Tables

Example 4: Who buys eggs from our supermarket?

- Contingency table and estimated probabilities:

		Y		
		yes $\equiv 1$	no $\equiv 0$	
X	f $\equiv 1$	49	317	366
	m $\equiv 0$	14	128	142
		63	445	508

		1	0	
X	1	49/508	317/508	366/508
	0	14/508	128/508	142/508
		63/508	445/508	1

13.1 Contingency Tables

Example 5: Gender and place of residence of Bilgi students.

- Variables:
  - X = sex (values: f / m) and
  - Y = place of residence (values: Asia / Europe)
- Now suppose we have a sample of 100 students.
- The marginal frequencies are as follows:

		place of residence		
		Asia	Europe	
sex	f			60
	m			40
		20	80	100

13.1 Contingency Tables

Example 5: Gender and place of residence of Bilgi students.

- Would we believe that the variables sex and $place\ of\ residence$ are dependent?

- Given the marginal frequencies:
  - How can we estimate the marginal probabilities?
  - How can we estimate the joint probabilities?
  - Which joint absolute frequencies would we $expect$?
  - An example of absolute frequencies we might actually $observe$?

13.1 Contingency Tables

Outlook on Chapter 13.

- 13.2 The Odds Ratio

  Measuring the degree of dependence in a 2×2-table; application in data mining; the odds ratio in inductive statistics

- 13.3 Testing Independence With χ^2

  Independence in an $r \times c$-table; the χ^2 statistic; the χ^2 test of independence

13.2 The Odds Ratio

The odds of an event.

- Model for a Bernoulli experiment:

$$X = \begin{cases} 1 & \text{if success occurs,} \\ 0 & \text{if failure occurs,} \end{cases}$$

  and $P(X = 1) = p$, $P(X = 0) = 1 - p$.

- Odds. . .
  . . . in favour of success: $\frac{p}{1-p}$, . . . against success: $\frac{1-p}{p}$

- "Odds in favour of success" means: How many times more probable is a success than a failure?

13.2 The Odds Ratio

Example: Rolling a fair die once.

- Let A be the event: "The die falls 6."
- Odds in favour of A:

$$\frac{\frac{1}{6}}{1-\frac{1}{6}} = 1:5$$

- Odds against A:

$$\frac{1-\frac{1}{6}}{\frac{1}{6}} = 5:1$$

13.2 The Odds Ratio

Example: Buying eggs from a supermarket.

- Event A: "A randomly selected customer buys eggs."
- We can estimate: $\hat{p}_A = 63/508 = 0.124$
- Odds in favour of A:

$$\frac{0.124}{1 - 0.124} = 1 : 7.06$$

- Odds against A:

$$\frac{1 - 0.124}{0.124} = 7.06 : 1$$

13.2 The Odds Ratio

A pair of Bernoulli experiments.

- Now consider a pair of Bernoulli experiments.
- Model for the Bernoulli experiments:

$$X = \begin{cases} 1 & \text{if success occurs in the } 1^{\text{st}} \text{ experiment,} \\ 0 & \text{if failure occurs in the } 1^{\text{st}} \text{ experiment,} \end{cases}$$

$$Y = \begin{cases} 1 & \text{if success occurs in the } 2^{\text{nd}} \text{ experiment,} \\ 0 & \text{if failure occurs in the } 2^{\text{nd}} \text{ experiment.} \end{cases}$$

13.2 The Odds Ratio

A pair of Bernoulli experiments: the probabilities.

- Probabilities were given like this:

		Y		
		1	0	
X	1	π_{11}	π_{10}	$\pi_{1\bullet}$
	0	π_{01}	π_{00}	$\pi_{0\bullet}$
		$\pi_{\bullet 1}$	$\pi_{\bullet 0}$	1

- Conditional probabilities:

$$P(Y = j|X = i) = \frac{P(X = i \text{ and } Y = j)}{P(X = i)} = \frac{\pi_{ij}}{\pi i\bullet}$$

13.2 The Odds Ratio

Definition of the odds ratio.

- Odds in favour of success in the 2nd experiment, given success in the 1st:

$$\text{odds}_1 = \frac{P(Y=1|X=1)}{P(Y=0|X=1)} = \frac{\pi_{11}/\pi_{1\bullet}}{\pi_{10}/\pi_{1\bullet}} = \frac{\pi_{11}}{\pi_{10}}$$

- Odds in favour of success in the 2nd experiment, given failure in the 1st:

$$\text{odds}_0 = \frac{P(Y=1|X=0)}{P(Y=0|X=0)} = \frac{\pi_{01}/\pi_{0\bullet}}{\pi_{00}/\pi_{0\bullet}} = \frac{\pi_{01}}{\pi_{00}}$$

- The odds ratio is defined as

$$\theta = \frac{\text{odds}_1}{\text{odds}_0} = \frac{\pi_{11} \cdot \pi_{00}}{\pi_{10} \cdot \pi_{01}}$$

13.2 The Odds Ratio

Properties of the odds ratio.

- It is a symmetric measure: Interchanging X and Y won't change its value.
- It is always a positive number.
- In words, the odds ratio means: "How many times more probable is a success in the 2$^{\text{nd}}$ experiment, if there was also a success in the 1$^{\text{st}}$ experiment?"
- In the case of independence: $\theta = 1$

13.2 The Odds Ratio

Computation of the odds ratio.

- Suppose we are given a 2×2 contingency table:

		Y		
		1	0	
X	1	n_{11}	n_{10}	$n_{1\bullet}$
	0	n_{01}	n_{00}	$n_{0\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 0}$	n

- Then the (empirical) odds ratio can be computed as:

$$\theta = \frac{n_{11}n_{00}}{n_{10}n_{01}}$$

13.2 The Odds Ratio

Example: Who buys eggs from our supermarket?

- The contingency table was:

		Y		
		yes	no	
X	f	49	317	366
	m	14	128	142
		63	445	508

- Odds ratio:

$$\theta = \frac{49 \cdot 128}{317 \cdot 14} = 1.41$$

- Thus: It is about 1.4 times as likely that we'll sell eggs to a female customer than to a male customer.

13.2 The Odds Ratio

Example: Gender and place of residence of Bilgi students.

- A fictitious contingency table:

		place of residence		
		Asia	Europe	
sex	f	12	48	60
	m	8	32	40
		20	80	100

- Odds ratio:

$$\theta = \frac{12 \cdot 32}{48 \cdot 8} = 1$$

- Knowledge of one variable does not contribute any information about the other.

13.2 The Odds Ratio

The odds ratio in data mining.

- Data mining is concerned with finding patterns in large data sets.
- "Patterns": relations between variables; regularities.
- Purpose: Obtain results useful for business.
- Another keyword: KDD = knowledge discovery in databases.
- The odds ratio is an important tool in data mining.
  (It can detect relations between pairs of categorical variables.)

13.2 The Odds Ratio

Example: Buying patterns of supermarket customers.

- Our example dataset has nine groups of items:

meat	cheese	bread
fruit	dairy	snacks
eggs	tea/coffee	drinks

(This is a real-world dataset. However, in statistical consulting datasets are usually much bigger.)

- Can we identify items which are often bought together?

13.2 The Odds Ratio

Example: Buying patterns of supermarket customers.

Raw data:

no.	sex	total	meat	cheese	bread	fruit	dairy	snacks	eggs	tea/coff	drinks
1	m	10.07	0	0	0	1	0	0	0	0	1
2	f	22.61	1	0	0	1	0	0	0	0	0
3	f	14.48	0	0	0	1	0	0	0	0	0
4	f	17.41	0	1	0	0	0	0	0	0	1
5	m	6.46	0	0	0	0	0	1	0	0	0
6	m	11.43	1	0	1	0	0	0	0	0	0
7	f	8.89	1	0	0	1	0	0	0	0	0
8	f	0.59	0	0	1	0	0	0	0	0	0
9	f	6.69	0	0	0	1	1	1	0	0	0
10	m	49.83	0	0	0	1	1	0	0	1	0
11	f	32.98	0	1	0	1	0	0	0	0	0
12	f	12.29	0	0	1	1	1	0	0	0	0
13	f	5.80	0	1	0	1	1	0	0	0	0
14	f	31.92	0	1	1	1	1	0	0	0	0
15	f	29.92	1	1	0	0	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

13.2 The Odds Ratio

Example: Buying patterns of supermarket customers.

- The combinations with the highest odds ratios:

prod1	prod2	odds ratio
meat	cheese	3.28
fruit	dairy	2.68
cheese	fruit	2.52
snacks	eggs	2.46
dairy	eggs	2.41
snacks	tea.coff	2.21
cheese	dairy	2.18

- How can we use this for managing our supermarket?

13.2 The Odds Ratio

Statistical properties of the odds ratio.

- Statistical properties are needed for statistical inference!
- Let X and Y describe Bernoulli experiments.
- $(X_1, Y_1), \ldots, (X_n, Y_n)$: a sample of size n from (X, Y)
- Let $n_{ij} = \#$ observations with $X = i$ and $Y = j$
- The odds ratio is:

$$\theta = \frac{n_{11} n_{00}}{n_{10} n_{01}}$$

13.2 The Odds Ratio

Statistical properties of the odds ratio.

- If X and Y are independent, it holds approximately for large n that:

$$\ln \theta \;\sim\; \mathrm{N}(0, s_\theta^2) \quad \text{with} \quad s_\theta^2 = \frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}$$

- Equivalently:

$$\frac{\ln \theta}{s_\theta} \;\sim\; \mathrm{N}(0, 1) \quad \text{approximately for large } n.$$

13.2 The Odds Ratio

Testing for independence in a 2×2-table.

- Test problem:
  - H_0: X, Y are independent.
    (That is: $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$ for all $i, j = 0, 1$.)
  - H_1: X, Y are not independent.
    (That is: $\pi_{ij} \neq \pi_{i\bullet} \cdot \pi_{\bullet j}$ for some $i, j = 0, 1$.)

- Test statistic:

$$T = \frac{\ln \theta}{s_\theta} \quad \text{with} \quad s_\theta = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}$$

  If H_0 is true, $T \sim \mathsf{N}(0, 1)$ approximately for large n.

- Critical for H_0: Too small and too large values of T.

13.2 The Odds Ratio

Example: Who buys meat from our supermarket?

- Is meat-purchasing gender-specific?
- Contingency table from our dataset:

	yes	no
f	60	306
m	25	117

- Then:

$$\theta = \frac{60 \cdot 117}{25 \cdot 306} = 0.918, \quad s_\theta^2 = 0.0685, \quad T = \frac{\ln \theta}{s_\theta} = -0.33$$

- Conclusions?

13.3 Testing Independence With χ^2

Example: Who buys eggs from our supermarket?

- Are X and Y independent?
- We can also ask in the opposite way: Given the marginal frequencies of X und Y, which contingency table would we expect if X and Y were indeed independent?

		Y		
		yes	no	
X	f	?	?	366
	m	?	?	142
		63	445	508

All frequencies are relevant!

13.3 Testing Independence With χ^2

Example: Who buys eggs from our supermarket?

- Are X and Y independent?
- If X and Y are independent, we expect a contingency table with. . .
  - identical relative frequencies in each row
    (in particular, the same as the marginal distribution of Y!),
  - identical relative frequencies in each column
    (in particular, the same as the marginal distribution of X!).

13.3 Testing Independence With χ^2

Example: Who buys eggs from our supermarket?

Are X and Y independent? If so, we expect to observe a contingency table with frequencies

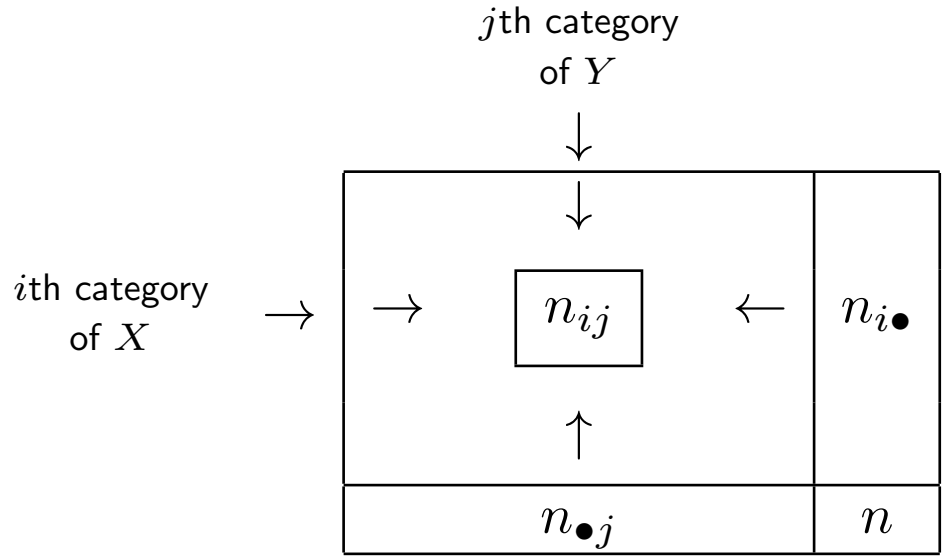
		Y		
		yes	no	
X	f	$\frac{63 \cdot 366}{508} \approx 45$	$\frac{445 \cdot 366}{508} \approx 321$	366
	m	$\frac{63 \cdot 142}{508} \approx 18$	$\frac{445 \cdot 142}{508} \approx 124$	142
		63	445	508

- This is not the same as our observed contingency table.
- On the other hand: The observed contingency table is based on a $random$ sample.

13.3 Testing Independence With χ^2

Are X and Y independent?

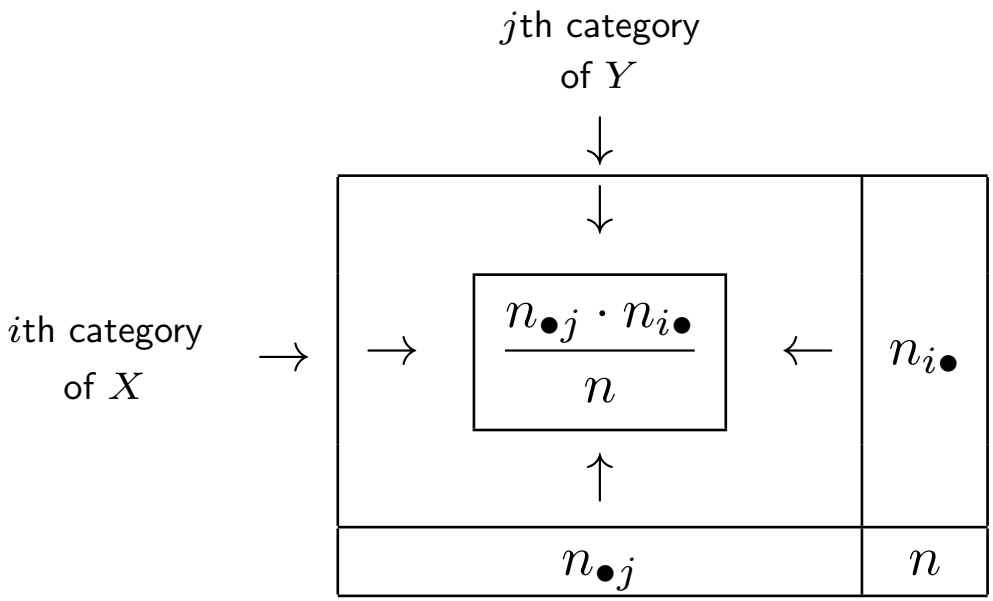
Observed frequencies:



13.3 Testing Independence With χ^2

Are X and Y independent?

Expected frequencies in case of independence:



13.3 Testing Independence With χ^2

Are X and Y independent?

- In case of independence, the distance between
  - the observed contingency table and
  - the contingency table computed assuming independence

  should be "small".

- The χ^2 statistic measures this distance:

$$\chi^2 = \sum \frac{\left(n_{ij} - \frac{n_{\bullet j} \cdot n_{i\bullet}}{n}\right)^2}{\frac{n_{\bullet j} \cdot n_{i\bullet}}{n}}$$

13.3 Testing Independence With χ^2

Are X and Y independent?

- In case of independence, the χ^2 statistic is approximately χ^2 distributed with $(r-1)(c-1)$ degrees of freedom.

  Here, r and c designate the number of rows (columns, resp.) of the contingency table. None of the expected frequencies should be smaller than 5.

- When will the null hypothesis of independence be rejected? — Critical: the too large values of χ^2!

  "Too large" means: exceeding the 95% quantile of the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

13.3 Testing Independence With χ^2

Testing for independence in an $r \times c$-table.

- Test problem:
  - H_0: X, Y are independent.
    (That is: $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$ for all i, j.)
  - H_1: X, Y are not independent.
    (That is: $\pi_{ij} \neq \pi_{i\bullet} \cdot \pi_{\bullet j}$ for some i, j.)
- Test statistic:

$$\chi^2 = \sum \frac{\left(n_{ij} - \frac{n_{\bullet j} \cdot n_{i\bullet}}{n}\right)^2}{\frac{n_{\bullet j} \cdot n_{i\bullet}}{n}}$$

  If H_0 is true, $\chi^2 \sim \chi^2_n$.
- Critical for H_0: Too large values of χ^2.

13.3 Testing Independence With χ^2

Example: direct marketing.

- The age of a person is important in identifying the target group.
- A usual practice is to estimate a person's age using the person's given name.
- Are there regional differences in people's given names?
- An example from Germany: A random sample of 3643 men named *Franz* were classified with respect to age and region (north/south Germany). The result:

	age group			
	18–40	40–60	60+	
north	312	517	298	1 127
south	690	1 027	799	2 516
	1 002	1 544	1 097	3 643

Is there evidence for a regional difference?

13.3 Testing Independence With χ^2

Example: direct marketing.

- Are the variables $X =$ region and $Y =$ age group independent?
- In case of independence, the expected frequencies are:

	age group			
	18–40	40–60	60+	
north	310.0	477.7	339.4	1127
south	692.0	1066.3	757.6	2516
	1002	1544	1097	3643

13.3 Testing Independence With χ^2

Example: direct marketing.

- Are the variables $X =$ region and $Y =$ age group independent?
- The value of the χ^2 statistic is

$$\begin{aligned}\chi^2 &= \frac{(312-310.0)^2}{310.0} + \frac{(517-477.7)^2}{477.7} + \frac{(298-339.4)^2}{339.4} \\ &+ \frac{(690-692.0)^2}{692.0} + \frac{(1027-1066.3)^2}{1066.3} + \frac{(799-757.6)^2}{757.6} \\ &= 12.01.\end{aligned}$$

- The value of the 95% quantile of the χ^2 distribution with $(2-1)(3-1) = 2$ degrees of freedom is 5.99.
- The null hypothesis of independence will be rejected.
- What does this imply for our direct marketing efforts?