

Bus 274: Further Statistics For Business

Harald Schmidbauer



Chapter 13:

Contingency Tables and Independence



13.1 Contingency Tables

Example: 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: 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: 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: 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 and 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!



13.1 Contingency Tables

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 $X!$).



13.1 Contingency Tables

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

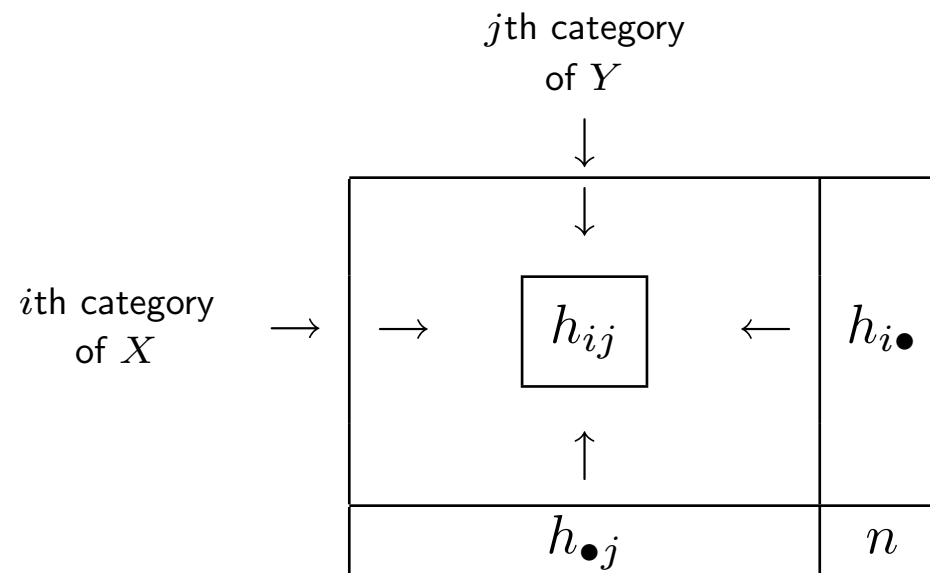
- This is not the same as our observed contingency table.
- On the other hand: The observed contingency table is based on a *random* sample.



13.2 Testing Independence With χ^2

Are X and Y independent?

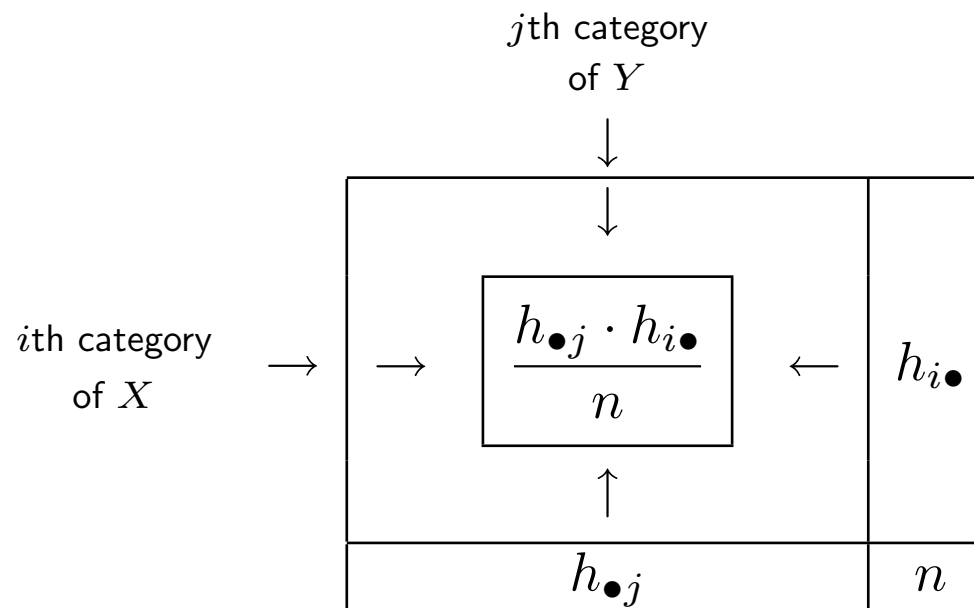
Observed frequencies:



13.2 Testing Independence With χ^2

Are X and Y independent?

Expected frequencies in case of independence:



13.2 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 independenceshould be “small”.
- The χ^2 statistic measures this distance:

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



13.2 Testing Independence With χ^2

Are X and Y independent?

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

Here, l and m 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 $(l - 1)(m - 1)$ degrees of freedom.



13.2 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?



13.2 Testing Independence With χ^2

Example: direct marketing.

- Are the variables $X = \text{region}$ and $Y = \text{age group}$ 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



13.2 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?

