Department of Statistics, University of Pretoria

ANALYSIS OF CATEGORICAL DATA

Compiled by Dr. EM Louw

Amended by SM Millard

STATISTICS 310

2005

Contents

ANALYSIS OF CATEGORICAL DATA

I TESTS FOR CONTINGENCY TABLES

Modern Statistics in Practice (Steyn, Smit, du Toit, Strasheim): p 547-564

ONE-WAY TABLE	Frequency distribution			
	PURPOSE	NAME OF TEST	TEST PROCEDURE	EXAMPLE
($1 \times k$) table	Can the frequency distribution be reconciled with an assumed theoretical distribution?	Chi-square goodness-of-fit test	Table 15.6 p 551	Ex 15.2 p 551
TWO-WAY TABLE	**One sample classified according to two characteristics**			
	PURPOSE	NAME OF TEST	TEST PROCEDURE	EXAMPLE
(2×2) table Table 15.8	Is there a relationship between the two factors?	2×2 independence test	Table 15.9 p 554	Ex 15.3 p 554
Table 15.11	Do the two characteristics occur with equal probability?	Mc Nemar test	Table 15.12 p 557	Ex 15.4 p 558
($I \times J$) table Table 15.13	Is there a relationship between factor A (I levels) and factor B (J levels)?	$I \times J$ independence test	Table 15.14 p 560	Ex 15.5 p 560
TWO-WAY TABLE	**J independent samples, each classified according to one characteristic with I levels**			
	PURPOSE	NAME OF TEST	TEST PROCEDURE	EXAMPLE
($I \times J$) table column totals known	Are the samples homogeneous with regard to this characteristic?	Test for homogeneity of samples	Table 15.14 p 560	Ex 15.6 p 563
special case: **($2 \times J$) table**	Test whether J proportions or percentages differ	Test for difference between proportions		

II STATISTICAL MODELLING OF CATEGORICAL DATA

Statistical modelling of categorical data (Crowther, N A S and Joubert, H M) HSRC report WS-41

1 LOGLINEAR MODELLING

1.1 Introductory example

1.1.1 The Cross Tabulation of Data

EXAMPLE

Modern Statistics in Practice (Steyn, Smit, du Toit, Strasheim): exercise 7 p 578

There is a suspicion that the academic achievements of house committee members of women's residences are unsatisfactory because they devote too much time to residence duties. 96 senior female students were involved in an investigation into the matter. The following **two-way contingency table** contains a classification of the academic achievements of the women who served on house committees and those who did not serve on house committees.

	HOUSE COMMITTEE MEMBER		
ACHIEVEMENT	Yes	No	Total
Good	8	14	22
Average	14	25	39
Poor	15	20	35
Total	37	59	96

The three levels of ACHIEVEMENT are known as the different **rows** of the contingency table, while the two levels of HOUSE COMMITTEE MEMBER specify the **columns** of the contingency table.

The intersections of these rows and columns are referred to as the various **cells** of the contingency table and contain the frequency of occurrence. This table is called a 3×2 contingency table.

The table indicates that a total of 96 women were encountered in this study. An interpretation of the **marginal frequencies** of HOUSE COMMITTEE MEMBER reveals that 37 women served on house committees, while 59 women did not serve on house committees. Given that a women has a poor achievement, the **partial frequencies** of HOUSE COMMITTEE MEMBER are 15 and 20 respectively.

1.1.2 Measures of Association

Consider the following SAS program and output to investigate the relationship between HOUSE COMMITTEE MEMBER and ACHIEVEMENT.

SAS PROGRAM

```
title1 'Academic achievements of house committee members';

proc format;
value aa 1='Good'
         2='Average'
         3='Poor';
value bb 1='Yes'
         2='No';

data duties;
input achieve member number @@;
label achieve='academic achievement';
label member='house committee member';
cards;
1 1 8      1 2 14
2 1 14     2 2 25
3 1 15     3 2 20
;

proc freq data=duties;
weight number;
tables achieve*member / chisq expected cellchisq;
format achieve aa. member bb.;
title3 'Chisquare test for independence';

run;
```

SAS OUTPUT

Chisquare test for independence:TABLE OF ACHIEVE BY MEMBER

ACHIEVE(academic achievement) Frequency / Expected / Cell Chi-Square / Percent / Row Pct / Col Pct	MEMBER(house committee member) Yes	No	Total
Good	8	14	22
	8.4792	13.521	
	0.0271	0.017	
	8.33	14.58	22.92
	36.36	63.64	
	21.62	23.73	
Average	14	25	39
	15.031	23.969	
	0.0708	0.0444	
	14.58	26.04	40.63
	35.90	64.10	
	37.84	42.37	
Poor	15	20	35
	13.49	21.51	
	0.1691	0.1061	
	15.63	20.83	36.46
	42.86	57.14	
	40.54	33.90	
Total	37	59	96
	38.54	61.46	100.00

Statistic	DF	Value	Prob
Chi-Square	2	0.434	0.805
Likelihood Ratio Chi-Square	2	0.432	0.806
Mantel-Haenszel Chi-Square	1	0.299	0.584
Phi Coefficient		0.067	
Contingency Coefficient		0.067	

The interpretation of the SAS OUTPUT follows.

Column Percentages

The specific relationship between HOUSE COMMITTEE MEMBER and ACHIEVEMENT may be described by means of the **column percentages**. From the frequency table in the output it is evident that 36.46% of the women have a poor achievement. If we investigate the **partial column percentages** it is possible to observe that a house committee member has a higher chance to perform poor (40.54%) than a woman who is not on the house committee (33.9%). The **row percentages** may be interpreted in a similar fashion.

The Pearson χ^2-test

Consider the following hypothesis:

H_0 : The two variables are independent ($i.e.$ there is no relationship between them)

H_1 : The two variables are related ($i.e.$ dependent)

Under the null hypothesis the cell frequencies in each cell would proportionally reflect the marginal frequencies. It is therefore expected under H_0 that the column percentages would be 22.92%, 40.63% and 36.46% for the two levels of HOUSE COMMITTEE MEMBER. It is also expected under H_0 that the composition of 38.54% women who serve on house committees and 61.46% women who do not serve on house committees is maintained over the three levels of ACHIEVEMENT.

Under this assumption the **expected frequency**, e, for each cell is as follows:

$$e = \frac{\text{row total} \times \text{column total}}{\text{overall total}}.$$

Significant deviations of the **observed frequencies** f from the expected frequencies e reflect a relationship between the two variables. The $degree\ of\ deviation$ may be determined by the **cell χ^2-value**. A cell χ^2-value that exceeds the value of 3.84 (the tabulated $\chi^2_{1,0.05}$-value) indicates a significant difference between the observed and the expected frequency on the 5% level of significance. The expected frequency, as well as the cell χ^2-value, for each cell are reported in the output as the second and third value in each cell of the table.

From the output it is clear that it is expected that 13.49 of the possible 37 women who serve on house committees, would perform poor. This is contrasted by the observed frequency of 15 women, indicating no possible association between the two variables.

According to the corresponding cell χ^2-value

$$\frac{(f-e)^2}{e} = \frac{(15-13.49)^2}{13.49} = 0.169$$

there is not a significant deviation between the observed frequency of 15 (f) and the expected frequency of 13.49 (e). Note that $0.169 < 3.48$.

In two-way contingency tables with multinomial sampling, the **null hypothesis of statistical independence could be evaluated by Pearsons' χ^2-statistic**

$$\chi^2 = \sum \frac{(f-e)^2}{e}.$$

The larger the value of this statistic, the larger the cell χ^2-values and the more evidence there is against the null hypothesis. The decision rule is to reject the null hypothesis at a 5% level of significance if the exceedance probability (p-value) is less than 0.05 .

The results of Pearsons' χ^2-test are generated by **PROC FREQ** in SAS by using the options CHISQ, EXPECTED and CELLCHISQ in the TABLES statement.

From the output we see that the value for Pearsons' χ^2-statistic is 0.805 with a corresponding p-value of 0.434. This indicates no significant relationship between HOUSE COMMITTEE MEMBER and ACHIEVEMENT.

1.1.3 Statistical Modelling

Now we are going to focus on the statistical modelling of the **structure of the contingency table**. The so-called **loglinear model** models the *natural log of the expected frequency for a specific cell* in the contingency table. In other words, it models $ln(e_{ij})$, the natural log of the expected frequency of cell (i,j) in the contingency table.

The Loglinear Model

The underlying structural relationship between ACADEMIC ACHIEVEMENT (with 3 categories) and HOUSE COMMITTEE MEMBER (with 2 categories) may be explained in terms of the **saturated** loglinear model

$$ln(e_{ij}) = \mu + \lambda_i^A + \lambda_j^H + \lambda_{ij}^{AH} \qquad \text{for} \qquad i = 1, 2, 3 \qquad \text{and} \qquad j = 1, 2$$

where

- μ : **overall mean effect** of the categories
- λ_i^A : effect of i^{th} **category of ACHIEVEMENT**
- λ_j^H : effect of j^{th} **category of HOUSE COMMITTEE MEMBER**
- λ_{ij}^{AH} : **interaction** effect between

  i^{th} category of ACHIEVEMENT and j^{th} category of HOUSE COMMITTEE MEMBER

Strictly speaking, μ is the **natural log of the geometric mean** of all the *expected cell frequencies in the table.* The more the λ-parameters deviates from zero, the higher the effect of the levels of ACHIEVEMENT and HOUSE COMMITTEE MEMBER.

The results of this loglinear analysis is generated by **PROC CATMOD** in SAS. Consider the following SAS program and output for the example:

SAS PROGRAM

```
options nodate linesize=64 pagesize=250;

title1 'Academic achievements of house committee members';

proc format;
value aa 1='Good'
         2='Average'
         3='Poor';
value bb 1='Yes'
         2='No';

data duties;
input achieve member number @@;
label achieve='academic achievement';
label member='house committee member';
cards;
1 1 8      1 2 14
2 1 14     2 2 25
3 1 15     3 2 20
;

proc catmod data=duties;
weight number;
model achieve*member = _response_ / ml nogls pred=freq noprofile;
loglin achieve member achieve*member;
format achieve aa. member bb.;
title3 'Loglinear model: saturated model is fitted';

proc catmod data=duties;
weight number;
model achieve*member = _response_ / ml nogls pred=freq noprofile;
loglin achieve member;
format achieve aa. member bb.;
title3 'Loglinear model: independence model is fitted';

run;
```

SAS OUTPUT

Academic achievements of house committee members

Loglinear model: saturated model is fitted

CATMOD PROCEDURE

Response: ACHIEVE*MEMBER	Response Levels (R)= 6
Weight Variable: NUMBER	Populations (S)= 1
Data Set: DUTIES	Total Frequency (N)= 96
	Observations (Obs)= 6

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
ACHIEVE	2	4.57	0.1016
MEMBER	1	4.80	0.0285
ACHIEVE*MEMBER	2	0.43	0.8052
LIKELIHOOD RATIO	0	.	.

****

Academic achievements of house committee members

Loglinear model: independence model is fitted

CATMOD PROCEDURE

Response: ACHIEVE*MEMBER	Response Levels (R)= 6
Weight Variable: NUMBER	Populations (S)= 1
Data Set: DUTIES	Total Frequency (N)= 96
	Observations (Obs)= 6

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
ACHIEVE	2	4.83	0.0896
MEMBER	1	4.95	0.0261
LIKELIHOOD RATIO	2	0.43	0.8056

RESPONSE_ MATRIX

	1	2	3
---	x1A	x2A	x1H
1	1	0	1
2	1	0	-1
3	0	1	1
4	0	1	-1
5	-1	-1	1
6	-1	-1	-1

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
RESPONSE	1	-0.3456	0.1619	4.55	0.0328
	2	0.2269	0.1401	2.62	0.1052
	3	-0.2333	0.1049	4.95	0.0261

NOTE: _RESPONSE_ = ACHIEVE MEMBER

ML PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES

		-----Observed-----		-----Predicted----		
Sample	Function Number	Function	Standard Error	Function	Standard Error	Residual
1	1	-0.91629	0.41833	-0.93093	0.343514	0.014634
	2	-0.35667	0.348466	-0.46431	0.272077	0.107631
	3	-0.35667	0.348466	-0.35841	0.313351	0.001731
	4	0.223144	0.3	0.108214	0.232836	0.11493
	5	-0.28768	0.341565	-0.46662	0.209705	0.178937
g,yes	F1	8	2.708013	8.479167	1.926996	-0.47917
g,no	F2	14	3.458082	13.52083	2.756737	0.479167
a,yes	F3	14	3.458082	15.03125	2.681919	-1.03125
a,no	F4	25	4.299952	23.96875	3.535439	1.03125
p,yes	F5	15	3.557562	13.48958	2.515188	1.510417
p,no	F6	20	3.979112	21.51042	3.379755	-1.51042

The interpretation of the SAS OUTPUT follows.

From the MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE in the output, it follows that this **saturated** loglinear model does not fit the data (a p-value of 0.8052 for the interaction term ACHIEVE*MEMBER), and therefore the loglinear **independence** model

$$ln(e_{ij}) = \mu + \lambda_i^A + \lambda_j^H \qquad \text{for} \qquad i = 1,2,3 \qquad \text{and} \qquad j = 1,2$$

is fitted where $\sum_{i=1}^{3} \lambda_i^A = 0$ and $\sum_{j=1}^{2} \lambda_j^H = 0$.

The estimation of the parameters of the loglinear independence model follows:

μ is estimated by

$$\begin{aligned}\hat{\mu} &= ln(\text{ geometric mean of the } \textbf{predicted frequencies})\\ &= ln(8.4792 \times 13.521 \times 15.031 \times 23.969 \times 13.49 \times 21.51)^{\frac{1}{6}}\\ &= ln(15.1278)\\ &= 2.7165 \quad .\end{aligned}$$

Note that the $\frac{rowtotal \times columntotal}{grandtotal}$ frequencies are given as $predicted\ frequencies\ F1, F2, \ldots, F6$ in the table named ML PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES in the output.

$\lambda_1^A, \lambda_2^A, \lambda_3^A, \lambda_1^H$ and λ_2^H are estimated by means of the regression coefficients of three dummy variables x_1^A, x_2^A and x_1^H in a regression.

The loglinear independence model for this example may be written as a GLM as follows:

$$\boxed{ln(e_{ij}) = \beta_0 + \beta_1^A x_1^A + \beta_2^A x_2^A + \beta_1^H x_1^H}$$

Note that $\beta_0 = \mu$. The data for the GLM analysis is in the output as the RESPONSE MATRIX.

From the ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES in the output, the regression coefficients of the GLM are

$$\begin{aligned}\hat{\beta}_1^A &= -0.3456\\ \hat{\beta}_2^A &= 0.2269\\ \hat{\beta}_1^H &= -0.2333 \quad .\end{aligned}$$

Note that $\hat{\beta}_0 = \hat{\mu} = 2.7165$ is not given in this output and must be calculated.

You will note that the parameter estimate of the **last category** of ACHIEVEMENT is omitted from the computer output. This is due to the fact that the λ^A-parameters for the variable ACHIEVEMENT add to zero, that is $\sum_{i=1}^{3} \lambda_i^A = 0$ and therefore
$\lambda_3^A = -(\lambda_1^A + \lambda_2^A) = -(-0.3456 + 0.2269) = 0.1187.$

Similarly the parameter estimate of the **last category** of HOUSE COMMITTEE MEMBER is omitted from the computer output and $\sum_{j=1}^{2} \lambda_j^H = 0$ so that $\lambda_2^H = -(\lambda_1^H) = -(-0.2333) = 0.2333.$

The **parameters of the loglinear model** are summarized in the following table.

Effect	Level	**Parameter**	Estimate
Overall mean effect		μ	2.7165
ACHIEVEMENT	Good	λ_1^A	-0.3456
	Average	λ_2^A	0.2269
	Poor	λ_3^A	0.1187
HOUSE COMMITTEE MEMBER	Yes	λ_1^H	-0.2333
	No	λ_2^H	0.2333

Note that the sum of parameters over the same subscript is equal to 0.

To investigate the specific effect of the various levels of ACHIEVEMENT and HOUSE COMMITTEE MEMBER, it is necessary to interpret the estimates of the parameters.

According to this table there is a **strong negative** estimate for the category GOOD ACHIEVEMENT. Note that $\lambda_1^A = -0.3456$ with an exceedance probability of 0.0328 (p-value $=0.0328$ in output). This implies that the *ln(geometric mean of the predicted frequencies* of **all the cells within the GOOD ACHIEVEMENT category**) is significantly lower than the *ln(geometric mean of the predicted frequencies* of **all the cells in the table**).

Note that the *ln(geometric mean of the predicted frequencies* of **all the cells within the GOOD ACHIEVEMENT category**) is

$$\widehat{ln(e_{1.})} = 2.7165 - 0.3456 = 2.3709$$

which is significantly lower than 2.7165, the *ln(geometric mean of the predicted frequencies* of **all the cells in the table**).

Due to the fact that we are currently working in the **natural log-scale**, the interpretation of results is a bit inconvenient.

The Loglinear Model in terms of Indices

By taking the *anti-log of the loglinear model*, the model becomes

$$\begin{aligned} e_{ij} &= e^{\mu+\lambda_i^A+\lambda_j^H} \\ &= e^{\mu} \times e^{\lambda_i^A} \times e^{\lambda_j^H} \\ &= i \times i_i^A \times i_j^H \quad . \end{aligned}$$

The loglinear model in terms of indices is

$$e_{ij} = i \times i_i^A \times i_j^H \qquad \text{for} \qquad i = 1, 2, 3 \qquad \text{and} \qquad j = 1, 2$$

where

i : the **overall mean effect** $= e^{\mu}$
(the geometric mean of the predicted frequencies of all the cells in the table)
i_i^A : an index for the i^{th} **category of ACHIEVEMENT** $= e^{\lambda_i^A}$
i_j^H : an index for the j^{th} **category of HOUSE COMMITTEE MEMBER** $= e^{\lambda_j^H}$.

The following table shows the **indices for the different levels** of ACHIEVEMENT and HOUSE COMMITTEE MEMBER.

Effect	Level	**Index**	
Overall mean effect		i	15.1278
ACHIEVEMENT	Good	i_1^A	0.7078
	Average	i_2^A	1.2547
	Poor	i_3^A	1.1260
HOUSE COMMITTEE MEMBER	Yes	i_1^H	0.7919
	No	i_2^H	1.2628

Note that products of indices over the same subscript are equal to 1. The interpretation of the indices in the loglinear model follows.

The **overall mean effect** is the *geometric mean of the predicted frequencies* of **all the cells in the table**. This value of 15.1278 may be verified from

$$\begin{aligned} & \text{geometric mean of the } \textbf{predicted frequencies} \quad F1, F2, \ldots, F6 \\ = \; & (8.4792 \times 13.521 \times 15.031 \times 23.969 \times 13.49 \times 21.51)^{\frac{1}{6}} \\ = \; & 15.1278 \quad . \end{aligned}$$

The index of 0.7078 for the GOOD ACHIEVEMENT category implies that the *geometric mean of the predicted frequencies* of **all the cells within the GOOD ACHIEVEMENT category** is 29.22% lower than 15.1278, the *geometric mean of the predicted frequencies* of **all the cells in the table**.

The geometric mean of the predicted frequencies of all the cells within the GOOD ACHIEVEMENT

category is

$$\begin{aligned}\hat{e}_{1.} &= i \times i_1^A \\ &= 15.1278 \times 0.7078 \\ &= 10.71 \quad .\end{aligned}$$

The index of 1.2628 for the NO category of HOUSE COMMITTEE MEMBER implies that the *geometric mean of the predicted frequencies* of **all the cells within the NO category of HOUSE COMMITTEE MEMBER** is 26.28% higher than 15.1278, the *geometric mean of the predicted frequencies* of **all the cells in the table**.

The geometric mean of the predicted frequencies of all the cells within this NO category is

$$\begin{aligned}\hat{e}_{.2} &= i \times i_2^H \\ &= 15.1278 \times 1.2628 \\ &= 19.10 \quad .\end{aligned}$$

The *expected cell frequency* for female students **with a good achievement and who are members of the house committee** is

$$\begin{aligned}\hat{e}_{11} &= i \times i_1^A \times i_1^H \\ &= 15.1278 \times 0.7078 \times 0.7919 \\ &= 8.4792 \quad .\end{aligned}$$

Similarly the *expected cell frequency* for female students **with a poor achievement and who are not members of house committee** is

$$\begin{aligned}\hat{e}_{32} &= i \times i_3^A \times i_2^H \\ &= 15.1278 \times 1.1260 \times 1.2628 \\ &= 21.5104 \quad .\end{aligned}$$

Note that the last two values correspond with the values of the *predicted frequencies* $F1$ and $F6$ in the table named ML PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES in the output.

1.2 Correspondence between Factorial Design and Contingency Table

In an $I \times J$ **factorial design**, the cell values are *observations of a continuous dependent variable* y on different **treatment combinations** (that is combinations of the I levels of factor A with the J levels of factor B). The **significance of factors A and B** is examined, as well as the **interaction** between the two factors. The cell values in an $I \times J$ **contingency table** are *observed frequencies* of the **factor combinations** $(A_1, B_1), (A_2, B_2), \ldots$. The **interdependence between factors A and B** is examined.

The **statistical ANOVA model with interaction** for a $I \times J$ factorial design is

$$E(y_{ij}) = \mu + \alpha_i^A + \alpha_j^B + \alpha_{ij}^{AB} \qquad \text{for} \qquad i = 1, 2, \ldots, I \qquad \text{and} \qquad j = 1, 2, \ldots, J$$

where $\sum_{i=1}^{I} \alpha_i^A = 0$, $\sum_{j=1}^{J} \alpha_j^B = 0$, $\sum_{i=1}^{I} \alpha_{ij}^{AB} = 0$ for each j and $\sum_{j=1}^{J} \alpha_{ij}^{AB} = 0$ for each i.

The parameters of the ANOVA model are

μ	=	the **overall mean** effect
α_i^A	=	the effect of i**th level of factor A**
α_j^B	=	the effect of j**th level of factor B**
α_{ij}^{AB}	=	**interaction** effect between ith level of factor A and jth level of factor B.

μ is estimated by $\overline{y}$ where

$$\overline{y} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}}{n} \quad .$$

$\alpha_i^A \; i = 1, 2, \ldots, I$, $\alpha_j^B \; j = 1, 2, \ldots, J$ and α_{ij}^{AB} are estimated by means of the **regression coefficients of** $(I-1) + (J-1) + (I-1).(J-1) = (I.J-1)$ **dummy variables** in a regression.

The **statistical ANOVA model without interaction** for a $I \times J$ factorial design is

$$E(y_{ij}) = \mu + \alpha_i^A + \alpha_j^B \qquad \text{for} \qquad i = 1, 2, \ldots, I \qquad \text{and} \qquad j = 1, 2, \ldots, J$$

where $\sum_{i=1}^{I} \alpha_i^A = 0$ and $\sum_{j=1}^{J} \alpha_j^B = 0$.

Note that the ANOVA model models the **expected** *value of the continuous variable* y *for cell (i,j)*, namely $E(y_{ij})$ in the factorial design. The so-called **loglinear model** models the **natural log of the expected** *frequency of cell (i,j)* in the contingency table, namely $ln(e_{ij})$.

1.3 The loglinear model for a two-way contingency table

1.3.1 The saturated loglinear model

The underlying structural relationship between factor A (with I categories) and factor B (with J categories) may be explained in terms of the loglinear model.

The **loglinear model** for a $I \times J$ contingency table is

$$ln(e_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad \text{for} \qquad i = 1, 2, \ldots, I \qquad \text{and} \qquad j = 1, 2, \ldots, J$$

where $\sum_{i=1}^{I} \lambda_i^A = 0$, $\sum_{j=1}^{J} \lambda_j^B = 0$, $\sum_{i=1}^{I} \lambda_{ij}^{AB} = 0$ for each j and $\sum_{j=1}^{J} \lambda_{ij}^{AB} = 0$ for each i.

The parameters of the loglinear model are

- μ = the **overall mean** effect of the categories
- λ_i^A = the effect of i**th category of factor A**
- λ_j^B = the effect of j**th category of factor B**
- λ_{ij}^{AB} = **interaction** effect between ith category of factor A and jth category of factor B.

μ is the arithmetic mean of $ln(e_{ij})$ namely

$$\begin{aligned} \overline{ln(e_{ij})} &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} ln(e_{ij})}{I \times J} \\ &= \frac{1}{I \times J} ln(\pi_{i=1}^{I}\pi_{j=1}^{J} e_{ij}) \\ &= ln(\pi_{i=1}^{I}\pi_{j=1}^{J} e_{ij})^{\frac{1}{IJ}} \\ &= ln(\text{ geometric mean of } e_{ij} \text{ 's}) \quad . \end{aligned}$$

so that

$$\mu = ln(\text{ geometric mean of the expected frequencies})$$

The number of unknown parameters in this loglinear model is $1+(I-1)+(J-1)+(I-1)(J-1) = I.J$. Since there are just as many unknown parameters in the model as cells in the contingency table, the model is known as the **saturated model**.

The saturated loglinear model always fits the data in the contingency table perfectly. Therefore the best estimator of e_{ij}, the so-called **ML-estimator** of e_{ij}, for the saturated loglinear model is

$$\text{the observed frequency } f_{ij} \quad .$$

Note that $\sum_{i=1}^{I}\sum_{j=1}^{J} e_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = n$, where n is the number of observations. Hence, in the **saturated loglinear model**, μ is estimated by

$$\boxed{\hat{\mu} = ln(\text{ geometric mean of the observed frequencies})}$$

$\lambda_i^A \; i = 1, 2, \ldots, I$, $\lambda_j^B \; j = 1, 2, \ldots, J$ and λ_{ij}^{AB} are estimated by means of the **regression coefficients of** $(I-1)+(J-1)+(I-1).(J-1) = (I.J-1)$ **dummy variables** in a regression.

1.3.2 The loglinear independence model

The **loglinear independence model** for a $I \times J$ contingency table is

$$\boxed{ln(e_{ij}) = \mu + \lambda_i^A + \lambda_j^B \qquad \text{for} \qquad i = 1, 2, \ldots, I \qquad \text{and} \qquad j = 1, 2, \ldots, J}$$

where $\sum_{i=1}^{I} \lambda_i^A = 0$ and $\sum_{j=1}^{J} \lambda_j^B = 0$.

The number of unknown parameters in the loglinear independence model is $1+(I-1)+(J-1) = (I+J)-1$.

The best estimator of e_{ij}, the so-called **ML-estimator** of e_{ij} for the loglinear independence model is

$$\frac{f_{i.} \times f_{.j}}{n} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}.$$

Hence, in the loglinear independence model, μ is estimated by

$$\boxed{\hat{\mu} = ln(\text{ geometric mean of } \frac{f_{i.} \times f_{.j}}{n} \text{ 's})}$$

$\lambda_i^A \; i = 1, 2, \ldots, I$ and $\lambda_j^B \; j = 1, 2, \ldots, J$ are estimated by means of the **regression coefficients of** $(I-1)+(J-1)$ **dummy variables** in a regression.

1.3.3 The loglinear model in terms of indices

The saturated loglinear model

$$ln(e_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

with $\sum_{i=1}^{I} \lambda_i^A = \sum_{j=1}^{J} \lambda_j^B = 0$ and $\sum_{i=1}^{I} \lambda_{ij}^{AB} = \sum_{j=1}^{J} \lambda_{ij}^{AB} = 0$ can be expressed as

$$\begin{aligned} e_{ij} &= e^{\mu+\lambda_i^A+\lambda_j^B+\lambda_{ij}^{AB}} \\ &= e^{\mu} \times e^{\lambda_i^A} \times e^{\lambda_j^B} \times e^{\lambda_{ij}^{AB}} \\ &= i \times i_i^A \times i_j^B \times i_{ij}^{AB} \quad . \end{aligned}$$

The **saturated loglinear model in terms of indices** is

$$e_{ij} = i \times i_i^A \times i_j^B \times i_{ij}^{AB} \quad \text{for} \quad i = 1, 2, \ldots, I \quad \text{and} \quad j = 1, 2, \ldots, J$$

where

i	:	an index for the **overall effect**	=	e^{μ}
		(the geometric mean frequency of all the cells in the table)		
i_i^A	:	an index for the i^{th} **category of factor A**	=	$e^{\lambda_i^A}$
i_j^B	:	an index for the j^{th} **category of factor B**	=	$e^{\lambda_j^B}$
i_{ij}^{AB}	:	an index for **cell** (i, j)	=	$e^{\lambda_{ij}^{AB}}$.

Note that from $\sum_{i=1}^{I} \lambda_i^A = 0$ it follows that

$$\begin{aligned} e^{\sum_{i=1}^{I} \lambda_i^A} &= e^0 \\ e^{\lambda_1^A}.e^{\lambda_2^A}.....e^{\lambda_I^A} &= 1 \\ i_1^A \times i_2^A \times \ldots \times i_I^A &= 1. \end{aligned}$$

Similarly $i_1^B \times i_2^B \times \ldots \times i_J^B = 1$. Therefore **products of indices over the same subscripts are equal to 1**.

These indices explain exactly the structure of the table. The index for any specific category of a factor measures how the geometric mean frequency of **all the cells within this category** increases

or decreases, compared to the geometric mean frequency of **all the cells in the contingency table**. An index greater than 1 denotes an increase in the geometric mean frequency and an index less than 1 a decrease in the geometric mean frequency.

1.4 Example where a loglinear model is fitted to a three-way table

Modern Statistics in Practice (Steyn, Smit, du Toit, Strasheim): table 15.4 p 548

Consider the following three-way $(2 \times 2 \times 2)$ contingency table of the number of unemployed black and coloured workers (in thousands) during May 1980, subdivided according to age group and gender.

AGEGROUP	COLOUREDS		BLACKS		
	male	female	male	female	total
less than 30 years	16	22	142	143	323
30 years or older	8	13	80	121	222
total	24	35	222	264	545

We want to describe the structure of this table. The underlying structural relationship among AGE (factor A with 2 categories), POPULATION GROUP (factor B with 2 categories) and GENDER (factor C with 2 categories) may be explained in terms of the saturated loglinear model

$$ln(e_{ijk}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \text{ for } i = 1, 2 \; j = 1, 2 \text{ and } k = 1, 2.$$

The SAS program and output follow:

SAS PROGRAM

```
options nodate linesize=64 pagesize=250;

title1 'Unemployed black and coloured workers,
        subdivided according to age and gender';

proc format;
value aa 1='Less than 30 years'
         2='30 years and older';
value bb 1='Coloureds'
         2='Blacks';
value cc 1='Male'
         2='Female';

data workers;
input age popgroup gender number @@;
label age='age in years';
label popgroup='population group';
cards;
1 1 1 16       1 1 2 22       1 2 1 142      1 2 2 143
2 1 1  8       2 1 2 13       2 2 1  80      2 2 2 121
;

proc freq data=workers;
weight number;
tables age*popgroup*gender;
format age aa. popgroup bb. gender cc.;
title3 'Three-way contingency table';

proc catmod data=workers;
weight number;
model age*popgroup*gender = _response_ / ml nogls pred=freq noprofile;
loglin age popgroup gender age*popgroup age*gender popgroup*gender age*popgroup*gender;
format age aa. popgroup bb. gender cc.;
title3 'Loglinear model: saturated model is fitted';

run;
```

SAS OUTPUT

Unemployed black and coloured workers, subdivided according to age and gender

Three-way contingency table

TABLE 1 OF POPGROUP BY GENDER:CONTROLLING FOR AGE=Less than 30 years

POPGROUP(population group)

Frequency Percent Row Pct Col Pct	GENDER Male	Female	Total
Coloureds	16 4.95 42.11 10.13	22 6.81 57.89 13.33	38 11.76
Blacks	142 43.96 49.82 89.87	143 44.27 50.18 86.67	285 88.24
Total	158 48.92	165 51.08	323 100.00

TABLE 2 OF POPGROUP BY GENDER:CONTROLLING FOR AGE=30 years and older

POPGROUP(population group)

Frequency Percent Row Pct Col Pct	GENDER Male	Female	Total
Coloureds	8 3.60 38.10 9.09	13 5.86 61.90 9.70	21 9.46
Blacks	80 36.04 39.80 90.91	121 54.50 60.20 90.30	201 90.54
Total	88 39.64	134 60.36	222 100.00

CATMOD PROCEDURE

Response: AGE*POPGROUP*GENDER — Response Levels (R)= 8
Weight Variable: NUMBER — Populations (S)= 1
Data Set: WORKERS — Total Frequency (N)= 545
Observations (Obs)= 8

Loglinear model: saturated model is fitted

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
AGE	1	11.15	0.0008
POPGROUP	1	214.06	0.0000
GENDER	1	4.35	0.0369
AGE*POPGROUP	1	0.66	0.4151
AGE*GENDER	1	0.96	0.3283
POPGROUP*GENDER	1	0.43	0.5139
AGE*POPGROUP*GENDER	1	0.17	0.6830
LIKELIHOOD RATIO	0	.	.

From the MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE in the output, it follows that this saturated loglinear model does not fit the data (a p-value of 0.683 for the interaction term AGE*POPGROUP*GENDER is found that is not significant). Therefore this term is deleted from the model and the SAS program is rerunned after deleting this highest order non-significant interaction.

Consider the following SAS program and output.

SAS PROGRAM

```
title1 'Unemployed black and coloured workers,
        subdivided according to age and gender';
proc format;
value aa 1='Less than 30 years'
         2='30 years and older';
value bb 1='Coloureds'
         2='Blacks';
value cc 1='Male'
         2='Female';
data workers;
input age popgroup gender number @@;
label age='age in years';
label popgroup='population group';
cards;
1 1 1 16       1 1 2 22       1 2 1 142      1 2 2 143
2 1 1  8       2 1 2 13       2 2 1  80      2 2 2 121
;

proc catmod data=workers;
weight number;
model age*popgroup*gender = _response_ / ml nogls pred=freq noprofile;
loglin age popgroup gender age*popgroup age*gender popgroup*gender;
format age aa. popgroup bb. gender cc.;
title3 'Loglinear model: fits model with three first order interactions';

proc catmod data=workers;
weight number;
model age*popgroup*gender = _response_ / ml nogls pred=freq noprofile;
loglin age popgroup gender age*popgroup age*gender;
format age aa. popgroup bb. gender cc.;
title3 'Loglinear model: fits model with two first order interactions';

proc catmod data=workers;
weight number;
model age*popgroup*gender = _response_ / ml nogls pred=freq noprofile;
loglin age popgroup gender age*gender;
format age aa. popgroup bb. gender cc.;
title3 'Loglinear model: fits model with one first order interaction';
run;
```

SAS OUTPUT

```
 Loglinear model: fits model with three first order interactions
         MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE
       Source                    DF   Chi-Square      Prob
       -------------------------------------------------
       AGE                        1        11.96    0.0005
       POPGROUP                   1       216.68    0.0000
       GENDER                     1         5.10    0.0239
       AGE*POPGROUP               1         0.84    0.3581
       AGE*GENDER                 1         4.68    0.0305
       POPGROUP*GENDER            1         0.65    0.4186

       LIKELIHOOD RATIO           1         0.17    0.6839
*********************************************************************
    Loglinear model: fits model with two first order interactions
         MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE
       Source                    DF   Chi-Square      Prob
       -------------------------------------------------
       AGE                        1        11.67    0.0006
       POPGROUP                   1       221.62    0.0000
       GENDER                     1         6.89    0.0087
       AGE*POPGROUP               1         0.72    0.3956
       AGE*GENDER                 1         4.56    0.0328

       LIKELIHOOD RATIO           2         0.83    0.6615
*********************************************************************

    Loglinear model: fits model with one first order interaction
                        CATMOD PROCEDURE
Response: AGE*POPGROUP*GENDER          Response Levels (R)=     8
Weight Variable: NUMBER                Populations     (S)=     1
Data Set: WORKERS                      Total Frequency (N)=   545
                                       Observations  (Obs)=     8

                       _RESPONSE_ MATRIX
                       1        2        3        4
    dummies ----------x1A------x1B------x1C-----x11AC---
              1        1        1        1        1
              2        1        1       -1       -1
              3        1       -1        1        1
              4        1       -1       -1       -1
              5       -1        1        1       -1
```

6	-1	1	-1	1
7	-1	-1	1	-1
8	-1	-1	-1	1

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
AGE	1	20.16	0.0000
POPGROUP	1	233.94	0.0000
GENDER	1	6.89	0.0087
AGE*GENDER	1	4.56	0.0328
LIKELIHOOD RATIO	3	1.56	0.6684

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
RESPONSE	1	0.1983	0.0442	20.16	0.0000
	2	-1.0543	0.0689	233.94	0.0000
	3	-0.1160	0.0442	6.89	0.0087
	4	0.0943	0.0442	4.56	0.0328

NOTE: _RESPONSE_ = AGE POPGROUP GENDER AGE*GENDER

ML PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES

		-----Observed-----		-----Predicted----		
Sample	Function Number	Function	Standard Error	Function	Standard Error	Residual
1	1	-2.0232	0.266016	-1.94392	0.181104	-0.07929
	2	-1.70475	0.231774	-1.90057	0.180361	0.195817
	3	0.160037	0.12372	0.164755	0.117438	-0.00472
	4	0.167054	0.123521	0.208106	0.11629	-0.04105
	5	-2.71635	0.365054	-2.52917	0.194507	-0.18717
	6	-2.23084	0.291869	-2.10867	0.137865	-0.12217
	7	-0.41376	0.144099	-0.4205	0.137209	0.006739
<30,C,M	F1	16	3.940847	17.10459	2.395161	-1.10459
<30,C,F	F2	22	4.594772	17.86239	2.484091	4.137615
<30,B,M	F3	142	10.24704	140.8954	9.676758	1.104587

```
<30,B,F   F4        143  10.27029  147.1376  9.813642  -4.13761
>=30,C,M  F5          8  2.807591  9.526606  1.495496  -1.52661
>=30,C,F  F6         13   3.56229  14.50642   2.08923  -1.50642
>=30,B,M  F7         80  8.261772  78.47339  7.749232  1.526606
>=30,B,F  F8        121  9.702359  119.4936   9.13995  1.506422
```

First a model with three first-order interaction terms is fitted. In this output, the interaction term POPGROUP*GENDER has the highest p-value of 0.4186, that again is not significant. Hence this term is now deleted from the model. In the second model's output the interaction term AGE*POPGROUP is not significant (a p-value of 0.3956). After deleting this term, the final model is reached. Note that all the terms in the MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE of the final's model output have p-values less than 0.05. Note also that the standard errors of the estimates of the parameters in the ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES in the output are small.

The **final loglinear model** is

$$\boxed{ln(e_{ijk}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} \text{ for } i = 1,2 \; j = 1,2 \text{ and } k = 1,2.}$$

Note that this final model is not a saturated model.

μ is estimated by

$$\begin{aligned} \hat{\mu} &= ln(\text{ geometric mean of predicted frequencies } \quad F1, F2, \ldots, F8) \\ &= ln(17.1 \times 17.9 \times 140.9 \times 147.1 \times 9.5 \times 14.5 \times 78.5 \times 119.5)^{\frac{1}{8}} \\ &= ln(41.1415) \\ &= 3.717 \quad . \end{aligned}$$

$\lambda_1^A, \lambda_1^B, \lambda_1^C$ and λ_{11}^{AC} are estimated by means of the regression coefficients of dummy variables x_1^A, x_1^B, x_1^C and x_{11}^{AC} in a regression.

The final loglinear model may be written as a GLM as follows:

$$\boxed{ln(e_{ijk}) = \beta_0 + \beta_1^A x_1^A + \beta_1^B x_1^B + \beta_1^C x_1^C + \beta_{11}^{AC} x_{11}^{AC}}$$

Note that $\beta_0 = \mu$.

The data for the GLM analysis is given in the output as the RESPONSE MATRIX.

From the ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES in the output, the regression

coefficients of the GLM are

$$\begin{aligned}\hat{\beta}_1^A &= 0.1983\\ \hat{\beta}_1^B &= -1.0543\\ \hat{\beta}_1^C &= -0.116\\ \hat{\beta}_{11}^{AC} &= 0.0943 \quad .\end{aligned}$$

Note that $\hat{\beta}_0 = \hat{\mu} = 3.717$ is not given in this output.

The **parameters of the loglinear model** are summarized in the following table.

Effect	Level	**Parameter**	Estimate
Overall effect		μ	3.717
AGE GROUP	Less than 30	λ_1^A	0.1983
	30 and older	λ_2^A	-0.1983
POPULATION GROUP	Coloured	λ_1^B	-1.0543
	Black	λ_2^B	1.0543
GENDER	Male	λ_1^C	-0.116
	Female	λ_2^C	0.116
AGE * GENDER	Less than 30, Male	λ_{11}^{AC}	0.0943
	Less than 30, Female	λ_{12}^{AC}	-0.0943
	30 and older, Male	λ_{21}^{AC}	-0.0943
	30 and older, Female	λ_{22}^{AC}	0.0943

Note that the sum of parameters over the same subscript is equal to 0.

According to this table there is a strong *negative* estimate for the category *coloureds*. Note that $\lambda_1^B = -1.0543$ with an exceedance probability of 0.0000 (p-value$=0.0000$ in output). This implies that the *ln(geometric mean frequency* of **all the cells within the COLOUREDS category**) is significantly **lower than the ln(geometric mean frequency of all the cells)** in the table.

Also there is a strong *positive* estimate for the category *less than 30 years, males*. Note that $\lambda_{11}^{AC} = 0.0943$ with an exceedance probability of 0.0328 (p-value$=0.0328$ in output).

By taking the anti-log of the loglinear model, the **loglinear model in terms of indices** is

$$\begin{aligned} e_{ijk} &= e^{\mu+\lambda_i^A+\lambda_j^B+\lambda_k^C+\lambda_{ik}^{AC}} \\ &= e^{\mu} \times e^{\lambda_i^A} \times e^{\lambda_j^B} \times e^{\lambda_k^C} \times e^{\lambda_{ik}^{AC}} \\ &= i \times i_i^A \times i_j^B \times i_k^C \times i_{ik}^{AC} \quad . \end{aligned}$$

The **indices of the loglinear model** are summarized in the following table.

Effect	Level	**Index**	
Overall effect		i	41.1415
AGE GROUP	Less than 30	i_1^A	1.2193
	30 and older	i_2^A	0.8201
POPULATION GROUP	Coloured	i_1^B	0.3484
	Black	i_2^B	2.87
GENDER	Male	i_1^C	0.8905
	Female	i_2^C	1.123
AGE * GENDER	Less than 30, Male	i_{11}^{AC}	1.0989
	Less than 30, Female	i_{12}^{AC}	0.91
	30 and older, Male	i_{21}^{AC}	0.91
	30 and older, Female	i_{22}^{AC}	1.0989

Note that the product of indices over the same subscript is equal to 1. The interpretation of the indices in the loglinear model follows.

The **overall index** of 41.1415 is the *geometric mean frequency* of **all the cells in the contingency table**.

The index of 1.2193 for an unemployed worker **less than 30 years** implies that the *geometric mean frequency* of **all the cells within the LESS THAN 30 category** is 22% higher than the geometric mean frequency of 41.1415 of all the cells in the table. This geometric mean frequency

is

$$\begin{aligned}\hat{e}_{1..} &= i \times i_1^A \\ &= 41.1415 \times 1.2193 \\ &= 50.1638 \quad .\end{aligned}$$

The index of 0.8905 for an unemployed **male** worker implies that the *geometric mean frequency* of **all the cells within the MALE category** is 11% lower than the geometric mean frequency of 41.1415 of all the cells in the table. This geometric mean frequency is

$$\begin{aligned}\hat{e}_{..1} &= i \times i_1^C \\ &= 41.1415 \times 0.8905 \\ &= 36.6365 \quad .\end{aligned}$$

The *geometric mean frequency* of **all the cells within the LESS THAN 30, MALE category** is

$$\begin{aligned}\hat{e}_{1.1} &= i \times i_1^A \times i_1^C \times i_{11}^{AC} \\ &= 41.1415 \times 1.2193 \times 0.8905 \times 1.0989 \\ &= 49.0888 \quad .\end{aligned}$$

The *expected cell frequency* of unemployed **black male workers less than 30 years** is

$$\begin{aligned}\hat{e}_{121} &= i \times i_1^A \times i_2^B \times i_1^C \times i_{11}^{AC} \\ &= 41.1415 \times 1.2193 \times 2.87 \times 0.8905 \times 1.0989 \\ &= 140.885 \quad .\end{aligned}$$

Note that this value corresponds with the value of the *predicted frequency* $F3$ in the table named ML PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES in the output.

2 LOGIT MODELLING

2.1 Introductory example

2.1.1 The Cross Tabulation of Data

EXAMPLE

Modern Statistics in Practice (Steyn, Smit, du Toit, Strasheim): exercise 9 p 579

An educational researcher wants to determine whether there is a relationship between scholastic achievement and number of hours per day that children spend watching television programmes. The results of a study on 90 pupils are displayed in the following **two-way contingency table**:

	HOURS			
ACHIEVEMENT	Less than two	Two to seven	More than seven	Total
Passed	13	32	15	60
Failed	4	8	18	30
Total	17	40	33	90

The two levels of ACHIEVEMENT are known as the different **rows** of the contingency table, while the three levels of HOURS specify the **columns** of the contingency table. The intersections of these rows and columns are referred to as the various **cells** of the 2×3 contingency table and contain the frequency of occurrence.

The table indicates that a total of 90 children were encountered in this study. An interpretation of the **marginal frequencies** of ACHIEVEMENT reveals that 60 children passed, while 30 children failed. Given that a child has watched TV less than two hours a day, the **partial frequencies** of ACHIEVEMENT are 13 and 4 respectively.

2.1.2 Measures of Association

Consider the following SAS program and output to investigate the relationship between ACHIEVEMENT and HOURS WATCHING TV.

SAS PROGRAM

```
options nodate linesize=64 pagesize=250;
title1 'Relationship between scholastic achievement and hours children watch TV';
proc format;
value aa 1='Passed'
         2='Failed';
value bb 1='Less than 2 hours'
         2='2 to 7 hours'
         3='More than 7 hours';

data tv;
input achieve hours number @@;
label achieve='scholastic achievement';
label hours='number of hours watching TV';
cards;
1 1 13      1 2 32     1 3 15
2 1  4      2 2  8     2 3 18
;

proc freq data=tv;
weight number;
tables achieve*hours / chisq expected cellchisq;
format achieve aa. hours bb.;
title3 'Chisquare test for independence';

run;
```

SAS OUTPUT

Relationship between scholastic achievement and hours children watch TV

Chisquare test for independence

TABLE OF ACHIEVE BY HOURS

ACHIEVE(scholastic achievement) HOURS(number of hours watching TV)

Frequency / Expected / Cell Chi-Square / Percent / Row Pct / Col Pct	Less than 2 hours	2 to 7 hours	More than 7 hours	Total
Passed	13	32	15	60
	11.333	26.667	22	
	0.2451	1.0667	2.2273	
	14.44	35.56	16.67	66.67
	21.67	53.33	25.00	
	76.47	80.00	45.45	
Failed	4	8	18	30
	5.6667	13.333	11	
	0.4902	2.1333	4.4545	
	4.44	8.89	20.00	33.33
	13.33	26.67	60.00	
	23.53	20.00	54.55	
Total	17	40	33	90
	18.89	44.44	36.67	100.00

STATISTICS FOR TABLE OF ACHIEVE BY HOURS

Statistic	DF	Value	Prob
Chi-Square	2	10.617	0.005
Likelihood Ratio Chi-Square	2	10.516	0.005
Mantel-Haenszel Chi-Square	1	7.088	0.008
Phi Coefficient		0.343	
Contingency Coefficient		0.325	
Cramer's V		0.343	

Sample Size = 90

The interpretation of the SAS OUTPUT follows.

Column Percentages

The specific relationship between ACHIEVEMENT and HOURS may be described by means of the **column percentages**. From the frequency table in the output it is evident that 66.67% of the children passed. If we investigate the **partial column percentages** it is possible to observe that a child watching TV more than seven hours a day has a lower chance to pass (only 45.45%) than a child watching TV up to seven hours a day (76.47% and 80% respectively for the two time periods *less than two hours* and *two to seven hours*). The **row percentages** may be interpreted in a similar fashion.

The Pearson χ^2-test

Consider the following hypothesis:

H_0 : The two variables are independent ($i.e.$ there is no relationship between them)

H_1 : The two variables are related ($i.e.$ dependent)

Under the null hypothesis the cell frequencies in each cell would proportionaly reflect the marginal frequencies. It is therefore expected under H_0 that the column percentages would be 66.67% and 33.33% for the categories *passed* and *failed*. It is also expected under H_0 that the composition of 19% children watching TV less than two hours, 44% children watching TV between two and seven hours and 37% watching TV for more than seven hours is maintained over the two levels of ACHIEVEMENT.

Under this assumption the **expected frequency**, e, for each cell is as follows:

$$e = \frac{\text{row total} \times \text{column total}}{\text{overall total}}.$$

Significant deviations of the **observed frequencies** f from the expected frequencies e reflect a relationship between the two variables. The *degree of deviation* may be determined by the **cell χ^2-value**. A cell χ^2-value that exceeds the value of 3.84 (the tabulated $\chi^2_{1,0.05}$-value) indicates a significant difference between the observed and the expected frequency on the 5% level of significance. The expected frequency, as well as the cell χ^2-value, for each cell are reported in the output as the second and third value in each cell of the table.

From the output it is clear that it is expected that 11 of the possible 33 children who watched TV more than seven hours per day, would fail. This is contrasted by the observed frequency of 18 children, indicating a possible association between the two variables. According to the corresponding cell χ^2-value

$$\frac{(f-e)^2}{e} = \frac{(18-11)^2}{11} = 4.45$$

there is a significant deviation between the observed frequency of 18 (f) and the expected frequency of 11 (e). Note that $4.45 > 3.48$.

In two-way contingency tables with multinomial sampling, the null hypothesis of statistical independence could be evaluated by Pearsons' χ^2-statistic

$$\chi^2 = \sum \frac{(f-e)^2}{e}.$$

The larger the value of this statistic, the larger the cell χ^2-values and the more evidence there is against the null hypothesis. The decision rule is to reject the null hypothesis at a 5% level of significance if the exceedance probability (p-value) is less than 0.05 .

The results of Pearsons' χ^2-test are generated by **PROC FREQ** in SAS by using the options CHISQ, EXPECTED and CELLCHISQ and is shown in the output. From the output we see that the value for Pearsons' χ^2-statistic is 10.617 with a corresponding p-value of 0.005. This indicates a significant relationship between ACHIEVEMENT and HOURS WATCHING TV.

Odds Ratio's

Most statistical analyses distinguish between a **response variable** (or dependent variable), the variable that we are trying to explain, and **explanatory variables** (independent variables). In the example the variable ACHIEVEMENT could be regarded as the response variable with HOURS the explanatory variable.

From the output it is evident that 60 of the 90 children will pass. Therefore, the **probability** that a child will pass is

$$\frac{60}{90} = 0.6667.$$

A fundamental concept to the analysis of tables of counts is the definition and use of **odds**. The probability, p, that a child who watches TV MORE THAN SEVEN HOURS per day, will pass is $p = \frac{15}{33} = 0.45$, while the corresponding $odds\ to\ pass$ are

$$odds = \frac{p}{1-p} = \frac{0.45}{1-0.45} = 0.82 \quad .$$

In this case the $odds\ to\ pass$ may also be referred to as 82 to 100. This tells us that for every 82 children that pass, 100 children will fail.

The following table summarizes the observed $odds\ to\ pass$ over the three categories of HOURS.

HOURS	Less than two	Two to seven	More than seven	Overall
$Odds\ to\ pass$	3.25	4	0.82	2

Given the $odds$ that an event occurs, the probability of the event is

$$p = \frac{odds}{1 + odds}.$$

A comparison of two categories with binary responses may be accomplished through the use of **odds ratio's**, which is the ratio of $odds$ in the particular categories.

The $oddsratio$ of the categories TWO TO SEVEN HOURS and MORE THAN SEVEN HOURS is $\frac{4}{0.82} = 4.88$, indicating that the $odds\ to\ pass$ are 4.88 **times higher** for children who watch TV TWO TO SEVEN HOURS per day than for children who watch TV MORE THAN SEVEN HOURS per day.

For the categories TWO TO SEVEN HOURS and LESS THAN TWO HOURS the $oddsratio$ is $\frac{4}{3.25} = 1.23$ and consequently the $odds\ to\ pass$ are 1.23 **times or** 23% **higher** for children who watch TV TWO TO SEVEN HOURS per day than for children who watch TV LESS THAN TWO HOURS per day.

For the categories LESS THAN TWO HOURS and MORE THAN SEVEN HOURS the $oddsratio$ is $\frac{3.25}{0.82} = 3.96$ and consequently the $odds\ to\ pass$ are approximately 4 **times higher** for children who watch TV LESS THAN TWO HOURS per day than for children who watch TV MORE THAN SEVEN HOURS per day.

2.1.3 Statistical Modelling

Now we are going to focus on the statistical modelling of a *binary response variable* in terms of a categorical independent (explanatory) variable.

The Logit Model

The *expected natural log of the odds* called **logit** is modelled for a specific category of the explanatory variable. The **logit model** for the example may be formulated as follows

$$ln(odds) = \mu + \lambda_j^H \qquad \text{for} \qquad j = 1, 2, 3$$

where

$odds$: *odds to pass* for a specific category of HOURS

μ : **overall mean effect** of all the categories

λ_j^H : effect of j^{th} **category of HOURS**

Strictly speaking, μ= ln(geometric mean of the *odds to pass* for the three categories of HOURS). The more the λ^H-parameters deviates from zero, the higher the variation in the *odds to pass*.

The results of this logit analysis is generated by **PROC CATMOD** in SAS. Consider the following SAS program and output for the example:

SAS PROGRAM

```
options nodate linesize=64 pagesize=250;
title1 'Relationship between scholastic achievement and
        number of hours that children spend watching TV';
proc format;
value aa 1='Passed'
         2='Failed';
value bb 1='More than 7 hours'
         2='2 to 7 hours'
         3='Less than 2 hours';

data tv;
input achieve hours number @@;
label achieve='scholastic achievement';
label hours='number of hours watching TV';
cards;
1 3 13      1 2 32     1 1 15
2 3  4      2 2  8     2 1 18
;

proc catmod data=tv;
weight number;
model achieve = hours / ml;
format achieve aa. hours bb.;
title3 'Logit model';

run;
```

SAS OUTPUT

Relationship between scholastic achievement and number of hours that children spend watching TV

Logit model

CATMOD PROCEDURE

Response: ACHIEVE	Response Levels (R)= 2
Weight Variable: NUMBER	Populations (S)= 3
Data Set: TV	Total Frequency (N)= 90
Frequency Missing: 0	Observations (Obs)= 6

POPULATION PROFILES

Sample	HOURS	Sample Size
1	More than 7 hours	33
2	2 to 7 hours	40
3	Less than 2 hours	17

RESPONSE PROFILES

Response	ACHIEVE
1	Passed
2	Failed

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	9.38	0.0022
HOURS	2	9.98	0.0068
LIKELIHOOD RATIO	0	.	.

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.7942	0.2594	9.38	0.0022
HOURS	2	-0.9765	0.3286	8.83	0.0030
	3	0.5921	0.3455	2.94	0.0866

According to the analysis-of-variance table in the computer output we see that the levels of HOURS differ significantly with regard to ACHIEVEMENT (p-value $=0.0068$).

To investigate the specific effect of the various levels of HOURS, it is necessary to interpret the estimates of the parameters of the logit model. λ_1^H, λ_2^H and λ_3^H are estimated by means of regression coefficients of 2 dummies x_1^H and x_2^H.

The logit model may be written as a GLM as follows:

$$\boxed{ln(odds) = \beta_0 + \beta_1^H x_1^H + \beta_2^H x_2^H}$$

Note that $\beta_0 = \mu$.

From the ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES in the output, the regression coefficients for the GLM are

$$\begin{aligned} \hat{\beta}_0 &= 0.7942 \\ \hat{\beta}_1^H &= -0.9765 \\ \hat{\beta}_2^H &= 0.5921 \quad . \end{aligned}$$

Note again that the parameter estimate of the last category of HOURS is omitted from the computer output. This is due to the fact that the λ^H-parameters for the variable HOURS add to zero, that is $\sum_{j=1}^{3} \lambda_H^j = 0$ and therefore $\lambda_3^H = -(\lambda_1^H + \lambda_2^H) = -(-0.9765 + 0.5921) = 0.3844$.

The parameters of the logit model are summarized in the following table.

Effect	Level	**Parameter**	Estimate
Overall mean effect		μ	0.7942
HOURS WATCHING TV	more than seven	λ_1^H	-0.9765
	two to seven	λ_2^H	0.5921
	less than two	λ_3^H	0.3844

Due to the fact that we are currently working in the **natural log-scale**, the interpretation of results is a bit inconvenient.

The Logit Model in terms of Indices

By taking the **anti-log of the logit model**, the logit model may be rewritten as

$$odds = e^{\mu+\lambda_j^H} = e^{\mu} \times e^{\lambda_j^H} = i \times i_j^H \qquad \text{for} \qquad j = 1, 2, 3$$

where

i : the overall $odds\ to\ pass$

(the geometric mean of the $odds\ to\ pass$ for the three categories of HOURS)

i_j^H : an index for the j^{th} **category of HOURS**

The following table shows the indices for the different levels of HOURS.

Effect	Level	**Index**	
Overall $odds\ to\ pass$		i	2.2127
HOURS WATCHING TV	More than seven hours	i_1^H	0.3766
	Between two and seven hours	i_2^H	1.8078
	Less than two hours	i_3^H	1.4687

The **overall** $odds\ to\ pass$ is the **geometric mean** of the $odds\ to\ pass$ for the three categories of HOURS. This value of 2.2127 may be verified from the table of the observed ODDS

$$\sqrt[3]{(0.82)(4)(3.25)} = 2.2008 \quad .$$

The interpretation of the indices follows:

The $odds\ to\ pass$ for category watch TV MORE THAN SEVEN HOURS per day are 0.3766 times as high (or 62.34% lower) as the **overall** $odds\ to\ pass$.

The $odds\ to\ pass$ for category TWO TO SEVEN HOURS are 1.8078 times higher (or 80.78% higher) than the **overall** $odds\ to\ pass$, while the $odds\ to\ pass$ for category LESS THAN TWO HOURS are 1.4687 times higher (or 46.87% higher) than the **overall** $odds\ to\ pass$.

By making use of these indices it is possible to **estimate** the $odds\ to\ pass$ for the three different levels of HOURS.

$$\begin{aligned} odds_1 &= 2.2127 \times 0.3766 = 0.8333 \\ odds_2 &= 2.2127 \times 1.8078 = 4.0001 \\ odds_3 &= 2.2127 \times 1.4687 = 3.2498 \end{aligned}$$

For this example the *estimated odds according to the logit model* are equivalent to the *observed odds* for the three categories of HOURS. This may be explained by the fact that we have a **saturated model** in the example. The number of parameters $(\mu, \lambda_1^H, \lambda_2^H)$ estimated by the logit model, equals the number of cells (three categories of HOURS). Therefore, the explanation of results is mainly incorporated to explain the basic formulation of the logit model.

2.2 The logit model for a two-way contingency table

2.2.1 The Logit Model

Consider factor A (with 2 categories) as the dependent variable and the problem is to **investigate the influence of the j-th category of factor B on** *each category of factor A*. The relationship between factor A (the dependent variable) and factor B (the predictor) may be expressed in terms of the so-called **logit model** where

$$\boxed{ln(\frac{e_{1j}}{e_{2j}}) \text{ is modelled}}$$

with

$$\begin{aligned} e_{1j} &= \text{the expected frequency in cell}(1,j) \\ e_{2j} &= \text{the expected frequency in cell}(2,j) \quad j = 1,2,\ldots,J. \end{aligned}$$

$$\boxed{\frac{e_{1j}}{e_{2j}}}$$

is called the **odds** for the *first category of factor A* against the *second category of factor A*, on the j**th level of factor** B.

The **logit model** for a $2 \times J$ contingency table is

$$\boxed{ln(odds) = \mu + \lambda_j^B \qquad \text{for} \quad j = 1,2,\ldots,J}$$

The parameters of the logit model are

$$\begin{aligned} \mu &= \text{the } \textbf{overall mean effect} \text{ of the categories} \\ \lambda_j^B &= \text{the effect of } j^{th} \textbf{ category of factor } B \quad . \end{aligned}$$

2.2.2 The Logit Model in terms of Indices

The logit model in terms of indices, called the **odds model**, is

$$odds = e^{\mu+\lambda_j^B} = i \times i_j^B \qquad \text{for} \quad j = 1, 2, \ldots, J$$

where

$$\begin{aligned} i &= \text{the } \textbf{geometric mean} \text{ of the } odds \text{ of the categories of factor } B = e^{\mu} \\ i_j^B &= \text{an index for the } j^{th} \textbf{ category of factor } B = e^{\lambda_j^B} \quad . \end{aligned}$$

A comparison between the first two categories of factor B, with regard to the first category of factor A, can be made by the so-called **odds ratio** where

$$oddsratio = \frac{odds \text{ for the first category of factor } A, \text{ on the first level of factor } B}{odds \text{ for the first category of factor } A, \text{ on the second level of factor } B}$$

2.3 Example where a logit model is fitted to a three-way table

Consider the following three-way contingency table ($2 \times 2 \times 2$ table):

ATTITUDE	PRIMARY SCHOOL		SECONDARY SCHOOL		
	rural	urban	rural	urban	total
negative	27	27	40	11	105
positive	9	1	8	7	25
total	36	28	48	18	130

We want to study the effect of **type of school** (primary or secondary) and **area** (rural or urban) on the **attitude of parents towards the school**.

The SAS program and output follow:

SAS PROGRAM

```
data school;
length attitude$ 12;
length type$ 9;
input attitude$ type$ area$ number;
cards;
Negative  Primary    Rural  27
Negative  Primary    Urban  27
Negative  Secondary  Rural  40
Negative  Secondary  Urban  11
Positive  Primary    Rural   9
Positive  Primary    Urban   1
Positive  Secondary  Rural   8
Positive  Secondary  Urban   7
;
proc freq  data=school;
weight number;
tables attitude*type*area;
run;

proc catmod data=school;
weight number;
model attitude=type area type*area /ml oneway;
run;
```

SAS OUTPUT

TABLE 1 OF TYPE BY AREA
CONTROLLING FOR ATTITUDE=Negative

TYPE / Frequency / Percent / Row Pct / Col Pct	AREA: Rural	Urban	Total
Primary	27	27	54
	25.71	25.71	51.43
	50.00	50.00	
	40.30	71.05	
Secondary	40	11	51
	38.10	10.48	48.57
	78.43	21.57	
	59.70	28.95	
Total	67	38	105
	63.81	36.19	100.00

TABLE 2 OF TYPE BY AREA
CONTROLLING FOR ATTITUDE=Positive

TYPE / AREA	Rural	Urban	Total
Primary	9	1	10
	36.00	4.00	40.00
	90.00	10.00	
	52.94	12.50	
Secondary	8	7	15
	32.00	28.00	60.00
	53.33	46.67	
	47.06	87.50	
Total	17	8	25
	68.00	32.00	100.00

CATMOD PROCEDURE

Response: ATTITUDE Response Levels (R)= 2
Weight Variable: NUMBER Populations (S)= 4
Data Set: SCHOOL Total Frequency (N)= 130
Frequency Missing: 0 Observations (Obs)= 8

ONE-WAY FREQUENCIES

Variable	Value	Frequency
ATTITUDE	Negative	105
	Positive	25
TYPE	Primary	64
	Secondary	66
AREA	Rural	84
	Urban	46

POPULATION PROFILES

Sample	TYPE	AREA	Sample Size
1	Primary	Rural	36
2	Primary	Urban	28
3	Secondary	Rural	48
4	Secondary	Urban	18

RESPONSE PROFILES

Response	ATTITUDE
1	Negative
2	Positive

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	26.56	0.0000
TYPE	1	3.47	0.0625
AREA	1	0.69	0.4065
TYPE*AREA	1	7.17	0.0074
LIKELIHOOD RATIO	0	.	.

```
ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES
                                          Standard    Chi-
Effect            Parameter  Estimate     Error     Square   Prob
-------------------------------------------------------------------
INTERCEPT                 1    1.6140    0.3131      26.56  0.0000
TYPE                      2    0.5833    0.3131       3.47  0.0625
AREA                      3   -0.2599    0.3131       0.69  0.4065
TYPE*AREA                 4   -0.8387    0.3131       7.17  0.0074
```

The interpretation of the SAS OUTPUT follows.

ATTITUDE is the dependent variable (factor A on two levels) with two categories $negative$ and $positive$, TYPE the first predictor or independent variable (factor B on two levels) with two categories $primary$ and $secondary$ and AREA the second predictor or independent variable (factor C on two levels) with two categories $rural$ and $urban$.

The **odds of a negative attitude** may be modelled, in terms of logs, as

$$\boxed{ln(odds) = \mu + \lambda_j^B + \lambda_k^C + \lambda_{jk}^{BC} \quad j = 1,2 \text{ and } k = 1,2}$$

where

μ = the **overall mean effect** of the categories

λ_j^B = the effect of the j^{th} **type of school** $j = 1,2$

λ_k^C = the effect of the k^{th} **area** $k = 1,2$

λ_{jk}^{BC} = the **interaction** effect between the j^{th} type of school and the k^{th} area.

From the MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE in the output, it follows that the interaction term TYPE*AREA is highly significant (a p-value of 0.0074) and therefore this **saturated logit model** fits the data well.

The estimation of the parameters of the logit model follows. This logit model may be written as a GLM as follows:

$$\boxed{ln(odds) = \beta_0 + \beta_1^B x_1^B + \beta_1^C x_1^C + \beta_{11}^{BC} x_{11}^{BC}}$$

Note that $\beta_0 = \mu$.

From the ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES in the output, the regression coefficients for the GLM are

$$\begin{aligned} \hat{\beta}_0 &= 1.6140 \\ \hat{\beta}_1^B &= 0.5833 \\ \hat{\beta}_1^C &= -0.2599 \\ \hat{\beta}_{11}^{BC} &= -0.8387 \quad . \end{aligned}$$

The parameters of the logit model are summarized in the following table.

Effect	**Level**	**Parameter**	**Estimate**
Overall mean effect		μ	1.6140
TYPE OF SCHOOL	Primary	λ_1^B	0.5833
	Secondary	λ_2^B	-0.5833
AREA	Rural	λ_1^C	-0.2599
	Urban	λ_2^C	0.2599
TYPE OF SCHOOL*AREA	Primary, rural	λ_{11}^{BC}	-0.8387
	Primary, urban	λ_{12}^{BC}	0.8387
	Secondary, rural	λ_{21}^{BC}	0.8387
	Secondary, urban	λ_{22}^{BC}	0.8387

From the estimated effects of the logit model, the estimated indices can be determined.

The logit model in terms of indices, called the **odds model of a negative attitude** is

$$\boxed{odds = e^{\mu+\lambda_j^B+\lambda_k^C+\lambda_{jk}^{BC}} = i \times i_j^B \times i_k^C \times i_{jk}^{BC} \quad j = 1,2 \text{ and } k = 1,2}$$

where

$$\begin{array}{lcll} i & = & \text{the } \textbf{geometric mean} \text{ of the odds of all the categories} & = & e^{\mu} \\ i_j^B & = & \text{an index for the } j^{th} \textbf{ type of school} & = & e^{\lambda_j^B} \\ i_k^C & = & \text{an index for the } k^{th} \textbf{ area} & = & e^{\lambda_k^C} \\ i_{jk}^{BC} & = & \text{an index for the } j^{th} \textbf{ type of school and the } k^{th} \textbf{ area} & = & e^{\lambda_{jk}^{BC}} \end{array}$$

The following table shows the indices for the different levels of TYPE OF SCHOOL and AREA.

Effect	**Level**	**Index**	
Overall effect		i	5.0229
TYPE OF SCHOOL	Primary	i_1^B	1.7919
	Secondary	i_2^B	0.5581
AREA	Rural	i_1^C	0.7711
	Urban	i_2^C	1.2968
TYPE OF SCHOOL*AREA	Primary, rural	i_{11}^{BC}	0.4323
	Primary, urban	i_{12}^{BC}	2.3134
	Secondary, rural	i_{21}^{BC}	2.3134
	Secondary, urban	i_{22}^{BC}	0.4323

The **odds of a negative attitude** of parents at $primary\ schools$ is estimated by

$$i \times i_1^B = 5.0229 \times 1.7919 = 9.0005 = \frac{9}{1} \quad .$$

For every 9 primary schools with negative parents, there is 1 primary school with positive parents.

The estimated **probability of a negative attitude** of parents at $primary\ schools$ is

$$\frac{odds}{1+odds} = \frac{9.0005}{1+9.0005} = 0.90 \quad .$$

The **odds of a negative attitude** of parents at $secondary\ schools$ is estimated by

$$i \times i_2^B = 5.0229 \times 0.5581 = 2.8 = \frac{28}{10} \quad .$$

For every 28 secondary schools with negative parents there will be 10 secondary schools with positive parents.

The estimated **probability of a negative attitude** of parents at $secondary\ schools$ is

$$\frac{odds}{1 + odds} = \frac{2.8}{1 + 2.8} = 0.74 \quad .$$

The **odds of a negative attitude** of parents at $urban\ primary\ schools$ is estimated by

$$i \times i_1^B \times i_2^C \times i_{12}^{BC} = 5.0229 \times 1.7919 \times 1.2968 \times 2.3134 = 27.0018 = \frac{27}{1} \quad .$$

For every 27 urban primary schools with negative parents there will be 1 urban primary school with positive parents.

The estimated **probability of a negative attitude** of parents at $urban\ primary\ schools$ is

$$\frac{odds}{1 + odds} = \frac{27.0018}{1 + 27.0018} = 0.96 \quad .$$

The **odds of a negative attitude** of parents at $urban\ secondary\ schools$ is estimated by

$$i \times i_2^B \times i_2^C \times i_{22}^{BC} = 5.0229 \times 0.5581 \times 1.2968 \times 0.4323 = 1.5715 = \frac{157}{100} \quad .$$

For every 157 urban secondary schools with negative parents there will be 100 urban secondary schools with positive parents.

The estimated **probability of a negative attitude** of parents at $urban\ secondary\ schools$ is

$$\frac{odds}{1 + odds} = \frac{1.5715}{1 + 1.5715} = 0.61 \quad .$$

Similarly the **odds of a negative attitude** of parents at $rural\ primary\ schools$ is estimated by 3.003 with associated probability 0.75 . The **odds of a negative attitude** of parents at $rural\ secondary\ schools$ is estimated by 5.0005 with associated probability 0.83 .

A **comparison between primary schools** in *urban and rural areas*, with regard to a negative attitude of parents towards the school, can be made by the

$$\begin{aligned} oddsratio &= \frac{\text{estimated odds of a negative attitude at urban primary schools}}{\text{estimated odds of a negative attitude at rural primary schools}} \\ &= \frac{27.0018}{3.0003} \\ &= 9 \quad . \end{aligned}$$

Thus the *odds of a negative attitude* is 9 **times higher at urban primary schools** than at rural primary schools. By comparing the *probabilities of a negative attitude* of parents at these schools, it follows that $\frac{0.96}{0.75} = 1.28$. There is 28% **more primary schools in the urban area** , with parents having a negative attitude towards the school, than in the rural area.

A **comparison between secondary schools** in *rural and urban areas*, with regard to a negative attitude of parents towards the school, can be made by the

$$\begin{aligned} oddsratio &= \frac{\text{estimated odds of a negative attitude at rural secondary schools}}{\text{estimated odds of a negative attitude at urban secondary schools}} \\ &= \frac{5.0005}{1.5715} \\ &= 3.18 \quad . \end{aligned}$$

Thus the *odds of a negative attitude* is 3 **times higher at rural secondary schools** than at urban secondary schools. By comparing the *probabilities of a negative attitude* of parents at these schools, it follows that $\frac{0.83}{0.61} = 1.36$. There is 36% **more secondary schools in the rural area** , with parents having a negative attitude towards the school, than in the urban area.

A **comparison between** *primary and secondary schools* **in the urban area**, with regard to a negative attitude of parents towards the school, can be made by the

$$\begin{aligned} oddsratio &= \frac{\text{estimated odds of a negative attitude at urban primary schools}}{\text{estimated odds of a negative attitude at urban secondary schools}} \\ &= \frac{27.0018}{1.5715} \\ &= 17.18 \quad . \end{aligned}$$

Thus the *odds of a negative attitude* in the urban area is 17 **times higher at primary schools** than at secondary schools. By comparing the *probabilities of a negative attitude* of parents at these schools, it follows that $\frac{0.96}{0.61} = 1.57$. There is 57% **more primary schools in the urban area** , with parents having a negative attitude towards the school, than secondary schools.

A **comparison between** *secondary and primary schools* **in the rural area**, with regard to a negative attitude of parents towards the school, can be made by the

$$\begin{aligned} oddsratio &= \frac{\text{estimated odds of a negative attitude at rural secondary schools}}{\text{estimated odds of a negative attitude at rural primary schools}} \\ &= \frac{5.0005}{3.0003} \\ &= 1.67 \quad . \end{aligned}$$

Thus the *odds of a negative attitude* in the rural area is 1.67 **times higher at secondary schools** than at primary schools. By comparing the *probabilities of a negative attitude* of parents at these schools, it follows that $\frac{0.83}{0.75} = 1.107$. There is 11% **more secondary schools in the rural area** , with parents having a negative attitude towards the school, than primary schools.