

Two way contingency table: Basic Five Models

1. Independence model

$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C$, where $\sum_i \lambda_i^R = 0$ and $\sum_j \lambda_j^C = 0$, R and C is superscript to indicate corresponding variables not powers.

How does this mathematical equation mean 'independence'?

Always think about it in terms of odds-ratio.

$$\ln(\theta) = \ln\left(\frac{F_{ij} F_{i'j'}}{F_{i'j} F_{ij'}}\right) = \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{i'j}) - \ln(\hat{F}_{ij'}) = 0.$$

Then, $\theta = 1$. So there is no association.

Number of parameters used: one for λ , (I-1) for λ_i^R , and (J-1) for λ_j^C .

So, degrees of freedom = $I*J - (1 + I - 1 + J - 1) = (I-1)*(J-1)$.

2. Uniform association model

$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + \beta * i * j$, (i and j is not symbols but they are just values, i.e., scores that are not estimated) $\rightarrow i$ and j are treated as interval variables just as linear-regression. So, at least they must be pre-ordered.

$$\begin{aligned} \ln(\theta) = \ln\left(\frac{F_{ij} F_{i'j'}}{F_{i'j} F_{ij'}}\right) &= \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{i'j}) - \ln(\hat{F}_{ij'}) \\ &= \beta * i * j + \beta * i' * j' \\ &\quad - \beta * i' * j - \beta * i * j' = \beta (i - i')(j - j') \end{aligned}$$

So in a uniform association model, log-odds are the same for any adjacent sub-tables. How many more parameters are introduced than independence model?

3. Row-effect model

$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + \mu_i * j$, j is a score (value of the column variable) not a parameter to be estimated.

$$\begin{aligned} \ln(\theta) = \ln\left(\frac{F_{ij} F_{i'j'}}{F_{i'j} F_{ij'}}\right) &= \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{i'j}) - \ln(\hat{F}_{ij'}) \\ &= \mu_i * j + \mu_{i'} * j' - \mu_{i'} * j - \mu_i * j' \\ &= (\mu_i - \mu_{i'})(j - j'). \end{aligned}$$

So in row-effect model, the adjacent sub-table odds-ratio is determined by difference of 'row' parameters. Here, we assume that column variable as interval variable. So, at least they must be ordinal.

How many more parameters are used than in uniform association model?

How is uniform association model nested within row-effect model?

4. Column-effect model

$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + v_j * i$, i is a score (value of the row variable) not a parameter to be estimated.

$$\begin{aligned} \ln(\theta) &= \ln\left(\frac{F_{ij} F_{i'j'}}{F_{ij'} F_{ij}}\right) = \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{ij'}) - \ln(\hat{F}_{ij}) = v_j * i + v_{j'} * i' - v_{j'} * i - v_j * i' \\ &= (v_j - v_{j'})(i - i'). \end{aligned}$$

So, in column-effect model, the adjacent sub-table odds-ratio is determined by difference of 'column' parameters. Here, we assume that row variable as interval variable. So, at least they must be ordinal.

How many more parameters are used than in uniform association model?

How is uniform association model nested within column-effect model?

4. Saturated model

$$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + \gamma_{ij}$$

$$\ln(\theta) = \ln\left(\frac{F_{ij} F_{i'j'}}{F_{ij'} F_{ij}}\right) = \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{ij'}) - \ln(\hat{F}_{ij}) = \gamma_{ij} + \gamma_{i'j'} - \gamma_{ij'} - \gamma_{ij'}$$

What is the degrees of freedom?

How can this model perfectly predict the observed frequencies?

```
. . tab degree sexfreq, ro ch
```

highest year of schooling r com	how often r had sex in last yea			Total
	1	2	3	
1	142 33.97	137 32.78	139 33.25	418 100.00
2	224 26.02	312 36.24	325 37.75	861 100.00
3	290 29.06	358 35.87	350 35.07	998 100.00
4	196 28.04	278 39.77	225 32.19	699 100.00
Total	852 28.63	1085 36.46	1039 34.91	2976 100.00

Pearson chi2(6) = 13.9371 Pr = 0.030

1: less than HS, 2: HS, 3: higher than HS without 4-year Bachelor's degree, 4: BA or higher than BA

1: about once a month (or less), 2: more than once a month or about once a week 3: more than once a week

```
set length=none/width=80.
data list list / deg sexfreq freq.
weight by freq.
begin data
1 1 142
1 2 137
1 3 139
2 1 224
2 2 312
2 3 325
3 1 290
3 2 358
3 3 350
4 1 196
4 2 278
4 3 225
end data.
compute d=deg.
compute f=sexfreq.
compute u=d*f.
loglinear deg (1,4) sexfreq (1,3) with u d f
/print=none
/design=deg, sexfreq
/design=deg, sexfreq, u
/design=deg, sexfreq, sexfreq by d
/design=deg, sexfreq, deg by f
/design=deg, sexfreq, deg by sexfreq.
```

***** LOG LINEAR ANALYSIS *****

DATA Information

12 unweighted cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
2976 weighted cases will be used in the analysis.

FACTOR Information

Factor	Level	Label
DEG	4	
SEXFREQ	3	

DESIGN Information

5 Designs/Models will be processed.

***** LOG LINEAR ANALYSIS *****

Correspondence Between Effects and Columns of Design/Model 1

Starting Column	Ending Column	Effect Name
1	3	DEG
4	5	SEXFREQ

*** ML converged at iteration 3.
Maximum difference between successive iterations = .00001.

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square =	13.75417	DF = 6	P = .033
Pearson Chi Square =	13.93708	DF = 6	P = .030

***** LOG LINEAR ANALYSIS *****

Correspondence Between Effects and Columns of Design/Model 2

Starting Column	Ending Column	Effect Name
1	3	DEG
4	5	SEXFREQ
6	6	U

*** ML converged at iteration 3.
Maximum difference between successive iterations = .00008.

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square = 13.74826 DF = 5 P = .017
Pearson Chi Square = 13.94705 DF = 5 P = .016

***** LOG LINEAR ANALYSIS *****

Correspondence Between Effects and Columns of Design/Model 3

Starting Column	Ending Column	Effect Name
1	3	DEG
4	5	SEXFREQ
6	7	SEXFREQ BY D

*** ML converged at iteration 3.
Maximum difference between successive iterations = .00002.

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square = 9.09512 DF = 4 P = .059
Pearson Chi Square = 9.03588 DF = 4 P = .060

***** LOG LINEAR ANALYSIS *****

Correspondence Between Effects and Columns of Design/Model 4

Starting Column	Ending Column	Effect Name
1	3	DEG
4	5	SEXFREQ
6	8	DEG BY F

*** ML converged at iteration 3.

Maximum difference between successive iterations = .00001.

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square = 5.94040 DF = 3 P = .115
Pearson Chi Square = 5.93175 DF = 3 P = .115

* * * * * L O G L I N E A R A N A L Y S I S * * * * *

Correspondence Between Effects and Columns of Design/Model 5

Starting Column	Ending Column	Effect Name
1	3	DEG
4	5	SEXFREQ
6	11	DEG BY SEXFREQ

Note: for saturated models .500 has been added to all observed cells.
This value may be changed by using the CRITERIA = DELTA subcommand.

*** ML converged at iteration 2.
Maximum difference between successive iterations = .00000.

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square = .00000 DF = 0 P = 1.000
Pearson Chi Square = .00000 DF = 0 P = 1.000

Model Comparison

	D.F.	Likelihood Ratio Chi-square	Pearson Chi- square	P-value
Independence	6	13.75	13.94	0.03
Uniform Association	5	13.75	13.95	0.02
Column-effect	4	9.1	9.04	0.06
Row-effect	3	5.94	5.93	0.12
Saturated	0	0	0	1

R vs. I: $\text{chiprob}(6-3, 13.75-5.94)=0.0501$

R vs. U: $\text{chiprob}(5-3, 13.75-5.94)=0.02$

R vs. C: ? Are these two are nested? How can we compare these two?

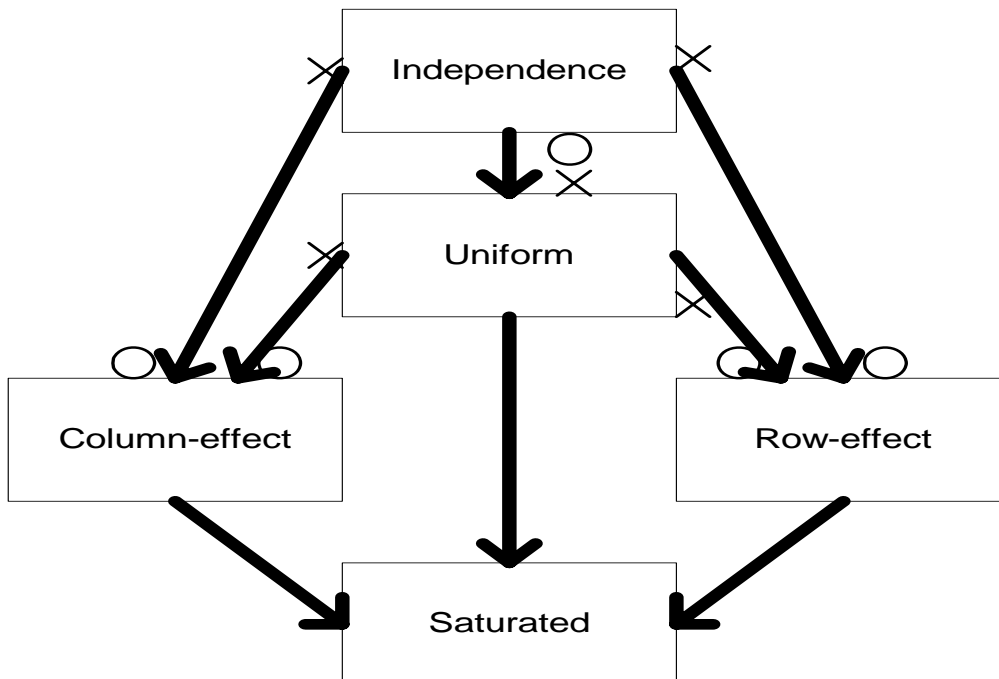
C vs. I: $\text{chiprob}(6-4, 13.75-9.1)=0.098$

C vs. U: $\text{chiprob}(5-4, 13.75-9.1)=0.03$

I vs. U: $\text{chiprob}(6-5, 0)=1$

What is the best model? (when $\alpha=10\%$)

Do we have to compare even when there is only one fitting model?



The simple chi-square test built in standard statistical program cannot give us accurate and rich information about association (degree, direction, etc.)

```

set length=none/width=80.
data list list / deg sexfreq freq.
weight by freq.
begin data
1 1 142
1 2 137
1 3 139
2 1 224
2 2 312
2 3 325
3 1 290
3 2 358
3 3 350
4 1 196
4 2 278
4 3 225
end data.
compute d=deg.
compute f=sexfreq.
compute u=d*f.
loglinear deg (1,4) sexfreq (1,3) with u d f
/print=none
/design=deg, sexfreq
/design=deg, sexfreq, u
/design=deg, sexfreq, sexfreq by d
/design=deg, sexfreq, deg by f
/design=deg, sexfreq, deg by sexfreq
/print=estim corr
/design=deg, sexfreq, deg by f.

```

***** LOG LINEAR ANALYSIS *****

Correspondence Between Effects and Columns of Design/Model 6

Starting Column	Ending Column	Effect Name
1	3	DEG
4	5	SEXFREQ
6	8	DEG BY F

*** ML converged at iteration 3.
Maximum difference between successive iterations = .00001.

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square =	5.94040	DF = 3	P = .115
Pearson Chi Square =	5.93175	DF = 3	P = .115

Estimates for Parameters

DEG

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.332752153	.10743	-3.09729	-.54332	-.12218
2	-.015777500	.08720	-.18093	-.18669	.15514
3	.3220825730	.08193	3.93123	.16150	.48266

SEXFREQ

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
4	-.132007810	.02804	-4.70857	-.18696	-.07706
5	.0964111857	.02546	3.78749	.04652	.14630

DEG BY F

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
6	-.095126347	.04982	-1.90941	-.19277	.00252
7	.1022430199	.03906	2.61777	.02569	.17880
8	.0111839165	.03724	.30035	-.06180	.08417

Covariance(below) and Correlation(above) Matrices of Parameter Estimates

Parameter	Parameter					
	1	2	3	4	5	6
1	.01154	-.41496	-.39068	.21034	.02524	-.92938
2	-.00389	.00760	-.20540	-.07490	-.03391	.38219
3	-.00344	-.00147	.00671	-.16878	-.00366	.36485
4	.00063	-.00018	-.00039	.00079	-.49576	-.23800
5	.00007	-.00008	-.00001	-.00035	.00065	-.02669
6	-.00497	.00166	.00149	-.00033	-.00003	.00248
7	.00166	-.00318	.00059	.00011	.00004	-.00082
8	.00149	.00059	-.00284	.00019	.00000	-.00075
	7	8				
1	.39568	.37227				
2	-.93487	.18259				
3	.18529	-.93242				
4	.09831	.18563				
5	.03658	.00421				
6	-.42044	-.40319				
7	.00153	-.18504				
8	-.00027	.00139				

```

set length=none/width=80.
data list list / deg sexfreq freq.
weight by freq.
begin data
1 1 142
1 2 137
1 3 139
2 1 224
2 2 312
2 3 325
3 1 290
3 2 358
3 3 350
4 1 196
4 2 278
4 3 225
end data.
compute d=deg.
compute f=sexfreq.
compute u=d*f.
loglinear deg (1,4) sexfreq (1,3) with u d f
/print=none
/design=deg, sexfreq
/design=deg, sexfreq, u
/design=deg, sexfreq, sexfreq by d
/design=deg, sexfreq, deg by f
/design=deg, sexfreq, deg by sexfreq
/print=estim corr
/design=deg, sexfreq, deg by f
/contrast(deg)=deviation(2)
/design=deg, sexfreq, deg by f.

```

* * * * * L O G L I N E A R A N A L Y S I S * * * * *

Correspondence Between Effects and Columns of Design/Model 7

Starting Column	Ending Column	Effect Name
1	3	DEG
4	5	SEXFREQ
6	8	DEG BY F

*** ML converged at iteration 3.
Maximum difference between successive iterations = .00001.

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square =	5.94040	DF = 3	P = .115
Pearson Chi Square =	5.93175	DF = 3	P = .115

Estimates for Parameters

DEG

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.332752153	.10743	-3.09729	-.54332	-.12218
2	.3220825730	.08193	3.93123	.16150	.48266
3	.0264470806	.09095	.29080	-.15181	.20470

SEXFREQ

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
4	-.132007810	.02804	-4.70857	-.18696	-.07706
5	.0964111857	.02546	3.78749	.04652	.14630

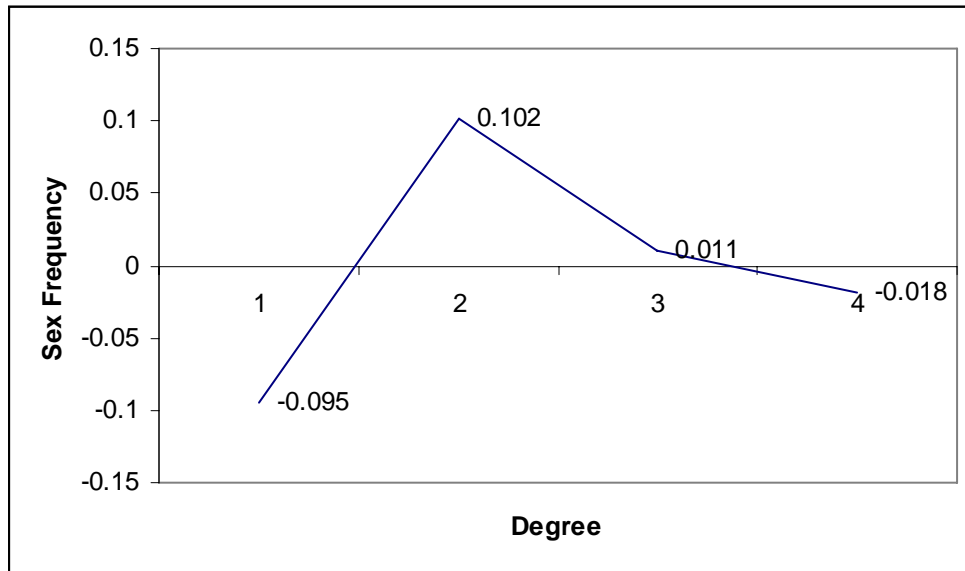
DEG BY F

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
6	-.095126347	.04982	-1.90941	-.19277	.00252
7	.0111839165	.03724	.30035	-.06180	.08417
8	-.018300590	.04152	-.44080	-.09967	.06307

Covariance(below) and Correlation(above) Matrices of Parameter Estimates

Parameter	Parameter					
	1	2	3	4	5	6
1	.01154	-.39068	-.43147	.21034	.02524	-.92938
2	-.00344	.00671	-.24241	-.16878	-.00366	.36485
3	-.00422	-.00181	.00827	-.02461	.00599	.40274
4	.00063	-.00039	-.00006	.00079	-.49576	-.23800
5	.00007	-.00001	.00001	-.00035	.00065	-.02669
6	-.00497	.00149	.00182	-.00033	-.00003	.00248
7	.00149	-.00284	.00076	.00019	.00000	-.00075
8	.00182	.00076	-.00352	.00003	-.00001	-.00092
	7	8				
1	.37227	.40912				
2	-.93242	.22416				
3	.22515	-.93173				
4	.18563	.02661				
5	.00421	-.00615				
6	-.40319	-.44284				
7	.00139	-.23899				
8	-.00037	.00172				

Interpretation



$$(0.102 - (-0.095)) = 0.197 \text{ and } \exp(0.197) = 1.2$$

Thus, people with HS grade are 1.2 times more likely to have frequent sex by one unit than people with less than HS grade.

Is this difference significant?

$$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2 * \text{cov}(X, Y)$$

We have to calculate $\frac{X - Y}{\sqrt{\text{var}(X - Y)}} = \frac{X - Y}{\sqrt{\text{var}(X) + \text{var}(Y) - 2 * \text{cov}(X, Y)}}$, which follows

standard normal distribution.

$$(0.102 - (-0.095)) / \text{square root of } (0.00248 + 0.00153 - 2 * -0.00082) = 2.62$$

$$\text{Thus, it is significant.. } p\text{-value} = 2 * (1 - \text{normprob}(2.62)) = 0.0088$$

Is the difference between people with more than HS (third category) and people with BA (last category) significant?