MEDSTA 2: Regression models in medical research

17 February 2014

Øystein Ariansen Haaland, PHD

Rolv Skjærven, PHD

Department of Global Public Health and Primary Care, University of Bergen

BINOMIAL DISTRIBUTION

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,\ldots,n$$

Contents

- 2x2 tables
- Case-Control data
- OR and RR with Confidence Intervals (C.I.)
- Extending 2x2 to more dimesions (2x2x2x2…)
- Logistic regression
- Relative Risk regression
- McNemar for matched pairs

3

Two alternative strategies

- The binomial distribution can be used to compare for instance differences in healing by two different medications.
- Example: Treatment of ulcer by the drugs Pirenzepine or Trihiozine

Drug	Healed	Not healed	Total	% healed
Pirenzepine	23 (a)	7 (c)	30 (r)	p1=76.7
Trithiozine	18 (b)	13 (d)	31 (s)	p2=58.1
Total	41 (m)	20 (n)	61 (N)	67.2

- Using the binomial distribution we can compare %healed, p1 and p2
- by testing 'equal treatment effect' or by using 95% C.I.

- ***We will now present an alternative strategy that has the capasity to extend to more factors, and to more levels for each factor.***

… especially well suited for case-control data (Kjuus-data)

- Lung Cancer (Case-Control material) and asbestos-exposure (Data from Kjuus, Skjærven, Langård, 1987)

ASBESTOS	Case	Control	Total
Exposed	105 (a)	74 (b)	179 (r)
Not exp.	71 (c)	102 (d)	173 (s)
Total	176 (m)	176 (n)	352 (N)

- We can see from the table that there are more exposed among the cases than among controls.
- How can we evaluate this effect, and conclude that the exposure causes lung cancer?
- Account for smoking habits?

… case-control data (cont.)

- Lung Cancer (Case-Control material) and asbestos-exposure (Data from Kjuus, Skjærven, Langård, 1987)

SMOKING	Case	Control	Total
Dose 1	36 (25%)	107	143
Dose 2	97 (62%)	59	156
Dose 3	43 (81%)	10	53
Total	176	176	352

- Account for asbestus exposure?

2x2 (or MxN) contingency-table

- We have two properties for each study unit, and we want to evaluate whether there is a statistical dependency between these two, or whether they distribute independent of each other.

- Null hypotesis specified as:

  H_0: Independence between exposure and disease.

Test statistic:

- We have two identical test statistics that can be used to test this hypothesis:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$ **Alternative 1**

$$\chi^2 = N(ad - bc)^2 / mnrs$$ **Alternative 2**

- O's represent observered value, E's expected
- The index gives the 'cell'-number
- a,b,c,d,m,n,r,s,N as in the table above
- The last formula is only for 2x2 tables, while the first formula is valid also for RxC tables, where both R and C can be larger than 2

Calculation, expected values (E_i)

- Assuming the null hypothesis, we can calculate the expected values using the marginal values: There are 173 unexposed individuals (49.15%), and 179 exposed (50.85%).
- Under the null hypothesis, we expect a simular distribution (unexposed/exposed) among the 176 cases. This gives us the 'expected values' 176*0.4915=86.5 and 176*0.5085=89.5.
- (Similarly for the controls, and here the proportions are equal since both marginals are 176).
- We get:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

$$= (71-86.5)^2 / 86.5 + (105-89.5)^2 / 89.5$$

$$+ (102-86.5)^2 / 86.5 + (74-89.5)^2 / 89.5 = 10.9$$

9

Chi square distribution

- Small values are in support of the null hypothesis.
- Large values of the chi square statistic is an indication for rejection of the null hypothesis.
- The test statistic follows a chi square distribution (under the null hypothesis).
- This allows us to calculate a p-value for the observed result.
- The critical values 3.84 corresponds to 5% (1 degree of freedom).

- Example: We find from tables that 10.9 corresponds to p=0.001.

10

Assumption for standard chi square tests

- 80% of the cells must have expected values >=5

- In 2x2 tables, all the cells must have expected values >=5

- If not: Use Fisher's exact test

Degrees of freedom

- The chi square distribution has one parameter (as for the t-distribution). This is called degrees of freedom
- In Norwegian: frihetsgrader
- For 2x2 tables, df=1
- In general, RxC tables have (R-1)x(C-1) df.

Odds Ratio (OR)

- The usual measure of effect in a 2x2 table is OR
- It is a ratio of two odds.
- In our example it tells us about excess of exposed cases, relative controls.
- Odds ratio is calculated as a cross product ratio in a 2x2 table
- (In a Rx2 table: choose a reference category)
- We find

$$OR = 102 \times 105 / (71 \times 74)$$
$$= a\ d / b\ c = 2.04$$

13

95% confidence interval (C.I.)

- We can calculate a 95% confidence interval for OR by focusing the logarithm of OR. We can show that (asymptotically):

$$SE(\ln(OR)) \approx \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

- We calculate first OR, thereafter ln(OR) and finally 95%C.I.
- We transform back to OR.
- Under H_0, OR will be equal 1, thus the hypothesis of independence is rejected if the CI does not include this value.
- Our example: ln(OR)=ln 2.04=0.713

95%CI for ln(OR): (0.713-1.96x0.217, 0.713+1.96x0.217)
= (0.713 - 0.425, 0.713 + 0.425) = (0.288 , 1.138) ,

og 95%CI for OR blir (exp(0.288), exp(1.138))=(1.33, 3.12)

Example:

(Exam medical students 10.04.2002; q 1-4)

- "The risk for preeclampsia (svangerskapsforgiftning) increase if the woman change partner between 1st and 2nd pregnancy".
- MBR data, case-control design: 1296 case, 7850 controls. 79 of the women with preeclampsia had new partner, compared to 449 in the control group.

a) Set up a 2x2 table for the situation

b) and c) Calculate an OR and a 95% confidence interval

d) Specify a null hypotesis and an alternative hypotesis and test the relation between 'new partner' and preeclampsia.

Published in New Engl J Med, 2002
Skjærven, Wilcox, Lie; 346, 33-36

preeclampsia (cont.) NEW DATA*

NP = New partner; SP=Same partner;

PR = Preeclampsia (Svangerskapsforgiftning)

	PR	notPR	Total
SP	602 (2.2%)	26222	26824
NP	2831 (1.6%)	174110	176941
Total	3433 (1.7%)	200332	203765

OR=1.41 , 95% C.I.: (1.29 – 1.54)

(*) 2nd birth 1999-2009, singletons 1st and 2nd pregnancies

16

More than two variables?

For example
2x2x2x2 tables?

17

Hypertension, relative smoking, obesity and snoring: 40+ years men (Altman, tab 12.19; s.353)

Smoking	Obesity	Snoring	Number of men	Hyper-tensive	Percent
0	0	0	60	5	8.3
1	0	0	17	2	11.8
0	1	0	8	1	12.5
1	1	0	2	0	0.0
0	0	1	187	35	18.7
1	0	1	85	13	15.3
0	1	1	51	15	29.4
1	1	1	23	8	34.8
			433	79	18.2

1=yes, 0=no

What lead to hypertension?

Some questions:

- What is the separate factor's contribution to disease?
- How do the different factors interrelate?
- How does this interrelation influence disease?

2x2 tables

	Hypertension 0	1	Total	OR-estimates (with 95% C.I.)
Smoke=0 (%)	250 (81.7)	56 (18.3)	306 (100)	
Smoke=1 (%)	104 (81.9)	23 (18.1)	127 (100)	1.0 (0.6-1.7)

	Hypertension 0	1	Total	
Obese=0 (%)	294 (84.2)	55 (15.8)	349 (100)	
Obese=1 (%)	60 (71.4)	24 (28.6)	84 (100)	2.1 (1.2-3.7)

	Hypertension 0	1	Total	
Snore=0 (%)	79 (90.8)	8 (9.2)	306 (100)	
Snore=1 (%)	275 (79.5)	71 (20.5)	127 (100)	2.6 (1.2-5.5)

Relation between risk factors

	Snore		
	1	0	Total
Obese=1 (%)	77 (22.1)	272 (77.9)	349 (100)
Obese=0 (%)	10 (11.9)	74 (88.1)	84 (100)
Total	87 (20.1)	346 (79.9)	433 (100)

	95% CI	
OR for obese (0/1)	Lower	Upper
2.095	1.033	4.249

21

Logistic regression

$$\text{logit(p)} = ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

- x is exposure (e.g., 0 = ‘non-smoker’, 1=‘smoker’, or x=age).
- α and β are unknown parameters.
- p is the risk /rate for disease.
- We assume p to depend on x.

22

Logistic regression

Binary/dichotomous x-variables

- x=1 for ’exposed’
- x=0 for ’unexposed’
- OR=exp(β)

23

Statistical evaluation of data using 2x2 tables or logistic regression:

- Are results comparable?
- Example: Obesity and hypertension
  - OR value
  - and Chi square (with p-value)

24

Hypertension and obesity (Altman Tab. 12.19).

- Results 2x2 table (χ^2-test)
  - OR=2.138
  - 95% CI: 1.229-3.721
  - Likelihood ratio: 6.826
- Results logistic regression
  - OR=2.138
  - 95% CI: 1.228-3.720
  - χ^2: 6.826

25

Logistic regression

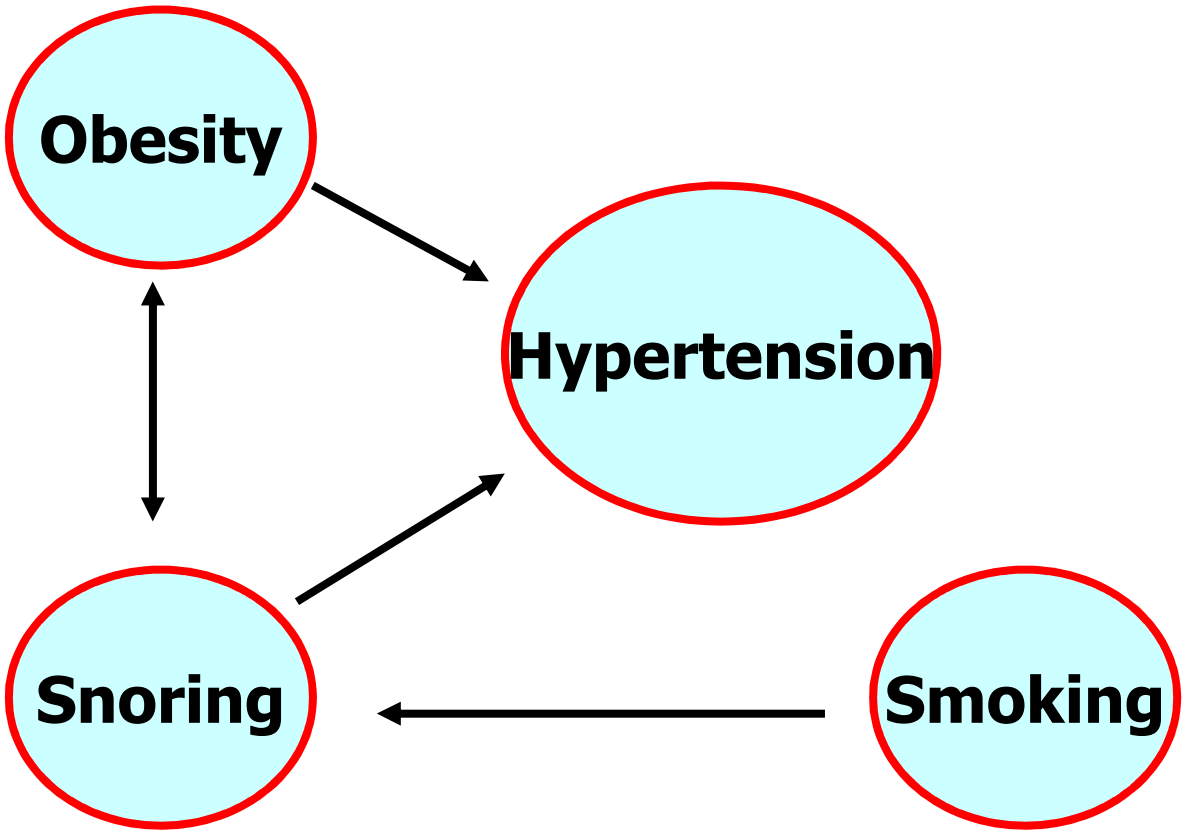
- Simultaneous estimation

  Effects of smoking, obesity and snoring on hypertension:

	OR	(95% C.I.)
Smoking	0.9	(0.5 – 1.6)
Snoring	2.4	(1.1 – 5.2)
Obese	2.0	(1.1 – 3.5)

26

Interpretation of results



27

Multiple logistic regression

A simple extention:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- $x_{1,\ldots,}$ x_k represent, e.g., obesity, snoring, and so on.
- Again, α and βs are unknown parameters that we need to estimate based on data.
- p is the proportion of, e.g., hypertension.

Intepretation of coefficients

- If x_1 represents obesity and x_2 represents snoring, β_2 is interpreted as the effect of snoring given that we have 'adjusted for obesity'.
- Alternatively, we can do a stratified analysis: one analysis on x_2 for each level of x_1 .

Logistic regression

- Stata
  - logit hypertension obese smoke
- Report OR
  - logit hypertension obese smoke, or
- If smoking has more than one category
  - logit hypertension obese i.smoke, or

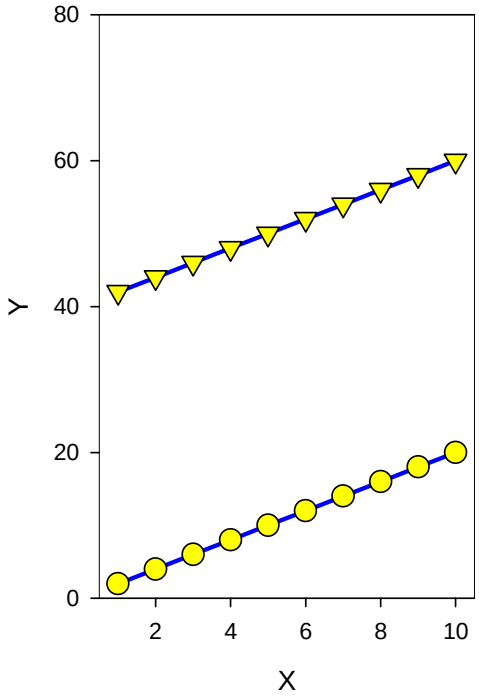
30

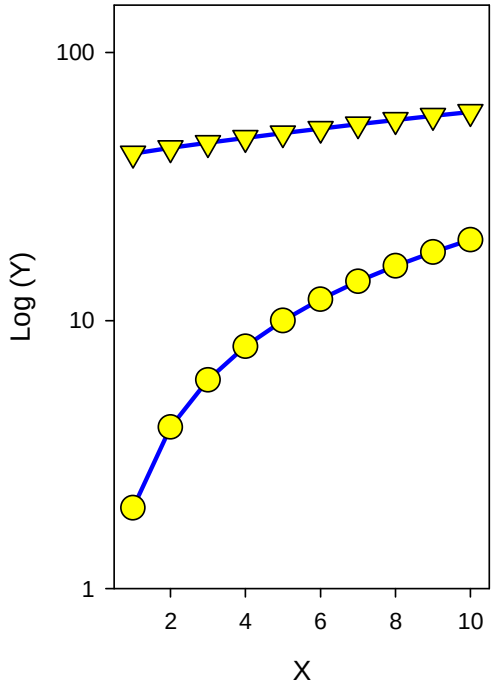
Log-binary regression

- Stata
  - glm hypertension obese, fam(bin) link(log)
  - Alternative: binreg hypertension obese
- Report relative risk
  - glm hypertension obese, fam(bin) link(log) eform
  - binreg hypertension obese, rr

31

INTERACTION and SCALE





These data are the same, but the Y variable is analysed in itself, and log-transformed.

32

Exam spring 2002 (cont.): Logistic regression

Question 2: Distance between the two first pregnancies

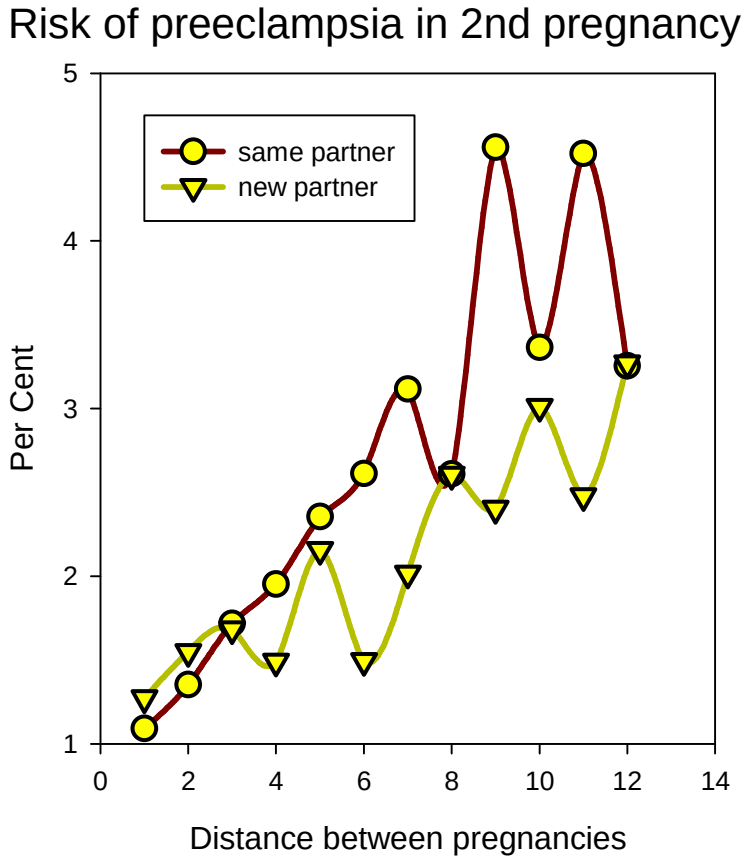
a) New partner: 8.1 years

b) Same partner: 3.2 years

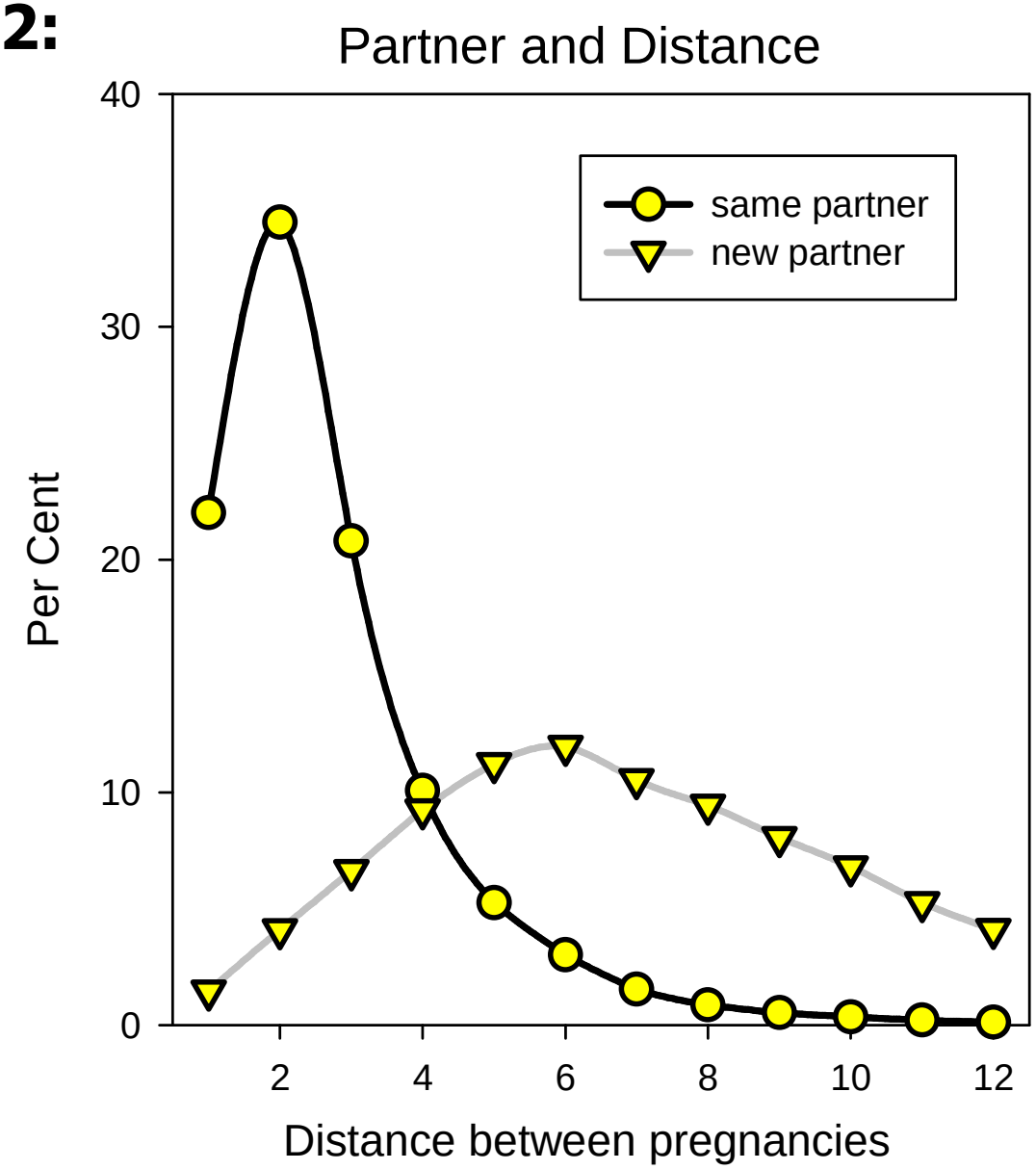
Question 3: Logistic regression,

- New partner (1), same partner (0)

- Distance between pregnancies (in whole years 1-12)



Question 2:


Partner and Distance
same partner
new partner
Per Cent
40
30
20
10
0
2
4
6
8
10
12
Distance between pregnancies

Results, logistic regression

	β	SE(β)
• New partner	-0.322	0.135
• Distance between pregnancies	0.121	0.013

Exercise 3 a) og c)

Calculate OR (with 95% C.I.) for effect of change of partner, adjusted for distance between pregnancies.

Answer: a) OR=exp(-0.322)=0.725

b) 95% C.I.: exp(-0.322±1.96*0.135) = exp(-0.322±0.265)
= (0.56-0.94)

PS: OR=exp(0.121)=1.13 is interpreted as the increase in occurrence of preeclampsia by one year increase in distance between pregnancies.

Preeclampsia by new partner, interval and smoking (New data)

	Unadjusted	Adjusted 1	Adjusted 2	Ajusted 3	Adjusted 4
New Partner	1.4 (1.3-1.5)	1.6 (1.4-1.7)	0.78 (.69 - .88)	0.84 (.74 - .96)	0.85 (.74 - .98)
Smoke	-	0.63(.55-.73)	-	0.60 (.52-.70)	0.61 (0.53-0.70)
Interval (*)	-	-	1.11 (1.10-1.13)	1.12 (1.10-1.13)	1.12 (1.10-1.13)
Maternal age	-	-	-	-	(see next)

(*) Per year

36

Categorical variable: Maternal age, 2nd pregn.

		Frequency	Parameter coding (1)	(2)	(3)	(4)	(5)
Mat. age, 2nd pregn.	1	448	1,000	,000	,000	,000	,000
	2	16765	,000	1,000	,000	,000	,000
	3	55023	,000	,000	,000	,000	,000
	4	61951	,000	,000	1,000	,000	,000
	5	22178	,000	,000	,000	1,000	,000
	6	2937	,000	,000	,000	,000	1,000

P-value			
0,76	OR	95% C.I.	
0,21	1,57	0,78	3,17
0,616	0,96	0,83	1,12
-	1,00	reference	
0,683	1,02	0,93	1,12
0,6	1,03	0,91	1,17
0,531	1,08	0,85	1,39

37

The assumption of linear relation between preeclampsia and interval between pregnancies?

We can test whether this linear relation can be justified by subtracting the goodness of fit between two models.

1) Linear trend, x=1,2, …,10 years (βx)

2) Use k=9 dummies (time 1,2,…,10 years, one ref.)

Test will have a χ^2_8-distribution (9-1)

Number of d.f. Model 1: n-2; Model 2: n-(1+k)

38

Relative Risk

- Odds Ratio (OR) is used as measure of 'effect' in retrospective studies (case-control studies).
- OR can also be used in prospective studies, but in these studies it is more correct to use the more intuitive Relative Risk (RR). RR is a measures for the relation between the risk for disease between two groups
-

	Disease		
Exposure	Yes	No	Total
+	a	b	a+b
-	c	d	c+d
Total	a+c	b+d	n

With the notations as in the table, RR is defined as:

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Relative Risk (RR)

- For the estimator RR we can calculate an asymptotic Standard Error:

$$SE(\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}$$

- and again, we can find a 95% C.I. for RR by assuming that ln(RR) is normally distributed.

Relative Risk (RR) models

- $\ln(p_x) = \alpha + \beta x$
- Why is this a RR model?
  - Assume x is binary
  - $\ln(p_1) = \alpha + \beta \times 1 = \alpha + \beta$
  - $\ln(p_0) = \alpha + \beta \times 0 = \alpha$
  - $\ln(\mathrm{RR}) = \ln\left(\frac{p_1}{p_0}\right)$

$$= \ln(p_1) - \ln(p_0) = \beta$$

41

Pairwise data

- Case-control data are often sampled pair-wise. For each case, a control is selected with identical values for typical factors like sex and age.
- The ’study units’ will therefore not be the individual, but rather the pairs of ’case-controls’.

42

Kjuus-data (cont.)

The material we have presented above is in fact such a matched material (therefore n1=n2=176). These 176 case-control pairs distribute (according to asbestos exposure) as follows:

		Case		
		Not exposed	Exposed	Total
Control	Not exp.	38 (e)	64 (f)	102
	Exposed	33 (g)	41 (h)	74
	Total	71	105	176

43

Kjuus-data (cont.)

We can extend to several factors, for instance asbestus (2 levels) and smoking (3 levels)

		Case		
		R1	R2	R3
Control	R1			
	R2			
	R3			
	Total			

44

Test statistic for matched data in a 2x2 table

- Concordant pairs (represented with e=38 and h=41) will not provide evidence for the relation between exposure and case/control-status.
- Only the discordant pairs (f=64 and g=33) contribute, and under the null hypotesis we will expect similar numbers of expose/not-exposed pairs as no-exposed/exposed pairs.
- Given the discription above, we will expect (f+g)/2 in each of these two cellls, and using the general formula for chi square evaluation in 2x2 tables we get Mc Nemar's test:

45

McNemar's test

- We can show that:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

$$\chi^2 = (f - g)^2 / (f + g)$$

$$\chi^2 = (64 - 33)^2 / (64 + 33) = 9.91$$

- The test has 1 degree of freedom

Using the chi square table we find that p<0.01, but p>0.001

46

Matched data (cont.)

- The effect measure, odds ratio (OR), will in this situation end up as a simple statistic, **OR=f/g.**
- In our material: OR=64/33=1.94.
- The relative similar results, given the two very different models and test statistics, indicate that the assumed dependence due to the matching of case with its control is weak (at least for the effect of asbestos).
- However, we can not know that apriori, therefore it will be obligatory to account for the matching in the analytical design.

Fisher's exact test

- When one of the expected values is smaller than 5, we can not use stanadard χ^2-tests for 2x2 tables. The solution is Fisher's exact test (...or a similar method).
- The test is based on the strategy of calculating the probability of all possible tables we can construct, given the same marginals as in the observed table.
- A p-value is calculated using the usual reasoning: What is the probability of observing this 'extreme' table or an even more extreme table under the null hypothesis?

Fisher's exact test

Example: Altman (s. 254) Spectacle wearers among juvenile delinquents who failes a vision test (Weindling et al.,1986)

We observe:

		Juvenile delinquents	Non-delinquents	Total
Spectacle wearers	Yes	1	5	6
	No	8	2	10
	Total	9	7	16

49

Assumption for standard χ^2-tests

- 80% of the cells must have expected values >=5

- In 2x2 tables, all the cells must have expected values >=5

50

Fisher's exact test

- In the 2x2 table on juveniles, we can not use the standard χ^2-test since the condition above is not present (Check this!).
- What is the probability for observing an extreme table, or a more extreme than the one we have observed, given that the null hypothesis is valid?

51

List of 'extreme' tables:

All possible tables we can observe, given the same row and column marginals:

1		**2**		3		4	
0	6	**1**	**5**	2	4	3	3
9	1	**8**	**2**	7	3	6	4

5		6		7	
4	2	5	1	6	0
5	5	4	6	3	7

Fisher's exact test

- Of these tables, we have observed table no 2, and there is one table that is more extreme: no 1.

  (... in the same tail)
- The probability for each of these tables occurring (given the null hypothesis) can be caluclated.
- Stata
  - tabulate delinquent spectacle, exact

Fisher's exact test

- The probabilities to observe the two first tables are 0.00087 and 0.02360, respectively.
- This gives us a p-value:

  p=2* (0.00087+0.02360)=0.049.
- This 'doubling' is discussed, and one may rather calculate the p-value based on the tables 1, 2 and 7. This gives us a 'two-sided' p-value=0.035.
- A χ^2-test would yield a two sided p-value of 0.013

Higher order tables

RxC-tables

- We can straight forward generalize much of the above to RxC tables, where R and C both can be larger than 2.
- Expected values are calculated using the same principles as for 2x2 tables.
- Chi square tests have (R-1) * (C-1) degrees of freedom.