

Kategoriske data - del 2 (Rosner, 10.3 - 10.7)

22 januar 2008

Stian Lydersen

1

Kategoriske data, del I:

- To behandlinger og to utfall. (generelt: 2 variable, 2 verdier). 2x2 tabell.
- Uavhengige observasjoner
- Sammenheng mellom behandling og utfall? (er de to variablene assosiert?)
- Sml to binomiske andeler (10.2.1), Pearson's χ^2 (10.2.2 – 10.2.4) eller Fisher's eksakte test (10.3)

2

Kategoriske data, del II:

- 2x2 tabell, parede data. McNemar's test (10.4)
- RxC tabell, uavh. obs. Pearson's χ^2 (10.6)
- 2xk tabell, k ordnede kategorier, test for trend (10.6.2)
- Litt om utvalgsstørrelse og styrkeberegninger (10.6)
- Litt om "Goodness of fit" (10.7)

3

2x2 tabell, parede data (Mc Nemar's test)

Avsnitt 10.4

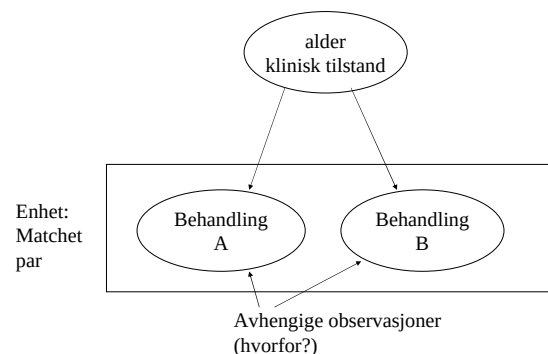
4

Table 10.13. 2x2 table based on 1242 patients

Count		Outcome		Total
		Survive for 5 years	Die within 5 years	
Treatment	A	526	95	621
	B	515	106	621
Total		1041	201	1242

Feil utgangspunkt: Ikke uavhengige observasjoner!

5



6

Table 10.14. Matched pairs as sampling unit, 621 matched pairs.

Count		Trt B patient		Total
		survive 5 years	die within 5 years	
Trt A patient	survive 5 years	510	16	526
	die within 5 years	5	90	95
Total		515	106	621

7

Typer par

- Concordant pair (overensstemmende par): Behandling A og B gir samme resultat
- Discordant pair (uoverensstemmende par): Behandling A og B gir forskjellig resultat. Antall n_D
- Type A uoverensstemmende par: Beh. A gir “hendelsen” og beh. B gir ikke “hendelsen”. Antall n_A

8

Table 10.14. Matched pairs as sampling unit, 621 matched pairs.

Count		Trt B patient		Total
		survive 5 years	die within 5 years	
Trt A patient	survive 5 years	510	16	526
	die within 5 years	5	90	95
Total		515	106	621

Type A

Type B

9

Hypotesetesting: McNemar's test

- $p = \Pr(\text{Et uoverensstemmende par er type A})$
- Da er $n_A \sim \text{bin}(n_D, p)$ (gitt n_D)
- $H_0: p=1/2$ versus $H_1: p \neq 1/2$
- Under H_0 er $E(n_A) = n_D/2$
- Forkast H_0 hvis n_A avviker “mye” fra $n_D/2$

10

Beregning av p-verdi McNemar's test

- Tilnærmet: Eqn 10.12
- Eksakt: Eqn 10.13
- SPSS:
 - Legg inn data:
 - En linje per case
 - eller Data -> Weight cases
 - Analyze -> Descriptive statistics -> Crosstabs. Velg Statistics: McNemar

11

Eks 10.24 – Brystkreft. $n_D=21$, $n_A=5$

Under H_0 er $N_A \sim \text{bin}(21, 0.5)$

p-verdi = $\Pr(|N_A - 10.5| \geq |5 - 10.5|)$

Tilnærmet (Eqn 10.12): $\chi^2 = 4.76$, p-verdi = 0.0291 (OK tilnærming hvis $n_D \geq 20$)

Eksakt (Eqn 10.13) p-verdi=0.0266

12

Estimert sannsynlighet for å overleve 5 år:

$$A: \frac{526}{621} = 0.847, \quad B: \frac{515}{621} = 0.829$$

$$\text{Differanse: } 0.847 - 0.829 = 0.018$$

95% konfidensintervall:

Se f.eks Agresti, A: An Introduction to Categorical data Analysis, 2nd edition, Wiley 2007. Page 244-247.)

$$0.018 \pm 1.96 \sqrt{(5+16) - (5-16)^2 / 621} / 621 = 0.018 \pm 0.014$$

Dvs 0.004 til 0.032

13

Utvalgsstørrelse og teststyrke

Avsnitt 10.5

14

10.5 Utvalgsstørrelse og teststyrke

- Uavhengige observasjoner i 2x2 tabell (Pearson's χ^2 eller Fischer's eksakt)
 - Utvalgsstørrelse n_1, n_2 ved gitt $p_1, p_2, \alpha, 1-\beta, k(=n_2/n_1)$: Eqn 10.14
 - Teststyrke $1-\beta$ ved gitt $p_1, p_2, \alpha, n_1, n_2$: Eqn 10.15
 - Ikke-perfekt compliance: Eqn 10.21
- Matchede par (McNemar's test):
 - Utvalgsstørrelse Eqn 10.16
 - Teststyrke: Eqn 10.17

15

Ikke-ordnet rxc tabell.

Avsnitt 10.6

16

Tabell 10.18

		age group					Total
		<20	20-24	25-29	30-34	>=35	
case/ control	case	320 18,4%	1206 21,4%	1011 25,9%	463 29,8%	220 35,1%	3220 23,9%
	control	1422 81,6%	4432 78,6%	2893 74,1%	1092 70,2%	406 64,9%	10245 76,1%
Total		1742 100,0%	5638 100%	3904 100%	1555 100%	626 100%	13465 100,0%

17

Hvis rader og kolonner er uavhengige
så er forventet antall i celle nr i,j

$$E_{ij} = \frac{\text{radsum} \times \text{kolonnesum}}{\text{totalsum}}$$

18

Pearson's kjikvadratobservator er definert som

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \dots + \frac{(O_{rc} - E_{rc})^2}{E_{rc}}$$

Under H_0 (rader og kolonner uavhengige) så er denne tilnærmet kjikvadratfordelt med $(r-1)(c-1)$ frihetsgrader dersom minst 80% av cellene har $E_{ij} > 5$ og alle cellene har $E_{ij} > 1$.

19

Eksempel 10.35

											sum
O	320	1206	1011	463	220	1422	4432	2893	1092	406	13465
E	416,6	1348	933,6	371,9	149,7	1325	4290	2970	1183	476,3	13465
O-E	-96,6	-142	77,4	91,1	70,3	96,6	142,3	-77,4	-91,1	-70,3	0
(O-E) ² /E	22,40	15,02	6,42	22,32	33,01	7,04	4,72	2,02	7,01	10,38	130,33

20

Ordnet 2xk tabell: Test for trend.

Avsnitt 10.6

21

Cochran-Armitage testen: Kjikkvadrat-test for trend i binomiske andeler

- Sett en score S for hver gruppe:
 - Numerisk egenskap ved gruppen, eller
 - Gruppenr
- Beregn størrelsen $X_1^2 = A^2/B$ ihht likning 10.24.
- Under H_0 er $X_1^2 \sim$ kjikkvadratfordelt med 1 fr.gr.

22

Cochran-Armitage test:

- $X_1^2 = nr^2$, hvor r er Pearsons korrelasjonskoeffisient mellom score og gruppe, og n er totalt antall
- Ekvivalent med score-test for score i logistisk regresjon

23

Mer generell test: Linear-by-linear test for association i dobbelt ordnet rxc tabell

- $X_1^2 = (n-1)r^2$, hvor r er Pearsons korrelasjonskoeffisient mellom score og gruppenr, og n er totalt antall
- Dette er "Linear-by-Linear Association" i SPSS. Analyze -> Descriptive statistics -> Crosstabs
- Tilnærmet lik test for trend i binomiske andeler.

24

Eks 10.37

Persons	Agegroup	Casecont
320	1	0
1422	1	1
1206	2	0
4432	2	1
1011	3	0
2893	3	1
463	4	0
1092	4	1
220	5	0
406	5	1

Hvis data ligger slik i SPSS Data Editor:
Data -> Weight Cases -> Weight cases by: Persons

25

Eks 10.37 - SPSS

		age group					Total
		<20	20-24	25-29	30-34	>=35	
case/control	case	Count	320	1206	1011	463	220
		Expected	416,6	1348,3	933,6	371,9	3220,0
		%	18,4%	21,4%	25,9%	29,8%	35,1%
control	Count	1422	4432	2893	1092	406	10245
	Expected	1325,4	4289,7	2970	1183	476,3	10245
	%	81,6%	78,6%	74,1%	70,2%	64,9%	76,1%
Total	Count	1742	5638	3904	1555	626	13465
	Expected	1742,0	5638,0	3904	1555	626,0	13465
	%	100%	100%	100%	100%	100%	100,0%

26

Eks 10.37 - SPSS

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	130,338 ^a	4	,000
Likelihood Ratio	127,385	4	,000
Linear-by-Linear Association	129,002	1	,000
N of Valid Cases	13465		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 149,70.

27

SPSS: Logistic regression

Variables not in the Equation

Step 0	Variables	Score	df	Sig.
	agegroup	129,012	1	,000
	Overall Statistics	129,012	1	,000

Identisk med Cochran-Armitage test for trend

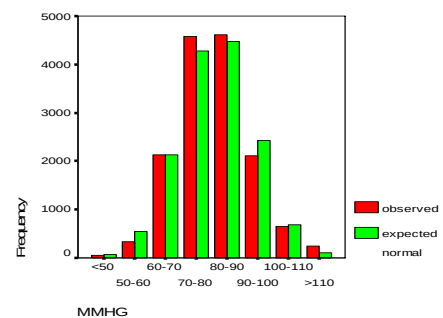
28

Chi square goodness-of-fit test

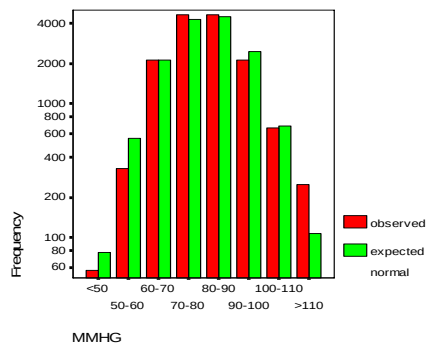
Avsnitt 10.7

29

Data fra tabell 10.21: Passer data med normalfordelingen?



30



31

Eksempel 10.41

gruppe	1	2	3	4	5	6	7	8	sum
O	57	330	2132	4584	4604	2119	659	251	14736
E	78	547	2127	4283	4479	2431	684	107	14736
O-E	-21	-217	5	301	125	-312	-25	144	0
(O-E) ² /E	5,65	86,09	0,01	21,15	3,49	40,04	0,91	193,79	351,14

32