NTNU
Det skapende universitet

Rosner, kap 13: Design and Analysis Techniques for Epidemiologic studies

Medisinsk statistikk del II
25 mars 2009

Stian Lydersen

Rosner, Chapter 13:

1. Common study designs in epidemiology
2. Measures of effect for categorical data
3. Assessment of disease-exposure relationship, controlling for confounding variables:
   - ~~Mantel-Haenzel methodology~~
   - Logistic regression (Neste forelesning)
4. Meta-analysis (KLMED 8006: Anvendt medisinsk statistikk)
5. ~~Alternative study designs~~
6. ~~Other techniques~~
   - ~~clustered binary data~~
   - ~~measurement error~~
7. Missing data (KLMED 8006: Anvendt medisinsk statistikk)

Epidemiology

- The study of how often diseases occur in different groups of people, and why. (Coggon, D, Rose, G, Barker, DJP: Epidemiology for the uninitiated, 4th ed, BMJ Publications, 1997)

Epidemiology

- A study of health and disease in populations, including aetiology[1], natural course and treatments. Clinical trials are considered by many to be one of the methods of epidemiology (Simon Day: Dictionary for Clinical Trials, 2nd ed, Wiley, 2007)

[1] Scientific account of the causes of any disease

Tabell 10.2

Count

		hjerteinfarkt innen 3 år		
		ja	nei	Total
bruker p-pille	ja	13	4987	5000
	nei	7	9993	10000
Total		20	14980	15000

Tabell 13.1

		Disease: Yes	Disease: No	
Exposure	Yes	a	b	$a+b=n_1$
	No	c	d	$c+d=n_2$
		$a+c=m_1$	$b+d=m_2$	

7

Forskjellige studiedesign

- Observasjonelle studier:
  - Prospektiv studie (kohort studie)
  - Retrospektiv studie (kasus-kontroll studie)
  - Tverrsnittstudie
- Eksperimentelle (intervensjons) studier
  - Randomiserte kontrollerte forsøk

Viktig å ta hensyn til bias / konfundering i observasjonelle studier!

8

Def. 13.9 Konfundering (confounding)

A **confounding variable** is a variable that is associated with both the disease and the exposure variable. Such a variable must usually be controlled [1] for before looking at the disease-exposure relationship.

[1] F.eks ved logistisk regresjon

9

Def. 13.1 Prospektiv studie (kohort studie)

A **prospective study** is a study in which a group of disease-free individuals are identified in one point in time and are followed over a period of time … The development of disease … is then related to …variables measured at baseline, generally referred to as the exposure variables. The study population …is often referred to as a **cohort**. …

10

Def. 13.2
Retrospektiv studie (kasus kontroll studie)

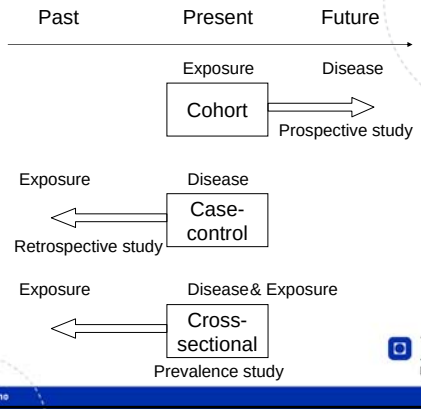
A **retrospective study** is a study in which two groups of individuals are initially identified: (1) a group that has the disease under study (the cases) and (2) a group that does not have the disease under study (the controls). …relate their *prior* health habits to their current disease status.

11

Def 13.3 Tverrsnitts-studie

A **cross-sectional study** is a study in which a study population is ascertained at one point in time. All the individuals … are asked about their current disease status and their current or past exposure status. … sometimes called a **prevalence study**, because the prevalence of disease … is compared between exposed and unexposed individuals. This contrasts to a prospective study, where one is interested in the incidence rather than the prevalence of disease.

12



13

Det finnes unntak

- Retrospektive kohortstudier
- Prospektive kasus-kontroll studier (sjeldne)

14

Tabell 10.2

Count

		hjerteinfarkt innen 3 år		
		ja	nei	Total
bruker p-pille	ja	13	4987	5000
	nei	7	9993	10000
Total		20	14980	15000

15

Tabell 10.1

Count

		alder v første fødsel		
		>= 30 år	< 30 år	Total
status	kasus (brystkeft)	683	2537	3220
	kontroll (ikke brystkreft)	1498	8747	10245
Total		2181	11284	13465

16

Def. 13.4

p_1 = sannsynlighet for at en eksponert person blir syk

p_2 = sannsynlighet for at en ueksponert person blir syk

Risikodifferense: $p_1 - p_2$

Risikoratio (relativ risiko): p_1 / p_2

Mer generelt:
p_1 , p_2 = sanns. for den aktuelle hendelsen i gruppe 1 og 2

17

Repetisjon fra kapitel 10

Tre metoder for analyse av 2x2 tabeller.

- To-utvalgstest for binomiske andeler:
  - Konfidensintervall for p_1-p_2 kan også beregnes (Avsnitt 13.3)
- Pearson's kjikvadrattest.
  - Generaliserbar til rxc tabeller (Avsnitt 10.6)
- Fisher's eksakte test.
  - Garanterer at reelt signifikansnivå ≤ nominelt signifikansnivå α
  - Men har noe lavere styrke enn asymptotisk metode uten kontinuitetskorreksjon

18

Repetisjon fra kapitel 10

To grupper av størrelse n_1 og n_2.

Observerer

$X_1 \sim \text{bin}(n_1, p_1)$ og $X_2 \sim \text{bin}(n_2, p_2)$

H_0: $p_1=p_2$ (eller $p_1-p_2=0$) mot H_1: $p_1 \neq p_2$.

Estimatorer for p_1 og p_2:

$$\hat{p}_1 = \frac{X_1}{n_1} \quad \text{og} \quad \hat{p}_2 = \frac{X_2}{n_2}$$

Forkaster H_0 hvis $\hat{p}_1 - \hat{p}_2$ avviker "mye" fra 0.

19

Repetisjon fra kapitel 10

Under H_0 er $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{Var(\hat{p}_1 - \hat{p}_2)}}$ tilnærmet standard normalfordelt.

$$Var(\hat{p}_1 - \hat{p}_2) \overset{\text{pga uavh.}}{=} Var(\hat{p}_1) + (-1)^2 Var(\hat{p}_2)$$
$$= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \overset{\text{Under } H_0}{=} \left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)$$

Dermed fås

$$z \approx \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} \quad \text{hvor} \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

NTNU Det skapende universitet

www.ntnu.no

20

Generelt:

$z = \dfrac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{Var(\hat{p}_1 - \hat{p}_2)}}$ er tilnærmet standard normalfordelt.

$$Var(\hat{p}_1 - \hat{p}_2) \overset{\text{pga uavh.}}{=} Var(\hat{p}_1) + (-1)^2 Var(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Dermed fås

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{1}{n_1} p_1(1-p_1) + \frac{1}{n_2} p_2(1-p_2)}} \approx \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{1}{n_1} \hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2} \hat{p}_2(1-\hat{p}_2)}}$$

NTNU Det skapende universitet

www.ntnu.no

21

Så

$$\Pr(-z_{1-\alpha/2} \le z \le z_{1-\alpha/2}) \approx 1-\alpha$$

$$\Pr\left(-z_{1-\alpha/2} \le \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{1}{n_1} \hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2} \hat{p}_2(1-\hat{p}_2)}} \le z_{1-\alpha/2}\right) \approx 1-\alpha$$

Løser den mhp $p_1 - p_2$ og får et tilnærmet 1-α konfidensintervall for $p_1 - p_2$

NTNU Det skapende universitet

www.ntnu.no

22

Tilnærmet 1-α konfidensintervall for $p_1 - p_2$ (Wald intervallet)

$$\hat{p}_1 - \hat{p}_2 \mp z_{1-\alpha/2} \sqrt{\frac{1}{n_1} \hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2} \hat{p}_2(1-\hat{p}_2)}$$

Tilnærmingen er OK hvis $n_1\hat{p}_1(1-\hat{p}_1) \ge 5$ og $n_2\hat{p}_2(1-\hat{p}_2) \ge 5$

Eqn 13.1 s 635 (582 i 5th ed) inneholder også en omdiskutert kontinuitetskorreksjon $\pm[1/(2n_1)+1/(2n_2)]$ som er ekvivalent med Yates' kontunitetskorreksjon i Pearsons χ^2 observator for 2x2 tabeller.

NTNU Det skapende universitet

www.ntnu.no

23

Eks 13.5 (tabell 10.2)

$\hat{p}_1 = 13/5000 = 0.0026, \quad \hat{p}_2 = 7/10000 = 0.0007$

95% konfidensintervall for risikodifferensen:

$$0.0026 - 0.0007 \mp 1.96 \sqrt{\frac{0.0026(1-0.0026)}{5000} + \frac{0.0007(1-0.0007)}{10000}}$$
$$= 0.0019 \mp 1.96(0.00077)$$
$$= (0.0004,\ 0.0034)$$

Merk at Rosner fikk (0.0002, 0.0033)
med den omdiskuterte kontinuitetskorreksjonen

NTNU Det skapende universitet

www.ntnu.no

24

Bedre konfidensintervall for p_1-p_2 (I):

- Det finnes bedre asymptotiske (tilnærmede) metoder enn ovennevnte. Newcombe's metode, kan lett programmeres eller beregnes f.eks med softwaren til Altman & al *Statistics with confidence 2.ed* (2000). 0.000588 til 0.00378 i eksempel 13.5.

NTNU Det skapende universitet

www.ntnu.no

25

Bedre konfidensintervall for p_1-p_2 (II):

- Det finnes eksakte metoder som garanterer at dekningssannsynligheten holder. Krever spesialsoftware. StatXact gir 0.000608 til 0.00379 i eksempel 13.5 (6 timers beregningstid på PCen!)

26

Agresti & Caffo (2000) konfidensintervall for p_1-p_2:

Beregn estimert risikodifferanse som før:

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

Legg til 1 i hver celle i 2x2 tabellen før du beregner "vanlig" asymptotisk konfidensintervall:

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2}, \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

Bedre tilnærmet konfidensintervall:

$$\tilde{p}_1 - \tilde{p}_2 \mp z_{1-\alpha/2}\sqrt{\frac{1}{n_1}\tilde{p}_1(1-\tilde{p}_1) + \frac{1}{n_2}\tilde{p}_2(1-\tilde{p}_2)}$$

27

Agresti & Caffo (2000) intervallet:

- Lett å beregne
- Gode egenskaper (dekningsgrad)
- Anbefalt i flere innføringsbøker i statistikk
- I eksempel 13.5 blir det 0.000496 til 0.00350

28

En sammenlikning

Metode	95% konfidensintervall		Anbefalt
	nedre	øvre	
Wald	0.00040	0.00340	nei
Wald med cc	0.00025	0.00325	NEI
Newcombe	0.00059	0.00378	ja
Agresti-Caffo	0.00050	0.00380	ja
Eksakt	0.00060	0.00378	ja
Asymp. eksakt	0.00062	0.00379	ja

29

Estimat for risikoratio (eqn 13.2): $\hat{RR} = \hat{p}_1 / \hat{p}_2$

1-α konfidensintervall for ln(RR):

$$\left[\ln(\hat{RR}) - z_{1-\alpha/2}\sqrt{\frac{b}{an_1} - \frac{d}{cn_2}},\quad \ln(\hat{RR}) + z_{1-\alpha/2}\sqrt{\frac{b}{an_1} - \frac{d}{cn_2}}\right]$$

1-α konfidensintervall for RR:

$$\left[e^{\ln(\hat{RR}) - z_{1-\alpha/2}\sqrt{b/(an_1)+d/(cn_2)}},\quad e^{\ln(\hat{RR}) + z_{1-\alpha/2}\sqrt{b/(an_1)+d/(cn_2)}}\right]$$

Tilnærmingen er OK hvis
Rosner: $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ og $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$
Price & Bonett, Statistics in medicine, 2008:
p_1 og p_2 mellom 0.1 og 0.9 samt $n_1 \geq 15$ og $n_2 \geq 15$

30

Eks 13.7 (Tabell 10.2)

$$\hat{p}_1 = 13/5000 = 0.0026, \quad \hat{p}_1 = 7/10000 = 0.0007$$

$$\hat{RR} = 0.0026/0.0007 = 3.71$$

$$c_1 = \ln\left(\frac{0.0026}{0.0007}\right) - 1.96\sqrt{\frac{4987}{13\times5000} + \frac{9993}{7\times10000}}$$
$$= 1.312 - 1.96\times0.4685 = 1.312 - 0.918 = 0.394$$

$$c_2 = 1.312 + 0.918 = 2.230$$

95% konfidensintervall for RR: $(e^{0.394}, e^{2.230}) = (1.48, 9.30)$

31

The Koopman (score) interval always works well.
Stata: Install "Koopman" first

```
. koopmani 13 5000 7 10000

                |          Event          |          Proportion
                |      Yes          No    |     Total        Yes
----------------+-------------------------+-------------------------
         Group1 |       13        4987    |      5000     0.0026
         Group2 |        7        9993    |     10000     0.0007
----------------+-------------------------+-------------------------
          Total |       20       14980    |     15000     0.0013
                |     Point estimate      |  [95% Conf. Interval]
----------------+-------------------------+-------------------------
     Odds Ratio |         3.714286        |  1.527265    9.032888
-------------------------------------------------------------------
```

NB! The printout erroneously writes `Odds Ratio` instead of RR

32

Hva er odds?

- Kjent begrep hos veddemålsagenter ("bookmakere")
- Odds er sannsynligheten for "utfallet" dividert på sannsynligheten for det motsatte.
- Odds = p/(1-p)
- Eksempel: Sannsynlighet 0.25 tilsvarer odds 0.25/0.75=0.33 (=1:3)
- Odds kan anta alle verdier mellom 0 og ∞.

33

Hva er Odds Ratio OR?

La p_1, p_2 være sannsynligheten i gruppe 1 og 2.

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

En tolkning av OR hvis $p_1 << 1$ og $p_2 << 1$:

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \approx \frac{p_1}{p_2} = RR$$

OR er alltid "mer ekstrem" enn RR.
(Lengre fra 1)

34

Estimat:

$$\hat{OR} = \frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)}$$

som alternativt kan skrives

$$\hat{OR} = \frac{[a/(a+b)]\times[d/(c+d)]}{[c/(c+d)]\times[b/(a+b)]} = \frac{ad}{bc}$$

35

Tabell 10.1

Count

		alder v første fødsel		
		>= 30 år	< 30 år	Total
status	kasus (brystkeft)	683	2537	3220
	kontroll (ikke brystkreft)	1498	8747	10245
Total		2181	11284	13465

36

Vi ønsker å sammenlikne $\Pr(D \mid E)$ og $\Pr(D \mid \bar{E})$

vha estimat, konfidensintervall eller hypotesetest for

risikodifferanse $\Pr(D \mid E) - \Pr(D \mid \bar{E})$

eller

relativ risiko $\dfrac{\Pr(D \mid E)}{\Pr(D \mid \bar{E})}$

eller

odds ratio $\dfrac{\Pr(D \mid E)/\Pr(\bar{D} \mid E)}{\Pr(D \mid \bar{E})/\Pr(\bar{D} \mid \bar{E})}$

37

Men i en kasus – kontroll studie observeres

$\Pr(E \mid D)$ og $\Pr(E \mid \bar{D})$.

Viktig resultat (Cornfield, 1956)

Sykdoms OR Eksponerings OR

$$\frac{\Pr(D \mid E)/\Pr(\bar{D} \mid E)}{\Pr(D \mid \bar{E})/\Pr(\bar{D} \mid \bar{E})} = \frac{\Pr(E \mid D)/\Pr(\bar{E} \mid D)}{\Pr(E \mid \bar{D})/\Pr(\bar{E} \mid \bar{D})}$$

Av interesse Observeres

38

Følgende 3 hypoteser er ekvivalente:

$$\Pr(D \mid E) - \Pr(D \mid \bar{E}) = 0$$

$$\frac{\Pr(D \mid E)}{\Pr(D \mid \bar{E})} = 1$$

$$\frac{\Pr(D \mid E)/\Pr(\bar{D} \mid E)}{\Pr(D \mid \bar{E})/\Pr(\bar{D} \mid \bar{E})} = 1$$

39

Mulig i en kasus-kontroll studie:

	estimat, konf.int	Hypotesetest (om ingen assosiasjon)
risiko diff.		OK
relativ risiko	≈OK ved lav prevalens	OK
OR	OK	OK

40

Estimat for odds ratio (eqn 13.11): $\hat{OR} = ad/bc$

Woolf 1-α konfidensintervall for ln(OR):

$$\left[\ln(\hat{OR}) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}},\quad \ln(\hat{OR}) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right]$$

1-α konfidensintervall for OR:

$$\left[e^{\ln(\hat{OR}) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}},\quad e^{\ln(\hat{OR}) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}\right]$$

Tilnærmingen er OK hvis $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ og $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$

Adjusted Woolf: Legg til ½ i alle 4 celler før beregning av konf.int.

41

Tabell 10.1

Count

		alder v første fødsel		
		>= 30 år	< 30 år	Total
status	kasus (brystkeft)	683	2537	3220
	kontroll (ikke brystkreft)	1498	8747	10245
Total		2181	11284	13465

42

Eksempel 13.11

$$\hat{OR} = \frac{683 \times 8747}{2537 \times 1498} = 1.572$$

$$\ln(1.572) \pm 1.96\sqrt{\frac{1}{683}+\frac{1}{2537}+\frac{1}{1498}+\frac{1}{8747}} = 0.452 \pm 0.101$$

dvs (1.42, 1.74)

95% konf.int. for OR: $(e^{0.352}, e^{0.553}) = (1.42, 1.74)$

Kan vi si noe om RR?