ANALYSE AV KATEGORISKE DATA-TABELLANALYSE

3. Mai 2005

Tron Anders Moger

Forrige gang:

- Snakket om *kontinuerlige* data, dvs data som måles på en kontinuerlig skala
- Hypotesetesting med t-tester evt. ikke-parametriske metoder for å se om gjennomsnitt er forskjellig i ulike grupper

I dag:

- Analyse av kategoriske data, teste om andeler er forskjellige i ulike grupper
- Krysstabeller og kji-kvadrat test

Binomisk fordeling

- Karakterisert ved:
  - Uavhengige forsøk
  - To mulige utfall, suksess og ikke-suksess
  - Sannsynligheten for suksess forandrer seg ikke fra forsøk til forsøk
  - Sannsynligheten for suksess + sannsynligheten for ikke-suksess er ALLTID lik 1
  - Finnes en matematisk formel for fordelingen som jeg ikke nevner her

Eksempel-blodtype

- Tre tilfeldig valgte personer; hvor mange har blodtype A?
  - Personene er uavhengige
  - De har enten blodtype A eller ikke A
  - Sannsynligheten for at en tilfeldig valgt person har A er 0.4, sannsynligheten for ikke A er 0.6 (1-0.4)

Binomisk fordeling forts.

- Sansynligheten for suksess kalles p
- Sannsynligheten for ikke-suksesses blir da 1- p
- Ønsker ofte å estimere p

Eksempel-gallup

- p =andelen velgere som vil stemme Høyre
- n=1500 tilfeldig valgte personer spørres
  - 1500 uavhengige forsøk
  - To mulige utfall: Høyre og ikke Høyre
- Ønsker å anslå p. Anslaget kalles et estimat. Finner at 313 vil stemme Høyre.

$$p=313/1500=0.209=20.9\%$$

- Hvor mye avviker p fra den sanne andelen (som er ukjent)?

Konfidensintervall

- Ønsker å finne et område rundt p som med stor sannsynlighet inneholder den sanne andelen
- Man forventer at p skal ha den sanne verdien, men siden man baserer seg på et utvalg får man en usikkerhet i estimatet
- Standardfeilen til p er se(p)= $\sqrt{\frac{p(1-p)}{n}}$
- se(p) sier hvor presist estimatet p er. se(p) går mot 0 når n vokser, dvs. presisjonen i p blir bedre når antall observasjoner øker.

Konfidensintervall forts.

- Et intervall på 1.96 standardavvik på hver side av forventningen dekker 95% av fordelingen (Fra tilnærming til normalfordelingen!)
- Konfidensintervallet blir da:

$$p \pm 1.96 * \sqrt{\frac{p(1-p)}{n}}$$

- Er 95% sikker på at den sanne verdien ligger i dette intervallet
- Konfidensintervall for en andel er alltid positive!

Eksempel-gallup forts.

- p var 0.209

$$se(p)=\sqrt{\frac{0.209(1-0.209)}{1500}}=0.011$$

- 95% konfidensintervall:

$$0.209\pm 1.96*0.011=(0.187,0.231)$$

$$=(18.7\%, 23.1\%)$$

Hypotesetesting – andel i en gruppe

- Vil teste om p har en eller annen bestemt verdi p_{exp}
- Starter med å sette opp en nullhypotese:

$$H_0: p=p_{exp}$$

- Nullhypotesen formuleres før data samles inn!
- Vi har at $z=\frac{p-p_{\exp}}{se(p)}$ er tilnærmet standard normalfordelt
- Siden vi nå vil teste nullhypotesen, er se(p) gitt ved $se(p)=\sqrt{\frac{p_{\exp}(1-p_{\exp})}{n}}$

To begreper fra forrige gang

- *P-verdi*: Sannsynligheten for at $p=p_{exp}$ etter at testen er gjort
- Hvis *p-verdien* er lavere enn *signifikansnivået* forkastes nullhypotesen
- *Signifikansnivået* gir den øvre sannsynligheten for å forkaste nullhypotesen hvis den er sann
- Vanlige signifikansnivåer: 5% eller 1%
- Signifikansnivå 5% betyr at man vil være minst 95% sikker på å ikke forkaste nullhypotesen hvis den er sann

Eksempel-gallup forts.

- Anta at Høyre fikk 18% oppslutning ved forrige stortingsvalg
- Vil teste om de fremdeles har 18% oppslutning med 5% sig.nivå; H_0: p=0.18

$$se(p)=\sqrt{\frac{0.18(1-0.18)}{1500}}=0.010 \quad z=\frac{0.209-0.18}{0.010}=2.9$$

- P-verdi på 0.0037 fra SPSS (ikke vist)
- Forkaster H_0 siden p-verdien er mindre enn 0.05

Sammenligning av sannsynligheter i to uavhengige grupper/analyse av tabeller

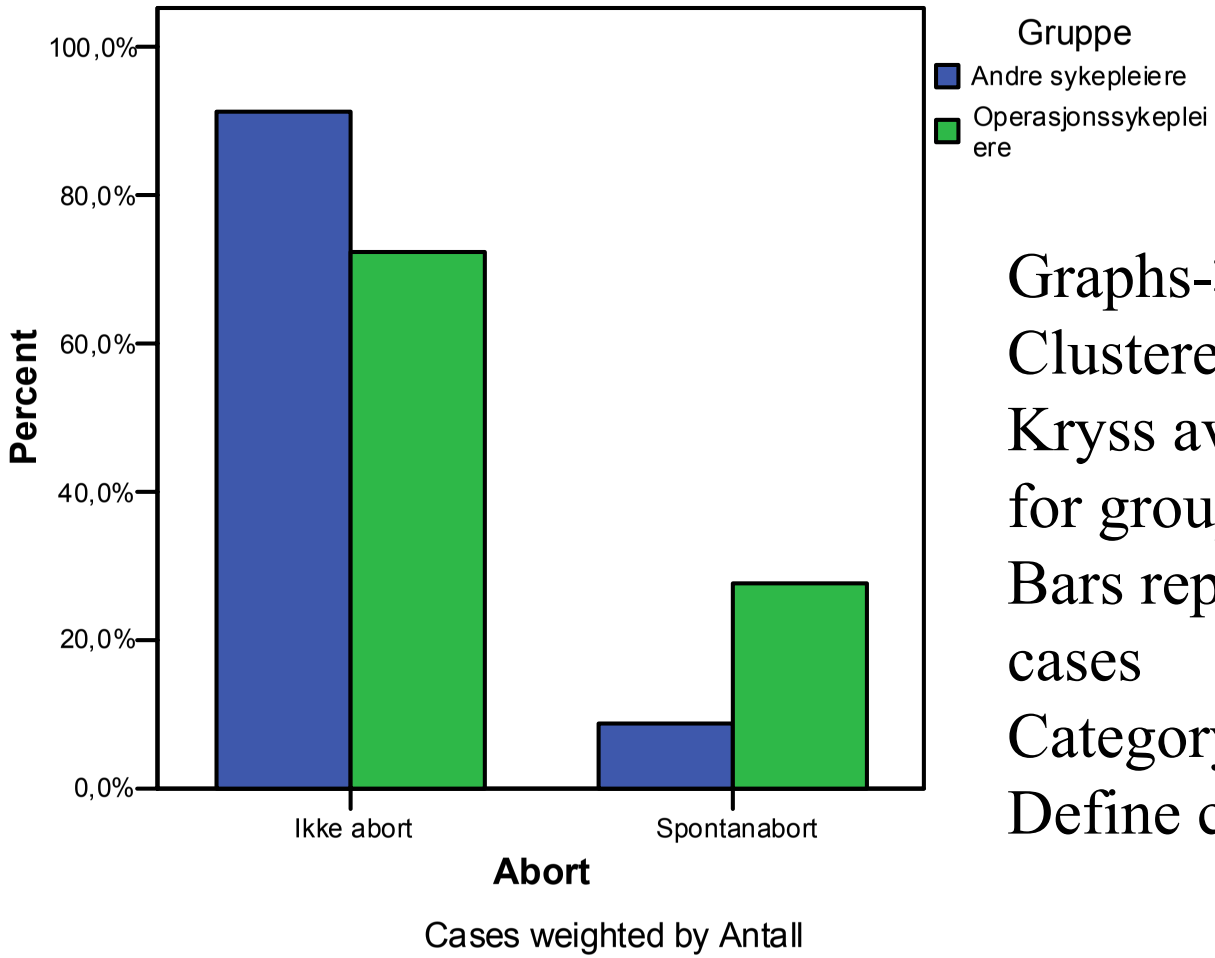
- Eksempel: Forekomst av spontane aborter blant kvinnelige sykepleiere: Opersjonssykepleiere og andre sykepleiere

	Op.sykepleiere	Andre sykepleiere
Antall intervjuet	67	92
Antall graviditeter	36	34
Antall spontanaborter	10	3
Prosent spontanaborter	27.8	8.8

Deskriptiv tabell i SPSS

- Er spontanabort hyppigere blant operasjonssykepleiere enn blant andre sykepleiere?
- Analyze->Descriptive statistics->Crosstabs
- Flytt gruppe-variabelen over i columns, og abort-variabelen over i rows
- Trykk på Cell, og merk av Percentages: Columns

Deskriptiv figur i SPSS:



Graphs->Bar->
Clustered
Kryss av Summaries for groups of cases
Bars represent: % of cases
Category axis: Abort
Define clusters by: Gruppe

To vanlige måter å uttrykke forskjeller på:

- Relativ risiko: Hva er risikoen for spontanabort blant op.sykepleiere sammenlignet med andre sykepleiere?
- Odds-ratio: Hva er oddsen for spontanabort blant op.sykepleiere sammenlignet med andre sykepleiere?

Op.sykepleier-eksempelet:

- Relativ risiko: Andelen spontanaborter blant operasjonssykepleiere dividert med tilsvarende andel i den andre gruppen

$$RR = \frac{10/36}{3/34} = 3.1$$

- Odds-ratio:
  Odds for abort blant op. sykepleiere: 10/26
  Odds for abort blant andre sykepleiere: 3/31
- Oddsratioen er forholdet mellom disse:

$$OR: \frac{10/26}{3/31} = 4.0$$

Analyse av tabeller

- Observerte hyppigheter

	Op.sykepleiere	Andre	Tot.
Ant. graviditeter m/abort	10	3	13
Ant. graviditeter u/abort	26	31	57
Total	36	34	70

- Generelt

	Eksponert	Ikke eksponert	Tot.
Syk	a	b	a+b
Frisk	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

Relativ risiko og odds-ratio

- Relativ risiko finnes slik i tabellen:

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

- Odds-ratio finnes slik:

$$\mathrm{OR}=\frac{ad}{bc}$$

Er forskjellen mellom gruppene reell?

- Kan anta uavhengighet mellom gruppene op.sykepleiere og andre sykeleiere
- Antall spontanaborter i de to gruppene er binomisk fordelte
- Vil teste H_0: $p_1=p_2$ på 5% sig.nivå. Ser på differansen p_1-p_2
- Vi har at $z=\frac{p_1-p_2}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\bar{p}(1-\bar{p})}}$ er tiln. st. normalfordelt
- Her er $\bar{p}$ gjennomsnittlig andel i de to gruppene

Eksempel-sykepleierne forts.

- p_1=0.278 p_2=0.088 n_1=36 n_2=34

$$\bar{p} = \frac{\text{Totalt antall aborter}}{\text{Totalt antall graviditeter}} = \frac{10+3}{36+34} = 0.186$$

$$z = \frac{0.278 - 0.088}{\sqrt{(\frac{1}{36} + \frac{1}{34})0.186(1-0.186)}} = 2.04$$

- P-verdi 0.0414=4.1%, forkaster H_0 på 5%-nivå
- 95% konfidensintervall for differansen p_1-p_2: $(p_1 - p_2) \pm 1.96 * \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = (0.015, 0.190)$

Konfidensintervall for relativ risiko og odds-ratio

- Et 95% konfidensintervall for relativ risiko er gitt ved:

$$(RR \times \exp(-1.96 s_{RR}), RR \times \exp(1.96 s_{RR}))$$

- Der s_{RR} er standardavviket, gitt ved

$$s_{RR} = \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{a+c} - \frac{1}{b+d}}$$

- Et 95% konfidensintervall for odds-ratio er gitt ved:

$$(OR \times \exp(-1.96 s_{OR}), OR \times \exp(1.96 s_{OR}))$$

- Der s_{OR} er gitt ved: $s_{OR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

Eksempel-sykepleierne forts.

- Konfidensintervallet for RR for sykepleierne blir da:

$$(3.1\times\exp(-1.96\sqrt{\frac{1}{10}+\frac{1}{3}-\frac{1}{10+26}-\frac{1}{3+31}}),3.1\times\exp(1.96\sqrt{\frac{1}{10}+\frac{1}{3}-\frac{1}{10+26}-\frac{1}{3+31}}))$$
$$=(0.95,10.47)$$

- Tilsvarende blir konfidensintervallet for OR for sykepleierne:

$$(4.0\times\exp(-1.96\sqrt{\frac{1}{10}+\frac{1}{3}+\frac{1}{26}+\frac{1}{31}}),4.0\times\exp(1.96\sqrt{\frac{1}{10}+\frac{1}{3}+\frac{1}{26}+\frac{1}{31}}))$$
$$=(0.99,16.08)$$

Vanligere test for forskjell i andeler: Kji-kvadrat testen

- SPSS bruker ikke tilnærming til normal-fordeling for å teste forskjeller i andeler
- Bruker derimot en mer generell test: Kji-kvadrat testen
- Gjelder generelt, ikke bare for 2*2-tabeller
- Nå følger en kort gjennomgang av bakgrunnen for testen

Forventede hyppigheter

- Hvilke tall ville vi forvente å få i tabellen hvis det ikke var noen sammenheng mellom gruppe og begivenhet (Dvs. at andelen spontanaborter blant Op.sykepleiere og andre sykepleiere var like)?
- For a: $\frac{(a+b)(a+c)}{N}$
- For b: $\frac{(a+b)(b+d)}{N}$
- For c: $\frac{(a+c)(c+d)}{N}$
- For d: $\frac{(b+d)(c+d)}{N}$

Eksempel-sykepleierne forts.

	Op. sykepl.	Andre sykepl.	Total
Graviditeter m/abort	6.69	6.31	13
Graviditeter u/abort	29.31	27.69	57
Total	36	34	70

- Vil teste om det er sammenheng mellom spontan abort og type sykepleier.
- Betegner observerte og forventede tall som hhv. O og E. Regner ut $(O-E)^2/E$ og summerer disse. Denne størrelsen blir tilnærmet kji-kvadratfordelt.
- Nullhypotesen er at andelen spontanaborter blant op.sykepleiere og andre sykepleiere er like ($p_1=p_2$)

Kji-kvadrat tester i SPSS

- Analyze->Descriptive statistics->Crosstabs
- Overfør gruppe-variabelen (op.sykepleiere) i rows og syk-variabelen (abort) i columns
- Kryss av på Expected under Cell
- Kryss av Chi-square og Risk under Statistics

Eksempel-sykepleierne forts.

- Antall frihetsgrader=
  (antall rader–1)*(antall kolonner-1)
- Lar SPSS summere $(O-E)^2/E$ for sykepleierne og får 4.15.
- Vi har én frihetsgrad, og får p-verdi 0.042 (fra SPSS).
- Forkaster H_0 på 5%-nivå

Hva hvis det er flere enn to grupper og to utfall?

- Analysen går på akkurat samme måte
- Husk at nå får man flere frihetsgrader (r utfall og c grupper: (r-1)*(c-1) frihetsgrader)
- SPSS regner ikke ut oddsratioer i det generelle tilfellet; det eneste man får ut er en p-verdi

2*2-tabell, parrede data:

- McNemars test (Finnes i crosstabs)

2*k tabell med grupper som er i en bestemt rekkefølge

- F.eks. noen som scorer bra, middels, dårlig
- Trend-test: Kun en frihetsgrad!
- Kommer automatisk i SPSS når man gjør kji-kvadrat tester