

SENIORKURSUS STATA OG BIOSTATISTIK

Aarhus Universitet juni 2011

DAGENS TEMA: SAMMENLIGNINGER

FORMIDDAG: KONTINUERTE DATA

EFTERMIDDAG: KATEGORISKE DATA

16. juni 2011

Michael Væth

1

TO STIKPRØVE PROBLEMET MED DIKOTOMT UDFALD.

Sammenligning af to uafhængige proportioner

Eksempel (skejby-cohort.dta)

På Skejby Sygehus fødte 425 gravide rygere og 1346 gravide ikke-rygere i 1994 deres første barn i graviditetsuge 37 eller senere.

18 spædbørn født af rygere og 21 spædbørn født af ikke-rygere vejede under 2500 gram ved fødslen (lav fødselsvægt, LBW).

Data: To stikprøver (rygere, ikke-rygere) med dikotomt respons med to kategorier "fødselsvægt < 2500 g" og "fødselsvægt ≥ 2500 g".

Sample	Birth weight		Total
	<2.5 kg	≥2.5 kg	
Non-smoker	21	1325	1346
Smoker	18	407	425
Total	39	1732	1771

Problem: Afhænger risikoen for lav fødselsvægt af moderens rygevaner? Hvis ja, hvordan kan vi kvantificere denne afhængighed?

16. juni 2011

Michael Væth

2

Statistisk model

To uafhængige binomial "eksperimenter": I hver gruppe er antallet af spædbørn med lav fødselsvægt – hhv. y_{NS} og y_S – en observation fra en binomial fordeling.

For ikke-rygere betragtes binomial parametrene n_{NS} og p_{NS} , og for rygere n_S og p_S .

Fortolkning af sandsynlighederne: p_{NS} er sandsynligheden for at en tilfældigt udvalgt, ikke-rygende gravid kvinde, som føder til terminen (fødselsuge ≥ 37) i 1994, får et LBW barn. p_S har en tilsvarende fortolkning.

Obs: Andre faktorer kan være associeret med risikoen for at føde et LBW barn, men sådanne faktorer tages der ikke hensyn til i modellen. Sandsynlighederne refererer til en population hvor disse faktorer er fordelt på samme måde som i stikprøven.

Model antagelser:

For hver stikprøve antages "uafhængighed", "to udfald", og "samme sandsynlighed".

16. juni 2011

Michael Væth

3

Estimation

For hver stikprøve estimeres sandsynligheden ved den tilsvarende relative hyppighed, dvs.

$$\hat{p}_{NS} = \frac{y_{NS}}{n_{NS}} = \frac{21}{1346} = 0.0156 \quad \hat{p}_S = \frac{y_S}{n_S} = \frac{18}{425} = 0.0424$$

Associationsmål

Associationen mellem rygevaner og LBW i stikprøverne kan estimeres med en **risiko differens** $\hat{p}_S - \hat{p}_{NS} = 0.0268$.

Alternative associationsmål er **relative risiko (RR)** og **odds ratio (OR)**

$$RR = \frac{p_S}{p_{NS}} \quad \text{og} \quad OR = \frac{p_S}{1-p_S} \bigg/ \frac{p_{NS}}{1-p_{NS}} = \frac{p_S(1-p_{NS})}{p_{NS}(1-p_S)}$$

Statistisk teori (middelværdi og varians af uafhængige binomialfordelinger) giver

$$E(\hat{p}_S - \hat{p}_{NS}) = p_S - p_{NS}$$

$$Var(\hat{p}_S - \hat{p}_{NS}) = p_S(1-p_S)/n_S + p_{NS}(1-p_{NS})/n_{NS}$$

16. juni 2011

Michael Væth

4

Hypotese

Risikoen for et LBW barn afhænger ikke af moderens rygevaner:

$$H: p_{NS} = p_S = p$$

Bemærk: Hypotesen formuleres som "ingen forskel". Hvis vi kan forkaste hypotesen, kan vi konkludere at den tilfældige stikprøve variation ikke kan forklare den observerede forskel.

Teststørrelse

Hvis hypotesen er sand vil

$$E(\hat{p}_S - \hat{p}_{NS}) = 0$$
$$Var(\hat{p}_S - \hat{p}_{NS}) = p(1-p) \left(\frac{1}{n_S} + \frac{1}{n_{NS}} \right)$$

Den fælles værdi af sandsynligheden p estimeres ved

$$\hat{p} = \frac{y_{NS} + y_S}{n_{NS} + n_S} = 0.0220$$

16. juni 2011

Michael Væth

5

Standard fejlen af forskellen mellem estimerne kan derfor estimeres ved

$$se(\hat{p}_S - \hat{p}_{NS}) = \sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_{NS}} + \frac{1}{n_S} \right)} = 0.0082$$

Hvis hypotesen er sand og stikprøverne ikke er alt for små, kan den tilfældige variation af forskellen $\hat{p}_S - \hat{p}_{NS}$, med god tilnærmelse beskrives med en **normal fordeling** med samme **middelværdi og varians**, så teststørrelsen

$$z = \frac{\hat{p}_S - \hat{p}_{NS}}{se(\hat{p}_S - \hat{p}_{NS})}$$

kan vurderes som en observation fra en standard normal fordeling.

Vi får

$$z = \frac{0.0424 - 0.0156}{0.0082} = 3.276$$

så p -værdien bliver $p\text{-value} = 2 \cdot P\{Z \geq 3.276\} = 0.0011$. Denne værdi tyder på at den observerede forskel mellem de to relative hyppigheder ikke kan betragtes som et tilfældigt fund.

16. juni 2011

Michael Væth

6

Sikkerhedsinterval for en risiko differens

Approksimative 95% sikkerhedsgrænser for risiko differensen fås som

$$(\hat{p}_S - \hat{p}_{NS}) \pm 1.96 \cdot se(\hat{p}_S - \hat{p}_{NS})$$

Hvor standard fejlen estimeres ved

$$se(\hat{p}_{NS} - \hat{p}_S) = \sqrt{\hat{p}_{NS}(1-\hat{p}_{NS})/n_{NS} + \hat{p}_S(1-\hat{p}_S)/n_S}$$

Indsættes de estimerede værdier fås

$$se(\hat{p}_{NS} - \hat{p}_S) = \sqrt{0.0156 \cdot (1 - 0.0156)/1346 + 0.0424 \cdot (1 - 0.0424)/425}$$
$$= 0.0103$$

så sikkerhedsgrænserne bliver

$$(0.0424 - 0.0156) \pm 1.96 \cdot 0.0103 = 0.0268 \pm 0.0203$$

dvs.

$$0.0065 < p_S - p_{NS} < 0.0470$$

16. juni 2011

Michael Væth

7

Sammenhængen mellem sikkerhedsintervaller og teststørrelser

Per definition vil et 95% sikkerhedsinterval indeholde de værdier af den betragtede parametre som ikke er statistisk signifikante.

Vi vil derfor forvente at

- p -værdien er større end 5% hvis den betragtede parameterværdi ligger inden i sikkerhedsintervallet
- p -værdien er mindre end 5% hvis parameterværdien ligger uden for sikkerhedsintervallet.

Dette er korrekt, hvis det er den **samme** test procedure, som bruges til beregning af teststørrelse og sikkerhedsinterval.

Ovenfor brugte vi to, en smule forskellige, estimer for standard fejlen, så test procedurerne er ikke helt identiske.

Begge metoder er valide og forskellen er typisk helt ubetydelig, men inkonsistente resultater kan ikke helt udelukkes.

16. juni 2011

Michael Væth

8

STATA

I Stata findes flere kommandoer som kan bruges til sammenligning af proportioner, og mere generelt til analyse af kategoriske data. Nogle kommandoer kræver data på individ niveau, mens andre (såkaldte immediate commands) kan bruges med aggregerede data som input

Kommandoer

`prtest`

`cs` "immediate" form: `csi`

`cc` "immediate" form: `cci`

`tabodds`

`tabulate`

`tabi`

Eksempler

```
gene small1=(bweight<2500)
```

```
label define slab 0 ">=2.5 kg" 1 "<2.5 kg"
```

```
label values small1 slab
```

```
gene include=(parity==0)&(gestage>36)&(year==1994)
```

* Og så selve sammenligningen

```
prtest small1 if include==1 , by(mtobacco)
```

16. juni 2011

Michael Væth

9

Output fra `prtest`

Two-sample test of proportion

smoker: Number of obs = 425

nonsmoke: Number of obs = 1346

Variable	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
smoker	.0423529	.009769			.0232061 .0614998
nonsmoke	.0156018	.0033779			.0089812 .0222224
diff	.0267512	.0103365			.0064919 .0470104
	under Ho:	.0081654	3.28	0.001	

Ho: proportion(smoker) - proportion(nonsmoke) = diff = 0

Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
z = 3.276	z = 3.276	z = 3.276
P < z = 0.9995	P > z = 0.0011	P > z = 0.0005

OBS: For at bruge `prtest` skal udfaldet, her `small1`, være en dikotom variabel, som har to værdier 0 (failure) og 1 (success)

16. juni 2011

Michael Væth

10

Kommandoen `cs` anvendes til simple analyse af kohorte studier. **Både** eksponering og udfald **skal** være dikotome 0-1 variable.

Eksponering: 0 ~ unexposed, 1 ~ exposed, Udfald: 0~noncase, 1~case

```
gene smoke=2-mtobacco
```

```
cs small1 smoke if include==1
```

	smoke Exposed	Unexposed	Total
Cases	18	21	39
Noncases	407	1325	1732
Total	425	1346	1771
Risk	.0423529	.0156018	.0220215
	Point estimate		[95% Conf. Interval]
Risk difference	.0267512	.0064919	.0470104
Risk ratio	2.714622	1.460259	5.046483
Attr. frac. ex.	.6316246	.3151899	.8018422
Attr. frac. pop	.291519		

chi2(1) = 10.73 Pr>chi2 = 0.0011

16. juni 2011

Michael Væth

11

Det samme output fås med kommandoen på "immediate form"

```
csi 18 21 407 1325
```

Bemærk:

- Rækkefølgen af de fire tal skal være: *exposed cases, unexposed cases, exposed noncases, unexposed noncases*.
- `cs` og `csi` angiver et **ch² test**. Dette test giver præcis den samme p-værdi som ovenfor. En **χ²-fordeling er per definition** fordelingen af **kvadratet på en normalfordelt variabel**, dvs.

$$Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2_{f=1}$$

P-værdierne er derfor identiske, da $3.276^2 = 10.73$.

- `cs` og `csi` kan også angive **relative risk** og **odds ratio**, begge med sikkerhedsintervaller, hvis `or` angives som option (efter kommaet)
- Kommandoerne `cc` og `cci` kan bruges til simple analyser af cases-control studier. Både eksponering og udfald **skal** være dikotome 0-1 variable. Association beskrives med en odds ratio.

16. juni 2011

Michael Væth

12

ANALYSE AF "TYNDE" TABELLER: FISHER'S EKSakte TEST

Problem: I analysen ovenfor benyttede vi at binomialfordelingen kan approksimeres med en normalfordeling. Hvis tallene i tabellen er små, vil denne approksimation være dårlig. *Hvad gør vi så?*

Løsning: Brug et eksakt test. For en sammenligning af to proportioner hedder det eksakte test: **Fisher's exact test**.

Eksempel: Malocclusion hos børn (fra Yates, 1934)

Data:	Barnets bid		Total
	Normal	Malocclusion	
Ammet	4	16	20
Flasket	1	21	22
Total	5	37	42

Spørgsmål: Tyder data på, at der er en sammenhæng mellem "bidtype" og "fodringsmetode"?

16. juni 2011

Michael Væth

13

For at besvare dette spørgsmål ser vi på alle de tabeller med **samme marginalsomme** (dvs samme række og søjlesomme). Hvis der ikke er en association vil den observerede tabel være en tilfældigt udvalgt tabel blandt disse tabeller.

Der er præcis 6 tabeller med samme marginaler som den observerede tabel:

0	20	20	1	19	20	2	18	20
5	17	22	4	18	22	3	19	22
5	37	42	5	37	42	5	37	42
						<i>observeret tabel</i>		
3	17	20	4	16	20	5	15	20
2	20	22	1	21	22	0	22	22
5	37	42	5	37	42	5	37	42

Hver tabel er entydigt bestemt ved værdien i det **øvre venstre hjørne**. Hvis risikoen for malocclusion er den samme i de to grupper, kan sandsynligheden for en given tabel beregnes vha. kombinatorik!

16. juni 2011

Michael Væth

14

Notation:

a	b	n₁
c	d	n₂
s₁	s₂	N

Sandsynligheden for denne tabel er $\frac{s_1!s_2!m_1!m_2!}{N!a!b!c!d!}$

Dette er sandsynligheden for hyppighederne i tabellen hypotesne om ingen sammenhæng er sand og hvis marginalerne er givne. Denne fordeling kaldes **den hypergeometriske fordeling**.

Vi har

	Værdi af a	sandsynlighed
	0	0.030957
	1	0.171982
	2	0.343965
	3	0.309568
observeret tabel →	4	0.125302
	5	0.018226
	sum	1.000000

16. juni 2011

Michael Væth

15

Den **to-sidede p-værdi** er sædvanligvis defineret som sandsynligheden for alle tabeller har en sandsynlighed som er højst så stor som den observerede tabel.

Vi får $p\text{-værdi} = 0.030957 + 0.125302 + 0.018226 = 0.1745$

Obs: Af og til kan man se den to-sidede p-værdi defineret som 2 gange den ensidede p-værdi i den samme retning, som den observerede afvigelse. Det vil her give 0.2871.

Konklusion: Data strider ikke mod hypotesen om uafhængighed mellem "bidtype" og "fodringsmetode"

STATA

Fisher's exact test beregnes, hvis option **exact** specificeres i kommandoerne **cs**, **csi**, **cc** og **cci**.

Eksempler:

```
cs small1 smoke if include==1 , exact
csi 18 21 407 1325 , exact
```

16. juni 2011

Michael Væth

16

TO STIKPRØVE PROBLEMET MED DIKOTOMT UDFALD.

Sammenligning af to *parrede proportioner*

Eksempel: Vejrtrækningssymptomer hos skolebørn (Bland, p 245-248)

Data: Selv-rapporterede vejrtrækningssymptomer hos 1319 skolebørn Som 12 årige og som 14 årige. Hver gang blev barnet spurgt om var alvorligt forkølet ("severe cold"). Svarmuligheder var "ja" eller "nej".

Børnene kan klassificeres i en af 4 katgorier afhængigt af svaret som 12 årig og som 14 årig

"severe cold"		antal
alder 12	alder 14	børn
Ja	Ja	212
Ja	Nej	144
Nej	Ja	256
Nej	Nej	707

Spørgsmål: Er prævalensen af de rapporterede symptomer den samme i de to aldersgrupper?

16. juni 2011

Michael Væth

17

Alternativ præsentation af data

		alder 14		
alder 12	Ja	Nej	Total	
Ja	212	144	356	
Nej	256	707	963	
Total	468	851	1319	

Prævalence proportionen af "severe cold" som 12-årig og som 14-årig er

$$\hat{p}_{12} = \frac{356}{1319} = 0.270 \quad \hat{p}_{14} = \frac{468}{1319} = 0.355$$

Sikkerhedsintervaller for disse proportioner beregnes som sædvanligt. Vi vil nu vurdere om forskellen mellem prævalenserne (differensen) er reel eller om den kan tilskrives tilfældig variation.

Bemærk: 2x2 tabellen ligner de tidligere tabeller, men den hypotese vi vil vurdere er en anden, så det sædvanlige χ^2 test kan ikke bruges

16. juni 2011

Michael Væth

18

Model overvejelser

Vi har $n = 1319$ observationspar på formen (ja, ja) , (ja, nej) , (nej, ja) eller (nej, nej) .

Lad $p_{yy}, p_{yn}, p_{ny}, p_{nn}$ angive sandsynligheden disse udfald og lad $x_{yy}, x_{yn}, x_{ny}, x_{nn}$ angive de tilsvarende hyppigheder i stikprøven.

Vi antager at udfald fra forskellige børn er uafhængige, og at sandsynlighederne $p_{yy}, p_{yn}, p_{ny}, p_{nn}$ beskriver forholdene for alle børn (er det rimelige antagelser?).

Sandsynligheden for "ja" som 12-årig og for "ja" som 14-årig bliver

$$P(\text{severe cold som 12-årig}) = P(12) = p_{yy} + p_{yn}$$

$$P(\text{severe cold som 14-årig}) = P(14) = p_{yy} + p_{ny}$$

Den relevante hypotese er $P(12) = P(14)$, som er ækvivalent med hypotesen $p_{yn} = p_{ny}$, dvs de **discordante par** (ja, nej) og (nej, ja) er **lige sandsynlige**.

16. juni 2011

Michael Væth

19

Hypotesen om ens prævalenser er derfor ækvivalent med hypotesen **blandt de discordante par** optræder hver type lige hyppigt, dvs. med sandsynlighed 0.5.

En sammenligning af prævalenser med parrede data er derfor ækvivalent med et en-stikprøve problem for dikotome data. Antallet af (ja, nej) -svar blandt alle discordante par følger en **binomial fordeling** med

$$n = x_{yn} + x_{ny} \quad \text{and} \quad p = \frac{p_{yn}}{p_{yn} + p_{ny}}$$

Og hypotesen er her $H: p = 0.5$. Vi får

Statistisk test. Antal discordante par er 400, så

$$p\text{-værdi} = 2 \cdot P\{Y \geq 256 \mid n=400, p=0.5\} = 2.3 \cdot 10^{-8}.$$

Approksimeres binomialsandsynligheden med en normalfordeling fås

$$z = \frac{|x_{yn} - 0.5 \cdot (x_{yn} + x_{ny})| - 0.5}{\sqrt{(x_{yn} + x_{ny}) \cdot 0.5 \cdot (1 - 0.5)}} = \frac{|x_{yn} - x_{ny}| - 1}{\sqrt{x_{yn} + x_{ny}}} \quad \text{kontinuitets korrektion}$$

16. juni 2011

Michael Væth

20

Man bruger ofte χ^2 -versionen af det approksimative test og i store stikprøver droppes kontinuitetskorrektionen. Teststørrelsen, som kaldes **McNemar's test** bliver da meget nem at beregne:

$$X^2 = \frac{(x_{yn} - x_{ny})^2}{x_{yn} + x_{ny}}$$

I eksemplet fås $X^2 = 31.36$, som er stærkt signifikant.

STATA

Kommandoerne **bitest** og **bittesti**, kan bruges til at løse enstikprøve problemer for binomialfordelingen. Stata har desuden kommandoen **mcc** med "immediate" form **mccci**.

Når man bruger **mcc skal** begge udfald være scoret som 0 eller 1.

Output fra **mcc** og **mccci** indeholder både sikkerhedsinterval og p-værdier, men terminologien, som bruges i output, afspejler at kommandoerne er beregnet til analyse af matchede case-control studier.

16. juni 2011

Michael Væth

21

mccci 212 144 256 707

		Alder 14		
		Ja	Nej	
Cases	Exposed	212	144	356
	Unexposed	256	707	963
Total		468	851	1319

chi-square test uden kontinuitets korrektion

McNemar's chi2(1) = 31.36 Prob > chi2 = 0.0000
Exact McNemar significance probability = 0.0000

Proportion with factor		Prævalens som 12-årig		
Cases	.2699014	[95% Conf. Interval]		
Controls	.3548143			
difference		-.0849128	-.1150345	-.0547911
ratio		.7606838	.6910295	.837359
rel. diff.		-.1316099	-.18061	-.0826098
odds ratio		.5625	.4553919	.6925884 (exact)

16. juni 2011

Michael Væth

22

ANALYSE AF KATEGORISKE DATA

En **kontingens tabel**, eller en **kryds-tabellering**, er en tabel med de hyppigheder man får, når en stikprøve af individer eller eksperimentelle enheder klassificeres efter én, to eller flere inddelingskriterier.

Eksempel 1: antal fødsler fordelt efter fødselsmåned (skejby-cohort.dta)
2118 førstegangsfødsler i 1993 inddelt efter fødselsmåned.

Måned	Hypp.	Procent	Måned	Hypp.	Procent
Jan	183	8.64	Jul	195	9.21
Feb	146	6.89	Aug	196	9.25
Mar	178	8.40	Sep	191	9.02
Apr	152	7.18	Okt	178	8.40
Maj	169	7.98	Nov	156	7.37
Jun	194	9.16	Dec	180	8.50

Spørgsmål: Er fødslerne jævnt fordelt over året?

16. juni 2011

Michael Væth

23

Eksempel 2: Lav fødselsvægt og moderens rygevaner (skejby-cohort.dta)

Førstegangs fødende mødre i 1994, mindst 40 år gamle inddelt efter rygevaner og fødselsvægt (+/-LBW)	Ryge vaner	fødselsvægt		Total
		2500-	2499	
	ryger	407	18	425
	Ikke-ryger	1325	21	1346
	Total	1732	39	1771

Eksempel 3: Køn og ABO blodtype

512 personer Opdelt efter køn og ABO blod type.	Blodtype					Total
	Køn	A	B	AB	O	
	Mand	114	15	10	104	243
	Kvinde	121	28	16	104	269
	Total	235	43	26	208	512

Spørgsmål: Er kategoriseringerne uafhængige af hinanden??

16. juni 2011

Michael Væth

24

Eksempel 4: moderens rygevaner og barnets fødselsår (skejby-cohort.dta)

FFirst parity births classified according to mother's smoking habits and year of birth.	ryge	År			Total
	vaner	1993	1994	1995	
	ryger	461	451	399	1,311
	Ikke-ryger	1,374	1,390	1,306	4,070
	Total	1,835	1,841	1,705	5,381

Example 5: moderens alder og paritet (skejby-cohort.dta)

Fødsler i december 1993 inddelt efter moderens alder og fødselsnummer (paritet).	Mors alder	Paritet			Total
		0	1	2-	
	-24	57	13	5	75
	25-29	70	40	20	130
	30-	53	52	33	138
	Total	180	105	58	343

Spørgsmål: Er kategoriseringerne uafhængige af hinanden?

16. juni 2011

Michael Væth

25

Disse eksempler minder alle om binomial problemer, men antallet af kategorier er større end 2 for mindst et af inddelingskriterierne. I eksempel 1 er der et enkelt inddelingskriterium, i de øvrige eksempler 2.

Den statistiske model som bruges i disse eksempler, og andre tilsvarende tabeller, er en såkaldt **multinomial fordeling**, som er en simpel generalisering af en binomialfordeling til situationer med **mere end to kategorier** for udfaldet.

I anvendelser vil man beskrive data med en eller flere multinomialfordelinger og hypoteserne vil specificere en eller anden restriktion for modellens sandsynligheder.

Eksemplerne illustrerer nogle typiske spørgsmål som kan være relevante at besvare på grundlag af en krydstabel. I analysen bliver spørgsmålet omformuleret til en hypotese, som kan evalueres ved et statistisk test. Der findes en omfattende teori for analyse af krydstabeller.

16. juni 2011

Michael Væth

26

De **teststørrelser** vi beregner her har alle formen

$$X^2 = \sum_{\text{cells}} \frac{(O_i - E_i)^2}{E_i}$$

Hvor O_i = **observeret** hyppighed i tabellens celle i .
 E_i = **forventet (expected)** hyppighed i tabellens celle i .

Statistisk teori: Teststørrelser på denne form vil approksimativt følge en χ^2 -fordeling hvis hypotesen er opfyldt. Antallet af frihedsgrader er afhængig af tabellens størrelse og den hypotesen som testes.

Store værdier af teststørrelsen betyder at data er uforenelige med hypotesen.

Approximationen bliver bedre og bedre, jo større tabellens hyppigheder er.

Antallet af frihedsgrader, f , for χ^2 -fordelingen kan beregnes som

$$f = (\text{antal parametre i modellen}) - (\text{antal parametre i modellen hvis hypotesen er sand})$$

16. juni 2011

Michael Væth

27

Example 1: antal fødsler fordelt efter fødselsmåned (skejby-cohort.dta)

Lad $p_{Jan}, p_{Feb}, \dots, p_{Dec}$ betegne sandsynligheder for at en tilfældigt udvalgt gravid føder i en given måned. Det forventede antal fødsler i et givet måned er $n \cdot p_{month}$, hvor n er det samlede antal fødsler.

Den relevante hypotese er en en fuldstændig specifikation af fordelingen, nemlig at hver af året dage har samme sandsynlighed. Hvis hypotesen er sand haves

$$p_{måned} = \frac{\text{dage i måneden}}{365} \quad \text{og} \quad E_{måned} = n \cdot \frac{\text{dage i måneden}}{365}$$

Måned	Obs.	Forv.	Month	Obs.	Forv.
Jan	183	179.9	Jul	195	179.9
Feb	146	162.5	Aug	196	179.9
Mar	178	179.9	Sep	191	174.1
Apr	152	174.1	Oct	178	179.9
May	169	179.9	Nov	156	174.1
Jun	194	174.1	Dec	180	179.9

16. juni 2011

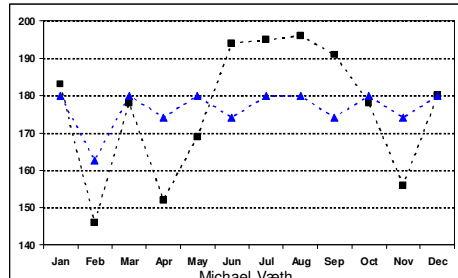
Michael Væth

28

Teststørrelse

$$X^2 = \frac{(183-179.9)^2}{179.9} + \frac{(146-162.5)^2}{162.5} + \dots + \frac{(180-179.9)^2}{179.9} = 13.74$$

Denne værdi skal vurderes i en χ^2 -fordeling med $(12-1)-0 = 11$ frihedsgrader: `display chi2tail(11, 13.74)` giver en p -værdi på 0.248, så disse data giver ikke anledning til at hypotesen forkastes. Et plot af de observerede og forventede hyppigheder, tyder dog på at der er systematiske afvigelser fra en uniform fordeling.



16. juni 2011

Michael Væth

29

TEST FOR UAFHÆNGIGHED I EN $R \times C$ TABEL

I eksempel 2,3,4, og 5 er den relevante hypotese en **hypotese om uafhængighed** mellem de to inddelingskriterier. Uafhængighed betyder at kendskab til kategorien for den ene faktor, ikke giver viden om kategorien for den anden faktor.

Hvis uafhængighedshypotesen er sand vil sandsynlighederne de **parret af udfald**, rækkekategorien og søjlekategorien, være et produkt af sandsynligheden for **rækkekategorien** og sandsynligheden for **søjlekategorien**, dvs.

$$p_{rc} = p_r \cdot p_c$$

hvor p_{rc} betegner sandsynligheden for et udfald i række r og søjle c , p_r sandsynligheden for et udfald i række r og p_c er sandsynligheden for et udfald i søjle c .

Lad y_{rc} angive hyppigheden i række r og søjle c og lad $y_{r\cdot}$ og $y_{\cdot c}$ angive hhv. Række- og søjlesummerne. Under uafhængighed kan de forventede antal da fås som

$$E_{rc} = n \cdot \hat{p}_{rc} = n \cdot \hat{p}_r \cdot \hat{p}_c = n \cdot \frac{y_{r\cdot}}{n} \cdot \frac{y_{\cdot c}}{n} = \frac{y_{r\cdot} \cdot y_{\cdot c}}{n}$$

16. juni 2011

Michael Væth

30

TEST FOR UAFHÆNGIGHED I EN $R \times C$ TABEL - fortsat

Antallet af frihedsgrader for et X^2 test af hypotesen om uafhængighed i en $R \times C$ tabel bliver

$$f = (R \cdot C - 1) - ((R-1) + (C-1)) = (R-1)(C-1)$$

Eksempel 2: Lav fødselsvægt og moderens rygevaner

rygning	2500-	-2499	Total	Forventet hypp.	
ryger	407	18	425	415.64	9.36
Ikke-ryger	1325	21	1346	1316.36	29.64
Total	1732	39	1771	$= 1732 \cdot 425 / 1771$	

Afvigelser og standardiserede residualer bliver

$O_i - E_i$		$R_i = (O_i - E_i) / \sqrt{E_i}$	
8.64	-8.64	0.42	-2.82
-8.64	8.64	-0.24	1.59

Teststørrelse: Antal frihedsgrader 1 og

$$X^2 = 0.42^2 + (-2.82)^2 + (-0.24)^2 + 1.59^2 = 10.73 \Rightarrow p\text{-value} = 0.0011$$

16. juni 2011

Michael Væth

31

Bemærk:

- De samme data blev også analyseret som et to-stikprøveproblem med dikotome data (side 2-12), og p -værdien er **identisk** med den p -værdi som tidligere blev beregnet. **Det er ikke en tilfældighed**. Man kan vise at $X^2 = z^2$ og i Stata's output fra kommandoen `cs` blev χ^2 -værdien også angivet.

Hypotesen om **uafhængighed** medfører af risikoen for et LBW barn er uafhængigt af moderens rygevaner, dvs. At risikoen for et LBW barn er **den samme i de to grupper**: rygere og ikke-rygere. Denne omformulering af uafhængighed viser at det er **præcis den samme hypotese**, som testes når data betragtes som et to-stikprøve problem med dikotome data.

- I store tabeller kan man med fordel bruge de standardiserede residualer til at undersøge om afvigelserne fra hypotesen udviser et systematisk mønster. Dette kan være vigtigt, da et χ^2 -test med mange frihedsgrader vil have relativ lav styrke.

16. juni 2011

Michael Væth

32

Analyse af RxC tabeller med STATA

Kommandoen `tabulate` viser tabellen med hyppigheder og χ^2 -testet udskrives, hvis `chi2` tilføjes som option. Tilføjes `exact` som option beregnes et eksakt test.

```
tabu mtobacco small1
    if parity==0 & year==1994 & gestage>36 , chi2
```

smoking habits of mother	small1		Total
	>=2.5 kg	<2.5 kg	
smoker	407	18	425
nonsmoker	1,325	21	1,346
Total	1,732	39	1,771

Pearson $\chi^2(1) = 10.7331$ Pr = 0.001

Stata har også en "immediate" version af `tabulate`. Kommandoen hedder `tabi`, og kan bruges hvis hyppighederne testes direkte ind på kommandolinien.

16. juni 2011

Michael Væth

33

Analyse af RxC tabeller med STATA - fortsat

Eksempel 3: køn og ABO blodtype

Køn	Blodtype				Total
	A	B	AB	O	
Mand	114	15	10	104	243
Kvinde	121	28	16	104	269
Total	235	43	26	208	512

```
tabi 114 15 10 104 \ 121 28 16 104 , chi2
```

row	col				Total
	1	2	3	4	
1	114	15	10	104	243
2	121	28	16	104	269
Total	235	43	26	208	512

Pearson $\chi^2(3) = 4.2139$ Pr = 0.239

16. juni 2011

Michael Væth

34

Analyse af RxC tabeller – diverse kommentarer

I **eksempel 4** er søjle kategorierne (fødselsår) *ordnede*, og i **eksempel 5** er både rækker (*mors alder*) og søjler (*paritet*) ordnede.

Hypotesen om uafhængighed kan også her testes med et χ^2 -test, men her er denne **teststørrelse tager ikke ordningen i betragtning**.

Søjlerne (eksempel 4) eller både rækker og søjler (eksempel 5) kan byttes rundt uden at teststørrelsen ændres.

Hvis en eller begge indeingskriterier er har ordnede kategorier vil vi typisk være interesseret i at se på afvigelser fra uafhængighed som viser en **trend med ordningen**.

χ^2 -test som inddrager ordningen vil derfor være at fortrække, da disse test er mere følsomme (har større styrke) over for afvigelser, som udviser en monoton trend i forhold til uafhængighed.

I 3- eller fler-dimensional tabeller, vil man ofte også være interesseret i hypoteser om **betinget uafhængighed**. **Mantel-Haenszel testet** som bruges i epidemiologi er et test af hypotesen om betinget uafhængighed.

16. juni 2011

Michael Væth

35