Analys av korstabeller

Analys av korstabeller hänvisar generellt till den situation, där vi betraktar flera kategoriska variabler samtidigt och vill dra slutsatser m.a.p. beroendestrukturen dem emellan. I den möjligast enkla situationen har vi då endast två variabler, vilka kan anta endast två olika värden (t. ex ja/nej, svart/vit osv.). Det gäller dock att vara försiktig när man utför dylika analyser. För att bilda oss en uppfattning om de väsentliga aspekterna för analys av korstabeller, studerar vi några exempel.

Tidningen 'The New York Times' publicerade den 11. mars, år 1979, följande korstabell:

Åtalad\Dom	Dödsdom	Annan dom	Totalt
Mörkhyad	59	2448	2507
Vithyad	72	2185	2257
Totalt	131	4633	4764

som innehåller uppgifter om 4764 mordfall, vilka hanterades i diverse domstolar i delstaten Florida under åren 1973-79. De två variablerna är den åtalades hudfärg (mörkhyad eller vithyad) och typ av dom (dödsdom eller en annan dom). Från tabellen kan vi uppskatta andelen mörk- respektive vithyade, som blev dömda till döden (2.4% respektive 3.2%).

Ett statistiskt test, som allmänt används i samband med analys av korstabeller, är ett sk. χ^2-test (uttalas 'khai'-square). Namnet kommer från den fördelning testkvantiteten följer approximativt under nollhypotesen. Olika former av χ^2-test används även generellt då man vill testa hur bra anpassning man får för en modell. Testet bygger på mängden diskrepans mellan observationerna och en hypotes, vilken i detta sammanhang motsvarar antangandet om ett visst oberoende mellan kategoriska variabler. I tabellen ovan har vi två variabler och kan bilda en nollhypotes: H_0 : 'Den åtalades hudfärg och typen av dom är oberoende av varandra'. Mothypotesen kunde

då vara t. ex. H_1 : 'Det finns någon form av samband mellan den åtalades hudfärg och typen av dom'.

Om vi betraktar noggrannare informationen i tabellen, identifierar vi följande parametrar under mothypotesen (okända kvantiteter som representerar variationen inom populationen av rättsfall gällande mord i Florida):

P(Hudfärgen är mörk & Domen är en dödsdom) = P_{11}
P(Hudfärgen är vit & Domen är en dödsdom) = P_{21}
P(Hudfärgen är mörk & Domen är ej en dödsdom) = P_{12}
P(Hudfärgen är vit & Domen är ej en dödsdom) = P_{22}

Dessa fyra parametrar är alltså sannolikheter för att ett slumpmässigt plockat rättsfall ur den givna populationen skulle ha de motsvarande egenskaperna. Eftersom sannolikheterna måste summera upp till ett, är de fyra parametrarnas möjliga värden bundna till varandra enligt $P_{11}+P_{21}+P_{12}+P_{22}=1$. Dvs. modellen har i själva verket endast tre fria parametrar. Vi antar dessutom att alla dessa sannolikheter är större än noll, dvs. $P_{ij}>0, i=1,2; j=1,2$.

Under en nollhypotes om oberoende mellan hudfärgen och typen av dom, kan variablerna betraktas skilt från varandra. Detta leder till följande parametrar:

P(Hudfärgen är mörk) = 1-P(Hudfärgen är vit) = P_1
P(Domen är en dödsdom) = 1-P(Domen är ej en dödsdom) = P_2

I motsats till den tidigare modellen, har vi nu endast två parametrar (P_1, P_2). Under nollhypotesen (oberoende) kan vi räkna fram sannolikheter för sammansatta händelser enligt följande produktregel:

P(Hudfärgen är mörk & Domen är en dödsdom) = P(Hudfärgen är mörk)P(Domen är en dödsdom) = P_1P_2
P(Hudfärgen är vit & Domen är en dödsdom) = P(Hudfärgen är vit)P(Domen är en dödsdom) = $(1-P_1)P_2$
P(Hudfärgen är mörk & Domen är ej en dödsdom) = P(Hudfärgen är mörk)P(Domen är ej en dödsdom) = $P_1(1-P_2)$
P(Hudfärgen är vit & Domen är ej en dödsdom) = P(Hudfärgen är vit)P(Domen är ej en dödsdom) = $(1-P_1)(1-P_2)$

De förväntade frekvenserna i varje cell av tabellen under oberoendet, erhålls genom att man multiplicerar de uppskattade sannolikheterna med totala antalet obervationer (= 4764 i morddatat). Punktskattningen av sannolikheterna skilt för de två variablerna enligt maximum likelihood -metoden ges av:

$\hat{\mathrm{P}}$(Hudfärgen är mörk) = 1-$\hat{\mathrm{P}}$(Hudfärgen är vit) = $\hat{P}_1 = 2507/4764$
$\hat{\mathrm{P}}$(Domen är en dödsdom) = 1-$\hat{\mathrm{P}}$(Domen är ej en dödsdom) = $\hat{P}_2 = 131/4764$

Detta leder till följande förväntade frekvenser under oberoendet:

Åtalad\Dom	Dödsdom	Annan dom	Totalt
Mörkhyad	68.9	2438.1	2507
Vithyad	62.1	2194.9	2257
Totalt	131	4633	4764

Teststatistikan för χ^2-testet bildas av summan av normerade kvadrerade skillnader mellan förvantade och observerade frekvenser:

$$T = \frac{(68.9-59)^2}{68.9} + \frac{(2438.1-2448)^2}{2438.1} + \frac{(62.1-72)^2}{62.1} + \frac{(2194.9-2185)^2}{2194.9} = 3.085\,6$$

Teststatistikans fördelning har för en 2 x 2 tabell en sk. frihetsgrad, vilket leder till ett p-värde lika med 0.08. Givet att datamaterialet är rätt stort, verkar det alltså inte finnas anledning att tro på ett samband mellan hudfärgen och typen av dom.

Uppfattningen om sambandet ändras, då vi tittar på samma observationer genom en utökad korstabell, där även offrets hudfärg är en variabel:

Mörkhyat offer :

Åtalad\Dom	Dödsdom	Annan dom
Mörkhyad	11	2209
Vithyad	0	111

Vithyat offer :

Åtalad\Dom	Dödsdom	Annan dom
Mörkhyad	48	239
Vithyad	72	2074

Om vi t. ex. testar oberoendet skilt för de två offergrupperna, får vi följande värden på χ^2-testkvantiteten: 0.55, p-värdet 0.54 (mörkhyat offer), 96.50, p-värdet < 0.0001 (vithyat offer). Analys av denna korstabell avslöjar att för samtliga par som kan bildas av de tre variablerna, finns det indikation på ett samband. Men för mörkhyade offer verkar den åtalades hudfärg och typen av dom vara oberoende, medan dessa två variabler har ett starkt

samband för vithyade offer. Då det finns fler än två variabler i en korstabell, brukar man använda sk. log-linjära eller grafiska modeller (en delgrupp av log-linjära modeller) för att studera strukturen på sambanden på ett mer meningsfullt sätt. Grafiska modeller är mycket användbara även för stora grupper av variabler och det finns flera datorprogram som hanterar sådana modeller (MIM, B-Course, Bayesian Network Toolbox (BNT), Hugin, mm.). Grafiska modeller kan intuitivt presenteras med hjälp av matematiska objekt som kallas grafer, där variablerna motsvarar noder och länkar samband mellan dem. Avsaknaden av en länk motsvarar vanligtvis ett betingat oberoende mellan de berörda variablerna.

Vi kan bilda oss en uppfattning om olika typer av grafiska modeller, som kunde beskriva beroendestrukturen mellan de tre variablerna. Låt oss beteckna variablerna enligt följande: Offrets hudfärg (U), Åtalades hudfärg (S), Typen av dom (T)). Det finns åtta möjliga beroendestrukturer som kan presenteras med hjälp av grafer. Om man ordnar dessa enligt grad av komplexitet, löper modellernas struktur från ett fullständigt oberoende ($P(U,S,T) = P(U)P(S)P(T)$) till en fullständigt beroende (inga begränsningar på sannolikheterna $P(U,S,T)$). Exempelvis, om en modell hävdar att U är oberoende från de övriga två variablerna, får vi begränsningen $P(U,S,T) = P(U)P(S,T)$. På ett motsvarande sätt, om vi antar betingat oberoende mellan S och T, får vi begränsningen

$$P(U,S,T) = P(S,T|U)P(U) = P(S|U)P(T|U)P(U).$$

Nedan finns en lista som omfattar alla åtta modellerna och representationen av deras strukturer.

$$
\begin{aligned}
M_1 &: P(U,S,T) = P(U)P(S)P(T) \\
M_2 &: P(U,S,T) = P(U,S)P(T) \\
M_3 &: P(U,S,T) = P(U,T)P(S) \\
M_4 &: P(U,S,T) = P(U)P(S,T) \\
M_5 &: P(U,S,T) = P(U|S)P(T|S)P(S) \\
&= P(U,S)P(S,T)/P(S) \\
M_6 &: P(U,S,T) = P(U|T)P(S|T)P(T) \\
&= P(U,T)P(S,T)/P(T) \\
M_7 &: P(U,S,T) = P(S|U)P(T|U)P(U) \\
&\quad P(U,S)P(U,T)/P(U) \\
M_8 &: P(U,S,T)
\end{aligned}
$$

För modellerna 5-7, får vi den senare formeln enligt definitionen på en betingad sannolikhet, t. ex.:

$$P(S|U) = \frac{P(U,S)}{P(U)}.$$

Graferna som motsvarar dessa modeller ser ut såhär:

$$
\begin{aligned}
M_1 &: \; U \;\; S \;\; T \\
M_2 &: \; U - S \;\; T \\
M_3 &: \; U - T \;\; S \\
M_4 &: \; U \;\; T - S \\
M_5 &: \; U - S - T \\
M_6 &: \; U - T - S \\
M_7 &: \; S - U - T \\
&\qquad\quad U \\
M_8 &: \quad \diagup \qquad \diagdown \\
&\qquad S \quad - \quad T
\end{aligned}
$$

Exemplet ovan, med två kontra tre dikotoma variabler, representerar en sk. **Simpson's paradox**, vilken innebär att variabler som inte finns med i betraktelsen, kan antingen dölja ett samband mellan variabler, eller introducera ett falskt samband mellan dem. Den senare typen av felaktiga slutsatser exemplifieras av följande datamaterial.

Joe Whittaker presenterar följande datamaterial i boken Graphical models in applied multivariate analysis (Wiley, 1990):

Grad av vård\Överlevnad	Dog	Överlevde	Totalt
Lite	20	373	393
Mycket	6	316	322
Totalt	26	689	715

Korstabellen innehåller uppgifter om 715 spädbarn under de tidiga veckorna efter födseln. Uppgifterna är samlade från två sjukhus som besöktes av vårdnashavarna till barnet. Uppskattningen av grad av vård har gjorts av sjukvårdspersonalen. Andelen barn som dog verkar vara betydligt högre bland de barn som fick endast lite vård (5.1% mot 1.9%). Detta resultat kunde leda till den slutsatsen att graden av vård påverkar dödligheten. Återigen ändras uppfattningen om sambandet, då vi känner till en annan variabel, nämnligen vilket sjukhus besöktes av vårdnashavarna. Följande tabell erhålls genom att denna variabel görs synlig (den sista kolumnen representerar andelen barn som dog):

Sjukhus 1:

Grad av vård\Överlevnad	Dog	Överlevde	(%)
Lite	3	176	1.7
Mycket	4	293	1.4

Sjukhus 2:

Grad av vård\Överlevnad	Dog	Överlevde	(%)
Lite	17	197	7.9
Mycket	2	23	8.0

I den nya tabellen finns det ingen nämnvärd skillnad m.a.p. dödligheten mellan de två nivåerna av vård, inom nåtdera sjukhuset. Det tidigare observerade sambandet har alltså försvunnit, då sjukhusindikatorn togs i beaktande. Fenomenet uppstår, eftersom det andra sjukhuset är förknippat med ett område där det finns högre spädbarnsdödlighet och fler barn som får sämre vård. En grafisk beskrivning av beroendestrukturen ser ut såhär:

$$\text{'Grad av vård'} - \text{'Sjukhus'} - \text{'Överlevnad'},$$

dvs. överlevnad är i detta datamaterial betingat oberoende av graden av vård, givet det sjukhus man besökt.