Stolpediagrammer for kategoriske data med -catplot-

[Revideret 4. oktober 2013]

Kim Mannemar Sønderskov

Institut for Statskundskab, Aarhus Universitet
ks@ps.au.dk

Denne note gennemgår, hvordan resultaterne fra deskriptive uni-, bi- og multivariate analyser af kategoriske variable kan illustreres i stolpediagrammer med Stata. Til dette formål introduceres kommandoen catplot. catplot er et såkaldt *user written program*, som kort fortalt er en uofficiel tilføjelse til Stata udarbejdet af en Stata-bruger. User written programs beskrives ikke grundigt her, men se evt. afsnit 8.3 i Sønderskov (2011). For at anvende catplot skal programpakken installeres; dette gøres ved at skrive følgende i kommandolinjen:

```
. *installerer programpakken catplot
. net install catplot.pkg, from(http://fmwww.bc.edu/RePEc/bocode/c)
```

1a): Installer programmet.

1b): Åbn datasættet 'Valg05Kap5.dta', der kan hentes fra http://ps.au.dk/soenderskov/stata/.

1.1 Illustration af univariate fordelinger i stolpediagrammer

Som beskrevet i Sønderskov (2011: 92ff) er det lettere kompliceret at illustrere univariate fordelinger for kategoriske variable i stolpediagrammer med Stata. Samme sted gennemgås en længere procedure, der ved hjælp af histogram-kommandoen producerer et univariat stolpediagram for kategoriske data. Konkret illustreres fordelinger af stemmer på danske politiske partier på baggrund af surveydata (Den danske valgundersøgelse fra 2005). Nedenstående kommando producerer stolpediagrammet i Figur 5.4 i Sønderskov (2011).

```
. *stolpediagram som Figur 5.4 i Sønderskov (2011)
. histogram partivalg, percent ///
  ytitle(Stemmefordeling (%)) yline(2, lpattern(dash)) ///
  xtitle("") ///
  xlabel(1(1)11, angle(forty_five) valuelabel) ///
  discrete gap(10) ///
  note(Den stiplede linje indikerer spærregrænsen (2%); Kilde: Valgundersøgelsen 2005)
```

Med catplot kan man med en væsentligt kortere kommando opnå et lignende plot. Der er ingen dialogboks til catplot (hvilket gælder for de fleste user writtens programs). Man skal i stedet bruge kommandolinjen. For at danne et simpelt univariate stolpediagram for variablen *katvar* skrives:

```
. *Stolpediagram for en kategorisk variabel katvar
. catplot katvar
```

Typiske vil man illustrere fordelingen med procenter frem for frekvenser; dette opnås ved at tilføje optionen percent:

```
. *Som ovenfor, men afrapportering af procentfordeling
. catplot katvar, percent
```

2: Producer et stolpediagram med catplot for den kategoriske variabel partivalg, der også blev brugt i ovennævnte eksempel med histogram. Brug procenter som målestok.

Diagrammet er ikke helt som Figur 5.4 i Sønderskov 2011; den væsentligste forskel er, at nærværende diagram er horisontalt frem for vertikalt, derudover mangler 2%-linjen, noten, en mere sigende forklaring på procent-aksen og at få fjernet den lettere forvirrende titel yderst til venstre. Diagrammet omdannes nemt til et vertikalt stolpediagram med optionen recast(bar):

```
*Som ovenfor, men i vertikal form
catplot katvar, percent recast(bar)
```

3: Dan et vertikalt stolpediagram med samme variabel som ovenfor.

Resultatet er ikke helt heldigt, da partinavnene står oven i hinanden og er ulæselige. Følgende kommando løser dette og tilføjer derudover noten, linjen ved 2% og fjerner titlen.

```
. * Som ovenfor, med tilpasninger
. catplot partivalg, percent recast(bar) ///
  ytitle(Stemmefordeling (%)) yline(2, lpattern(dash)) ///
  b1title("") ///
  yvaroptions(label(angle(45))) ascategory ///
  note(Den stiplede linje indikerer spærregrænsen (2%); Kilde: Valgundersøgelsen 2005)
```

1. linje blev forklaret ovenfor. 2. og 5. linje er identiske med linje 2. og 6. linje i histogram-kommandoen ovenfor. 3. linje er næsten identisk med 3. linje i histogram-kommandoen, men referer til x-aksen med b1. i stedet. 4. linje er nødvendig for at få drejet partinavnene.

Det fremgår at catplot er yderst anvendeligt til hurtigt at lave et groft univariat stolpediagram, men catplot er næsten lige så kompliceret som histogram-kommandoen, såfremt diagrammet skal være lidt anderledes end standarddiagrammet. catplot er dog yderst anvendeligt til illustration af resultater fra bi- og multivariate analyser, hvor histogram kommer til kort.

1.2 Illustration af bivariate sammenhænge i stolpediagrammer

Multinominel afhængig – nominal/ordinal uafhængig

Et eksempel på en analyse af sammenhængen mellem en multinominel afhængig variabel – altså en nominalskaleret afhængig variabel med mere end to kategorier – og en nominal eller ordinal afhængig variabel kunne være om mænd og kvinder stemmer forskelligt ved folketingsvalg. Dette undersøges her med dummyvariablen mand og den samme variabel med oplysninger om partivalg som ovenfor (partivalg). Man kan starte med at foretage en krydstabulering af de to variable, hvor der procentures på den uafhængige variabel mand (se eventuelt afsnit 6.1 i Sønderskov, 2011).

4: Brug tabulate til at krydstabulere partivalg og mand; procentuer på mand.

Tabellen viser, at stikprøvens mænd og kvinder i det store hele stemmer rimeligt ens, men også at der er relativt flere kvinder end mænd, der støtter Socialdemokraterne og SF, mens relativt flere mænd end kvinder støtter Venstre og Dansk Folkeparti.

Man kunne supplere analysen med beregning af et sammenhængsmål (fx *lambda*) og en test af om forskellen mellem kønnenes stemmeadfærd er statistisk signifikant (fx med *chi2* eller *multinomial logistisk regression*). Her fokuseres der dog på at illustrere sammenhængen i stikprøven, hvilket kan være en hjælp til fortolkning af sammenhængen eller i visse situationer et alternativ til tabelafrapportering. catplot kan bruges til dette, og i situationer som denne med en multinominal afhængig giver det ofte det mest illustrative diagram, hvis man placerer det, man vil sammenligne (kategorierne på den uafhængige variabel; her mænd og kvinder) inderst i plottet, således at det man sammenligner, står ved siden af hinanden. For at få den relative fordeling skal det også specificeres, at man vil procentuere på den uafhængige variabel. Dette gøres således:

```
. *Søjlediagram, procentuering på uafhængig variabel xvar,
. *kategorierne på den uafhængige variabel holdes sammen
. catplot xvar yvar, percent(xvar)
```

5a): Dan et horisontalt søljediagram, der viser stemmefordelingen for stikprøvens mænd og kvinder på baggrund af variablene mand og partivalg.

5b) Sammenhold diagrammet med krydstabuleringen, der blev dannet ovenfor.

Dikotom/ordinal afhængig - nominal/ordinal uafhængig

Når den afhængige variabel kan ordnes i en naturlig rækkefølge giver det oftest mest mening at holde kategorierne på den afhængige variabel samlet og vise fordelingen for hver kategori på den uafhængige variabel. I catplot-kommandoen gøres dette ved at placere den afhængige variabel først i kommandoen:

```
. *Som ovenfor, men kategorierne på den afhængige variabel (yvar) holdes sammen
. catplot yvar xvar, percent(xvar)
```

Som øvelse kan du illustrere sammenhængen mellem respondenternes boligform (hhv. lejet bolig, andelsbolig. villa/gård eller ejerlejlighed) målt med v335 og tilbøjeligheden på at støtte partier fra 'blå' blok ved det dengang netop overståede folketingsvalg (målt med blå). Boligform behandles som den uafhængige variabel.

6a) Lav en krydstabulering mellem de to variable; procentuer på den uafhængige variabel.

6b) Fortolk resultatet

6c) Illustrer sammenhængen med catplot. Husk at procentuere på den uafhængige variabel.

Diagrammet indeholder otte søjler, hvilket ikke fylder meget. Havde hvis begge variable havde indeholdt eksempelvis fire kategorier ville diagrammet derimod fylde en del. I de tilfælde, hvor den afhængige variabel er dikotom eller ordinalskaleret kan man illustrere fordelingen på den afhængige variabel med én søjle for hver kategori på den uafhængige variabel uden tab af mening, og på den måde komprimere plottet. Dette gøres med optionerne stack og asyvars:

```
. *Som ovenfor, men kategorierne på den afhængige variabel "stackes"
. catplot yvar xvar, percent(xvar) asyvars stack
```

7) Producer diagrammet igen med ovenstående options.

Dette giver en mindre diagram, og havde den afhængige variabel haft mange kategorier ville dette diagram fylde væsentlig mindre end det foregående. Diagrammet har dog én hage – i hvert fald i min opsætning af Stata. Søjlen, der skal præsentere røde stemmer, er tilfældigvis blå og omvendt. Med følgende kommando fås mere retvisende farver. Logikken er, at man aktivt specificerer hvilken farve specifikke søjler skal have. Årsagen til, at søjle 1 er den der skal være rød kræver en længere forklaring, der ikke gives her – men man kan altid prøve sig frem.

```
. *Som ovenfor, men med andre farver
. catplot blå v335, percent(v335)  asyvars stack  bar(1, color(red)) bar(2,color(blue))
```

Du kan nu prøve et andet eksempel, hvor det giver rigtig god mening at "stacke" kategorierne på den afhængige variabel. Du skal analysere sammenhængen mellem alder og frygt for immigration. Til formålet bruges variablene alder4kat og v186, og alder behandles som uafhængig variabel.

8a) Inspicer variablene alder4kat og v186 med codebook.

8b) Omkod v186 så den bliver ordinalskaleret, dvs. "ved ikke" omkodes til missing; overskriv v186.

Kodningen af alder4kat fremgår ikke af kodebogen og variablen har ingen *value labels*, hvilket nedsætter læsbarheden af søjlediagrammet. Du skal derfor tilføje value labels. Af opgave 80 i Sønderskov (2011) fremgår det, at alder4kat's fire kategorier indikerer følgende aldersintervaller (målt i år): 1) <30, 2) 30-49, 3) 50-64, 4) 65+.

9a) Brug kommandoerne label define til at oprette en value label, der knytter værdierne 1-4 til ovennævnte labels. Kald labelen 'alder4LB'.

9b) Tilknyt dernæst den oprettede label til alder4kat med kommandoen label values. Se evt. afsnit 3.5 i Sønderskov (2011).

10a) Brug tabulate til at krydstabulere de to variable; procentuer på den uafhængige og beregn også gamma-statistikken.

10b) Fortolk tabellen og gamma-statistikken.

10c) Illustrer sammenhængen med catplot – først i et diagram, hvor kategorierne på den afhængige ikke "stackes" og siden i et diagram, hvor de "stackes".

1.3 Illustration af multivariate sammenhænge i stolpediagrammer

Multivariate sammenhænge kan også illustreres i stolpediagrammer, selvom det er mindre oplag end i ovenstående tilfælde, da det nemt bliver et omfattende diagram i det multivariate tilfælde (hvorved diagrammet måske ikke hjælper til at anskueliggøre sammenhængen).

Multivariate sammenhænge, hvor en sammenhæng mellem to variable kontrolleres for én eller flere kontrolvariable illustreres med catplot ved at indsætte kontrolvariablene i optionen over(). Ydermere skal det specificeres, at der skal procentueres på den primære uafhængige variabel for

hvert kategori af kontrolvariablene; dette gøres ved at indsætte kontrolvariablene i optionen percent() sammen med den uafhængige variabel:

```
. *Som ovenfor, men opsplittet over kontrolvariablen zvar:
. catplot yvar xvar, percent(xvar zvar) over(zvar)
. //optionerne stack og asyvars kan også med fordel bruges her
```

11a) Undersøg samme sammenhæng som ovenfor (alder og frygt for immigration), men kontroller for køn vha. variablen mand. Start med at krydstabulere med kommandoen tabulate og prefixet bysort (jf. Sønderskov 2012).

11b) Illustrer derefter sammenhængen med catplot – kollaps kategorierne på den afhængige til én sølje vha. af stack.

Løsninger

```
*Opgave 1
net install catplot.pkg, from(http://fmwww.bc.edu/RePEc/bocode/c) //1a
use http://ps.au.dk/uploads/media/Valg05Kap5.dta //1b
*Opgave 2
catplot partivalg, percent
*Opgave 3
catplot partivalg, percent recast(bar)
*Opgave 4
tabulate partivalg mand, column
*Opgave 5a
catplot mand partivalg, percent(mand)
*Opgave 6
tabulate blå v335,colum //6a
catplot blå v335, percent(v335) //6c
*Opgave 7
catplot blå v335, percent(v335) asyvars stack
*Opgave 8
codebook alder4kat v186 //8a
recode v186 (8=.) //8b
*Opgave 9
label define alder4LB 1 "<30" 2 "30-49" 3 "50-64" 4 "65+" //9a
label value alder4kat alder4LB //9b
*Opgave 10
tabulate v186 alder4kat, colum gamma //10a
catplot v186 alder4kat, percent(alder4kat) asyvars //10c1
catplot v186 alder4kat, percent(alder4kat) asyvars stack //10c2
*opgave 11
bysort mand: tabulate v186 alder4kat, colum gamma //11a
catplot v186 alder4kat, percent(alder4kat mand) over(mand) asyvars stack //11b
```

Referencer

Sønderskov, K.M. (2011). *Stata – En praktisk introduktion*. København: Hans Reitzels.

Sønderskov, K.M. (2012). *Multivariat analyse af kategoriske variable med tabelanalyse og -tabulate-*. Aarhus: Institut for Statskundskab.