

Extra datorövning

Statistikens Grunder 1

Syfte

1. Lära sig sätta text på utfallen av numeriska variabler
2. Lära sig klassindela

Exempel

Sätta text på utfallen av numeriska variabler

Betrakta data-setet med 12 observationer där individens kön och dryckesvana noteras

Dryck	Kön
1	0
2	0
1	1
1	0
2	1
2	1
3	0
3	1
1	0
2	1
1	1
3	1

För variabeln "*dryck*" motsvarar de numeriska värdena

- 1 = Coca-Cola
- 2 = Fanta
- 3 = Varken eller

För variabeln "*kön*" motsvarar de numeriska värdena

- 0 = Kvinna
- 1 = Man

Vi vill nu sätta ett format få data-set som anger dessa värden i text istället för i nummer.

Vi börjar med att läsa in datat i SAS

```
data work.colafanta;
input dryck kon;
datalines;
1 0
2 0
1 1
1 0
2 1
2 1
3 0
3 1
1 0
2 1
1 1
3 1
;
run;
```

Nu ska vi använda "*proc format*" för ändra formatet på variabeln kön

```
proc format;
value konf
0 = 'kvinna'
1 = 'man'
;
data work.colafanta;
set work.colafanta;
run;
```

Vi skriver "*value*" för att tala om vad formatet på variabeln ska heta. I detta exempel heter den "*konf*". Ni kan döpa formatet till vad ni vill, men här använder vi "*konf*". Detta för att det är enkelt att komma ihåg att den hör till variabeln "*kon*" samt att vi har lagt på ett format - alltså "*konf*".

Efter det specificerar vi att 0 ska vara kvinna och att 1 ska vara man. I slutet använder vi "*data...set*" för att tala om vilket data-set det är vi ska formatera, samt i vilket data-set vi ska spara formatet i.

Vi gör nu samma sak för variabeln *"dryck"*.

```
proc format;  
value dryckf  
1 = 'Cola'  
2 = 'Fanta'  
3 = 'Varken eller'  
;  
data work.colafanta;  
set work.colafanta;  
run;
```

Nu har vi endast sparat formatet i SAS. När vi vill använda det måste vi också ange det i den procedur vi vill köra. Vi kan till exempel välja att skriva ut datat med *"proc print"* där vi använder det nya formatet. Vi skriver

```
proc print data=colafanta;  
format kon konf. dryck dryckf.;  
run;
```

Vi lägger alltså till en rad mellan *"proc print"* och *"run"*. Där skriver vi *"format"* för att vi vill att SAS ska applicera det nya formatet. Efter det skriver vi *"kon"* *"konf."*. När vi gör det talar vi om att variabeln *"kon"* ska ha formatet *"konf"*. **Observera** att det måste komma en punkt efter format-variabelnamnet. Vi gör samma sak för variabeln *"dryck"*. Detta genererar en utskrift som ser ut som följer

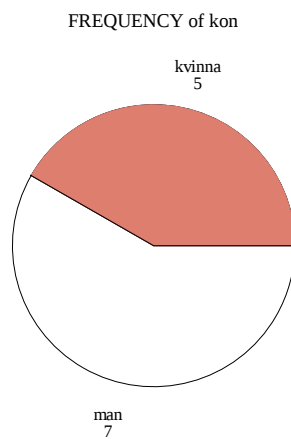
Obs	dryck	kon
1	Cola	kvinna
2	Fanta	kvinna
3	Cola	man
4	Cola	kvinna
5	Fanta	man
6	Fanta	man
7	Varken eller	kvinna
8	Varken eller	man
9	Cola	kvinna
10	Fanta	man
11	Cola	man
12	Varken eller	man

Istället för siffror har vi nu ändrat formatet så att SAS skriver ut text.

Vi kan även applicera formatet när vi ritar diagram. För att skapa ett cirkeldiagram med format skriver vi

```
proc gchart data=work.colafanta;  
pie kon/discrete;  
format kon konf.;  
run;  
quit;
```

vilket genererar ett diagram som ser ut så här



Klassindela datamaterialet

Vi betraktar filen "*rokdata.xls*" som finns på hemsidan. Vi importerar datat till SAS. När vi använder "*proc freq*" för att göra en tabell över variabeln ålder får vi en utskrift som ser ut som följer

The FREQ Procedure

age

age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
15	3	5.00	3	5.00
20	1	1.67	4	6.67
21	1	1.67	5	8.33
22	1	1.67	6	10.00
23	2	3.33	8	13.33
27	1	1.67	9	15.00
28	5	8.33	14	23.33
29	1	1.67	15	25.00
30	3	5.00	18	30.00
33	4	6.67	22	36.67
34	2	3.33	24	40.00
36	2	3.33	26	43.33
37	1	1.67	27	45.00
39	4	6.67	31	51.67
40	1	1.67	32	53.33
41	1	1.67	33	55.00
44	1	1.67	34	56.67
45	1	1.67	35	58.33
47	5	8.33	40	66.67
52	1	1.67	41	68.33
53	4	6.67	45	75.00
55	1	1.67	46	76.67
56	3	5.00	49	81.67
58	1	1.67	50	83.33
60	1	1.67	51	85.00
61	1	1.67	52	86.67
62	1	1.67	53	88.33
63	2	3.33	55	91.67
66	1	1.67	56	93.33
67	1	1.67	57	95.00
68	1	1.67	58	96.67
69	2	3.33	60	100.00

vilket inte är en speciellt överskådlig tabell över ålder.

Vi ska nu klassindela variabeln ålder. Vi skriver

```
proc format;  
value ageclassf  
15-25 = '15-25'  
26-30 = '26-30'  
31-45 = '31-45'  
46-50 = '46-50'  
51-55 = '51-55'  
56-60 = '56-60'  
61-65 = '61-65'  
65-70 = '65-70'  
;  
data work.rokdata;  
set work.rokdata;  
run;
```

Formatnamnet ovan är "*ageclassf*". Vi hämtar data från "*work.rokdata*" och sparar formatet i samma data-set. Observera att hur man väljer att klassindela materialet är frivilligt. Ovanstående klassindelning är endast till för att exemplifiera.

När vi vill göra en tabell över variabeln ålder med det nya formatet skriver vi koden

```
proc freq data=work.rokdata;  
format age ageclassf.;  
tables age;  
run;
```

(Den ursprungliga, numeriska, variabeln ålder får automatiskt namnet "*age*" när vi skriver "getnames" i "*proc import*" koden.)

Ovan talar vi om att variabeln "*age*" ska ha formatet som heter "*ageclassf*". Vi skriver "*tables age*" för att det är för variabeln "*age*" vi vill göra en tabell över. Observera att det inte går att ange formatnamnet efter

"*tables*". Tabellen får utseendet

The FREQ Procedure

age

age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
15 - 25	8	13.33	8	13.33
26 - 30	10	16.67	18	30.00
31 - 45	17	28.33	35	58.33
46 - 50	5	8.33	40	66.67
51 - 55	6	10.00	46	76.67
56 - 60	5	8.33	51	85.00
61 - 65	4	6.67	55	91.67
65 - 70	5	8.33	60	100.00

Om vi vill göra en korstabell mellan ålder och rökvana, där vi använder det nya formatet, skriver vi

```
proc freq data=work.rokdata;  
format age ageclassf.;  
tables age*smoke;  
run;
```

vilket ger en korstabell som ser ut så här

The FREQ Procedure

Table of age by smoke

age(age)	smoke(smoke)		
	0	1	Total
Frequency			
Percent			
Row Pct			
Col Pct			
15-25	0	8	8
	0.00	13.33	13.33
	0.00	100.00	
	0.00	22.22	
26-30	5	5	10
	8.33	8.33	16.67
	50.00	50.00	
	20.83	13.89	
31-45	4	13	17
	6.67	21.67	28.33
	23.53	76.47	
	16.67	36.11	
46-50	2	3	5
	3.33	5.00	8.33
	40.00	60.00	
	8.33	8.33	
Total	24	36	60
	40.00	60.00	100.00

(Continued)

The FREQ Procedure

Table of age by smoke

age(age)		smoke(smoke)		
Frequency				
Percent				
Row Pct				
Col Pct	0	1	Total	
51-55	4	2	6	
	6.67	3.33	10.00	
	66.67	33.33		
	16.67	5.56		
56-60	2	3	5	
	3.33	5.00	8.33	
	40.00	60.00		
	8.33	8.33		
61-65	3	1	4	
	5.00	1.67	6.67	
	75.00	25.00		
	12.50	2.78		
65-70	4	1	5	
	6.67	1.67	8.33	
	80.00	20.00		
	16.67	2.78		
Total	24	36	60	
	40.00	60.00	100.00	

Tabellen är inte speciellt lättöverskådlig. En grövre klassindelning kan behövas. Vi skulle även kunna specificera ett format till variabeln "*smoke*" för att skriva ut "*röker*" och "*röker inte*" istället för 1 och 0. Detta ges som en övning nedan.

Uppgifter

1. Gör en grövre klassindelning av variabeln "*age*" i data-setet work.rokdata (Ni behöver importera det från Excel).
2. Ge svarsalternativen i variabeln "*smoke*" ett textformat.
3. Skapa en korstabell mellan "*age*" och "*smoke*" där ni använder era nya format.