

Datorövning 1

Statistikens Grunder 2

Syfte

1. Lära sig göra betingade frekvenstabeller
2. Lära sig beskriva en variabel numeriskt med "*proc univariate*"
3. Lära sig rita histogram
4. Lära sig rita diagram med avseende på en annan variabel
5. Lära sig beskriva en variabel numeriskt med avseende på en annan variabel

Exempel

Betingade korstabeller

För att beskriva en kategorisk variabel numeriskt använder vi frekvenstabeller. Om vi vill beskriva en kategorisk variabel med avseende på en annan variabel kan vi göra det i SAS med "*proc freq*"-kommandot.

Som exempel använder vi återigen rökdatat från hemsidan. Data-setet är utökat och innehåller nu tre variabler; kön, rökvana och ålder. (Variabeln ålder ska vi använda längre fram.) För att göra en frekvenstabell för variabeln rökvana betingat på kön använder vi koden

```
proc freq data=work.rokdata;  
tables smoke;  
by gender;  
run;
```

Vi kommer inte att få ut någon utskrift. Om vi tittar i loggen finner vi följande:

```

71  proc freq data=work.rokdata;
72  tables smoke;
73  by gender;
74  run;

```

ERROR: Data set WORK.ROKDATA is not sorted in ascending sequence. The current BY group has gender

= 1 and the next BY group has gender = 0.

NOTE: The SAS System stopped processing this step because of errors.

NOTE: There were 32 observations read from the data set WORK.ROKDATA.

NOTE: PROCEDURE FREQ used (Total process time):

real time 0.01 seconds

cpu time 0.00 seconds

Vi får ett felmeddelande om att datamaterialet måste vara sorterat med avseende på den variabeln vi vill betinga på. I detta fall måste vi sortera materialet med avseende på kön. Vi sorterar datat med koden

```

proc sort data=work.rokdata;
by gender;
run;

```

Vi kan nu ta fram en tabell med hjälp av koden ovan. Den ser ut så här

----- gender=0 -----					

The FREQ Procedure					
smoke					
smoke	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
0	14	42.42	14	42.42	
1	19	57.58	33	100.00	
----- gender=1 -----					

The FREQ Procedure					
smoke					
smoke	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
0	10	35.71	10	35.71	
1	18	64.29	28	100.00	

Givet att gender=0 motsvarar "kvinna" och smoke=0 motsvarar "röker ej", ser vi att 14 kvinnor av 33 inte röker och att 19 kvinnor av 33 röker.

Beskriva en variabel numeriskt med "proc univariate"

Vi har lärt oss att vi kan använda "proc means" för att beskriva en variabel numeriskt. För en lite mer utförligare numerisk beskrivning kan vi använda "proc univariate". Låt oss betrakta exemplet från Datorövning 1, Statistikens Grunder 1, med personen som simmar. Data-setet ges som

Tid	Puls
34.12	152
35.72	124
34.72	140
34.05	152
34.13	146
35.72	128
36.17	136
35.57	144
35.37	148
35.57	144
35.43	136
36.05	124
34.85	148
34.70	144
34.75	140
33.93	156

Vi vill nu beskriva dessa båda variabler med "proc univariate". Vi skriver koden

```
proc univariate data=work.simma;  
run;
```

Den utskrift som den här koden genererar ser ut enligt följande:

För variabeln "tid":

The UNIVARIATE Procedure				
Variable: tid				
Moments				
N	16	Sum Weights	16	
Mean	35.053125	Sum Observations	560.85	
Std Deviation	0.74237204	Variance	0.55111625	
Skewness	-0.1497197	Kurtosis	-1.3344082	
Uncorrected SS	19667.8119	Corrected SS	8.26674375	
Coeff Variation	2.11784839	Std Error Mean	0.18559301	
Basic Statistical Measures				
Location		Variability		
Mean	35.05313	Std Deviation	0.74237	
Median	35.11000	Variance	0.55112	
Mode	35.57000	Range	2.24000	
		Interquartile Range	1.23000	
Note: The mode displayed is the smallest of 2 modes with a count of 2.				
Tests for Location: Mu0=0				
Test	-Statistic-		-----p Value-----	
Student's t	t	188.8709	Pr > t	<.0001
Sign	M	8	Pr >= M	<.0001
Signed Rank	S	68	Pr >= S	<.0001
Quantiles (Definition 5)				
	Quantile	Estimate		
	100% Max	36.170		
	99%	36.170		
	95%	36.170		
	90%	36.050		
	75% Q3	35.645		
	50% Median	35.110		
	25% Q1	34.415		
	10%	34.050		
	5%	33.930		
	1%	33.930		
	0% Min	33.930		
Extreme Observations				
-----Lowest----		---Highest---		
Value	Obs	Value	Obs	
33.93	16	35.57	10	
34.05	4	35.72	2	
34.12	1	35.72	6	
34.13	5	36.05	12	
34.70	14	36.17	7	

För variabeln "puls"

```

The UNIVARIATE Procedure
Variable:  puls

Moments
N          16      Sum Weights          16
Mean       141.375  Sum Observations     2262
Std Deviation 9.70824392  Variance          94.25
Skewness    -0.5776568  Kurtosis         -0.4376614
Uncorrected SS 321204  Corrected SS      1413.75
Coeff Variation 6.86701603  Std Error Mean    2.42706098

Basic Statistical Measures

Location          Variability
Mean      141.3750  Std Deviation      9.70824
Median    144.0000  Variance          94.25000
Mode      144.0000  Range             32.00000
                        Interquartile Range    12.00000

Tests for Location: Mu0=0

Test          -Statistic-    -----p Value-----
Student's t    t    58.24946  Pr > |t|    <.0001
Sign          M         8    Pr >= |M|    <.0001
Signed Rank    S        68    Pr >= |S|    <.0001

Quantiles (Definition 5)

Quantile      Estimate
100% Max      156
99%           156
95%           156
90%           152
75% Q3        148
50% Median    144
25% Q1        136
10%           124
5%            124
1%            124
0% Min        124

Extreme Observations

----Lowest----      ----Highest---
Value      Obs      Value      Obs
124         12      148         9
124          2      148        13
128          6      152          1
136         11      152          4
136          7      156         16

```

Bland annat ger "proc univariate"

- antalet observationer
- medelvärde
- standardavvikelse och varians
- standardfel

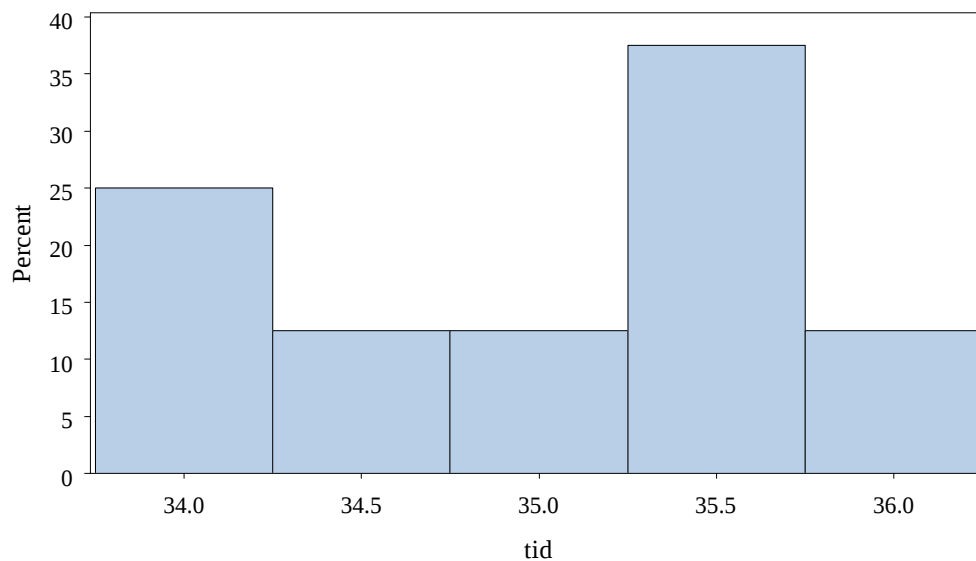
- median
- typvärde

Rita histogram

För att beskriva en kontinuerlig variabel grafiskt använder vi ofta histogram. Vi ritar ett histogram i SAS med "proc univariate". Låt oss använda samma exempel som i exemplet ovan. För att rita ett histogram över variabeln "tid", skriver vi koden

```
proc univariate data=work.simma;
  histogram tid;
run;
```

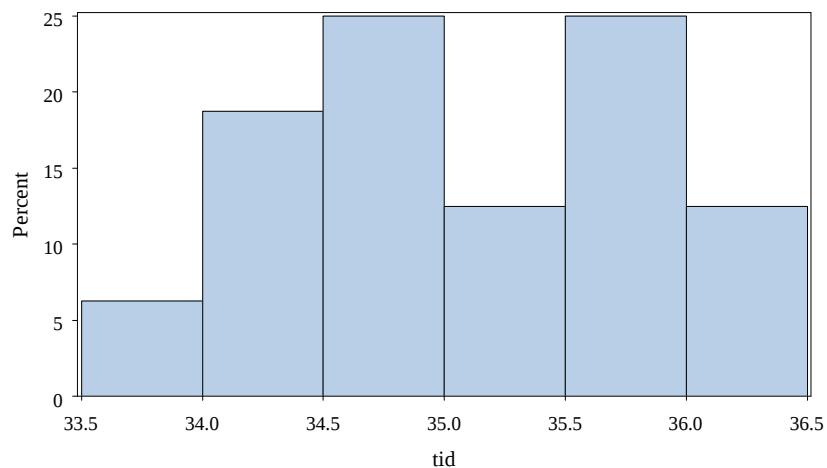
Detta genererar följande histogram:



Här bestämmer SAS klassindelningen själv samt att det mittersta värdet anges för varje stapel på x-axeln, vilket kallas "midpoint". Vi kan ändra klassindelningen genom att skriva

```
proc univariate data=work.simma;
  histogram tid / endpoints = 33.5 to 36.5 by 0.5;
run;
```

Detta ger histogrammet



Ovan har vi bytt från "midpoints" till "endpoints", samt angett vilka värden som ska sätta gränserna. Vilka värden man väljer här beror helt på hur datamaterialet ser ut.

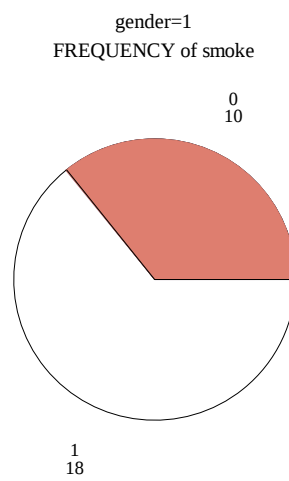
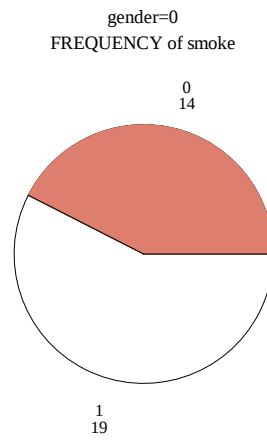
Rita diagram med avseende på en annan variabel

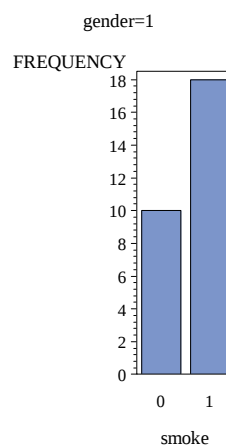
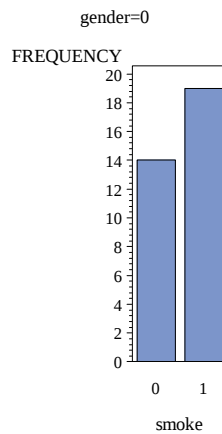
Betrakta exemplet "rokdata". För att rita ett cirkel- och ett stapeldiagram för variabeln "smoke" med avseende på variabeln "kön" skriver vi koden

```
proc gchart data=work.rokdata;  
  pie smoke / discrete;  
  vbar smoke / discrete;  
  by gender;  
run;
```

Observera att vi även här måste sortera datamaterialet efter variabeln som vi anger efter kommandot "by". Eftersom vi redan sorterat datat (se ovan)

behöver vi inte göra det igen. Koden ger oss 4 diagram





I diagrammen ovan anges frekvenser. För att visa procent skriver man

```
proc gchart data=work.rokdata;
pie smoke / discrete percent=outside;
vbar smoke / discrete type=percent;
by gender;
run;
```

För att rita ett histogram uppdelat på en annan variabel använder vi liknande proc univariate-kod som ovan. Vi skriver

```
proc univariate data=work.simma;
histogram tid;
by variabelnamn;
run;
```

Här har vi inte angett något "riktigt" variabelnamn efter "by" eftersom det inte finns någon variabel att dela upp datamaterialet på. Nedan, under uppgifter, ska vi träna på att rita histogram för en variabel, uppdelat med avseende på en annan variabel.

Beskriva en variabel numeriskt med avseende på en annan variabel

När vi kör "*proc univariate*"-koden ovan får vi ut histogram som är betingade på variabeln som vi anger efter kommandot "*by*". Vi får också ut en numerisk beskrivning av *alla variabler* i data-setet, betingade på samma variabel. Vill vi inte ha något histogram, utan bara en numerisk beskrivning uppdelat på en annan variabel tar vi bort kommandoraden "*histogram...*".

Uppgifter

Basuppgifter

1. Använd datat från Datorövning 1, Statistikens Grunder 1

	Clas Ohlson	H&M	Teknikmagasinet	MQ
Kvinnor	11	57	6	26
Män	46	4	32	18

och skapa en frekvenstabell för variabeln butik som är betingad på kön.

2. Använd "Rökdata" från hemsidan. Beskriv variabeln "age" numeriskt med "proc univariate". Vad blir medelvärdet, medianen och typvärdet ("mode")? Vilket är det minsta värdet och vilket är det största värdet?
3. Fortsätt med data-setet "Rökdata". Rita ett histogram för åldern uppdelat på rökvana. Ser man någon tydlig skillnad mellan de båda fördelningarna? Observera att när ni gör ett histogram uppdelat på rökvana får ni även en numerisk beskrivning av variabeln ålder uppdelat på rökvana. Jämför den grafiska och den numeriska beskrivningen för att övertyga er om att de överensstämmer.
4. Vi utökar datamaterialet "simma" genom att införa ännu en variabel,

kön, och ytterligare observationer.

Tid	Puls	Kön
34.12	152	1
35.72	124	0
34.72	140	1
34.05	152	0
34.13	146	1
35.72	128	0
36.17	136	1
35.57	144	0
35.37	148	1
35.57	144	0
35.43	136	1
36.05	124	0
34.85	148	1
34.70	144	0
34.75	140	1
34.93	136	0
34.85	148	1
35.57	147	0
34.05	146	1
36.17	124	0
36.05	120	1
35.20	137	0

Läs in datat i SAS. Skapa histogram över variablerna "tid" och "puls" uppdelat på kön.

5. Använd datamaterialet i basuppgift 1. Rita stapeldiagram för variabeln butik uppdelat på kön. Obs! Eftersom datat är inläst med en frekvensvariabel behövs ett tillägg i koden. Se Datorövning 2, Statistikens Grunder 1.
6. Använd datamaterialet "Rökdata". Rita cirkeldiagram över variabeln "smoke" uppdelat på kön. Specificera att vi vill ha både frekvens och procent i diagrammet. Kom ihåg att lägga till kommandot "discrete".