Datorövning 2

Statistikens Grunder 1

Syfte

1. Lära sig presentera data i tabeller
2. Lära sig beskriva data numeriskt
3. Lära sig presentera data i grafer
   (a) Lära sig beräkna sannolikheter för binomial- och normalfördelade variabler
   (b) Lära sig presentera binomial- och normalfördelningen grafiskt

Exempel

Presentera data i tabeller

För att demonstrera hur man kan skapa tabeller i SAS använder vi oss av exemplet från Datorövning 1, Statistikens Grunder 1.

	Studenter röker	Studenter röker ej
Båda föräldrarna röker	400	1380
En förälder röker	416	1823
Ingen förälder röker	188	1168

När vi vill presentera datat i en tabell för varje variabel använder vi koden

```
proc freq data=work.smoke;
weight frequency;
tables students parents;
run;
```

Efter komandot "*weight*" ska vi ange vilken variabel vi har frekvenserna i. Eftersom vi döpt den till "*frequency*" skriver vi det. (Ett annat tänkbart variabel namn är antal. Har ni döpt variabeln till det, skriver ni in det efter "weight".) Efter weight kommer kommandot "*tables*". Här ska vi ange vilka variabler vi vill göra tabellen för. I koden ovan anges variablerna "*students*" och "*parents*". Notera att det är ett mellanslag mellan de båda variabelnamnen. Koden vi skriver genererar en utskrift som ser ut så här

The FREQ Procedure

students	Frequency	Percent	Cumulative Frequency	Cumulative Percent
not_smoke	4371	81.32	4371	81.32
smoke	1004	18.68	5375	100.00

parents	Frequency	Percent	Cumulative Frequency	Cumulative Percent
both	1780	33.12	1780	33.12
none	1356	25.23	3136	58.34
one	2239	41.66	5375	100.00

Vill man istället göra en korstabell använder man koden

```
proc freq data=work.smoke;
weight frequency;
tables students*parents;
run;
```

Skillnaden är att när man skapar en korstabell sätter man en asterix mellan de variabler man vill skapa korstabellen för. Denna kod genererar en utskrift som ser ut så här

Frequency Percent Row Pct Col Pct	both	none	one	Total
not_smok	3203	1168	0	4371
	59.59	21.73	0.00	81.32
	73.28	26.72	0.00	
	88.90	86.14	0.00	
smoke	400	188	416	1004
	7.44	3.50	7.74	18.68
	39.84	18.73	41.43	
	11.10	13.86	100.00	
Total	3603	1356	416	5375
	67.03	25.23	7.74	100.00

Förklaring till bilden ovan: Det är 3203 studenter som inte röker och har föräldrar som båda röker. Dessa 3203 utgör 59.59% av hela urvalet. 73.28% av de studenter som inte röker har två föräldrar som båda röker. 88.90% av alla studenter som har två föräldrar som båda röker, röker *inte* själva.

Beskriva data numeriskt

Det finns olika procedurer som beskriver datat numeriskt i SAS. Vi börjar med att lära oss använda "*proc means*". I exemplet nedan använder vi oss av datat "*work.nummer*" som går att hitta i Datorövning 1, Statistikens Grunder 1. I data-setet hade vi tre variabler; X, X^2 och $\ln X$. För att beskriva dessa variabler numeriskt använder vi koden

```
proc means data=work.nummer;
run;
```

och vi får utskriften

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
x	3	4.0000000	3.0000000	1.0000000	7.0000000
xsquare	3	22.0000000	24.5560583	1.0000000	49.0000000
lnx	3	1.1107348	1.0017941	0	1.9459101

De mått som "*proc means*" ger är alltså

- antal observationer
- medelvärde
- standardavvikelse
- minimum
- maximum

(Det går att välja vilka mått som ska anges, men här lär vi oss bara default-värdena.)

Om man har ett data-set med många variabler vill man kanske bara beskriva en eller några stycken av dem. Då använder man koden

```
proc means data=work.nummer;
var x;
run;
```

Anger vi "*var* x" kommer "*proc means*" bara beräkna mått på variabeln x.

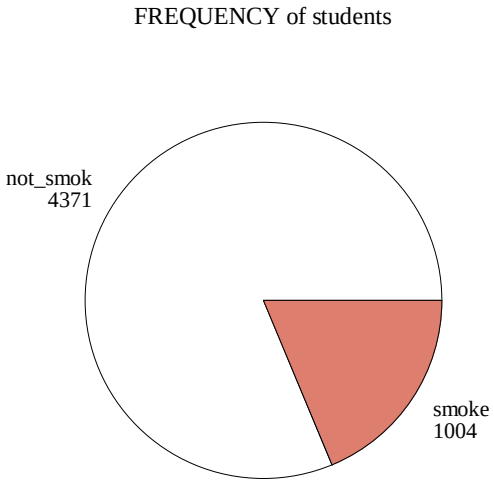
Presentera data i grafer

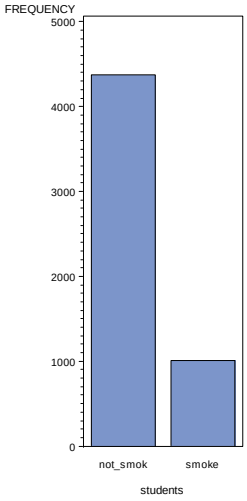
Detta exempel handlar om att göra cirkel- och stapeldiagram. För att illustrera använder vi oss av data-setet "*work.smoke*". Vi ritar graferna med koden

```
proc gchart data=work.smoke;
pie students / freq = frequency;
vbar students / freq = frequency;
run;
```

Proceduren som gör denna typ av grafer är alltså "*proc gchart*". Kommandot "*pie*" anges om vi vill rita ett cirkeldiagram och kommandot "*vbar*" om vi vill rita ett vertikalt stapeldiagram. Eftersom vi har angett antalet i variabeln "frequency" måste vi lägga till kommandot "*freq*".

Koden genererar grafer som ser ut som följer





Om vi använder ett data-set som inte är skrivet med variabeln "frequency", utan har datat uppräknat observation för observation utesluter man helt enkelt koden "*/ freq=frequency*". Det kan se ut så här

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

Vi har dock lagt till ett annat kommando till denna kod. Eftersom den kvantitativa variabeln "*gender*" är kodad 0 och 1 specificerar vi att variabeln ska behandlas som en diskret variabel, och inte som en kontinuerlig.

Beräkna sannolikheter för en binomialfördelad variabel samt beskriva fördelningen grafiskt

För att presentera frekvensfunktionen för en viss fördelning grafiskt måste vi först skapa ett data-set som innehåller värden på variabeln som har den fördelningen samt beräkna sannolikheter för just dessa värden. Vi gör detta för binomialfördelningen genom koden

```
data work.binomial;
do x = 0 to 20 by 1;
probability = pdf('binomial', x, 0.1, 20);
output work.binomial;
end;
run;
```

För att generera tal använder vi en "*do-sats*". En "*do-sats*" ska alltid avslutas med "*end*".

Här låter vi x vara variabelvärdena som går från 0 till 20 i steg om 1 (fördelningen är diskret). Sedan skapar vi en variabel som heter "*probability*" vilken anger sannolikheten för just det värdet som x antar. Vi skriver "*pdf*" för att det är en frekvensfunktion vi vill göra (jämför "*cdf*"). Därefter anger vi vilken fördelning X har. Här gäller

$$X \sim bin(n = 20,\ p = 0.1).$$

Efter det skriver vi "*output*" för att specificera i vilket data-set vi vill spara dessa variabler. Här anger vi samma filnamn som ovan. När "*do-satsen*" är slut skriver vi "*end*" och när hela koden är klar skriver vi "*run*".

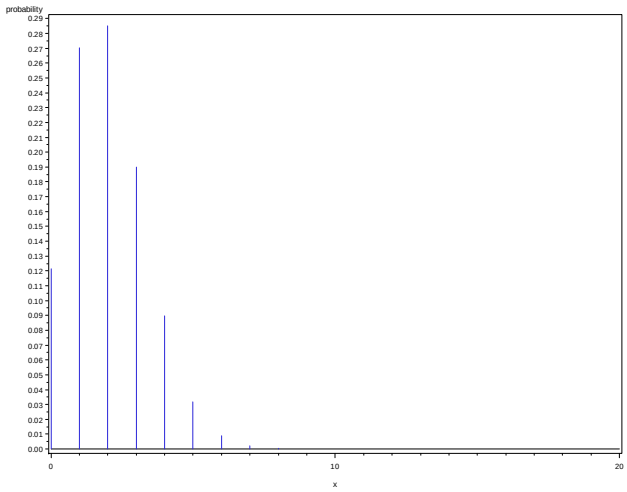
Om vi granskar datat med "*proc print*" får vi följande

Obs	x	probability
1	0	0.12158
2	1	0.27017
3	2	0.28518
4	3	0.19012
5	4	0.08978
6	5	0.03192
7	6	0.00887
8	7	0.00197
9	8	0.00036
10	9	0.00005
11	10	0.00001
12	11	0.00000
13	12	0.00000
14	13	0.00000
15	14	0.00000
16	15	0.00000
17	16	0.00000
18	17	0.00000
19	18	0.00000
20	19	0.00000
21	20	0.00000

För att plotta upp frekvensfunktionen använder vi koden

```
proc gplot data=work.binomial;
plot probability*x;
symbol i=needle;
run;
```

Efter kommandot "*plot*" skriver vi in variablerna vi vill plotta. Den variabel som ska vara på y-axeln skriver man först. Kommandot "*symbol*" är ett kommando som man använder när man vill ändra utseendet på plotten. Skriver vi inget kommer observationerna representeras av plus-tecken. När vi skriver "$i = needle$" betyder det att vi vill representera observationerna med stolpar. Plotten ser ut så här



Nu vill vi rita upp fördelningsfunktionen för binomialfördelningen. Vi börjar med att skapa variabelvärden samt att beräkna sannolikheter för dessa

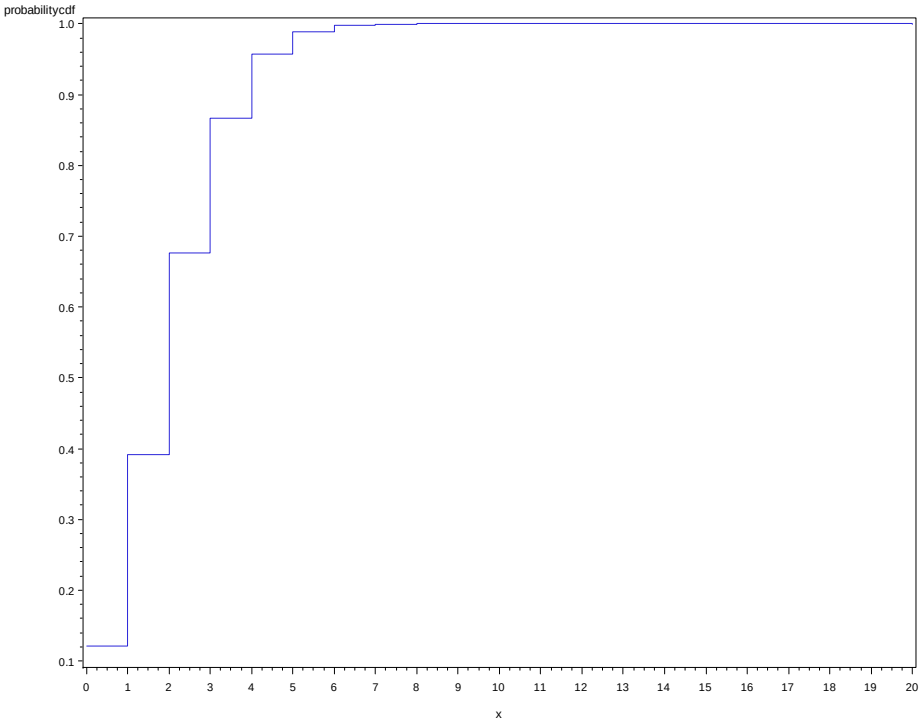
```
data work.binomial;
do x = 0 to 20 by 1;
probabilitycdf = cdf('binomial', x, 0.1, 20);
output work.binomial;
end;
run;
```

Variabeln som innehåller sannolikheter kallar vi här "*probabilitycdf*". Detta för att kunna skilja de olika sannolikheterna åt. Vi ser ovan att koden för att generera sannolikheter som är kumulativa är densamma som för de enskilda sannolikheterna, förutom att man skriver "*cdf*" istället för "*pdf*" efter likamed tecknet.

För att rita upp fördelningsfunktionen använder vi återigen samma kod som ovan, men byter ut variabeln "*probability*" mot "*probabilitycdf*".

```
proc gplot data=work.binomial;
plot probabilitycdf*x / haxis = 0 to 20 by 1;
symbol i=stepJ ;
run;
```

Vi lägger till två valbara kommandon. Det första är "*haxis = 0 to 20 by 1*", vilket gör att vi ser alla värdena på $x-$axeln. Det andra är "*symbol=stepJ*" för att vi vill rita ett trappstegsdiagram. Vi lägger till "J" för att vi vill binda ihop "trappstegen". Prova gärna att köra koden utan J och haxis kommandot. Detta trappstegsdiagram ser ut så här



Beräkna sannolikheter hos en normalfördelad variabel samt beskriva fördelningen grafiskt

Nu ska vi rita täthetsfunktionen hos en normalfördelning. Vi börjar med att skapa en variabel med värden och sannolikheter.

```
data work.normal;
do x = -12 to 18 by 0.05;
density = pdf('normal', x, 3, 5);
output work.normal;
end;
run;
```

Här är fördelningen för variabeln

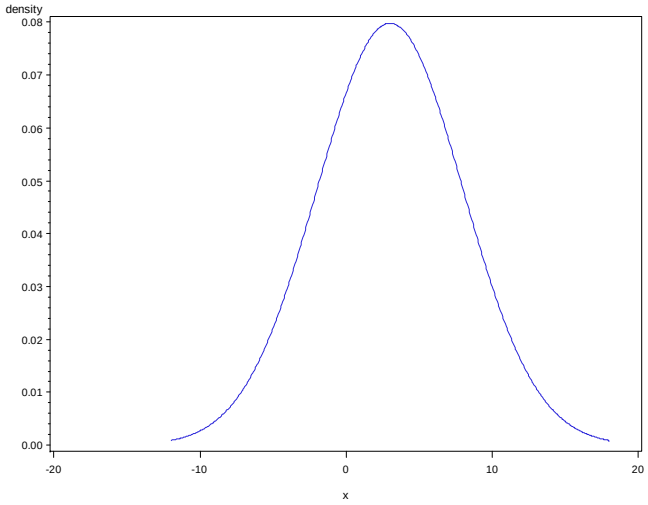
$$X \sim Normal(\mu = 3,\ \sigma = 5)$$

Eftersom vi vet att sannolikhetsytan är nästan 0 när vi befinner oss $3 \times \sigma$ steg från μ åt båda hållen väljer vi att X ska gå från -12 till 18.

Nu ritar vi plotten med koden

```
proc gplot data=work.normal;
plot density*x;
symbol i=join;
run;
```

Här ser vi att vi väljer ännu ett nytt utseende hos plotten. Genom att skriva "*i=join*" efter "*symbol*" binder vi ihop punkterna. Plotten får utseendet



Vi kan sammanfatta de olika valmöjligheterna vi har när vi specificerar "*symbol*" kommandot.

- "i = needle" ritar stolpar
- "i = stepJ" ritar ett trappstegsdiagram
- "i = join" binder ihop punkterna
- skriver vi inget får vi plus-tecken

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 korstabell för variablerna kön och butik. Hur många procent av alla kvinnor handlar på Clas Ohlson?

2. Använd rokdata.xls från Datorövning 1, Statistikens Grunder 1 för att rita ett stapeldiagram och ett cirkeldiagram för de båda variablerna.

3. Tentaresultatet hos 10 klasskamrater har registrerats

45	57	59	97	83
72	74	29	49	56

   Läs in datat och beräkna medelvärde och standardavvikelse för tentaresultatet.

4. Rita **fördelningsfunktionen** för normalfördelningen i exemplet ovan där

$$X \sim Normal(\mu = 3,\ \sigma = 5).$$

5. Rita **täthetsfunktionen** hos en normalfördelning där

$$X \sim Normal(\mu = 7,\ \sigma = 3).$$

   Skapa värden på X så att så gott som hela fördelningen representeras.

6. Rita en **frekvensfunktion** för en binomialfördelning som uppfyller kraven för en normalapproximation. Prova att använda både "symbol i=join" och "symbol i=needle".

Extra uppgifter

1. Använd datat i basuppgift 1. Gör en separat tabell för variabeln kön och en separat tabell för variabeln butik.

2. Beskriv datat från extrauppgift 1, Datorövning 1, Statistikens Grunder 1 numeriskt.