Datorövning 4

Statistikens Grunder 2

Syfte

1. Lära sig utföra goodness of fit-test
2. Lära sig utföra test av homogenitet
3. Lära sig utföra prövning av hypoteser om oberoende

Exempel

Utföra goodness of fit-test

Utgå från exempel 18.1 i kompendiet som handlar om tärningskast. Vi kan läsa in datat på följande sätt:

```
data work.dice;
input number count;
datalines;
1 18
2 23
3 16
4 21
5 18
6 24
;
run;
```

Vi vill nu testa om tärningen är symmetrisk eller ej. Om tärningen är symmetrisk, och vi kastar den 120 gånger förväntar vi oss att varje nummer ska komma upp 20 gånger. Vi får hypoteserna

$$\begin{aligned} H_0 &: \text{Lika fördelning mellan nummer} \\ H_A &: \text{Ej lika fördelning} \end{aligned}$$

För att göra ett goodness of fit-test skriver vi koden

```
proc freq data=work.dice;
weight count;
tables number / testf=(20 20 20 20 20 20) chisq;
run;
```

Eftersom datat är inläst med en variabel som innehåller frekvenser, måste vi ange "*weight count*". Efter "*tables*" anger vi den variabel som vi vill testa frekvenser för. "*testf*" står för testfrekvens, men det går även bra att skriva

"*testp*" för att istället ange testprocent. "*chisq*" skriver vi för att vi vill att SAS ska räkna ut chi-två värdet.

Vi får utskriften

The FREQ Procedure

number	Frequency	Test Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	18	20	15.00	18	15.00
2	23	20	19.17	41	34.17
3	16	20	13.33	57	47.50
4	21	20	17.50	78	65.00
5	18	20	15.00	96	80.00
6	24	20	20.00	120	100.00

Chi-Square Test for Specified Frequencies

Chi-Square	2.5000
DF	5
Pr > ChiSq	0.7765

Sample Size = 120

Det observerade chi-två värdet blev 2.5, vi har 5 frihetsgrader och p-värdet är 0.7765. Bör vi förkasta H_0?

Utföra test av homogenitet

Utgå från exempel 18.2 i kompendiet som handlar om ätvanor framför TV:n inom olika länder. Vi kan läsa in datat på följande sätt:

```
data work.ex182 ;
input land$ intensitet$ antal;
datalines;
sverige ofta 512
sverige ibland 125
sverige sällan 212
italien ofta 395
italien ibland 128
italien sällan 404
england ofta 685
england ibland 130
england sällan 32
usa ofta 743
usa ibland 131
usa sällan 71
kina ofta 532
kina ibland 184
kina sällan 181
;
run;
```

Vi vill nu testa om fördelningen av intensiteten är densamma för länderna. Vi får hypoteserna

$$\begin{aligned} H_0 &: \text{Fördelning av intensitet är samma över länderna} \\ H_A &: \text{Fördelning av intensitet är ej samma över länderna} \end{aligned}$$

För att testa detta i SAS skriver vi koden

```
proc freq data=work.ex182 order=data;
weight antal;
tables land*intensitet / chisq expected cellchi2 norow nocol;
run;
```

Kommandot "*order=data*" används för att vi vill att SAS ska skriva ut tabellen i samma ordning som datat är inläst. Vi skriver "*chisq*" för att få ut chi-två värdet, det vill säga

$$\sum_{i=1}^{n} \frac{(n_i - E[n_i])^2}{E[n_i]}.$$

Vi skriver "*expected*" för att vi vill få ut förväntat värde i varje cell, det vill säga $E[n_i]$. Vi skriver "*cellchi2*" för att få chi-två värdet i varje cell, det vill säga

$$\frac{(n_i - E[n_i])^2}{E[n_i]},$$

för varje i. Vi skriver "*norow*" och "*nocol*" för att vi INTE vill ha ut rad- och kolumnprocent, detta för att minimera antalet värden i varje cell.

Vi får dels en korstabell med olika värden i varje cell,

The FREQ Procedure

Table of land by intensitet

land	intensitet			
Frequency Expected Cell Chi-Square Percent	ofta	ibland	sällan	Total
sverige	512	125	212	849
	545.15	132.72	171.13	
	2.0155	0.4492	9.7602	
	11.47	2.80	4.75	19.01
italien	395	128	404	927
	595.23	144.92	186.85	
	67.356	1.9744	252.35	
	8.85	2.87	9.05	20.76
england	685	130	32	847
	543.86	132.41	170.73	
	36.626	0.0438	112.73	
	15.34	2.91	0.72	18.97
usa	743	131	71	945
	606.79	147.73	190.48	
	30.576	1.8944	74.946	
	16.64	2.93	1.59	21.16
kina	532	184	181	897
	575.97	140.23	180.81	
	3.3565	13.665	0.0002	
	11.91	4.12	4.05	20.09
Total	2867	698	900	4465
	64.21	15.63	20.16	100.00

och dels en tabell med chi-två värden och p-värden.

The FREQ Procedure

Statistics for Table of land by intensitet

Statistic	DF	Value	Prob
Chi-Square	8	607.7415	<.0001
Likelihood Ratio Chi-Square	8	628.2416	<.0001
Mantel-Haenszel Chi-Square	1	93.3486	<.0001
Phi Coefficient		0.3689	
Contingency Coefficient		0.3461	
Cramer's V		0.2609	

Sample Size = 4465

Vi är bara intresserade av den första raden utskriften. Chi-två värdet blev

607.7415 och tillhörande p-värde är mindre än 0.0001. Vi bör alltså förkasta H_0. Övertyga er om att ni kan få fram chi-två värdet genom att summera alla cellers chi-två värden.

Utföra prövning av hypoteser om oberoende

Utgå från exempel 18.3 i kompendiet som handlar om civilstånd och typ av bostad. Vi kan läsa in datat på följande sätt:

```
data work.ex183;
input bostad$ civilstand$ antal;
datalines;
villa gift 25
villa ogift 5
villa frånskild 10
hyreslägenhet gift 15
hyreslägenhet ogift 25
hyreslägenhet frånskild 20
;
run;
```

Vi vill nu testa om det råder oberoende mellan "*civilstånd*" och "*bostadstyp*", det vill säga

$$\begin{aligned} H_0 &: \text{oberoende mellan "civilstånd" och "bostadstyp"} \\ H_A &: \text{beroende mellan "civilstånd" och "bostadstyp"} \end{aligned}$$

För att testa skriver vi koden

```
proc freq data=work.ex183 order=data;
weight antal;
tables bostad*civilstand / chisq expected cellchi2 norow nocol;
run;
```

Som vi ser använder vi samma kod som för exemplet "*utföra test av homogenitet*". Vi får en liknande utskrift

The FREQ Procedure

Table of bostad by civilstand

bostad / Frequency / Expected / Cell Chi-Square / Percent	civilstand: gift	ogift	frånskil	Total
villa	25	5	10	40
	16	12	12	
	5.0625	4.0833	0.3333	
	25.00	5.00	10.00	40.00
hyresläg	15	25	20	60
	24	18	18	
	3.375	2.7222	0.2222	
	15.00	25.00	20.00	60.00
Total	40	30	30	100
	40.00	30.00	30.00	100.00

Statistics for Table of bostad by civilstand

Statistic	DF	Value	Prob
Chi-Square	2	15.7986	0.0004
Likelihood Ratio Chi-Square	2	16.4528	0.0003
Mantel-Haenszel Chi-Square	1	7.2337	0.0072
Phi Coefficient		0.3975	
Contingency Coefficient		0.3694	
Cramer's V		0.3975	

Sample Size = 100

p-värdet är så pass litet att vi kan förkasta H_0 på alla signifikansnivåer som är större än $\alpha = 0.0004$. Det råder alltså beroende mellan variablerna "*civilstånd*" och "*bostadstyp*".

Uppgifter

Basuppgifter

1. Utgå från övning 18.1 i kompendiet. Läs in datat och testa

   $$H_0 \quad : \quad \text{Lika fördelning mellan märke}$$
   $$H_A \quad : \quad \text{Ej lika fördelning}$$

   Använd signifikansnivå $\alpha = 0.05$. Bör vi förkasta H_0?

2. Utgå från övning 18.8 i kompendiet. Läs in datat och testa

   $$H_0 \quad : \quad \text{Fördelning av "butik" är samma för båda könen}$$
   $$H_A \quad : \quad \text{Fördelning av "butik" är ej samma för båda könen}$$

   Använd signifikansnivå $\alpha = 0.05$. Bör vi förkasta H_0?

3. Utgå från övning 18.10 i kompendiet. Läs in datat och testa

$$\begin{aligned} H_0 &: \text{Oberoende mellan "uppfattning om rökning" och "rökvana"} \\ H_A &: \text{Beroende mellan "uppfattning om rökning" och "rökvana"} \end{aligned}$$

Använd signifikansnivå $\alpha = 0.05$. Bör vi förkasta H_0?

Extra uppgifter

1. Utgå från tentamen 2009-10-29, fråga 4. Läs in datat och testa

$$\begin{aligned} H_0 &: \text{proportionen över inkomstklasser är } \{0.18\ 0.13\ 0.19\ 0.20\ 0.3\} \\ H_A &: \text{proportionen är ej } \{0.18\ 0.13\ 0.19\ 0.20\ 0.3\} \text{ över inkomstklasserna} \end{aligned}$$

Använd signifikansnivå $\alpha = 0.05$. Bör vi förkasta H_0?

2. Utgå från tentamen 2009-11-23, fråga 4. Läs in datat och testa

$$\begin{aligned} H_0 &: \text{oberoende mellan "inställning till småbilar" och "personlighetstyp"} \\ H_A &: \text{beroende mellan "inställning till småbilar" och "personlighetstyp"} \end{aligned}$$

Använd signifikansnivå $\alpha = 0.01$. Bör vi förkasta H_0?

3. Utgå från övning 18.2a. Läs in datat och testa

$$\begin{aligned} H_0 &: \text{proportionen är } \{9{:}3{:}3{:}1\} \\ H_A &: \text{proportionen är inte } \{9{:}3{:}3{:}1\} \end{aligned}$$

Använd signifikansnivå $\alpha = 0.05$. Bör vi förkasta H_0?