Dr. Quapp: Statistik für Mathematiker mit SPSS

Hinweise zur 9. Übung – Kontingenztafel + Likelihood-Schätzer + Zufallszahlen erzeugen

1.] Die Verteilungsdichte eines 2-dimensionalen Zufallsvektor (X, Y) ist folgendermaßen definiert:

P(X=x,Y=y)	1	2	3
0	0,3	0,2	0,1
1	0,2	0,1	0,1

a) Denken Sie sich eine Datentabelle in SPSS aus, die dieser Kontingenztafel entspricht. Bestimmen Sie die Randverteilungen von X und Y.
b) Berechnen Sie den Korrelationskoeffizienten von X und Y.
c) Bestimmen Sie die Verteilung, die Erwartung und die Varianz von $2X+Y$.

Die Aufgabe ist so reduziert, daß der theoretische Teil mit Hand gelöst werden kann. Dies sollte man auch einmal durchrechnen!
zu a) In SPSS erzeugt man eine Datentabelle mit 10 Zeilen und zwei Spalten für X und Y: Sie enthält die Paare (0,1), (0,1), (0,1), (1,1), (1,1), (0,2), (0,2), (1,2), (0,3) und (1,3).
Will man sich, auch bei grösseren Problemen, das Aufzählen aller einzelnen Fälle sparen, kann man mit Gewichten arbeiten: Das Paar (0,1) zum Beispiel bekommt Gewicht 3, und dies muss in einem extra Fenster → DatenWichten dann eingestellt werden. Die 6 Paare haben Gewichte: 3,2,1,2,1,1. Mit BeschreibenderStatistik, Kreuztabellen kann die gegebene Tabelle erzeugt werden.
Die Randverteilungen sind entsprechende Zeilen- und Spaltensummen; also 0.6 und 0.4 für x, sowie 0.5, 0.3, und 0.2 für y.
zu b) Es ergibt sich: $E\,X = 0.4, E\,Y = 1.7$, und es ist:
$Cov(X,Y) = E[(X - E\,X)(Y - E\,Y)] = E\,XY - E\,X\;E\,Y$ mit $E\,XY = 1\,0.2 + 2\,0.1 + 3\,0.1 = 0.7$, also $Cov(X,Y) = 0.7 - 0.4\,1.7 = 0.02$
$E\,X^2 = 0.4$, $Var\,X = 0.4 - 0.4^2 = 0.24$,
$E\,Y^2 = 3.5$, $Var\,X = 3.5 - 1.7^2 = 0.61$,
also
$\rho = Cov(X,Y)/\sqrt{Var\,X\;Var\,Y} = 0.052$. Dies bedeutet, daß X und Y praktisch unabhängig sind. Aus den Produkten der Randverteilung ergibt sich eine Tabelle mit unabhängigen X und Y, vergleiche die Aufgabe:

P(X=x,Y=y)	1	2	3	$\sum$
0	0,3	0,18	0,12	0.6
1	0,2	0,12	0,08	0.4
$\sum$	0.5	0.3	0.2	$\sum$=1

zu c) Für $Z = 2\,X + Y$ ergibt sich durch Ausrechnen die Tabelle: Also wird

z	1	2	3	4	5
P(Z=i)	0.3	0.2	0.3	0.1	0.1
z^2	1	4	9	16	25

$E\,Z = 2.5 = 2\,E\,X + E\,Y$ und $E\,Z^2 = 7.9$. Damit dann die Varianz $Var\,Z = 7.9 - 2.5^2 = 1.65$. Man beachte wieder, daß SPSS die "empirischen" Varianzen berechnet: $Var_{emp}\,X$=0.26, $Var_{emp}\,Y$=0.68, $Var_{emp}\,Z$=1.83. Zum Vergleich muß man diese mit (n-1)/n=0.9 multiplizieren. Siehe auch Befehle in spss091.sps .

2.] Erzeugen Sie in Form einer Tabelle zur hypergeometrischen Verteilung h(r,N,n,R) mit N=13 und n=6 die Wahrscheinlichkeitsdichten h(r_i,N,n,R_j) für r_i aus dem Intervall $[0, 6]$ und R_j aus dem Intervall $[1, 13]$. Deuten Sie Zeilen und Spalten dieser Tabelle als Wahrscheinlichkeitsdichten bzw. als Likelihood-Funktion.
Überlegen Sie sich Konfidenzintervalle entsprechender Likelihood- Schätzer zum Niveau α=0.1 .
Die hypergeometrische Verteilung beschreibt die Wahrscheinlichkeiten beim Urnenmodell ohne Zurücklegen. Es seien
N die Anzahl der Kugeln, n eine Stichprobe,
R die Anzahl der roten Kugeln in N, und r die roten Kugeln in der Stichprobe. Dann gibt

$$P(X = r) = h(r; N, n, R) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

als Funktion von r mit den Parametern N, n, R die Wahrscheinlichkeit an, daß genau r rote Kugeln in der Stichprobe sind. Die Verteilungsfunktion wird kumulativ

$$H(r; N, n, R) = \sum_{k=0}^{r} h(k; N, n, R) \quad \text{mit} \quad H(n; N, n, R) = 1.$$

Im Parameter R ist $H(r; N, n, R)$ monoton fallend.
Ein Schätzproblem besteht darin, aus einer Stichprobe zurück auf R zu schätzen. Die Maximum-Likelihood-Idee sucht dabei jenen Parameterwert, für den das beobachtete Testresultat am wahrscheinlichsten ist. (Bei einer bekannten Art der Verteilungsdichte ist "nur" der zugehörige Parameter gesucht.) Sei bei einer Realisierung der Stichprobe für ein festes r die Parameterfunktion dann $L_r(R) = h(r; N, n, R)$ – bei festem N und n. Wir suchen den Maximum-Likelihood-Schätzer $L_r(\hat{R}) = \sup_R L_r(R)$. Für $R < r$ ist $L_r(R) = 0$. Wir betrachten den Quotient für zwei aufeinanderfolgende R-Werte:

$$\frac{L_r(R+1)}{L_r(R)} = ... = \frac{(R+1)(N-R-n+r)}{(R+1-r)(N-R)} < 1 \; ? \; ?$$

Die Kleinerrelation ist für

$$\frac{r(N+1)}{n} < R + 1$$

erfüllt. Also ist die Grenze $\hat{R} = \left[\frac{r(N+1)}{n}\right]$. Davor gilt die Ungleichung nicht, bei $\hat{R}$ ist das Maximum überschritten. $\hat{R}$ ist der Maximum-Likelihood-Schätzer für R auf Grund des gefundenen r. Im Grenzfall $r = n$ folgt $\hat{R} = N$ was nicht unlogisch ist, und wenn $\hat{R}$ ganzzahlig ist, so ist auch noch $(\frac{r(N+1)}{n} - 1)$ ein Schätzer.

Sei nun N=13 und n=6. In SPSS können dann 7 Zeilen für r_i von 0 bis 6 belegt werden, und dazu sind dann 13 Variable Hy_j in 13 Spalten einzeln berechenbar:
$Hy_j = PDF.HYPER(r, 13, 6, Rj)$ mit je R_j=1,2,...,13 mit der Dichte der hypergeometrischen Verteilung von SPSS. (In SPSS ist die Variable Rj etwas schräg als "Treffer" bezeichnet.) Der Vollständigkeit halber sollte noch eine Hy_0-Spalte eingefügt werden: Wenn $R = 0$, dann ist sicher auch $r = 0$, also steht in der ersten Zeile dort eine 1, sonst Nullen. Die Zeilen dieser Hy-Tabelle ergeben die Likelihood-Funktion zu einer Realisierung der Stichprobe für r .
Durch einen Syntax-Zyklus ist die Rechnung wieder zu vereinfachen:

```
/* Ein ZyklusHypergeometricus fuer Zahlen R_j in (1,13)  */
Input Program .
LOOP #I=1 to 7 .
 Compute r=#I -1 .
 Compute Rgr=1 .
   DO REPEAT
     B=Hy1 to Hy13 .
     COMPUTE B= PDF.HYPER(r,13,6,Rgr) .
     Compute Rgr=Rgr+1 .
   END REPEAT .
END CASE .
END LOOP .
END FILE .
END INPUT PROGRAM .
EXECUTE .
/* Hy_i: Tabelle der Dichten der Hypergeometrischen Verteilung fuer N=13, n=6 */
/* Die Zeilen sind fuer r, die Spalten fuer R */
```

Zu Konfidenzbereichen: Betrachtet man eine Spalte der Tabelle für einen festen Wert von R, so ist die Summe der Wahrscheinlichkeiten über P(R=r) Eins (wie zu erwarten war). Die Zeilen dagegen bilden die jeweilige Likelihood-Funktion. Bei $r = 4$ etwa hat diese ihr Maximum bei $\hat{R}$=9, dies ist somit der Maximum-Likelihood-Schätzer für R, was auch aus obiger Formel herauskommt! Bei gegebenem r kann man natürlich nicht mit 100%-iger Sicherheit auf ein R schließen, nur $r \leq R \leq N-(n-r)$ ist sicher. Grenzt man aber die "Sicherheit" ein auf einen Wert 1-α, so kann man grob gesprochen den kumulativen Bereich der entsprechenden Wahrscheinlichkeiten ablesen, wenn man die Zeilen noch "normiert". Bei α=0.1 bleibt die Rest-Summe 0.9 übrig, und bei $R = 9$ etwa sind das die Fälle für r=3, 4, 5. Betrachtet man nun ein erhaltenes r, so kann man diese R-Intvalle abzählen. Das Resultat $c(r)$ sind die folgenden Intervalle für R.

r	0	1	2	3	4	5	6
$\hat{R}$	0	2	4	6, 7	9	11	13
c(r)	0-3	1-6	2-8	4-9	5-11	7-12	10-13

Es ist klar, daß der Schätzer im Konfidenzintervall liegt.

3.] Zur Erzeugung gleichmäßig in $(0,1)-$verteilter Zufallszahlen benutzt man häufig lineare Kongruenzen: Zunächst werden Zufallszahlen über der Menge {0, 1, 2, ... ,m} gemäß $x_{n+1} = (a\,x_n) mod m$ erzeugt, wobei a eine ganze Zahlen ist. Dann sind $u_n = x_n/m$ Zufallszahlen aus $(0,1)$. Erzeugen Sie mit dieser Vorschrift 200 Zufallszahlen mit $a = 2^{16} + 3$ und $m = 2^{31}$. Analysieren Sie die Daten durch Vergleich mit Zufallszahlen von SPSS. (Variieren Sie auch a und m !)

Folgendes SPSS-Programm löst die Aufgabe (Syntax-Datei in D: spss093.sps):

```
/* Berechnung von Zufallszahlen aus (0,1)  */
/* Ein erstes Feld im Datenfenster muss aktiviert sein */
/* Startwerte  */
  COMPUTE y=100000.
  COMPUTE A=2**16 +3.
  COMPUTE m=2**31.
EXECUTE.
/*   Ein Zyklus fuer Zahlen in (0,1)  */
/*   Die Hilfsvariable Y liegt dabei in (0,m)  */
Do Repeat
B=x1 to x200.
COMPUTE Y=MOD(Y*A,m).
COMPUTE B=Y/m.
END REPEAT.
EXECUTE.

/* Transponieren der berechneten Zeile in eine Spalte,
   die neue Variable dieser Spalte wird var001 */
FLIP Variables= x1 to x200 .

/* Berechne Vergleichsvariable uu */
COMPUTE
uu=RV.UNIFORM(0,1) .
EXECUTE.
 GRAPH
  /HISTOGRAM=var001 .
 GRAPH
  /HISTOGRAM=uu .
 EXECUTE.
```

Die Histogramme sind im Allgemeinen vom SPSS-Programm mit verschieden-anzahligen Balken berechnet, so daß die Bilder nicht direkt vergleichbar sind. Dies kann man mit Hand verstellen. Aber unsere Zufallszahlen liegen sozusagen theoretisch gut im Trend: Der Mittelwert war bei einem Test mit 0.49 sogar besser als der von SPSS mit 0.47. Es ist ja $E\,var001 = \frac{1}{2}(B - A)$, wenn $(A,\ B)$ das Intervall ist, und $\sigma = (B - A)/\sqrt{12}$, also hier 0.2887. Die Streuung von $var001$ ist 0.28, die von uu 0.29; also beides sehr ordentliche Werte.