Dr. Quapp: Statistik für Mathematiker mit SPSS

Hinweise zur 12. Übung – Regression mit Konfidenzintervallen und χ^2-Anpassung –

1.] Der "Schiefe Turm" von Pisa wurde im Dezember 2001 nach langer Behandlung wieder für Besucher freigegeben. Folgende Meßreihe schildert das dramatische Vorgeschehen. Angegeben ist die Schiefe y in Addition zu 2,90 m in cm. x sind die Jahreszahlen. 1975 betrug die Schiefe also 2,9642 m.

x	75	76	77	78	79	80	81	82	83	84	85	86	87
y	6,42	6,44	6,56	6,67	6,73	6,88	6,96	6,98	7,13	7,17	7,25	7,42	7,57

(a) Finden Sie die Regressionsgrade. Welcher prozentuale Anteil der Änderung der Meßwerte ist durch die Gerade erklärbar?

Die Regressionsgrade ist $y = -0.611 + 0.09319\,x$. Durch Einsetzen ergibt sich sofort der Wert für (c) von y(1997)= $8.4292cm$. Beim Verwenden der →Methode Kurvenanpassung (Linear) wird auch der Korrelationskoeffizient ausgegeben: 0.99397, sowie sein quadrierter Wert: R Square= 0.98798. Dieser gibt den Anteil an, den die Regressionsgrade an der Entwicklung der y-Werte hat, nämlich 98.798%. Nur 1.202% sind demnach zufällige Jahresschwankungen. Dies kann man auch gut in einem Diagramm sehen: Die Meßpunkte schmiegen sich eng an die Regressionsgrade an.

(b) Geben Sie ein 95% Konfidenzintervall CI für den wahren Anstieg β_1 an! Interpretieren Sie dieses Intervall.

Dieses CI wird im direkten → Linearen Fit →Button Statistik mitgeliefert, es muß dort extra angeklickt werden. Das Intervall ist $\beta_1 \in (0.086, 0.1)$. Im ungünstigen Fall könnte die Schiefe also mit dem Anstieg 0.1 wachsen. Bei dieser Methode kann das Konfidenzintervall für die y-Werte noch variabel eingestellt werden, und außerdem kann unterschieden werden zwischen individuellem CI per Regressor, und mittlerem CI.

(c) Die mit dem Problem des "Schiefen Turms" befaßten Bauleute waren 1987 naturgemäß an einer Aussage interessiert, um wie viele cm sich der Turm bis z.B. 1997 weiter neigen würde, wenn keinerlei Korrekturen am Bau vorgenommen werden würden. Nutzen Sie die Regressionsgerade für eine derartige Vorhersage.

Es ist $y(97) = 8.4292\,cm$, und z.B. $y(104) = 9.08076cm$, also im abgelaufenem Jahr 2004 wäre dieser Wert erwartet worden. (Aber 1990 hat man mit ernsten Korrekturen begonnen, die ein langsames Zurückkippen von etwa $40cm$ bewirkt haben.)

(d) Bestimmen Sie die Fehlergrenze für die vorhersagbare mittlere Entwicklung der Schiefe im Jahr 1997 für ein 95% CI und ein 99% CI.

In →Methode Kurvenanpassung (Linear) →Speichern →Vorhergesagte Werte, Vorhersageintervalle ist eine Lösung dieser Teilaufgabe "automatisch" möglich. Man hat als x-Achse zuerst die Variable x einzusetzen, dann aber die weitere Möglichkeit →Zeit anzuklicken. SPSS verwendet dann die bisherigen Zeilen der Datentabelle als Zeitschritte,

also hier 13. Damit kann im Fenster →Speichern die Anzahl der weiteren Zeitschritte für eine Vorhersage angegeben werden. Wir haben Daten für 13 Jahre bis 1987, also sind bis 1997 dann 23 Zeitschritte einzugeben. Auch Konfidenzintervalle können in diesem Fenster angeklickt werden, und das CI-Niveau ist stufenweise einstellbar.

2.] (a) Erzeugen Sie die χ^2-Verteilung mit 4 Freiheitsgraden: Zeichnen Sie die Dichte der theoretischen χ^2 Verteilung über einer Achse zum Intervall (0,15).
(b) Bei der Untersuchung über den Schädlingsbefall von Apfelbäumen wurden drei verschiedene Apfelesorten (A,B,C) überprüft. Es wurden insgesamt $n = 100$ Bäume untersucht. Es ergab sich folgende Kontingenztafel:

Sorte \ Befall	**gering**	**mittel**	**stark**
A	22	6	2
B	11	12	7
C	17	12	11

Man prüfe die Unabhängigkeit von Schädlingsbefall und Sorte mit einem geeigneten Testverfahren zum Niveau $\alpha = 0,05$.

Das Problem besteht darin zu vergleichen, ob gleichzeitig die neun Werte B_1 bis B_9 des Befalls einer entsprechenden Unabhängigkeits-Bedingung, hier für eine Zähldichte p_i, $i = 1, ..., 9$, genügen. Die p_i ergeben sich bei einer Kreuztabelle als Produkte der jeweiligen Randverteilungen. Wie bei anderen Tests sucht man nach einer eindimensionalen Testgröße, mit der man die Gesamtschwankung adäquat wiedergeben kann. Betrachtet man etwa

$$\sum_{i=1}^{9}(B_i - n\,p_i)^2\,,$$

so kann man sich darauf stützen, daß die Zufallsvariablen B_i einzeln binomialverteilt mit den Parametern (n,p_i) sind. (Im Falle der Gültigkeit der Nullhypothese.) Für entsprechend großes n ist also

$$(B_i - n\,p_i)/\sqrt{n\,p_i(1-p_i)}$$

näherungsweise N(0,1)-verteilt. Wie in vorigen Aufgaben könnte man also wieder die Summe von quadrierten normalverteilten Zufallszahlen als χ^2-Verteilung auffassen:

$$\sum_{i=1}^{9}(B_i - n\,p_i)^2/(n\,p_i(1-p_i)) \sim \chi^2_9$$

Dies trifft jedoch nicht ganz zu, weil die Schwankungssumme der $(B_i - n\,p_i)^2$ wegen der Abhängigkeit $\sum_{i=1}^{9} B_i = n$ gegenüber dem stochastisch unabhängigen Fall verkleinert ist. Hat man es nur mit k Ausprägungen B_i zu tun wie in Aufgabe 3 unten, so ist die Zahl der Freiheitsgrade durch $(k-1)$ zu ersetzen. Bei einer Kreuztabelle wie in dieser Aufgabe ist die Zahl der Freiheitsgrade für Zeilen und Spalten je um Eins zu reduzieren: Hier ergeben sich $2 * 2$ Freiheitsgrade. Die Testgröße wird somit

$$Z_9 = \sum_{i=1}^{9}(B_i - n\,p_i)^2/(n\,p_i)$$

und diese ist mit dem $(1-\alpha)$-Quantil der χ^2_4-Verteilung zu vergleichen. Der Faktor $(1-p_i)$ im Nenner ist dabei zur Vereinfachung noch weggelassen.

zu (a) Die 151.Zeile ist anzuklicken, und die Achsenvariable x=($casenum-1)/10 zu setzen. Mit chi4=CDF.CHISQ(x,4) haben wir die Verteilungsfunktion, und mit chi4di=Diff(chi4) in →Berechnen, Zeitreihen ergibt sich die Dichte. In der kumulativen Verteilungsfunktion kann man sofort den Wert ablesen, der bei dem gängigen 95% CI gilt: Es ist 9.48 . Analoges kann man für No.3 unten mit 5 Freiheitsgraden (df) rechnen.
zu (b) Zuerst muß das Problem gelöst werden, wie man SPSS die Kreuztabelle beibringt! Dazu sind alle $n = 100$ Fälle einzeln als Zeilen einzugegeb. (Hinweis: Datei in D: spss122.sav leistet dies. – Wenn man mit gewichteten Fällen arbeitet, ginge es etwas schneller: Datei in D: apfel9.sav).
Bei →Deskriptive Statistik, Kreuztabelle kann dann die in der Aufgabe gegebene Tabelle erzeugt werden. Im Fenster →Statistik muß der Chi-Square Test angeklickt werden. Es resultieren je nach Methode zwei Werte: 10.7 oder 11.3, beide sind größer als der kritische Wert 9.48, also kann die Nullhypothese einer Unabhängigkeit des Befalls von der Sorte nicht bestätigt werden. Die Zahl der Freiheitsgrade DF ist dabei (m-1)(n-1) mit m Zeilen und n Spalten in der Kreuztabelle. Der Test gibt die genaue "Signifikanz" aus, die mit 2.96%, oder 2.33% auch "hinter" den 5% des geforderten CI's liegt.
Im Falle der Unabhängigkeit von Schädlingsbefall und Sorte hätte man die folgende Tabelle vorgefunden (mit den gleichen Randverteilungen /100 wie oben)

$Sorte \setminus Befall$	*gering*	*mittel*	*stark*	$p_{i.}$
A	15	9	6	0.3
B	15	9	6	0.3
C	20	12	8	0.4
$p_{.j}$	0.5	0.3	0.2	1.0

Abschließende Frage: "Wo kaufen Sie Ihre Äpfel?"
Antwort: "Wer keine weiche Birne hat, kauft harte Äpfel aus Halberstadt."

3.] (a) Erzeugen Sie die χ^2-Verteilung mit 5 Freiheitsgraden: Zeichnen Sie die Dichte der theoretischen χ^2 Verteilung über dem Intervall $x \in$(0,15).
(b) Erzeugen Sie für das Merkmal X $\sim$ Würfel je 600 Zufallszahlen. (Anleitung: Gleichverteilung im Intervall (1.0, 6.999) auf ganze Zahlen reduzieren.) Zeichnen Sie ein Histogramm für die sechs Kategorien von X, und überprüfen Sie die Gleichheit der Ausprägungen der einzelnen Augenzahlen mit dem χ^2-Anpassungstest.

Zur Theorie: Siehe No.2.
Zu (a) siehe ebenfalls No.2; man kann beide Kurven auch in einem Diagramm darstellen, um die Änderung bei mehr Freiheitsgraden zu sehen.
Zu (b) Mit W=TRUNC(RV.UNIFORM(1,6.9999)) erhalten wir den zufälligen Würfel. In den Tests sollte dann der →Chi-Quadrat-Test angeklickt werden (in → Nichtparametrische Tests), und in diesem ist nur noch darauf zu achten, daß beim voreingestellten Button angeklickt ist, daß alle Kategorien gleich sein sollen (Nullhypothese): D.h. hier soll gelten $p_i = 1/6$ für i=1,...,6. Es resultieren bei mehreren Tests Chi-Quadrat-Werte von 1.84 bis 8.21, alle Werte sind kleiner gewesen als der kritische Wert von 11.07 für 95% CI bei 5 Freiheitsgraden. Die ausgegebene Signifikanz ist mit 0.87 (entsprechend 87% > 5% CI) bis 0.18 sehr hoch bis hoch. Also kann die Nullhypothese nicht verworfen werden, das unser SPSS-Würfel "ehrlich" arbeitet.