Helge Toutenburg
Christian Heumann

Induktive Statistik

mit Beiträgen von

Michael Schomaker und Malte Wißmann

Eine Einführung mit R und SPSS für Windows.

Vierte, aktualisierte und erweiterte Auflage

13. November 2007

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Inhaltsverzeichnis

13. Fehlende Daten

In der Praxis taucht häufig das Problem auf, dass trotz aller Bemühungen bei der Erhebung der Daten die Ausprägungen eines oder mehrerer Merkmale an einigen, oft auch an vielen Untersuchungseinheiten, nicht erhoben werden konnten. Wir sind also **nach Erhebung der Daten** in der Situation, die wahren Merksmalsausprägungen nicht immer beobachtet zu haben. Die folgenden Beispiele sollen die Problematik verdeutlichen:

- Nichtantworter in statistischen Befragungen: Sensible Fragen nach z.B. Einkommen, Sexualverhalten oder Alkohol- und Drogenkonsum werden oft nicht beantwortet. Auch bei anonymen Befragungen ist dies ein häufiges Problem.
- Drop-out in klinischen (Längsschnitt-)Studien. Patienten kommen nach gewisser Zeit nicht mehr zu den Kontrolluntersuchungen. Die Gründe können vielfältig sein, z.B. weil sie die Behandlung nicht gut vertragen oder weil es ihnen sehr gut geht oder weil sie weggezogen sind. Ab einem gewissen Zeitpunkt hat man also keine Kenntnis über die Ausprägungen der Variablen bei diesen Patienten (z.B. Blutdruck, Blutwerte).
- Zensierte Daten in der Lebensdaueranalyse sind unvollständige Daten, da die genaue Lebensdauer nicht bekannt ist.
- Experimentelle Versuche: Ergebnisse fehlen z.B. wegen fehlerhafter Messgeräte.
- Geplantes Fehlen: Bestimmte Merkmale werden nur in einer Teilstichprobe erhoben, z.B. weil die Erhebung sehr teuer ist (z.B. aufwendiger medizinischer Test). Ein weiteres Beispiel für geplantes Fehlen sind in Fragebögen eingebaute Verzweigungen, die dazu führen, dass bestimmte Fragen nur dann beantwortet werden können, wenn eine andere Frage zuvor mit einer bestimmten Merkmalsausprägung beantwortet wurde. Als einfaches Beispiel diene die Frage nach Anzahl und Alter der Kinder, die nur dann sinnvoll beantwortet werden kann, wenn die Frage „Haben Sie Kinder?“ zuvor mit „ja“ beantwortet wurde.
- Schlechtes Design in einer Meinungsumfrage: „Bevorzugen Sie Kandidat A oder Kandidat B?“ Die Antwort wird vermutlich fehlen, wenn die Person keinen der beiden Kandidaten bevorzugt (deshalb sollte man dies als weitere Kategorie zulassen).

Teil IV

Einführung in statistische Software

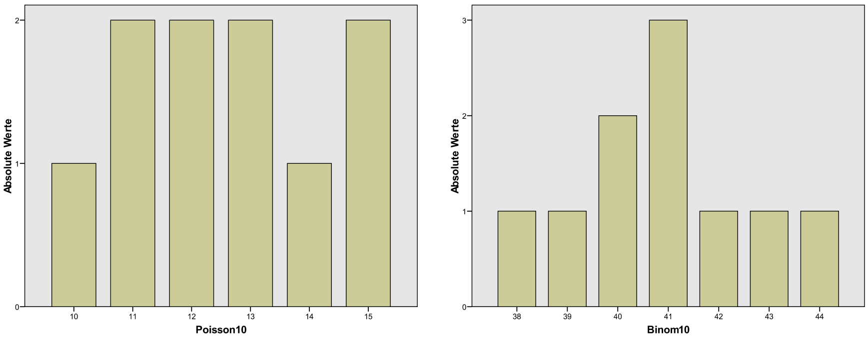

Abb. 14.5. Veranschaulichung von Zufallszahlen einer Poissonverteilung ($\lambda = 14$) und einer Binomialverteilung ($n = 100, p = 0.4$) bei einer $N = 10$-fachen Realisierung

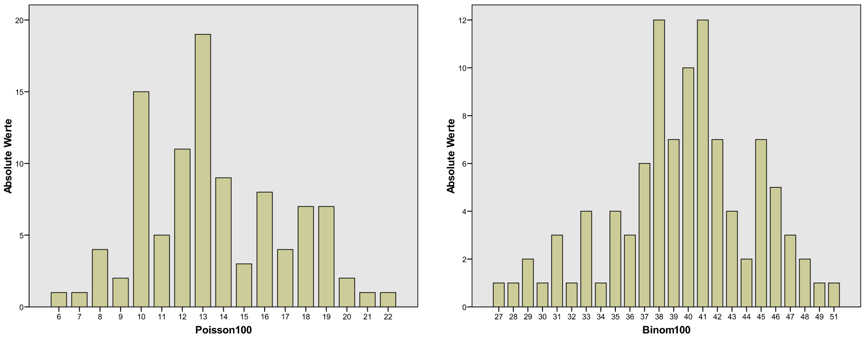

Abb. 14.6. Veranschaulichung von Zufallszahlen einer Poissonverteilung ($\lambda = 14$) und einer Binomialverteilung ($n = 100, p = 0.4$) bei einer $N = 100$-fachen Realisierung

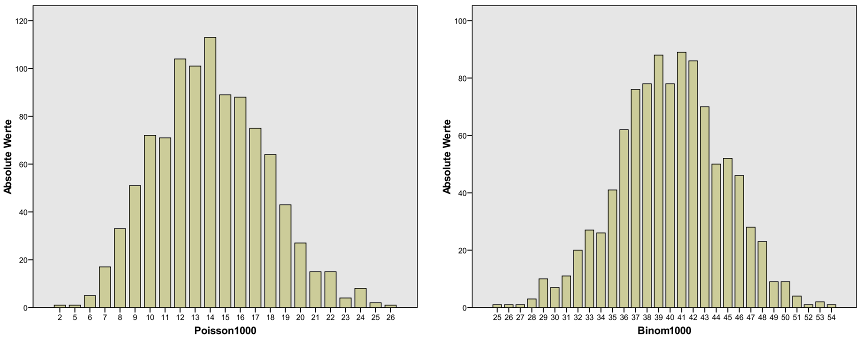

Abb. 14.7. Veranschaulichung von Zufallszahlen einer Poissonverteilung ($\lambda = 14$) und einer Binomialverteilung ($n = 100, p = 0.4$) bei einer $N = 1000$-fachen Realisierung

Es ist leicht zu sehen, dass sowohl bei der Binomial- als auch bei der Poissonverteilung für $N = 1000$ die empirische Verteilung einer Normalverteilung ähnelt - im Gegensatz zu den geringeren Stichprobengrößen von $N = 10$ bzw. $N = 100$. Man kann erahnen, dass sich für noch größere N die Verteilungen immer stärker einer Normalverteilung mit $\mu = 14$, $\sigma^2 = 14$ bzw. $\mu = 40$, $\sigma^2 = 24$ annähern.

In Kapitel 15.2.1 wollen wir den zentralen Grenzwertsatz mit Hilfe des statistischen Programmpakets *R* noch einmal direkt veranschaulichen. Das dortige Beispiel betrachtet die standardisierten Summen am Beispiel einer Poissonverteilung.

14.2.2 Induktive Statistik - Verwendung von statistischen Tests

In diesem Abschnitt wollen wir die Anwendung statistischer Tests mit SPSS demonstrieren. Dazu verwenden wir den von SPSS mitgelieferten Datensatz *'Employee.sav'* (ehemals *'bank.sav'*), der die Einkommensstruktur der Mitarbeiter einer Bank beschreibt. Tabelle 14.2 zeigt einen Auszug aus dem Datensatz. Interessierende Variablen sind v.a. 'gehalt' und 'agehalt', die das

Tabelle 14.2. Auszug des SPSS-Datensatzes 'Employee.sav'

id	geschl	ausbild	tätig	gehalt	agehalt	dauer	erfahr	mind
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	1	8	1	21900	13200	98	190	0
5	0	15	1	45000	21000	98	138	0
6	0	15	1	32100	13500	98	67	0
7	0	15	1	36000	18750	98	114	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

aktuelle Gehalt bzw. das Einstiegsgehalt (in US$/Jahr) der Bankangestellten beschreiben. Potentielle Einflussfaktoren sind das Geschlecht ('geschl'), die Ausbildungslänge in Jahren ('ausbild'), die Art der Tätigkeit ('tätig'), das erste Arbeitsjahr ('dauer'), die Berufserfahrung in Monaten ('erfahr'), sowie die Zugehörigkeit zu einer Minderheit ('mind').

Tests auf Mittelwerte; 1-Stichproben t-Test. Zur Durchführung des 1-Stichproben t-Tests folgt man dem Pfad

Analysieren → Mittelwerte vergleichen → T-Test bei einer Stichprobe.

Im Menüpunkt 'Testvariable' gibt man durch einfachen Mausklick die interessierende Variable ein, unter 'Testwert' kann μ_0 angegeben werden. Wir wollen nun testen, ob man von einem mittleren Gehalt von 35000 US$/Jahr unter den Mitarbeitern ausgehen kann. Dazu verwenden wir den 1-Stichproben t-Test und testen die Hypothese $\mu = 35000$. Abbildung 14.8 zeigt die Ergebnisse des Tests.

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Gehalt	474	34.419,57	17.075,661	784,311

	Testwert = 35000					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
Gehalt	-.740	473	.460	-580,432	-2.121,60	960,73

Abb. 14.8. SPSS-Ausgabe zum 1-Stichproben t-Test

Wir sehen, dass das mittlere Gehalt in der Bank bei 34419.57 US$/Jahr liegt. Der t-Wert von -0.740 bzw. der p-Wert von 0.460 zeigen, dass die Nullhypothese zum Niveau $\alpha = 0.05$ nicht verworfen werden kann. Wir können also die Nullhypothese von $\mu = 35000$ nicht verwerfen.

Tests auf Mittelwerte; 2-Stichproben t-Test. Zur Durchführung des 2-Stichproben t-Tests folgt man dem Pfad

Analysieren → Mittelwerte vergleichen → T-Test bei unabhängigen Stichproben.

Wir interessieren uns bei unserem Datensatz *'Employee.sav'* dafür ob sich das mittlere Gehalt bei Männern und Frauen unterscheidet. Abbildung 14.9 zeigt die Ergebnisse von SPSS. Die deskriptiven Statistiken sprechen ein ein-

	Geschlecht	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Gehalt	Männlich	258	41.441,78	19.499,214	1.213,968
	Weiblich	216	26.031,92	7.558,021	514,258

		Levene-Test der Varianzgleichheit	
		F	Signifikanz
Gehalt	Varianzen sind gleich	119.669	.000
	Varianzen sind nicht gleich		

		T-Test für die Mittelwertgleichheit		
		T	df	Sig. (2-seitig)
Gehalt	Varianzen sind gleich	10.945	472	.000
	Varianzen sind nicht gleich	11.688	344.262	.000

Abb. 14.9. SPSS-Ausgabe zum 2 Stichproben t-Test

deutiges Bild: Die männlichen Angestellten haben ein mittleres Gehalt von über 41000 US$/Jahr, während dies bei den weiblichen Angestellten gerade mal 26331 US$/Jahr beträgt. Die Varianzen sind nicht bekannt, deshalb

führt SPSS den 'Levene-Test auf Varianzgleichheit' durch, um festzustellen, ob von einer Gleichheit der Varianzen ausgegangen werden kann oder nicht. Entsprechend der Entscheidung wird dann ein 'doppelter t-Test' bzw. der 'Welch-Test' durchgeführt. In unserem Beispiel lehnt der Levene Test die Nullhypothese gleicher Varianzen ab (p-Wert = 0.000), dementsprechend ist die untere Zeile in der Ausgabe zum t-Test relevant. Der dortige Wert der Teststatistik (T=11.688) bzw. des p-Wertes (p=0.000) lehnen die Nullhypothese eines gleichen mittleren Gehalts ab, Männer scheinen in dieser Bank also signifikant mehr zu verdienen.
Um eine endgültige Aussage bezüglich dieser Thematik zu treffen, sollten jedoch weitere Analysen zu Rate gezogen werden; möglicherweise liegt ein Schichtungseffekt vor, und Frauen sind vor allem in Segmenten tätig, die schlechter bezahlt werden, während Männer Führungspositionen besetzen.

Tests auf Mittelwerte; paired t-Test. Zur Durchführung des paired t-Testes folgt man dem Pfad

Analysieren → Mittelwerte vergleichen → T-Test bei gepaarten Stichproben.

Die einzigen abhängigen Variablen in unserem Datensatz sind die beiden Variablen 'Gehalt' und 'Anfangsgehalt'. Ein gepaarter t-Test soll nun die (möglicherweise triviale) Fragestellung klären, ob sich mittleres Gehalt und mittleres Anfangsgehalt unterscheiden, ob es also zu Gehaltserhöhungen gekommen ist oder nicht. Abbildung 14.10 zeigt die Ausgabe von SPSS.

		Mittelwert	N	Standardabweichung	Standardfehler des Mittelwertes
Paaren 1	Gehalt	34.419,57	474	17.075,661	784,311
	Anfangsgehalt	17.016,09	474	7.870,638	361,510

		Gepaarte Differenzen				
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	95% Konfidenzintervall der Differenz	
					Untere	Obere
Paaren 1	Gehalt - Anfangsgehalt	17.403,481	10.814,620	496,732	16.427,407	18.379,555

		T	df	Sig. (2-seitig)
Paaren 1	Gehalt - Anfangsgehalt	35.036	473	.000

Abb. 14.10. SPSS-Ausgabe zum gepaarten t-Test

Die deskriptiven Statistiken zeigen ein mittleres Anfangsgehalt von etwas über 17000 US$/Jahr, sowie ein aktuelles Gehalt von über 34000 US$/Jahr. Es ist nicht verwunderlich, dass der gepaarte t-Test die Nullhypothese gleicher mittlerer Gehälter verwirft (p=0.000).

Tests auf Varianz. Der von uns vorgestellte χ^2-Test auf Varianz sowie der F-Test zum Vergleich zweier Varianzen sind in SPSS leider nicht enthalten. Zur Durchführung dieser Tests können andere Programmpakete wie z.B. R (siehe auch Kapitel 15.2.2) oder MINITAB verwendet werden. Der Test von Levene, der die Varianz von zwei Stichproben vergleicht, wird dagegen in entsprechendem Zusammenhang ausgegeben: so z.B. beim 2-Stichproben t-Test bzw. bei der ANOVA.

Tests auf Korrelation. Zur Durchführung des Tests auf Korrelation folgt man dem Pfad

$$\textit{Analysieren} \rightarrow \textit{Korrelation} \rightarrow \textit{Bivariat}.$$

Uns interessiert nun der Zusammenhang zwischen der 'Berufserfahrung in Monaten' und dem 'Gehalt'. Abbildung 14.11 enthält die Ergebnisse von SPSS. Der Korrelationskoeffizient von Bravais-Pearson beträgt $r_{BP} = -0.097$

		Berufserfahrung in Monaten	Gehalt
Berufserfahrung in Monaten	Korrelation nach Pearson	1	-.097*
	Signifikanz (2-seitig)		.034
	N	474	474
Gehalt	Korrelation nach Pearson	-.097*	1
	Signifikanz (2-seitig)	.034	
	N	474	474

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Abb. 14.11. SPSS-Ausgabe zum Test auf Korrelation

und deutet auf einen leicht negativen Zusammenhang hin. Der p-Wert von 0.034 zeigt, dass die Korrelation zum Niveau $\alpha = 0.05$ signifikant von Null verschieden ist.
Dass die Berufserfahrung einen negativen Effekt auf das Gehalt haben soll, ist auf den ersten Blick verwunderlich. Ein Blick auf Abbildung 14.12 liefert hierfür jedoch eine Erklärung. Eine lange Berufserfahrung haben im Wesentlichen Mitarbeiter des Sicherheitsbereiches, die jedoch im unteren Gehaltsgefüge der Bank anzutreffen sind.

Tests auf Binomialwahrscheinlichkeiten. Zur Durchführung des Binomialtests für p folgt man dem Pfad

$$\textit{Analysieren} \rightarrow \textit{Nichtparametrische Tests} \rightarrow \textit{Binomial}.$$

Uns interessiert, ob der Anteil der Männer und der Frauen in der Bank gleich groß ist, oder anders formuliert, ob der Anteil der Männer bei $p = 0.5$ liegt oder nicht. SPSS liefert uns folgende Ausgabe (Abbildung 14.13):

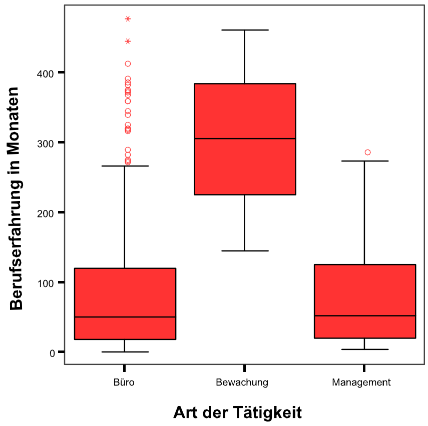

Abb. 14.12. Boxplot der Berufserfahrung (in Monaten) in Abhängigkeit der 'Art der Tätigkeit'

		Kategorie	N	Beobachteter Anteil	Testanteil	Asymptotische Signifikanz (2-seitig)
Geschlecht	Gruppe 1	männlich	258	.54	.50	.060[a]
	Gruppe 2	weiblich	216	.46		
	Gesamt		474	1.00		

a. Basiert auf der Z-Approximation.

Abb. 14.13. SPSS-Ausgabe zum Test auf Binomialwahrscheinlichkeiten

Der Männeranteil in der Stichprobe beträgt 54%. Da der p-Wert 0.06 beträgt, können wir die Nullhypothese von einem Anteil von 50% zum Niveau $\alpha = 0.05$ jedoch nicht verwerfen.

Zum Vergleich zweier Binomialwahrscheinlichkeiten haben wir den (approximativen) Binomialtest, den Test von Fisher, sowie den Test von McNemar kennengelernt. Die Durchführung des Binomialtests ist in SPSS nicht möglich, wohl aber die der Tests von Fisher und McNemar. Zur Durchführung des Tests von Fisher folgt man dem Pfad

$$\textit{Analysieren} \rightarrow \textit{Deskriptive Statistiken} \rightarrow \textit{Kreuztabellen}$$

und wählt im Menüpunkt 'Statistik' den Button 'Chi-Quadrat' aus. Der Test von Fisher wird dann automatisch mitausgegeben.
In unserem Datensatz *'Employee.sav'* definieren wir uns zuerst eine neue Variable 'Gehaltkat', die angibt, ob ein Gehalt niedrig ($<$ 28880 US\$/Jahr) oder hoch ($\geq$ 28880 US\$/Jahr) ist. Wir wollen nun die beiden Binomialwahrscheinlichkeiten für ein niedriges Gehalt in den beiden Gruppen 'männlich' und 'weiblich' gegeneinander testen. Wir haben schon bei der Durchführung

des 2-Stichproben t-Tests bereits gesehen, dass sich das Gehalt geschlechtsspezifisch unterscheidet und erwarten bei dieser Analyse ein ähnliches Ergebnis. Abbildung 14.14 zeigt die Ergebnisse von SPSS.
In der Kreuztabelle erkennen wir bereits, dass der Anteil weiblicher Beschäftig-

		Geschlecht		
		männlich	weiblich	Gesamt
Gehalt_kat	niedrig	73	164	237
	hoch	185	52	237
Gesamt		258	216	474

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	106.694[b]	1	.000		
Kontinuitätskorrektur[a]	104.797	1	.000		
Likelihood-Quotient	111.280	1	.000		
Exakter Test nach Fisher				.000	.000
Zusammenhang linear-mit-linear	106.469	1	.000		
Anzahl der gültigen Fälle	474				

a. Wird nur für eine 2x2-Tabelle berechnet

b. 0 Zellen (.0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 108.00.

Abb. 14.14. SPSS-Ausgabe zum Test von Fisher

ter mit niedrigem Gehalt deutlich höher ist als der der männlichen Angestellten. Der p-Wert von 0.000 in der unteren Ausgabe bestätigt dies noch einmal und zeigt, dass Frauen in dieser Bank weitaus weniger im Mittel verdienen als Männer.

Zur Durchführung des Tests von McNemar folgt man dem Pfad

Analysieren → Nichtparametrische Tests → Tests bei zwei verbundenen Stichproben

und wählt im unteren Bereich des sich öffnenden Fensters den Test von McNemar aus. Analog zur Vorgehensweise beim Test von Fisher definieren wir eine neue Variable 'Agehaltkat', die angibt, ob das Anfangsgehalt niedrig ($<$ 15000 US$/Jahr) oder hoch ($\geq$ 15000 US$/Jahr) war. Die beiden Variablen 'Gehaltkat' und 'Agehaltkat' bilden nun eine verbundene Stichprobe. Wir wollen nun den Anteil eines hohen Gehaltes in den beiden Gruppen 'niedriges Anfangsgehalt' und 'hohes Anfangsgehalt' miteinander vergleichen. Es stellt sich also die Frage, ob ein hohes Anfangsgehalt auch ein hohes Gehalt zum aktuellen Zeitpunkt garantiert und analog, ob ein niedriges Anfangsgehalt auch ein niedriges aktuelles Gehalt bedeutet. Die SPSS-Ausgabe zum Test von McNemar ist in Abbildung 14.15 dargestellt.