Prof. H.R. Künsch   Wahrscheinlichkeitsrechnung und Statistik   Abt. IX, SS99

Musterlösung zur Übung 12

1. **a)** Die Stichproben sind ungepaart. (Es gibt keine natürliche Zuordnung zweier Messungen bei verschiedenen Typen zu einer gemeinsamen Beobachtungseinheit. Die Anzahlen geprüfter Kugellager jedes Typs könnten verschieden sein.)

b)
1. Modell: $(X_i)_{1\le i\le n=10}$ i.i.d. $\sim \mathcal{N}(\mu_1,\sigma^2)$, $(Y_j)_{1\le j\le m=10}$ i.i.d. $\sim \mathcal{N}(\mu_2,\sigma^2)$, alle unabhängig, $\sigma^2 = \mathrm{Var}(X_i) = \mathrm{Var}(Y_j)$ unbekannt
2. H_0: $\mu_1 = \mu_2$
3. H_A: $\mu_1 \ne \mu_2$ (zweiseitige Fragestellung)
4. Teststatistik:
$$T = \frac{(\overline{X}-\overline{Y})\sqrt{n+m-2}}{\sqrt{(\frac{1}{n}+\frac{1}{m})(\sum_{i=1}^n (X_i-\overline{X})^2 + \sum_{j=1}^m (Y_j-\overline{Y})^2)}}$$
5. Unter H_0 gilt: $T \sim t_{n+m-2}$. Grosse Beträge $|T|$ sprechen gegen H_0.
6. H_0 wird auf dem 5%-Niveau verworfen, falls $|T| > t_{n+m-2,0.975}$. Die kritischen Werte $t_{n+m-2,0.975}$ sind tabelliert. Für $n = m = 10$ haben wir $t_{18,0.975} = 2.10$.
7. Mit unseren Daten beträgt die Teststatistik $T = |T| = 2.07 < t_{18,0.975} = 2.10$. Der P-Wert liegt damit minim über 5% (exakt 0.053). Die Nullhypothese wird (gerade noch) beibehalten. Die beiden Kugellagertypen können als gleich zuverlässig betrachtet werden.

c)
1. Modell: $(X_i)_{1\le i\le n=10}$ i.i.d. $\sim F$, $(Y_j)_{1\le j\le m=10}$ i.i.d. $\sim G$, alle unabhängig
2. H_0: $F = G$
3. H_A: $F \ne G$ (zweiseitige Fragestellung)
4. Teststatistik:
$$W = \text{Anzahl Paare } (i,j) \text{ mit } X_i < Y_j$$
5. H_0 wird auf dem 5%-Niveau verworfen, falls $W \le 23$ oder $W \ge 100-23 = 77$. (Kritische Werte aus der Tabelle!)
6. Mit unseren Daten beträgt die Teststatistik $W = 25$, also $24 < W < 76$. Die Nullhypothese wird beibehalten. Die beiden Kugellagertypen können als gleich zuverlässig betrachtet werden.

d) Da die Messungen beim Typ 1 einen grösseren Mittelwert haben als bei Typ 2 und da der Entscheid mit dem t-Test sehr knapp ausgefallen ist, könnte man vermuten, dass die Vergrösserung eines X_i's nun zur Verwerfung führt. Erstaunlicherweise ist das nicht der Fall. Durch die Modifikation von X_{10} wird die Streuungsschätzung von X_i und damit auch von $\overline{X}-\overline{Y}$ grösser. Der Teststatistik T bewegt sich gegen die Null hin und beträgt neu 1.85. Der P-Wert beträgt 0.081. H_0 wird deutlicher akzeptiert!
Beim Wilcoxontest hat die Änderung keine Auswirkung, da der alte Wert 16.84 schon grösser als alle Y_j war.

2. Die interessierende Frage ist natürlich, ob sich die Verteilung der Anzahl Unfälle auf die verschiedenen Sternzeichen von der Verteilung der Sternzeichen in der Bevölkerung unterscheidet.

Für eine etwas mathematischere Formulierung indizieren wir die Sternzeichen mit i, bezeichnen den Anteil des Sternzeichens i in der Bevölkerung mit p_i (also zum Beispiel $p_3 = 0.0904$) und die beobachtete Anzahl Unfälle von Fahrern mit Sternzeichen i mit n_i. Die Anzahl untersuchter Personen in der Studie war $n = 25003$. Die Nullhypothese H_0 lautet dann: Die Anzahl Unfälle in den zwölf Klassen besitzt eine Multinomialverteilung mit den Parametern $(p_1,\dots,p_{12})$.

Unter H_0 erwartet man np_i Unfälle in der Klasse i. Es ergibt sich folgende Tabelle:

Sternzeichen	in Bevölkerung	beobachtete Anz.	erwartete Anz.	$\frac{(n_i-np_i)^2}{np_i}$
Widder	8.90%	2135	2225.27	3.66
Stier	8.58%	2440	2145.26	40.50
Zwillinge	9.04%	2258	2260.27	0.00
Krebs	8.56%	2120	2140.26	0.19
Löwe	8.60%	1963	2150.26	16.31
Jungfrau	8.45%	2240	2112.75	7.67
Waage	7.94%	2043	1985.24	1.68
Skorpion	7.70%	1778	1925.23	11.26
Schütze	7.48%	1835	1870.22	0.66
Steinbock	8.01%	2053	2002.74	1.26
Wassermann	8.32%	2030	2080.25	1.21
Fische	8.42%	2108	2105.25	0.00
Total	100.00%	25003		84.40

Die Teststatistik beim Chiquadrattest beträgt $\chi^2 = 84.40$ bei 11 Freiheitsgraden. Das bedeutet einen P-Wert von weniger als 0.0005. (Das grösste tabellierte Quantil ist das 99.95%-Quantil bei 33.14.) Die Unterschiede in den Unfallhäufigkeiten bei den verschiedenen Sternzeichen lassen sich also nicht durch den Zufall erklären.

Damit ist zwar ein systematischer Effekt statistisch nachgewiesen. Dass dieser aber mit dem Einfluss der Sterne zu tun haben muss, akzeptieren wir wohl nicht ohne weiteres. Hat die im Sternbild Stier gegründete Versicherungsgesellschaft beispielsweise vor zwei Jahren allen Stieren ein Jubiläumsangebot gemacht, zählt sie heute etwas mehr Stiere zu ihren Kunden gemessen am Anteil der Stiere in der Bevölkerung. Auch wenn die Verteilung der Unfälle auf die verschiedenen Sternzeichen mit den Anteilen der Sternzeichen unter den Kunden der Versicherung übereinstimmt, fällt in diesem Fall der obige Test signifikant aus. Es ist eben so, dass bei derart grossen Stichprobeumfängen schon ein sehr kleiner Anteil von Beobachtungen die Nullhypothese kippen kann! — Welche Erklärungen hast Du?

3. **a)** Unter der Nullhypothese der Unabhängigkeit gilt (beispielsweise)
$$P[\text{Mann leptsom und Frau athletisch}] = P[\text{Mann leptsom}] \cdot P[\text{Frau athletisch}]\,.$$
Wir erwarten also
$$\begin{aligned}\widehat{n}_{12} &= n\,\widehat{P}[\text{Mann leptsom}]\,\widehat{P}[\text{Frau athletisch}]\\ &= \;= n\frac{n_{1\cdot}}{n}\frac{n_{\cdot 2}}{n} = \frac{n_{1\cdot}n_{\cdot 2}}{n} = \frac{53\cdot 116}{224} = 27.45\end{aligned}$$
Paare in Zelle $(1,2)$. Mit den gleichen Überlegungen findet man die übrigen Tabelleneinträge:

Körperbautyp des Ehemannes	Körperbautyp der Ehefrau leptosom	athletisch	pyknisch	insges.
leptosom	12.30	27.45	13.25	53
athletisch	31.56	70.43	34.00	136
pyknisch	8.13	18.13	8.75	35
insgesamt	52	116	56	224

b) Zuerst berechnen wir die Grössen $(n_{ij} - \widehat{n}_{ij})^2/\widehat{n}_{ij}$.

Körperbautyp des Ehemannes	Körperbautyp der Ehefrau leptosom	athletisch	pyknisch
leptosom	15.26	11.09	1.06
athletisch	7.67	15.06	8.50
pyknisch	0.43	12.63	20.06

Seien k die Anzahl der Zeilen und l die Anzahl der Spalten der Kontingenztafel:

$$\text{Teststatistik: } \chi^2 = \sum_{i=1}^{k=3} \sum_{j=1}^{l=3} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}} = 91.76$$

Freiheitsgrade: $\nu = (k-1)\cdot(l-1) = 2\cdot 2 = 4$. (Wir haben gerade 4 Zahlen benutzt, um das Innere der Kontigenztafel zu ergänzen.) Wegen $\chi^2 = 91.76 > \chi^2_{4,0.95} = 9.488$ wird die Annahme der Unabhängigkeit auf dem 5%-Signifikanzniveau verworfen. Zwischen den Körperbautypen der Ehepartner in der hessischen Landbevölkerung besteht ein statistisch gesicherter Zusammenhang.

c) Unter der Nullhypothese ist $n_{ij} \sim \mathrm{Bin}(n, \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n})$ mit $\mathbf{E}[n_{ij}] = \frac{n_{i\cdot}n_{\cdot j}}{n} = \widehat{n}_{ij}$ und $\mathrm{Var}(n_{ij}) = \widehat{n}_{ij}(1 - \frac{n_{i\cdot}n_{\cdot j}}{n^2}) \approx \widehat{n}_{ij}$. Grob kann man sagen, dass die Abweichung zwischen dem beobachteten Wert n_{ij} und dem erwarteten Wert $\widehat{n}_{ij}$ in der Zelle (i,j) signifikant ist, wenn der beobachtete Wert um mehr als zwei Standardabweichungen $(\sqrt{\widehat{n}_{ij}})$ vom erwarteten Wert abweicht. Das liefert die Faustregel

$$|n_{ij} - \widehat{n}_{ij}| \geq 2\sqrt{\widehat{n}_{ij}} \quad \text{bzw.} \quad \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}} \geq 4\,.$$

Die stärksten signifikanten Abweichungen liegen auf der Diagonalen. Es ist damit für die vorliegenden Daten erwiesen, dass Heiraten unter Personen mit gleichem Körperbautyp häufiger vorkommen. In zweiter Linie kommen Ehen zwischen Partnern mit „naheliegenden“ Körperbautypen vor.

4. a) Wir gehen von folgendem Modell aus. X_i i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. μ ist also der wahre aber unbekannte Erwartungswert von X_i. Dann gilt

$$\frac{\overline{X} - \mu}{s_x/\sqrt{n}} \sim t_{n-1} \quad \forall \mu, \sigma\,.$$

Daraus folgen

$$P\Big[\,\Big|\frac{\overline{X} - \mu}{s_x/\sqrt{n}}\Big| < t_{n-1,1-\frac{\alpha}{2}}\Big] = 1 - \alpha \quad \forall \mu, \sigma$$

und

$$P\Big[\mu \in \underbrace{\Big]\overline{X} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s_x}{\sqrt{n}}\Big[}_{\text{Konfidenzintervall}}\,\Big] = 1 - \alpha \quad \forall \mu, \sigma\,.$$

Hier ist α jeweils 5%. Aus $\overline{X} = 20.44$, $n = 10$, $t_{9,0.975} = 2.262$ und $s_x/\sqrt{10} = 0.257$ erhält man das Konfidenzintervall

$$]20.44 \pm 2.262 \cdot 0.257[\, = \,]19.86, 21.02[\,.$$

b) Diesmal betrachten wir das Modell X_i i.i.d. $\sim F$ mit wahrem aber unbekanntem Median m. Es gilt

$$\sum_{i=1}^{n} \mathbf{I}_{\{X_i > m\}} \sim \mathrm{Bin}\big(n, \frac{1}{2}\big) \quad \forall m, F\,.$$

Daraus folgen

$$P\Big[b_{n,\frac{1}{2},\frac{\alpha}{2}} \leq \sum_{i=1}^{n} \mathbf{I}_{\{X_i > m\}} \leq b_{n,\frac{1}{2},1-\frac{\alpha}{2}}\Big] \geq 1 - \alpha \quad \forall m, F$$

und

$$P\Big[m \in \underbrace{\Big\{m' \Big| b_{n,\frac{1}{2},\frac{\alpha}{2}} \leq \sum_{i=1}^{n} \mathbf{I}_{\{X_i > m'\}} \leq b_{n,\frac{1}{2},1-\frac{\alpha}{2}}\Big\}}_{\text{Konfidenzintervall}}\Big] \geq 1 - \alpha \quad \forall m, F\,,$$

wobei $b_{n,\frac{1}{2},\frac{\alpha}{2}}$ das $\frac{\alpha}{2}$-Quantil der Binomialverteilung mit Parametern n und $\frac{1}{2}$ bedeutet.

Um zu entscheiden, ob ein bestimmtes m' zum Konfidenzintervall gehört, muss man also nachzählen, ob weder zu viele noch zu wenige Beobachtungen X_i grösser als m' sind. Diese implizite Definition des Konfidenzintervalles ist darum etwas mühsam in der Handhabung. Es gibt aber eine einfach anzuwendende Faustregel: Für grosse n $(n > 6)$ kann man die $\mathrm{Bin}(n, \frac{1}{2})$-Verteilung durch die Normalverteilung mit gleichen Momenten, also durch $\mathcal{N}(\frac{n}{2}, n\frac{1}{2}\frac{1}{2} = \frac{n}{4})$ approximieren. Die Quantile der Binomialverteilung sind dann ungefähr gleich den Quantilen der Normalverteilung, also $b_{n,\frac{1}{2},0.975} \approx \frac{n}{2} + 2\sqrt{\frac{n}{4}} = \frac{n}{2} + \sqrt{n}$. Es dürfen folglich höchstens $\frac{n}{2} + \sqrt{n}$ Beobachtungen grösser als m' sein. In anderen Worten: m' muss mindestens so gross sein wie die $n - (\frac{n}{2} + \sqrt{n}) \approx (\frac{n+1}{2} - \sqrt{n})$-t-kleinste Beobachtung: $m' \geq X_{(\frac{n+1}{2} - \sqrt{n})}$. Analog findet man die obere Schranke. Man erhält als Faustregel für das 95%-Konfidenzintervall für den Median

$$[X_{(\frac{n+1}{2} - \sqrt{n})}, X_{(\frac{n+1}{2} + \sqrt{n})}] \quad (n > 6)\,.$$

Hier gelten $n = 10$, $\frac{n+1}{2} - \sqrt{n} \approx 2$ und $\frac{n+1}{2} + \sqrt{n} \approx 9$. Das Konfidenzintervall ist damit

$$[X_{(2)}, X_{(9)}] = [20.1, 21.4]\,.$$

Dieses Intervall ist gegenüber dem Konfidenzintervall aus Aufgabe 4.a) etwas nach rechts verschoben. Das liegt daran, dass der Ausreisser 19.2 den Mittelwert verkleinert und damit das ganze Vertrauensintervall nach links schiebt. Der Median hingegen bleibt von einem einzelnen Ausreisser unbeeinflusst.