Lösungen zur Übung (5)

(1) (a) Man hätte die Population zu präzisieren, etwa: Alle Studienabsolventen von Universitäten und Technischen Hochschulen in der BRD der Abgangsjahre ..., weiter wäre festzuhalten, ob das Einkommen zum Berufseinstieg oder erst eine Weile später genommen werden soll oder ob diesbezüglich keine genaueren Bestimmungen getroffen werden. Ebenso wäre selbstverständlich festzulegen, ob man von Jahres- oder Monatseinkommen, brutto oder netto spricht. Natürlich bereitet auch 'Studienfach' als eindeutige Zuordnung Probleme: Wenn es mehr als eines gibt, das absolviert wurde, so wäre naheliegend das für den Beruf wichtigere zu nehmen. Natürlich sind die Variablen abhängig, da die Einkommen bekanntlich verschieden sind. Es genügt, etwa ein Beispiel der folgenden Art herauszugreifen: Die Absolventen der Germanistik verdienen sicher mit kleinerer relativer Häufigkeit über 40000 Euro jährlich als die Absolventen der Medizin.

(b) Auch hier wäre die Population genauer zu fixieren: Alle Studenten in Europa? Welcher Jahrgänge? Es sollte hier jedem Teilnehmer klar sein, dass sicherlich etwa der Frauenanteil bem Studienfach Psychologie (Romanistik und Kunstgeschichte wären da noch radikaler) ein anderer ist als sagen wir beim Studienfach Elektrotechnik. (Studienfach: Etwa als 'Erstes Fach' zu präzisieren.)

(c) Tageszeit: Feiner oder gröber zu messen. Demgemäß könnte man einem beliebigen Augenblick die genaue Tageszeit und den momentanen Stromverbrauch zuordnen, oder aber einer Zeitspanne (gemäß Einteilung der Tageszeiten) eben den Namen der Tageszeit und den mittleren Stromverbrauch (etwa einer Stadt oder auch der ganzen BRD usw.) während dieser Spanne zuordnen. Es sollte klar sein, dass der Stromverbrauch 'Spitzenzeiten' am Morgen und am Abend kennt und daher Abhängigkeit vorliegt.

(d) (Natürlich wären auch bei der Population der Bücher weitere Präzisierungen anzubringen wie 'zu einer bestimmten Zeit auf dem Markt' oder 'in einem gewissen Zeitraum erschienen' usw.) Wir können annehmen (ohne es genau zu wissen!), dass der Anteil von 'e' an allen Vokalen völlig unabhängig vom Inhaltstyp ist (den man wieder geeignet zu präzisieren hätte, etwa 'Literatur - Trivialliteratur - Trivialsachbuch - Wissenschaft', oder auch ganz anders!) Denn 'e' ist nun einmal bei weitem der häufigste Vokal in deutschen Wortklassen jeglicher Art. Anders steht es mit 'y', das kommt mit Sicherheit in wissenschaftlicher Literatur wesentlich häufiger vor als in sonstiger - man denke an Fremdwörter griechischen Ursprungs wie 'Psychologie', 'Physiologie', 'Ethymologie', 'Hypothese' etc. Anteil des Buchstabens 'y' und Art des Buches werden also sicher abhängig sein, wenn man die Unterscheidung von 'wissenschaftlich' - 'nicht wissenschaftlich' bei der Typologie der Bücher eingebaut hat.

(e) Hier scheiden sich die Geister, natürlich sind die Variablen unabhängig, wenn es auch eine Vielzahl von Spökenkiekern gibt, die bei all solchen Anlässen das Gegenteil vermuten. Allenfalls könnte man diskutieren darüber, ob es vielleicht wegen tendenziell schlechterer ärztlicher Versorgung der Neugeborenen am Wochenende eine messbare Benachteiligung der am Wochenende Geborenen gibt. (Die traditionelle Spökenkiekermeinung würde natürlich umgekehrt die 'Sonntagskinder' bevorzugt sehen.)

Es ist sehr wichtig festzuhalten, dass statistische (bzw. wahrscheinlichkeitstheoretische) Abhängigkeit nichts bedeutet wie 'Verursachen', 'Beeinflussen' usw. Allenfalls können statistische Abhängigkeiten auf Wirkzusammenhängen (teilweise) beruhen, aber umgekehrt ist nichts Derartiges zu schließen: Ist etwa 'weiblich sein' Ursache dafür, dass Romanistik studiert wird? Lächerlich!

(2) Zunächst 'mit Zurücklegen' - Wir nennen die Variable 'Zahl der roten Kugeln' X:

$$P(X=2) = \binom{6}{2}\left(\frac{4}{11}\right)^2\left(\frac{7}{11}\right)^4 \approx 0.32527. \text{ (Binomialverteilung ist zuständig.)}$$

Das Resultat ändert sich überhaupt nicht bei der erhöhten Kugelzahl, da $p = 4/11$ dasselbe bleibt. Nunmehr 'ohne Zurücklegen', nennen wir hier die (andere!) Variable 'Zahl der roten Kugeln' Y:

$$P(Y=2) = \frac{\binom{4}{2}\binom{7}{4}}{\binom{11}{6}} \approx 0.455. \text{ (Hypergeometrische Verteilung!)}$$

Für die größere Zahl der Kugeln in der Urne kann man eine Annäherung an das Resultat bei der Binomialverteilung erwarten, und so ist es auch:

$$P(Y_1=2) = \frac{\binom{12}{2}\binom{21}{4}}{\binom{33}{6}} \approx 0.357.$$

Bei denselben Verhältissen in einer Urne mit 11000 Kugeln würde man schon nicht mehr bemerken, ob man mit Zurücklegen oder ohne zieht.

(3) Zunächst ein paar nützliche Bezeichnungen: A für 'Bewerbung bei Universität A', analog B. E für 'erfolgreiche Bewerbung', W und M für die Geschlechter. Man hat folgende Resultate:

$$\begin{aligned} P(E|W) &= \frac{190}{1100} = \frac{19}{110}, \\ P(E|M) &= \frac{820}{1300} = \frac{82}{130}. \end{aligned}$$

Das scheint für eine klare Bevorzugung männlicher Bewerber zu sprechen. Aber das beinahe verrückt Anmutende:

$$\begin{aligned} P(E|W\cap A) &= \frac{1}{10} > \frac{1}{15} = P(E|M\cap A) \\ P(E|W\cap B) &= \frac{9}{10} > \frac{4}{5} = P(E|M\cap B), \end{aligned}$$

also waren an beiden Universitäten die Bewerbungen der Frauen erfolgreicher. Da $A\cup B = \Omega$ in unserem Falle, wirkt das wie ein Widerspruch. Aber eine vielleicht intuitiv naheliegende Formel der Art: '$P(E|M\cap A) + P(E|M\cap \overline{A}) = P(E|M)$' ist nämlich falsch. Das Phänomen erklärt sich leicht, wenn man bedenkt, dass Universität A einen viel kleineren Anteil aufnahm als B, während sich der Großteil der Frauen bei A, der Männer bei B bewarb. Natürlich wäre es wiederum eine klare Benachteiligung, wenn die Eigenschaften dieser Hochschulen sich ändern würden, je nach dem, ob mehr Frauen oder Männer sich jeweils bewerben.

(4) Die Anzahl X der Sechsen bei 1000 Würfen ist binomialverteilt mit $n = 1000$, $p = \frac{1}{6}$, also $\mu(X) = 1000/6$, $\sigma(X) = \sqrt{\frac{5}{36}\cdot 1000}$, also

$$P(X \le 190) \approx \Phi_{0,1}\left(\frac{190.5 - 1000/6}{\sqrt{\frac{5}{36}\cdot 1000}}\right) \approx \Phi_{0,1}(2.02) \approx 0.9783.$$

Das ist sehr genau, man erhält bei exakter Rechnung 0.9770 (gerundete Angabe). Natürlich macht die Stetigkeitskorrektur hier nicht mehr viel Unterschied, das ist für kleinere n wichtiger.

(5) Wir haben für die Augensumme X bei 1000 Würfen: $\mu(X) = 3500$, $\sigma(X) = \sigma_1\cdot\sqrt{1000}$, mit der Streuung σ_1 für die Augenzahl bei einmaligem Würfeln, die wir schnell so ausrechnen:

$$\sigma_1^2 = \frac{1}{3}\left(\left(1-\frac{7}{2}\right)^2 + \left(2-\frac{7}{2}\right)^2 + \left(3-\frac{7}{2}\right)^2\right) = \frac{35}{12}. \text{ Also } \sigma_1 = \sqrt{\frac{35}{12}}.$$

Also gewinnt man das zugehörige Vertrauensintervall näherungsweise so:

$$\left[3500 - 1.96\sqrt{\frac{35000}{12}}; 3500 + 1.96\sqrt{\frac{35000}{12}}\right] \approx [3394.15; 3605.85]$$

Man würde hier $[3394; 3606]$ angeben. (Eine Stetigkeitskorrektur, an die man hier denken könnte, würde nichts ändern an diesen ganzzahligen Angaben. Sie sind übrigens aberwitzig genau - man beachte, dass man mit einer diskreten Verteilung ein Vertrauensintervall im Allgemeinen nicht ganz genau treffen kann, sondern immer nur das kleinste Intervall (betreffender Art, hier zweiseitig) angeben kann, dass mindestens die vorgeschriebene Wahrscheinlichkeit enthält.)

(6) Es sei hier nur noch einmal für c,d erinnert: Zwei Ereignisse A, B im Rahmen eines Experiments (sonst ist das Unfug!) sind genau dann unabhängig - im statistischen bzw. wahrscheinlichkeitstheoretischen Sinne (!), wenn $P(A \cap B) = P(A) P(B)$, wobei es inhaltlich verständlicher ist, für $P(B) \neq 0$ zu formulieren: A und B sind genau dann unabhängig, wenn $P(A|B) = P(A)$. Das ist die Grundlage für die Formulierung der Unabhängigkeit für ein Paar von Zusallsvariablen (im Rahmen derselben Population mit derselben Wahrscheinlichkeitsfunktion !): X, Y sind unabhängig genau dann, wenn alle Ereignispaare $(X \leq \alpha)$, $(Y \leq \beta)$ es sind. Man kann es auch in Worten so sagen: X, Y sind genau dann unabhängig, wenn jedes über X allein formulierbare Ereignis unabhängig von jedem über Y allein formulierbaren Ereignis ist.