Einführung in die Statistik
Aufgabenserie 4

Andreas Kuhn
Soziologisches Institut der Universität Zürich
kuhn@soziologie.uzh.ch

Sommersemester 2007

4 Wahrscheinlichkeitstheorie

1. **Bestimmung von Wahrscheinlichkeiten: Geburten**
   Die empirische Wahrscheinlichkeit für die Geburt eines Jungen beträgt circa 51,5% (Siehe VL–Unterlagen).

   (a) Bestimmen Sie die Wahrscheinlichkeit für vier Jungen bei vier Geburten.

   (b) Bestimmen Sie die Wahrscheinlichkeit für zwei Mädchen bei vier Geburten.
   Hinweis: Beachten Sie, dass unterschiedliche Anordnungen von Mädchen und Jungen zum gesuchten Ereignis führen können.

   (c) Welche beiden Annahmen sind zentral für die Bestimmung dieser Wahrscheinlichkeiten?

2. **Panini–Bilder**
   Wie in der VL erwähnt gibt es 596 unterschiedliche Panini–Bilder. Treffen Sie dieselben Annahmen wie in den VL–Unterlagen, d.h. gehen Sie davon aus, dass man einzelne Bilder kaufen kann und nehmen Sie an, dass alle Bilder gleich häufig vorkommen.

   (a) Reproduzieren Sie Abbildung 1 aus den VL–Unterlagen. D.h. bestimmen Sie die Anzahl Bilder, die Sie kaufen müssen um das n–te Bild zu erhalten, wenn Sie bereits $n - 1$ unterschiedliche Bilder besitzen.
   Hinweis: Bestimmen Sie dazu zunächst die jeweilige Wahrscheinlichkeit beim n–ten Kauf ein "neues" Bild zu kaufen, wenn Sie bereits $n - 1$ Bilder bestitzen.

   (b) Reproduzieren Sie Abbildung 2 aus den VL–Unterlagen. D.h. bestimmen Sie die Anzahl Bilder, die Sie insgesamt kaufen müssen, um alle 596 Bilder zu erstehen.

3. **Das Paradox des Chevalier de Méré**
   Im 17. Jahrhundert wurden in Frankreich verschiedene Würfelspiele gespielt – schauen wir uns zwei Würfelspiele an.
   In einem ersten Spiel wurde vier Mal ein einzelner Würfel geworfen und man gewann das Spiel, wenn mindestens eine "1" in den vier Würfen geworfen wurde. In einem zweiten Spiel wurden 24 Mal jeweils zwei Würfel geworfen und man gewann, wenn mindestens einmal eine doppelte "1" geworfen wurde.
   Ein Adliger aus dieser Zeit, der Chevalier de Méré, überlegte sich dazu Folgendes:

   - In einem einzelnen Wurf habe ich eine Chance von $1/6$, eine "1" zu würfeln. Demzufolge habe ich bei vier Würfen eine Chance von $4 \times 1/6 = 2/3$, mindestens eine "1" zu würfeln.

- Im zweiten Spiel beträgt die Chance 1/36, bei einem Wurf mit zwei Würfeln eine doppelte "1" zu werfen. Also habe ich bei 24 Würfen mit zwei Würfeln eine Chance von $24 \times 1/36 = 2/3$, mindestens eine doppelte "1" zu werfen.

Da der Chevalier (sehr) viel spielte vermutete er allerdings, dass die Gewinnchance beim ersten Spiel grösser ist als beim zweiten Spiel – obwohl gemäss seinen Überlegungen beide Spiele dieselbe Gewinnchance aufweisen. Des Rätsels Lösung gaben dann Blaise Pascal und Pierre de Fermat und zeigten, dass die Überlegungen des Chevalier nicht richtig sind.

(a) Bestimmen Sie die tatsächliche Wahrscheinlichkeit, beim viermaligen Werfen eines einzelnen Würfels mindestens einmal eine "1" zu werfen.

(b) Bestimmen Sie die tatsächliche Wahrscheinlichkeit, bei 24 Würfen von jeweils zwei Würfeln mindestens eine doppelte "1" zu werfen.

Hinweis zu den beiden Teilaufgaben: Bestimmen Sie zunächst die jeweilige Gegenwahrscheinlichkeit, ansonsten sind die gesuchten Wahrscheinlichkeiten sehr schwierig zu berechnen.

4. **Randomized Response**

Mit den Hilfsmitteln der Wahrscheinlichkeitstheorie können wir nun auch das "Rätsel" um die Frage nach dem (illegalen) Drogenkonsum aus dem Fragebogen lösen. Die Frage ist ein Beispiel für eine Fragetechnik, welche man als "randomized response" bezeichnet[1]. Man setzt diese Technik bei Fragestellungen ein, bei denen ansonsten kaum mit ehrlichen Antworten gerechnet werden kann, wie eben z.B. Fragen nach dem Drogenkonsum.

Wir bezeichnen im folgenden den ersten Münzwurf als Ereignis W_1 und den zweiten als W_2[2]. Es ist somit $\Omega_{1,2} = \{\text{Kopf}, \text{Zahl}\}$. π sei die Wahrscheinlichkeit, schon einmal illegale Drogen konsumiert zu haben.

(a) Bestimmen Sie die auf W_1 bedingten Wahrscheinlichkeiten von W_2, d.h. bestimmen Sie $Pr(W_2|W_1 = \text{Kopf})$ sowie $Pr(W_2|W_1 = \text{Zahl})$.

(b) Erstellen Sie aus den Angaben aus Teilaufgabe (a) eine Kreuztabelle, in welcher W_1 und W_2 abgetragen sind. Bestimmen Sie die gemeinsame Verteilung von (W_1, W_2) als auch die beiden Randverteilungen von W_1 bzw. W_2.

(c) Zeigen Sie, dass man aus der Randverteilung von W_2 den Parameter π schätzen kann. Schätzen Sie den Parameter entsprechend aus den Daten unter `rand_response`. Was stellen Sie fest?

(d) Welches (praktische) Problem taucht hier auf?

(e) Wie können Sie vorgehen, wenn mehr als 50% der Beobachtungen die Frage beantworten sollen (beispielsweise 70%)?

5. **Randomisierung**

Verwenden Sie für diese Aufgabe die Daten, welche unter `randomisierung` abgelegt sind (Es handelt sich wiederum um einen Teil der Daten, welche in der Veranstaltung erhoben wurden). Wir werden diese Daten wiederholt verwenden, um Eigenschaften von Randomisierung und das Schätzen von kausalen Effekten zu illustrieren (siehe Aufgabenserie 5).

[1] Siehe dazu die Artikel von Goodstadt und Gruson (1975), sowie von Tracy und Fox (1981).

[2] Zur Erinnerung: Zeigt der erste Wurf "Kopf", sollte man die Frage mit "Zahl" (= "nein") bzw. "Kopf" (= "ja") beantworten; bei "Zahl" beim ersten Wurf sollte man die Münze ein zweites Mal werfen.

(a) Zeigen Sie zunächst, dass unter Randomisierung des treatments D folgende Beziehungen gelten:

$$Pr(D=1|X=x) = Pr(D=1)$$
$$Pr(X=x|D=1) = Pr(X=x)$$

(b) Als Illustration wollen wir ein binäres treatment $D = 0, 1$ randomisieren. Verwenden Sie zur Randomisierung den Geburtstag der Personen: Weisen Sie denjenigen Personen das treatment zu, welche an einem geraden Tag Geburtstag haben. Erläutern Sie, ob und weshalb diese Art der Randomisierung funktioniert.

(c) Welche anderen (praktischen) Möglichkeiten sind denkbar, um D zu randomisieren?

(d) Teilaufgabe (a) impliziert verschiedene Möglichkeiten zu überprüfen, ob das treatment tatsächlich randomisiert wurde. Erläutern Sie, wie man dies tun kann (in den Beispieldaten).

6. **Das "Geburtstagsproblem"**
Nehmen Sie an, dass alle Geburtstage des Jahres sind gleich wahrscheinlich und vernachlässigen Sie den 29. Februar (siehe dazu auch Diaconis und Mosteller, 1989).

(a) Wieviele Personen müssen (zufällig) zusammen kommen damit die Wahrscheinlichkeit, dass mindestens zwei der anwesenden Personen am gleichen Tag Geburtstag haben, grösser oder gleich 50% ist?
Hinweis: Es gibt keine explizite Lösung (= Formel) für diese Aufgabe, d.h. Sie müssen das Problem numerisch lösen, indem Sie die Wahrscheinlichkeit für verschiedene Gruppengrössen berechnen, bis die gesuchte Wahrscheinlichkeit erreicht ist.

(b) Wieviele Personen müssen zusammen kommen damit die Wahrscheinlichkeit, dass mindestens eine zweite Person am selben Tag Geburtstag hat wie Sie, grösser oder gleich 50% ist?

(c) Weshalb unterscheiden sich Ihre Antworten aus (a) und (b)?

(d) Wie gross sind die Wahrscheinlichkeiten für die Ereignisse aus (a) und (b), wenn 50 (100) Personen zusammen kommen? Stellen Sie die beiden Wahrscheinlichkeiten aus (a) und (b) über die Anzahl Personen $n = 1, \ldots 365$ grafisch dar. Was stellen Sie fest?

(e) Gibt es unter den Teilnehmern Personen mit demselben Geburtstag? Schauen Sie sich dazu die Daten unter `geburtstag` an.

7. **Schweizer Zahlenlotto**
Die Wahrscheinlichkeit, im Schweizer Zahlenlotto (6 aus 45) eine "Sechs" zu tippen, beträgt nach dem klassischen Wahrscheinlichkeitsbegriff:

$$Pr(\text{Sechser}) = \frac{1}{\binom{45}{6}} = \frac{1}{8'145'060}$$

D.h. theoretisch kommt ein richtiger Tipp auf 8'145'060 ausgefüllte Lottozettel.

(a) Heisst das nun, dass Sie mit 100% Wahrscheinlichkeit einen Sechser tippen, wenn Sie 8'145'060 mal Lotto spielen? Wie gross ist die Wahrscheinlichkeit, dass Sie auch nach dieser Anzahl von Spielen keinen Treffer landen?

(b) Dennoch gibt es gewissermassen eine "goldene" Regel für Lottospieler. Welche ist das? Hinweis: Wovon hängt der allfällige Lottogewinn ab?

(c) Ergibt es (statistisch) gesehen Sinn, sich die relativen Häufigkeiten der Zahlen aus den bisherigen Ziehungen anzuschauen (und entsprechend Zahlen anzukreuen, welche bisher weniger häufig gezogen worden sind)? Weshalb (nicht)?

8. **Reliabilität von Testverfahren**
Ein Aids–Test weise folgende Reliabilität auf:

$$\begin{aligned} Pr(+|\text{HIV}) &= 1,00 \\ Pr(-|\text{kein HIV}) &= 0,99 \end{aligned}$$

Ein + bezeichnet ein positives Testergebnis an (d.h. der Test deutet auf Ansteckung mit dem Virus hin), − dementsprechend ein negatives Testergebnis.

(a) Nach einem Test in einer grossen Bevölkerungsgruppe, in der jeder Tausendste tatsächlich mit dem Virus infiziert ist, erhält Herr X Post vom Gesundheitsamt mit der Nachricht, das Testergebnis sei "positiv". Herr X klammert sich an die Hoffnung, das Testergebnis könnte eventuell falsch–positiv sein (d.h. positives Ergebnis ohne Infektion). Frage: Wie hoch ist die Wahrscheinlichkeit dafür?

(b) Berechnen Sie die Chancen (odds) einer Infektion, gegeben das Testergebnis fällt positiv aus, um das Ergebnis aus Teilaufgabe (a) zu verstehen und interpretieren Sie die resultierenden beiden Terme. Wie erklärt sich somit das Resultat aus (a)?
Hinweis: Die Chancen eines dichotomen Ereignisses Y sind definiert als Verhältnis aus $Pr(Y = 1)$ und $Pr(Y = 0)$. Dasselbe gilt, wenn wir auf ein zweites Ereignis $X = x$ bedingen. D.h. bestimmen Sie das Verhältnis von $Pr(\text{HIV}|+)$ und $Pr(\text{kein HIV}|+)$.

9. **Einfache Zufallsvariablen**[3]
Sechs Personen nehmen an einem Kurs teil. Im Verlauf des Kurses werden fünf Aufgaben besprochen, wobei jede Aufgabe jeweils zufällig durch einen der Teilnehmer gelöst werden muss. Die Auslosung ist unabhängig davon, wieviele Aufgaben ein Teilnehmer bereits gelöst hat.

(a) Teilnehmer A hat bereits die ersten vier Aufgaben vorlösen müssen. Wie gross ist die Wahrscheinlichkeit, dass A auch noch die letzte Aufgabe lösen muss? Begründen Sie das Ergebnis formal und argumentativ.

(b) Teilnehmer B hat noch keine Aufgabe lösen müssen. Wie gross ist die Wahrscheinlichkeit, dass B die fünfte Aufgabe lösen muss? Begründen Sie das Ergebnis formal und argumentativ.

(c) Y sei die Anzahl der Aufgaben, welche ein Teilnehmer im Verlauf des gesamten Kurses lösen muss. Bestimmen Sie $Pr(Y = y)$ für $y = 0, 1, \ldots, 5$.
Hinweis: Sie benötigen zur Lösung den sog. Binomialkoeffizienten:

$$\binom{n}{y} = \frac{n!}{(n-y)!y!}$$

Der Binomialkoeffizient gibt dabei die Anzahl Möglichkeiten an, aus n Objekten y auszuwählen (wenn die Anordnung der Objekte keine Rolle spielt).

(d) Stellen Sie die Häufigkeitsverteilung und die Verteilungsfunktion von Y grafisch dar.

[3] Eine Zufallsvariable ist eine Variable mit numerischen Ausprägungen, welche unterschiedliche Ergebnisse eines zufälligen Vorganges beschreiben. Mehr dazu in Teil 5 der Vorlesung.

(e) Sie möchten wissen, wie lange ein Teilnehmer durchschnittlich warten muss, bis er das erste Mal eine Aufgabe vorlösen muss. X sei die Aufgabe, bei der zum ersten Mal vorgelöst werden muss, d.h. $X = 1, 2, \ldots, 5$. Bestimmen Sie $Pr(X = x)$.

(f) Stellen Sie die Häufigkeitsverteilung und die Verteilungsfunktion von X grafisch dar.

10. **Das "Monty–Hall" Paradox**[4]

Stellen Sie sich das folgende Fernseh–Quiz vor: Es gibt drei Türen, hinter denen zufällig ein "Treffer" und zwei "Nieten" versteckt sind. Nachdem Sie als Kandidat eine der drei Türen ausgewählt haben, öffnet der Moderator eine der beiden verbleibenden Türen (eröffnet allerdings nur eine Tür, wenn sich dahinter eine "Niete" versteckt). Sie haben nun – nachdem eine der drei Türen offen ist – die Möglichkeit, bei Ihrer ersten Wahl zu bleiben oder aber die Türe zu wechseln. Es stellt sich damit die Frage, was Sie tun sollten: Wechseln oder bleiben? Und weshalb?

(a) Zeigen Sie dass es sich unter den obigen Annahmen lohnt, die Türe immer zu wechseln. Hinweis: Verwenden Sie Bayes' Theorem zur Lösung, und führen Sie die Berechnung unter der Bedingung durch, dass Sie zunächst z.B. die erste Türe gewählt haben.

(b) Dieses Beispiel eignet sich gut, um die Situation am Computer zu simulieren (dies bezeichnet man als Monte–Carlo–Simulation). In diesem Beispiel ist die Simulation einfacher als die formale Lösung aus (a).
Erläutern Sie, wie man die Situation am Computer (z.B. mit Excel) simulieren kann und führen Sie diese dann 1000–mal durch. Vergleichen Sie das Ergebnis mit Teilaufgabe (a).

Hinweis zum Vorgehen in Excel: Verwenden Sie die Funktion `RAND`, mit welcher Sie Zufallszahlen im Intervall $[0, 1]$ generieren können, um die Zuteilung des Preises auf die drei Türen zu simulieren. Dann simulieren Sie die folgenden zwei Situationen: (i) Jemand wählt immer z.B. die erste Tür und wechselt nie. (ii) Jemand wählt immer die erste Tür und wechselt aber danach immer (auf die zweite oder die dritte Tür, je nachdem welche Türe der Moderator geöffnet hat).

Literatur

[1] Diaconis, P. and F. Mosteller (1989). Methods for Studying Coincidences. *Journal of the American Statistical Association*, 408, 853–861.

[2] Friedman, Daniel (1998). Monty Hall's Three Doors: Construction and Deconstruction of a Choise Anomaly. *The American Economic Review*, 88(4), 933–946.

[3] Goodstadt, Michael S. and Valerie Gruson (1975). The Randomized Response Technique: A Test on Drug Use. *Journal of the American Statistical Association*, 70(352), 814–818.

[4] Morgan, J. P., N. R. Chaganty, R.C. Dahiya and M.J. Doviak (1991). Let's Make a Deal: The Player's Dilemma. *The American Statistician*, 45(4), 284–287.

[5] Tracy, Paul E. and James Alan Fox (1981). The Validity of Randomized Response for Sensitive Measurements. *American Sociological Review*, 46(2), 187–200.

[4] Dies ist wahrscheinlich eines der bekanntesten Beispiele aus der Wahrscheinlichkeitstheorie, das im übrigen auch gestandene Mathematiker ins Schwitzen bringt. Der Name des Beispiels stammt vom damaligen Moderator Monty Hall der Fernsehsendung "Let's make a deal". Siehe dazu auch die Artikel von Friedman (1998) und Morgan et al. (1991).