Technische Universität München
Fakultät für Informatik
Lehrstuhl für Effiziente Algorithmen
Prof. Dr. Ernst W. Mayr
Dr. Werner Meixner

Sommersemester 2011
Lösungsblatt 9
15. Juli 2011

# Diskrete Wahrscheinlichkeitstheorie

## Hausaufgabe 1 (5 Punkte)

Sei $X$ eine kontinuierliche Zufallsvariable.

1. Zeigen Sie: Falls $X \sim \mathcal{N}(2, \frac{1}{2})$, dann gilt $2X + 1 \sim \mathcal{N}(5, 2)$.

2. Seien $d_1, d_2, c \in \mathbb{R}$ mit $d_1 < d_2$ und $c > 0$. Berechnen Sie $a, b \in \mathbb{R}$, so dass für $Y = aX + b$ gilt
$$\Pr[d_1 \leq X \leq d_2] = \Pr[-c \leq Y \leq c]\,.$$

**Lösung**

1. Seinen $\mu$ und $\sigma$ der Erwartungswert bzw. die Varianz von $X$, d. h. $\mu = 2$ bzw. $\sigma^2 = \frac{1}{2}$. Nach Satz der Vorlesung ist $2X + 1$ normalverteilt mit Erwartungswert $2\mu + 1 = 5$ bzw. Varianz $2^2\sigma^2 = 2$. W. z. b. w.

2. Sei $a > 0$. Dann gilt
$$d_1 \leq X \leq d_2 \quad \Longleftrightarrow \quad ad_1 + b \leq Y \leq ad_2 + b\,.$$
Wir lösen für $a, b$ die Gleichungen
$$ad_1 + b = -c \quad \text{und} \quad ad_2 + b = c\,.$$
Lösung:
$$a = \frac{2c}{d_2 - d_1}\,, \qquad b = \frac{d_1 + d_2}{d_1 - d_2} \cdot c\,.$$

## Hausaufgabe 2 (5 Punkte)

Seien $X_1, X_2, \ldots$ kontinuierliche Zufallsvariable, die identisch verteilt und unabhängig sind mit $E[X_i] = 2$ und $\text{Var}[X_i] = 4$. Wir betrachten $Y_n = \sum_{i=1}^{n} X_i$. Berechnen Sie die folgenden Grenzwerte und begründen Sie die Korrektheit Ihrer Herleitung.

1. $\lim_{n\to\infty} \Pr[\frac{Y_n}{n} = 2]$.
   Benutzen Sie für Ihre Herleitung <u>nicht</u> den Zentralen Grenzwertsatz.

2. $\lim_{n\to\infty} \Pr[1{,}9 < Y_n < 2{,}1]$.

3. $\lim_{n\to\infty} \Pr[1{,}99 < \frac{Y_n}{n} < 2{,}01]$.

**Lösung**

1. Für stetige Zufallsvariable $X$ gilt $\Pr[X = r] = 0$ für jeden Wert $r \in \mathbb{R}$. Da $Y_n/n$ stetig ist, erhält man $\Pr[Y_n/n = 2] = 0$ für alle $n$. Trivialerweise ist der Grenzwert für $n \to \infty$ gleich 0 und wir brauchen nichts zu berechnen.

2. Auch hier ist der Grenzwert gleich Null, allerdings muss er nun wie folgt bestimmt werden.

   Es gilt $\mathbb{E}[Y_n] = 2n$ und $\mathrm{Var}[Y_n] = 4n$.

   Falls $1{,}9 < Y_n < 2{,}1$ gilt, dann folgt für alle $n \geq 3$ die Ungleichung

   $$|Y_n - 2n| \geq n, \text{ d. h. } |Y_n - \mathbb{E}[Y_n]| \geq n\,.$$

   Es folgt

   $$\Pr[1{,}9 < Y_n < 2{,}1] \;\leq\; \Pr[|Y_n - \mathbb{E}[Y_n]| \geq n]\,.$$

   Nun wenden wir die Ungleichung von Chebyshev an und erhalten

   $$\Pr[1{,}9 < Y_n < 2{,}1] \;\leq\; \frac{\mathrm{Var}[Y_n]}{n^2} = \frac{4}{n}\,.$$

   Es folgt

   $$\lim_{n\to\infty} \Pr[1{,}9 < Y_n < 2{,}1] = 0\,.$$

3. Rechnung nach Transformation $Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}$:

   $$\begin{aligned}
   \Pr[1{,}99 < \tfrac{Y_n}{n} < 2{,}01] &= \Pr[1{,}99n < Y_n < 2{,}01n] \\
   &= \Pr\left[\frac{1{,}99n - 2n}{2\sqrt{n}} < Z_n < \frac{2{,}01n - 2n}{2\sqrt{n}}\right] \\
   &= F_{Z_n}\left(\frac{2{,}01n - 2n}{2\sqrt{n}}\right) - F_{Z_n}\left(\frac{1{,}99n - 2n}{2\sqrt{n}}\right) \\
   &= F_{Z_n}\left(0{,}005\sqrt{n}\right) - F_{Z_n}\left(-0{,}005\sqrt{n}\right)\,.
   \end{aligned}$$

   Nun folgt

   $$\begin{aligned}
   \lim_{n\to\infty} \Pr[1{,}99 < \tfrac{Y_n}{n} < 2{,}01] &= \lim_{n\to\infty} \left(F_{Z_n}\left(0{,}005\sqrt{n}\right) - F_{Z_n}\left(-0{,}005\sqrt{n}\right)\right) \\
   &= 1 - 0 = 1\,.
   \end{aligned}$$

## Hausaufgabe 3 (5 Punkte)

Sei $X \sim \mathrm{Bin}(10^6, \frac{1}{10})$ eine binomialverteilte Zufallsvariable. Sei $k = 500$.

1. Zeigen Sie $\Pr[10^5 - k \leq X \leq 10^5 + k] \geq 0{,}64$.

2. Approximieren Sie $\Pr[10^5 - k \leq X \leq 10^5 + k]$ unter der Annahme, dass der Satz von de Moivre anwendbar ist.

**Lösung**

1. *Bemerkung*: Leider ist unbemerkt geblieben, dass die Richtung der Ungleichung falsch gegeben war. Die ursprüngliche Ungleichung ist unerfüllbar. Die Teilaufgabe entfällt für die Bewertung.

   Es gilt $\mathbb{E}[X] = 10^5$. Deshalb gilt $10^5 - k \leq X \leq 10^5 + k$ genau dann, wenn $|X - \mathbb{E}[X]| \leq k$ gilt. Mit $\mathrm{Var}[X] = 10^6 \cdot \frac{1}{10} \cdot \frac{9}{10}$ erhalten wir

$$\begin{aligned} \Pr[10^5 - k \leq X \leq 10^5 + k] &= \Pr[|X - \mathbb{E}[X]| \leq k] \\ &= 1 - \Pr[|X - \mathbb{E}[X]| > k] \\ &\geq 1 - \frac{\mathrm{Var}[X]}{(k+1)^2} \\ &= 1 - \frac{9 \cdot 10^4}{(501)^2} \\ &\geq 1 - 0{,}36 = 0{,}64\,. \end{aligned}$$

2. Sei

$$X^* = \frac{X - np}{\sqrt{np(1-p)}} = \frac{X - 10^5}{300}\,.$$

   Wir approximieren $X^* \sim N(0,1)$ und erhalten

$$\begin{aligned} \Pr[10^5 - k \leq X \leq 10^5 + k] &= \Pr\left[-\frac{k}{300} \leq X^* \leq \frac{k}{300}\right] \\ &= \Pr\left[-\frac{5}{3} \leq X^* \leq \frac{5}{3}\right] \\ &= \Phi\left(\frac{5}{3}\right) - \Phi\left(-\frac{5}{3}\right) \\ &= 2 \cdot \Phi(1.6666\ldots) - 1 \approx 2 \cdot 0{,}9522 - 1 \\ &\approx 0{,}9044\,. \end{aligned}$$

## Hausaufgabe 4 (5 Punkte)

An der dänischen Grenze werden Grenzkontrollen durchgeführt. Im Schnitt treffen alle 30 Sekunden an der Grenzstation Personen ein, die zu kontrollieren sind. Die Zeit zwischen zwei Kontrollen sei exponentialverteilt mit Parameter $\frac{1}{30}$. Wenn 2 Minuten lang kein Kontrollfall eingetroffen ist, dann machen die Grenzbeamten Ruhepause.

Seien $T_1, T_2, \ldots$ die Zeitspannen zwischen dem Eintreffen von zu kontrollierenden Personen und $W$ die Wartezeit bis zur nächsten Ruhepause.

1. Berechnen Sie $\mathbb{E}[T_1 \mid T_1 \geq 120]$.

2. Berechnen Sie $\mathbb{E}[W]$.

**Lösung**

1. Sei $\lambda = \frac{1}{30}$. Dann gilt $\mathbb{E}[T_1] = 30$.
   Es gilt $\mathbb{E}[T_1 \mid T_1 \geq 120] = 120 + \mathbb{E}[T_1] = 150\,.$

2. Wir wählen die Bezeichnung $W'$ wie in Vorbereitungsaufgabe 2 von Blatt 8.
   Es gilt $W = 120 + \sum_{j=1}^{N-1} T_j$ oder $W = 120 + W'$ mit $W' = \sum_{j=1}^{N-1} T_j\,.$
   $N$ ist eine geometrisch verteilte Zufallsvariable mit Erfolgswahrscheinlichkeit $p = \Pr[T \geq 120]$. Es gilt
   $$p = \Pr[T \geq 120] = e^{-4}\,.$$
   Berechnung von $\mathbb{E}[W']$:
   Wir setzen $T = T_1$.
   $$\begin{aligned}
   \mathbb{E}[W'] &= \sum_{n=1}^{\infty} \mathbb{E}[W'|N=n] \cdot p(1-p)^{n-1} \\
   &= \sum_{n=1}^{\infty} \mathbb{E}[T|T \leq 120] \cdot (n-1) \cdot p(1-p)^{n-1} \\
   &= \mathbb{E}[T|T \leq 120] \cdot \sum_{n=1}^{\infty} (n-1) \cdot p(1-p)^{n-1} \\
   &= \mathbb{E}[T|T \leq 120] \cdot \mathbb{E}[N-1].
   \end{aligned}$$
   Es gilt
   $$\mathbb{E}[N-1] = e^4 - 1\,.$$
   $\mathbb{E}[T|T \leq 120]$ erhalten wir aus den Gleichungen
   $$\begin{aligned}
   \mathbb{E}[T] &= \mathbb{E}[T|T \leq 120] \cdot \Pr[T \leq 120] \;+\; \mathbb{E}[T|T \geq 120] \cdot \Pr[T \geq 120] \\
   &= \mathbb{E}[T|T \leq 120] \cdot (1 - e^{-4}) + 150 \cdot e^{-4} \\
   &= 30\,.
   \end{aligned}$$
   Es folgt
   $$\mathbb{E}[T|T \leq 120] \;=\; \frac{30 - 150 \cdot e^{-4}}{1 - e^{-4}} \;=\; \frac{30 \cdot e^4 - 150}{e^4 - 1}\,.$$
   Ergebnis:
   $$\begin{aligned}
   \mathbb{E}[W'] &= 30 \cdot e^4 - 150\,, \\
   \mathbb{E}[W] &= 30 \cdot e^4 - 150 + 120 \\
   &\approx 26,8 \text{ (Minuten)}\,.
   \end{aligned}$$

**Hinweis:** *Auf den Übungsblättern in diesem Semester wird es grundsätzlich die drei Aufgabentypen Vorbereitungsaufgabe, Tutoraufgabe und Hausaufgabe geben. Die als Vorbereitung bezeichneten Aufgaben dienen der häuslichen Vorbereitung der Tutoraufgaben. Tutoraufgaben werden in den Übungsgruppen bearbeitet. Dabei wird die Lösung der Vorbereitungsaufgaben vorausgesetzt. Die Vorbereitungsaufgaben werden in der Zentralübung unterstützt.*

## Vorbereitung 1

Seien $X_1, \ldots, X_n$ unabhängige Stichproben einer Zufallsvariablen $X$ und sei $\bar{X}$ das arithmetische Mittel der $X_i$. Wir verwenden die Zufallsvariable

$$V = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

als Schätzer für die Varianz von $X$.

Berechnen Sie den Bias von $V$! Welche Aussage gilt für $n \to \infty$?

**Lösung**

Sei $S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Dann folgt

$$V = \frac{n-1}{n} \cdot S\,.$$

Berechnung des Bias von $V$:

$$\begin{aligned}
\mathbb{E}[V - \mathrm{Var}[X]] &= \mathbb{E}[V] - \mathrm{Var}[X] \\
&= \frac{n-1}{n} \cdot \mathbb{E}[S] - \mathrm{Var}[X] \\
&= \frac{n-1}{n} \cdot \mathrm{Var}[X] - \mathrm{Var}[X] \\
&= -\frac{\mathrm{Var}[X]}{n}\,.
\end{aligned}$$

Der Bias strebt gegen 0 für $n \to \infty$.

## Vorbereitung 2

Eine Werbeagentur möchte am letzten Tag der Fußballweltmeisterschaft mit einer Blitzumfrage schätzen, welcher Anteil $\vartheta$ der per Bahn anreisenden Fußballfans einen Platz im Stadion hat. Jeder der 12 Mitarbeiter befragt so lange zufällig ausgewählte Fans, bis er einen Fan gefunden hat, der eine Karte für das Stadion besitzt. Die Anzahl der vom Mitarbeiter $i$ befragten Fans sei $X_i$.

Wir nehmen an, dass alle $X_i$ die gleiche geometrische Verteilung besitzen mit

$$\Pr_\vartheta[X_i = k] \;=\; (1-\vartheta)^{k-1} \cdot \vartheta\,,\;\; \forall k \in \mathbb{N}\,.$$

1. Man bestimme auf der Basis der ermittelten Stichprobenwerte

   $$3,\ 1,\ 2,\ 3,\ 2,\ 1,\ 2,\ 1,\ 5,\ 3,\ 2,\ 3$$

   einen Maximum-Likelihood-Schätzwert für $\vartheta$.

2. Man gebe mit Hilfe des Zentralen Grenzwertsatzes ein konkretes 95%-Konfidenzintervall für $\vartheta$ an.

**Lösung**

1. Für die Likelihood-Funktion $L(\vec{x};\vartheta)$ der Stichprobe $\vec{x} = (x_1, \ldots, x_{12})$ gilt mit $n = 12$

$$\begin{aligned} L(\vec{x};\vartheta) &= \prod_{i=1}^{n} \Pr_\vartheta[X_i = x_i] \\ &= \prod_{i=1}^{n} \left[(1-\vartheta)^{x_i-1} \cdot \vartheta\right] . \end{aligned}$$

Wir schreiben $\sum_{i=1}^n x_i - n = n\bar{x} - n$ und erhalten

$$L(\vec{x};\vartheta) = (1-\vartheta)^{n\bar{x}-n} \cdot \vartheta^n .$$

Gesucht ist der Wert $\vartheta = \bar{\vartheta}$ mit $0 \le \vartheta \le 1$, für den $L(\vec{x};\vartheta)$ das Maximum in dem $[0,1]$-Intervall annimmt.

Falls $\bar{x} = 1$, dann folgt sofort $\vartheta = 1$. Dies bedeutet, dass mit maximaler Wahrscheinlichkeit (die der Test erlaubt) jeder Fan eine Stadionkarte besitzt.

Falls $\bar{x} \ne 1$, dann bestimmen wir das Maximum von $\vartheta$ zwischen 0 und 1 im ersten Schritt durch Bestimmung einer Nullstelle der Ableitung von $L$ nach $\vartheta$ wie folgt.

$$\begin{aligned} \frac{\mathrm{d}\, L(\vec{x};\vartheta)}{\mathrm{d}\,\vartheta} &= (n\bar{x}-n)(1-\vartheta)^{n\bar{x}-n-1}(-1)\vartheta^n + (1-\vartheta)^{n\bar{x}-n} n\vartheta^{n-1} \\ &= (1-\vartheta)^{n\bar{x}-n-1}\vartheta^{n-1}[(n\bar{x}-n)(-\vartheta) + n(1-\vartheta)] \\ &= (1-\vartheta)^{n\bar{x}-n-1}\vartheta^{n-1}[n - n\bar{x}\vartheta] \end{aligned}$$

Als Nullstellen der Ableitung der Likelihood-Funktion erhalten wir $\vartheta = 0$, $\vartheta = 1$ und $\vartheta = \frac{1}{\bar{x}}$. Man beachte, dass $0 < \frac{1}{\bar{x}} < 1$ gilt. Da

$$L(\vec{x};0) = L(\vec{x};1) = 0 \quad \text{und} \quad L\left(\vec{x};\frac{1}{\bar{x}}\right) > 0$$

gilt, scheiden $\vartheta = 0$ und $\vartheta = 1$ als Maximumstellen aus. Es folgt

$$\bar{\vartheta} = \frac{1}{\bar{x}} .$$

Nun berechnen wir $\frac{1}{\bar{x}}$ aus der Stichprobe mit $\bar{x} = \frac{7}{3}$ und erhalten

$$\bar{\vartheta} = \frac{3}{7} .$$

Es haben also bei maximaler Testwahrscheinlichkeit $\frac{3}{7}$ der Fans Stadionkarten.

2. Die Stichprobe liefert den Wert $\bar{x} = \frac{7}{3}$ für die Zufallsvariable $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ mit $n = 12$. Für das Konfidenzintervall für $\vartheta$ zum Konfidenzniveau $1 - \alpha = 0.95$ setzen wir an

$$\Pr\left[-c \le \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \le c\right] \approx 0.95$$

mit dem Quantil $c = x_{1-\frac{\alpha}{2}} = x_{0.975} \approx 1.96$ und den Gleichungen $\mu_{\bar{X}} = \frac{1}{\vartheta}$ und $\sigma^2_{\bar{X}} = \frac{1-\vartheta}{n\vartheta^2}$ für $n = 12$. Wir bestimmen nun die Menge aller $\vartheta$, für die die Ungleichung $-c \leq \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq c$ gilt.

$$\begin{aligned}
-c \leq \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq c \quad &\Longleftrightarrow \quad (\bar{X}-\mu_{\bar{X}})^2 \leq c^2\sigma^2_{\bar{X}} \\
&\Longleftrightarrow \quad (\bar{X}-\frac{1}{\vartheta})^2 \leq c^2 \cdot \frac{1}{n} \cdot \frac{1-\vartheta}{\vartheta^2} \\
&\Longleftrightarrow \quad (\vartheta\bar{X}-1)^2 \leq \frac{c^2}{n} \cdot (1-\vartheta) \\
&\Longleftrightarrow \quad \vartheta^2\bar{X}^2 + (\frac{c^2}{n} - 2\bar{X})\vartheta + (1-\frac{c^2}{n}) \leq 0
\end{aligned}$$

Nun setzen wir sämtliche Zahlenwerte ein und erhalten (wegen $c \approx 1.96$ näherungsweise)

$$-c \leq \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq c \quad \Longleftrightarrow \quad 49\vartheta^2 - 39.1188\vartheta + 6.1188 \leq 0$$

Wenn wir nach bekannter Formel die Nullstellen der quadratischen Gleichung

$$49\vartheta^2 - 39.1188\vartheta + 6.1188 = 0$$

mit $\alpha_1$, $\alpha_2$ bezeichnen, wobei wir $\alpha_1 < \alpha_2$ annehmen können, dann ergibt sich als Lösung

$$\alpha_1 \leq \vartheta \leq \alpha_2$$

mit $\alpha_1 \approx 0.2135$ und $\alpha_2 \approx 0.5848$.

## Tutoraufgabe 1

Die tatsächlich benötigte CPU-Zeit einer Benutzersitzung an einer Workstation werde als eine Zufallsvariable mit unbekanntem Erwartungswert $\mu$ und der Varianz $\sigma^2 = 6{,}25$ [sec$^2$] angenommen.

Seien $X_i$ unabhängige Stichproben der CPU-Zeit und $\bar{X}$ das arithmetische Mittel der $X_i$. Wie viele unabhängige Stichproben sollten mindestens gemessen werden, damit

$$\Pr[|\bar{X} - \mu| < 0{,}1] \geq 0{,}9$$

gilt. Man verwende zur Beantwortung der Frage

1. die Ungleichung von Chebyshev,
2. den Zentralen Grenzwertsatz.

**Lösung**

1. $\Pr[|\bar{X} - \mu| < 0{,}1] \geq 0{,}9$ gilt genau dann, wenn $\Pr[|\bar{X} - \mu| \geq 0{,}1] \leq 0{,}1$. Wir benutzen die letztere Form.

   Es gilt für alle $i$ $\mathbb{E}[\bar{X}] = \mathbb{E}[X_i] = \mu$ und $\mathrm{Var}[\bar{X}] = \frac{1}{n}\mathrm{Var}[X_i]$. Chebyshev kann nun direkt für alle $t > 0$ wie folgt angewandt werden.

   $$\Pr[|\bar{X} - \mu| \geq 0{,}1] \leq \frac{\mathrm{Var}[X_i]}{n \cdot 0{,}1^2} = \frac{6,25}{n \cdot 0{,}1^2}\,.$$

   Wir betrachten $n$ so, dass gilt $\frac{6{,}25}{n \cdot 0{,}1^2} \leq 0{,}1$. Diese Ungleichung gilt genau dann, wenn gilt

   $$n \geq \frac{6,25}{0{,}1^3} = 6250\,.$$

   Es reicht aus, $n = 6250$ zu setzen.

2. Wir setzen $\bar{\sigma}^2 = \mathrm{Var}[\bar{X}]$. Dann gilt $\bar{\sigma}^2 = \frac{\mathrm{Var}[X_i]}{n} = \frac{6{,}25}{n}$ und $\bar{\sigma} = \frac{2{,}5}{\sqrt{n}}$.

   $$\begin{aligned} \Pr[|\bar{X} - \mu| < 0{,}1] &= \Pr[-0{,}1 < \bar{X} - \mu < 0{,}1] \\ &= \Pr\left[-\frac{0{,}1}{\bar{\sigma}} < \frac{\bar{X} - \mu}{\bar{\sigma}} < \frac{0{,}1}{\bar{\sigma}}\right] \\ &\approx \Phi\left(\frac{0{,}1}{\bar{\sigma}}\right) - \Phi\left(-\frac{0{,}1}{\bar{\sigma}}\right) \\ &= 2 \cdot \Phi\left(\frac{0{,}1}{\bar{\sigma}}\right) - 1\,. \end{aligned}$$

   Es gilt $2 \cdot \Phi\left(\frac{0{,}1}{\bar{\sigma}}\right) - 1 \geq 0,9$ genau dann, wenn

   $$\Phi\left(\frac{0{,}1}{\bar{\sigma}}\right) \geq 0,95\,.$$

   Das Quantil $z_{0{,}95}$ bestimmt man mit Interpolation aus der Tabelle

   $$z_{0{,}95} \approx 1,645\,.$$

Damit erhalten wir für $n$ die Ungleichung

$$\frac{0{,}1}{\bar{\sigma}} = \frac{\sqrt{n}}{25} \geq 1,645$$

mit der Lösung

$$n \geq 25^2 \cdot 1,645^2 = 1691,26\ldots \quad .$$

## Tutoraufgabe 2

Ziel dieser Aufgabe ist es, einen ML-Schätzer für die Erfolgswahrscheinlichkeit $p \in (0,1)$ einer negativ binomialverteilten Zufallsvariablen zu bestimmen. Hierfür seien $X_1, \ldots, X_n$ unabhängige Zufallsvariable, wobei jedes $X_i$ negativ binomialverteilt mit Erfolgswahrscheinlichkeit $p$ bei $m$ zu erzielenden Erfolgen sei, d. h., jedes $X_i$ hat die Dichte

$$f_{X_i}(k) = \binom{k-1}{m-1} p^m (1-p)^{k-m} \quad (\text{mit } k \geq m)\,.$$

Der Parameter $m$ sei bekannt. Zu schätzen ist $p$.

1. Es sei $\vec{k} = (k_1, \ldots, k_n)$ ein Stichprobenvector (mit $k \geq m$). Stellen Sie die Likelihood-Funktion $L(\vec{k}; p)$ auf.

2. Maximieren Sie $L(\vec{k}; p)$ und bestimmen Sie den entsprechenden ML-Schätzer für $p$.

3. Zeigen Sie, dass der hergeleitete ML-Schätzer i. A. nicht erwartungstreu ist.

   *Hinweis*: Verwenden Sie $\ln(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x^i}{i}$ für $x \in (-1, 1]\,.$

**Lösung**

1.

$$\begin{aligned}
L(\vec{k}; p) &= \prod_{i=1}^{n} f_{X_i}(k_i) \\
&= \prod_{i=1}^{n} \left( \binom{k_i - 1}{m-1} p^m (1-p)^{k_i - m} \right) \\
&= \left( \frac{p}{1-p} \right)^{mn} \cdot (1-p)^{\overbrace{\sum_{i=1}^{n} k_i}^{=:K}} \cdot \underbrace{\prod_{i=1}^{n} \binom{k_i - 1}{m-1}}_{=:C} \\
&= p^{mn} \cdot (1-p)^{K-mn} \cdot C\ .
\end{aligned}$$

2. Wir bestimmen das Maximum von $L$ analog wie in der VA 1 durch Bildung der Ableitung und eine Begründung, dass die 2. Ableitung an der gefundenen Stelle negativ sein muss.

$$\begin{aligned}
\frac{\mathrm{d}\, L(\vec{k}; p)}{\mathrm{d}\, p} &= Cmn \cdot p^{mn-1} \cdot (1-p)^{K-mn} - Cp^{mn} \cdot (K - mn) \cdot (1-p)^{K-mn-1} \\
&= Cp^{mn-1} \cdot (1-p)^{K-mn-1} \cdot [mn(1-p) - p(K-mn)] \\
&= Cp^{mn-1} \cdot (1-p)^{K-mn-1} \cdot [mn - pK]\,.
\end{aligned}$$

Die Nullstellen sind $p = 0$, $p = 1$ und $p = \frac{mn}{K}$, wobei die mittlere Nullstelle das Maximum von $L$ liefert. Der gesuchte Schätzer für die Erfolgswahrscheinlichkeit $p$ ist also

$$U = \frac{mn}{\sum_{i=1}^{n} X_i} \,.$$

3. $U$ ist erwartungstreu, falls $E[U] = p$ gilt. $U$ ist aber nicht einmal für $m = n = 1$ erwartungstreu, wie die folgende Rechnung zeigt.

   Seien $m = n = 1$. Dann gilt $U = \frac{1}{X_1}$ und $X_1$ ist geometrisch verteilt:

$$\mathbb{E}[U] = \mathbb{E}\left[\frac{1}{X_1}\right] = \sum_{i=1}^{\infty} \frac{1}{i} \cdot p \cdot (1-p)^{i-1} = \frac{p}{1-p} \sum_{i=1}^{\infty} \frac{(1-p)^i}{i} \,.$$

   Wir berechnen die unendliche Reihe (siehe Hinweis):

$$-\ln p = -\ln(1 + (p-1)) = -\sum_{i=1}^{\infty} (-1)^{i+1} \frac{(p-1)^i}{i} = \sum_{i=1}^{\infty} \frac{(1-p)^i}{i} \,.$$

   Damit folgt

$$\mathbb{E}[U] = p \cdot \frac{-\ln p}{1-p} \neq p \,.$$

## Tutoraufgabe 3

Beim Testen von Hypothesen bezeichnet man die zu überprüfende Hypothese (Nullhypothese) generell mit $H_0$ und die Alternative mit $H_1$.

Ein Tierhändler erhält ein Paket mit 100 Frettchen. Er will testen, ob weniger als zehn ($< 10$) dieser Frettchen aggressiv und bissig sind. Dazu hält er zehn Frettchen seinen Finger hin und nimmt das Paket nur an, wenn ihn keines davon beißt (wir nehmen an, dass ein aggressives Frettchen sofort zubeißen würde).

Wie lauten die Hypothesen des Händlers? Was ist das Signifikanzniveau des Tests?

**Lösung**

Die Zahl $X$ der aggressiven Frettchen in der Stichprobe vom Umfang 10 ist hypergeometrisch verteilt mit $N = 100$ und $S$= Anzahl der aggressiven Frettchen im Paket. Der Händler testet die Hypothese

$$H_0 : S \leq 9$$

gegen

$$H_1 : S \geq 10$$

und verwirft die Nullhypothese dabei im Fall $X \geq 1$. Für $S = 9$ ist

$$\Pr_S[X > 0] = 1 - \Pr_S[X = 0] = 1 - \frac{\binom{9}{0}\binom{91}{10}}{\binom{100}{10}} = 62.9\%.$$

Das ist auch das effektive Niveau des Tests, denn für $S < 9$ ist die Verwerfungswahrscheinlichkeit kleiner.