Technische Universität München (I7) **Sommer 2012**
Prof. J. Esparza / Dr. M. Luttenberger / M. Schlund **5. Juli 2012**

Lösung

Diskrete Wahrscheinlichkeitstheorie – Hausaufgabe 10

Abgabe bis zum 4.7. bis 8:30.

Alle Antworten sind unter Angabe des Rechenwegs zu begründen, soweit nicht anders gefordert! Fragen gerne im infler-Forum posten :).

Aufgabe 10.1 Abzugeben 2P

Die Zeit, bis Deutschland das erste Tor schießt, sei $X \sim \exp(\lambda)$. Entsprechend sei $Y \sim \exp(\mu)$ die Zeit, bis Italien das erste Tor schießt. Wir nehmen an, dass X, Y unabhängig sind (was sicherlich eine grobe Vereinfachung darstellt).

Natürlich gilt $\lambda^{-1} < \mu^{-1}$.

Drücken Sie die W'keit $\Pr[X < Y]$, dass Deutschland das erste Tor schießt, allein durch λ und μ aus.

Lösung:

$$\Pr[X < Y] = \int_{\{(x,y)\in[0,\infty)^2:\ x<y\}} f_{X,Y}(x,y)dxdy \overset{X,Y \text{ unabh.}}{=} \int_0^\infty f_Y(y) \int_0^y f_X(x)dxdy = \int_0^\infty f_Y(y)F_X(y) =$$

$$= \int_0^\infty \mu e^{-\mu y}(1 - e^{-\lambda y})dy = 1 - \frac{\mu}{\lambda + \mu} \int_0^\infty (\lambda + \mu)e^{-(\lambda+\mu)y}dy = 1 - \frac{\mu}{\lambda + \mu}$$

Aufgabe 10.2 Abzugeben: a) 2P

Bei einem Bürgerentscheid gebe es die zwei Wahlmöglichkeiten "Pro" und "Contra". Unter den 1000000 Wahlberechtigten gebe es eine Minderheit von 2000 Personen, die geschlossen für "Pro" stimmen, während die restlichen 998000 Wähler unentschlossen sind und sich unabhängig voneinander, durch Werfen einer fairen Münze für eine der beiden Wahlmöglichkeiten entscheiden. Wie immer gilt der Entscheid als angenommen, wenn mehr als 50% der Wähler dafür stimmen.

(a) Berechnen Sie approximativ (mittels Approximation durch die Normalverteilung) die Wahrscheinlichkeit, dass der Bürgerentscheid angenommen wird.

(b) Wie groß muss allgemein bei n Wählern diese "entschlossene Minderheit" mindestens sein, um die Wahl mit einer W'keit von mehr als 0.95 für sich zu entscheiden (wenn der Rest der Wähler wie vorher zufällig dafür bzw. dagegen stimmt)?

Lösung:

(a) Sei X die ZV, welche die "Pro"-Stimmen zählt und $n = 10^6$ die Gesamtzahl der Wähler. $Y := X - 2000$ ist $\text{bin}(n - 2000, 1/2)$-verteilt, also hat Y den Erwartungswert $\mu_Y := \mathbb{E}[Y] = \frac{10^6-2000}{2} = 499000$. $\text{Var}[Y] = (n - 2000) \cdot 1/2 \cdot 1/2 = 249500$ und $\sigma_Y := \sqrt{\text{Var}[Y]} = 499.49975\ldots$.

$$\Pr[X > 500000] = \Pr[Y > 498000] = \Pr\left[\frac{Y - \mu_Y}{\sigma_Y} > \frac{498000 - \mu_Y}{\sigma_Y}\right] \approx 1-\Phi(\frac{498000 - \mu_Y}{\sigma_Y}) \approx 1-\Phi(-2.00) = \Phi(2.00) = 0.9772$$

(b) Ähnliche Rechnungen wie vorher—nur diesmal allgemein. Sei k die gesuchte Größe der entschlossenen Minderheit. Setze $Y = X - k$. $Y \sim \text{bin}(n-k, 1/2)$. $\mu_Y := \mathbb{E}[Y] = \frac{n-k}{2}$, $\sigma_Y = \sqrt{(n-k)/4}$

$$\Pr[X > n/2] = \Pr\left[\frac{Y-\mu_Y}{\sigma_Y} > \frac{n/2-k-\mu_Y}{\sigma_Y}\right] \approx 1 - \Phi(\frac{n/2-k-\mu_Y}{\sigma_Y}) \overset{!}{\geq} 0.95$$

$$\Leftrightarrow \frac{n-2k-2\mu_Y}{2\sigma_Y} \leq z_{0.05} \Leftrightarrow n - 2k - (n-k) \leq z_{0.05}\sqrt{n-k}$$

$$\Leftrightarrow k^2 + z_{0.05}^2 k - z_{0.05}^2 n \geq 0 \Leftarrow k \geq \frac{-z_{0.05}^2 + \sqrt{z_{0.05}^4 + 4z_{0.05}^2 n}}{2}$$

Aus einer Tabelle: $z_{0.05} = -1.645$

Größenordnungsmäßig reicht also $k \sim \sqrt{n}$.

Aufgabe 10.3 **Abzugeben: (a),(c)** **2P+2P**

Die Zeit, bis ein Student eine Klausur abgibt, sei als $\exp(\lambda)$-verteilt angenommen mit $\lambda^{-1} = 90$ Minuten. Die Abgabezeiten von n Studenten sind dann durch unabhängige $\exp(\lambda)$-verteilte ZV $X_1, \ldots, X_n$ beschrieben. Wir interessieren uns für die zeitlichen Abstände $\Delta_i \geq 0$. Das heißt, die i-te Abgabe findet zum Zeitpunkt $T_i := \sum_{j=1}^{i} \Delta_j$ statt. Offensichtlich gilt $T_1 = \Delta_1 = \min(X_1, \ldots, X_n)$ und $T_n = \max(X_1, \ldots, X_n)$.

(a) Bestimmen Sie $\mathbb{E}[T_1]$.

(b) Argumentieren Sie (informell) anhand der Unabhängigkeit der X_i und der Gedächtnislosigkeit der Exponentialverteilung, dass Folgendes gilt:

Die ZVen $\Delta_1, \ldots, \Delta_n$ sind unabhängig mit $\Delta_i \sim \exp((n-i+1)\lambda)$.

Hinweis: Es ist keine explizite Rechung verlangt. Eine kurze, schlüssige Begründung genügt.

(c) Bestimmen Sie den erwarteten Zeitpunkt der letzten Abgabe $\mathbb{E}[T_n]$.

Lösung:

(a) Nach Vorlesung ist $T_1 = \min(X_1, \ldots, X_n)$ exponentialverteilt mit Parameter $n\lambda$.

Somit $\mathbb{E}[T_1] = \frac{1}{n\lambda}$.

(b) Informelle Argumentation:

Es sei $T_i = t_i$ der Zeitpunkt der i-ten Abgabe. Da $\Pr[X_i = X_j] = 0$ für $i \neq j$ auf Grund der Unabhängigkeit der X_i, können wir annehmen, dass $T_{i+1}, \ldots, T_n > t_i$. Auf Grund der Gedächtnislosigkeit der Exponentialverteilung $\Pr[X > t+s \mid X > t] = \Pr[X > s]$, bedeutet dies, dass das ,,Abgabeexperiment" mit $n-i$,,frichen" Studenten $X'_1, \ldots, X'_{n-i}$ neugestartet wird. Die nächste Abgabe findet daher zum Zeitpunkt $T_{i+1} = T_i + \min(X'_1, \ldots, X'_{n-i})$ statt. Damit muss $\Delta_{i+1} = \min(X'_1, \ldots, X'_{n-i}) \sim \exp((n-i)\lambda)$ gelten. Weiterhin müssen die Δ_i unabhängig sein, da bei jeder Abgabe das Experiment auf Grund der Gedächtnislosigkeit vergisst, was bisher geschehen ist.

Formale Rechnung:

Auf Grund der Unabhängigkeit ist auch $X_i - X_j$ stetig verteilt für $i \neq j$. Daher geben zwei Studenten mit W'keit $\Pr[X_i = X_j] = \Pr[X_i - X_j = 0] = F_{X_i - X_j}(0) - F_{X_i - X_j}(0) = 0$ zum selben Zeitpunkt ab. Entsprechend folgt $\Pr\left[\bigvee_{i \neq j} X_i = X_j\right]$ 0 bzw. $\Pr\left[\bigwedge_{i \neq j} X_i \neq X_j\right] = 1$. Seien $d_1, \ldots, d_n \geq 0$ gegeben.

Wir bestimmen zunächst die gemeinsame Dichte der Δ_i für den Fall $n = 2$:

$$
\begin{aligned}
&\Pr[\Delta_1 \leq d_1, \Delta_2 \leq d_2] \\
=&\Pr[T_1 \leq d_1, T_2 - T_1 \leq d_2] \\
=&\Pr[T_1 \leq d_1, T_2 - T_1 \leq d_2, X_1 < X_2] + \Pr[T_1 \leq d_1, T_2 - T_1 \leq d_2, X_1 > X_2] + \Pr[T_1 \leq d_1, T_2 - T_1 \leq d_2, X_1 = X_2] \\
=&\Pr[X_1 \leq d_1, X_2 - X_1 \leq d_2, X_1 < X_2] + \Pr[X_2 \leq d_1, X_1 - X_2 \leq d_2, X_1 > X_2] \\
=&2 \cdot \Pr[X_1 \leq d_1, X_2 - X_1 \leq d_2, X_1 < X_2] \\
=&2 \cdot \Pr[X_1 \leq d_1, X_1 < X_2 \leq X_1 + d_2] \\
=&2 \int_{x_1=0}^{d_1} \lambda e^{-\lambda x_1} \int_{x_2=x_1}^{x_1+d_2} \lambda e^{-\lambda x_2} dx_2 dx_1 \\
=&2 \int_0^{d_1} \lambda e^{-\lambda x_1} e^{-\lambda x_1} (1 - e^{-\lambda d_2}) dx_1 \\
=&(1 - e^{-2\lambda d_1})(1 - e^{-\lambda d_1}).
\end{aligned}
$$

Damit folgt

$$
\begin{aligned}
\Pr[\Delta_1 \leq d_1] &= 1 - e^{-2\lambda d_1} \\
\Pr[\Delta_2 \leq d_2] &= 1 - e^{-\lambda d_2} \\
\Pr[\Delta_1 \leq d_1] \cdot \Pr[\Delta_2 \leq d_2] &= \Pr[\Delta_1 \leq d_1] \cdot \Pr[\Delta_2 \leq d_2]\,.
\end{aligned}
$$

Somit sind Δ_1, Δ_2 unabhängig mit $\Delta_i \sim \exp((n + 1 - i)\lambda)$ für $n = 2$.

Für allgemeines $n \geq 2$ folgt analog (mittels Induktion):

$$
\begin{aligned}
&\Pr\left[\bigwedge_{i\in[n]} \Delta_i \leq d_n\right] \\
=&n!\Pr\left[X_1 \leq d_1 \wedge \bigwedge_{i=2}^{n} X_{i-1} < X_i \leq X_{i-1} + d_i\right] \\
=&\int_{x_1=0}^{d_1} \int_{x_2=x_1}^{x_1+d_1} \cdots \int_{x_n=x_{n-1}}^{x_{n-1}+d_n} n!\lambda^n e^{-\lambda(x_1+x_2+\ldots+x_n)} dx_1 dx_2 \ldots dx_n \\
=&(1 - e^{-n\lambda d_1})(1 - e^{-(n-1)\lambda d_2}) \cdot \ldots \cdot (1 - e^{-\lambda d_n}).
\end{aligned}
$$

Somit $\Pr[\Delta_i \leq d_i] = 1 - e^{-(n-i+1)\lambda d_i}$, womit die Behauptung folgt.

(c) $\mathbb{E}[T_n] = \sum_{i=1}^n \mathbb{E}[\Delta_i] = \sum_{i=1}^n \frac{1}{(n-i+1)\lambda} = \frac{1}{\lambda} \sum_{i=1}^n \frac{1}{i} = \frac{1}{\lambda} H_n$ mit H_n die n-te harmonische Zahl.

Aufgabe 10.4 Abzugeben: (a),(b) 2P+2P

Die zeitlichen Abstände Δ_i zwischen den Zeitenpunkten, an denen die deutsche Fussballnationalmannschaft Tore gegen Italien erzielt, sind unabhängig expontentialverteilt mit Parameter $\lambda^{-1} = 5$ Minuten. Das heißt, das k-te Tor fällt $T_k = \sum_{i=1}^k \Delta_i$ Minuten nach Anstoß (mit $T_0 := 0$).

Da Prof. E. am Donnerstag die Vorlesung erst später beendet, verpassen Sie die ersten $t > 0$ Minuten des Spiels.

(a) Bestimmen Sie zunächst die Verteilung und den Erwartungswert der ZV $N(t)$, welche die Tore zählt, welche Sie verpasst haben.

Sie können somit erst das $N(t) + 1$-te Tor am Fernseher direkt miterleben, womit Sie $W(t) = T_{N(t)+1} - t$ Minuten auf das nächste Tor warten müssen.

(b) Leiten Sie (informell) auf Grund der Gedächtnislosigkeit der Exponentialverteilung und der Unabhängigkeit der Δ_i die Verteilung von $W(t)$ her.

Hinweis: Ihre Argumentation sollte schlüssig sein, muss jedoch keine Rechnungen beinhalten.

Lösung:

(a) Nach Vorlesung ist $N(t) = \max\{n \in \mathbb{N}_0 \mid T_n < t \leq T_{n+1}\}$ Poisson-verteilt mit Parameter λt.

Somit $\mathbb{E}[N(t)] = \lambda t = t/5$ Tore.

(b) Wir nehmen an, dass das letzte Tor zum Zeitpunkt $t_n < t$ gefallen ist. Das Experiment ,,nächstes Tor" ist somit durch Δ_{n+1} beschrieben. Da bis vor dem Eintreffen zum Zeitpunkt t das $n+1$-te Tor noch nicht gefallen ist, gilt auf Grund der Gedächtnislosigkeit der Exponentialverteilung

$$\Pr[W_t > s] = \Pr[\Delta_{n+1} \geq (t-t_n) + s \mid \Delta_{n+1} \geq (t-t_n)] = \Pr[\Delta_{n+1} \geq s],$$

d.h. W_t ist ebenfalls exponentialverteilt mit Parameter λ.

Formaler Beweis (nicht verlangt):

Es bezeichne $f_n(t) := f_{T_n}(t)$ die Dichte von $\sum_{i=1}^n \Delta_i$. (Nach Zentralübung gilt $f_n(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} I_{[0,\infty)}(t)$, was aber nicht benötigt wird.)

$$\begin{aligned}
&\Pr[W > s] \\
=& \sum_{n\in\mathbb{N}_0} \Pr[T_n < t, T_{n+1} > t+s] = \sum_{n\geq 0} \Pr[T_n < t, T_n + \Delta_{n+1} > t+s] = \sum_{n\geq 0} \Pr[t+s-\Delta_{n+1} < T_n < t] \\
=& \sum_{n\geq 0} \int_{\{(\delta,\tau)\in\mathbb{R}^2 \mid t+s-\delta<\tau<t\}} f_{\Delta_{n+1},T_n}(\delta,\tau) d\delta d\tau = \sum_{n\geq 0} \int_{\{(\delta,\tau)\in\mathbb{R}^2 \mid t+s-\delta<\tau<t, \delta\geq s\}} f_{\Delta_{n+1}}(\delta) f_{T_n}(\tau) d\delta d\tau \\
=& \sum_{n\geq 0} \int_{\{(\delta,\tau)\in\mathbb{R}^2 \mid t+s-\delta<\tau<t, \delta\geq s\}} f_1(\delta) f_n(\tau) d\delta d\tau = \sum_{n\geq 0} \int_{\delta=s}^{\infty} f_1(\delta) \int_{\tau=t+s-\delta}^{t} f_n(\tau) d\tau d\delta \\
=& \sum_{n\geq 0} \int_{\delta=s}^{\infty} f_1(\delta) \left(F_n(t) - F_n(t+s-\delta)\right) d\delta = \sum_{n\geq 0} F_n(t)\left(1-F_1(s)\right) - \int_{\delta=s}^{\infty} f_1(\delta)\cdot F_n(t+s-\delta) d\delta \\
=& \sum_{n\geq 0} F_n(t) e^{-s\lambda} - \int_{x=0}^{\infty} f_1(s+x) F_n(t-x) dx = \sum_{n\geq 0} F_n(t) e^{-s\lambda} - e^{-s\lambda} \int_{x=0}^{\infty} f_1(x) F_n(t-x) dx \\
=& \sum_{n\geq 0} F_n(t) e^{-s\lambda} - e^{-s\lambda} \int_{x=-\infty}^{\infty} f_1(x) F_n(t-x) dx = \sum_{n\geq 0} F_n(t) e^{-s\lambda} - e^{-s\lambda} F_{n+1}(t) \\
=& e^{-s\lambda} \sum_{n\geq 0} \Pr[T_n < t \leq T_{n+1}] = e^{-s\lambda} \sum_{n\geq 0} \Pr[N(t) = n] \\
=& e^{-s\lambda}
\end{aligned}$$

Aufgabe 10.5 **Abzugeben: (a), (b), (d), (e)** **2P+2P+2P+2P**

In der Vorlesung wurden Schätzer für den Parameter M einer diskreten Gleichverteilung auf $\{1, 2, \ldots, M\}$ diskutiert.

Wir betrachten den einfacheren Fall einer stetigen Gleichverteilung auf dem Intervall $[0, M]$ für zu schätzenden Parameter M.

Hierzu seien $X_1, \ldots, X_n$ unahbängige ZVen, jeweils gleichverteilt auf $[0, M]$.

Wir betrachten die Schätzer:

$$T_1 := \max\{X_1, \ldots, X_n\} \quad \text{und} \quad T_2 = 2\overline{X} = \frac{2}{n}\sum_{i=1}^{n} X_i.$$

(a) Zeigen oder widerlegen Sie, dass T_1 der Maximum-Likelihood-Schätzer für M ist.

(b) Entscheiden Sie für jeden der beiden Schätzer, ob er erwartungstreu bzgl. M ist.

(c) Zeigen Sie dass, $\mathrm{MSE}(T) = \mathrm{Var}[T] + (\mathbb{E}[T-M])^2$ für einen Schätzer T des Parameter M.

(d) Vergleichen Sie die beiden Schätzer anhand des mittleren quadratischen Fehlers.

Welcher Schätzer ist konsistent im quadratischen Mittel?

Hinweis: Sie dürfen das Resultat der letzten Teilaufgabe verwenden

(e) Bestimmen Sie (approximativ) für T_2 ein möglichst kleines $\delta > 0$, so dass

$$\Pr[|T_2 - M| \geq \delta] \leq 0.05 =: \alpha$$

gilt.

Hinweis: Verwenden Sie eine Approximation mittels der Standardnormalverteilung.

Lösung:

(a) ML-Funktion: $L = L(x_1, \ldots, x_n) = \prod_{i=1}^n f_X(x_i) = M^{-n}$ für $M \geq \max(x_1, \ldots, x_n)$.

Somit L maximal für $M = \max(x_1, \ldots, x_n)$, womit T_1 der ML-Schätzer ist.

(b) Verteilungsfunktion von T_1 für $t \in [0, M]$:

$$\Pr[T_1 \leq t] = \Pr[X_1 \leq t, \ldots, X_n \leq t] = F_X(t)^n = \frac{1}{M^n} t^n$$

Somit hat T_1 die Dichte $f(t) = \frac{n}{M^n} t^{n-1} I_{[0,M]}(t)$.

$$\mathbb{E}[T_1] = \frac{n}{M^n} \int_0^M t^n dt = \frac{n}{n+1} M.$$

T_1 ist also nicht erwartungstreu für M.

T_2 ist erwartungstreu für M, da

$$\mathbb{E}[T_2] = \frac{2}{n} \sum_{i=1}^n \mathbb{E}[X_i] = 2\mathbb{E}[X_1] = 2\frac{M}{2} = M.$$

(c)

$$\begin{aligned}\mathsf{mse}(T) =& \mathbb{E}[(T-\theta)^2] = \mathbb{E}[(T - \mathbb{E}[T] + \mathbb{E}[T] - \theta)^2] \\ =& \mathrm{Var}[T] + 2(\mathbb{E}[T] - \theta)\mathbb{E}[T - \mathbb{E}[T]] + (\mathbb{E}[T] - M)^2 = \mathrm{Var}[T] + (\mathbb{E}[T - M])^2.\end{aligned}$$

(d)

$$\begin{aligned}\mathsf{mse}(T_1) =& \mathrm{Var}[T_1] + (\frac{n}{n+1}M - M)^2 = \mathbb{E}[T_1^2] - \left(\frac{n}{n+1}M\right)^2 + \left(\frac{1}{n+1}M\right)^2 \\ =& \frac{n}{n+2}M^2 - \frac{n^2-1}{(n+1)^2}M^2 = \frac{n}{n+2}M^2 - \frac{n-1}{n+1}M^2 = \frac{n^2+n-n^2-n+2}{(n+2)(n+1)}M^2 = \frac{2}{(n+2)(n+1)}M^2. \\ \mathrm{mse}(T_2) =& \mathrm{Var}[T_2] = \frac{4}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{4}{n}\mathrm{Var}[X_1] = \frac{1}{3n}M^2.\end{aligned}$$

Damit sind beide Schätzer konsistent im quadratischen Mittel, allerdings konvergiert T_1 quadratisch, T_2 nur linear mit n.

(e)

$$\begin{aligned}\Pr[|T_2 - M| \leq \delta] =& \Pr[|T_2 - \mathbb{E}[T_2]| \leq \delta] = \Pr\left[-\frac{\sqrt{3n}}{M}\delta \leq \frac{T_2 - \mathbb{E}[T_2]}{\sqrt{\mathrm{Var}[T_2]}} \leq \frac{\sqrt{3n}}{M}\delta\right] \\ \approx& \Phi(\frac{\sqrt{3n}}{M}\delta) - \Phi(-\frac{\sqrt{3n}}{M}\delta) = 2\Phi(\frac{\sqrt{3n}}{M}\delta) - 1 \overset{!}{\geq} 0.95. \\ \rightsquigarrow& \frac{\sqrt{3n}}{M}\delta \overset{!}{\geq} \Phi^{-1}(0.975) \approx 1.96 \\ \rightsquigarrow& \delta \geq \frac{1.132}{\sqrt{n}}M.\end{aligned}$$

Zum Vergleich (nicht verlangt):

Man beachte, dass $T_1 \leq M$ immer gilt:

$$\begin{aligned}\Pr[|T_1 - M| \geq \delta] = \Pr[T_1 \leq M - \delta] = \left(1 - \frac{\delta}{M}\right)^n \overset{!}{\leq} 0.05 \\ \rightsquigarrow \delta \geq M(1 - 0.05^{1/n}).\end{aligned}$$

Da $\frac{1.132}{\sqrt{n}} \geq 1 - 0.05^{1/n}$, erhält man somit für T_1 wiederum das genauere Konfidenzintervall.