Übungen zur Vorlesung „Stochastik für die Informatik"

Abgabe der Lösungen zu den beiden S-Aufgaben dieses Blattes per E-Mail bis Montag, 18. Januar 2021, 10 Uhr.
Die beiden S-Aufgaben W2 und W6 bieten die Möglichkeit zum Erarbeiten zusätzlicher Bonuspunkte, sodass Sie die über das ganze Semester maximal möglichen 12 BP noch leichter erreichen können. Für die Aufgaben W2 und W6 gibt es nicht wie sonst üblich je 6, sondern je 9 Übungspunkte.

Leitfaden zur Abgabe der Übungsaufgaben: siehe "Ankündigungen" im Moodle Kurs

W1. ***Ein approximatives Konfidenzintervall für den Erwartungswert bei bekannter Varianz.*** $X_1, X_2, \ldots$ seien unabhängige, identisch verteilte reellwertige Zufallsvariable mit Standardabweichung $\sigma = 5$ und Erwartungswert μ. Wir setzen $M_n := \frac{1}{n}(X_1 + \cdots + X_n)$.[1]

(i) Aufgrund welchen Satzes ist es gerechtfertigt anzunehmen, dass M_n für hinreichend großes n approximativ normalverteilt ist? Was ist die Standardabweichung von M_n?

(ii) Sei $n = 25$. Wie ist die Zahl c zu wählen, damit das um M_n zentrierte zufällige Intervall $\mathcal{I}_n := [M_n - c, M_n + c]$ die Zahl μ mit Wahrscheinlichkeit $\approx 95\%$ überdeckt? (Übungsaufgabe A27 ist hilfreich.)

(iii) Wie groß muss n sein, damit μ vom zufälligen Intervall $[M_n - 0.2, M_n + 0.2]$ mit Wahrscheinlichkeit $\approx 95\%$ überdeckt wird?

W2.S ***Ein approximatives Konfidenzintervall für den Populationsmittelwert bei bekannter Populationsvarianz.*** S sei eine endliche Menge bestehend aus g Elementen, J sei uniform verteilt auf S und h sei eine Abbildung von S nach $\mathbb{R}$. Wie in F7a3.12 bezeichnen wir

$$\mu = \mathbf{E}[h(J)] = \tfrac{1}{g}\sum_{a\in S} h(a)$$

als Populationsmittelwert und

$$\sigma^2 = \mathbf{Var}[h(J)] = \tfrac{1}{g}\sum_{a\in S}(h(a)-\mu)^2$$

als Populationsvarianz des individuellen Merkmals h. Für $n \le g$ sei $J_1, \ldots, J_n$ ein rein zufälliges Ziehen ohne Zurücklegen aus S, und $X_i := h(J_i)$ das Merkmal die i-ten gezogenen Individuums. Ähnlich wie in W1 geht es um den zufälligen Stichprobenmittelwert $M_n = \frac{1}{n}(X_1 + \cdots + X_n)$ und dessen Standardabweichung σ_{M_n}.

a) Drücken Sie σ_{M_n} durch σ, n und g aus. (Die Übungsaufgabe A24 ist hilfreich!)

b) Bestimmen Sie unter Verwendung von Teil a) und der Aussage in F7a3.15 die approximative Verteilung von M_n, wenn sowohl der Stichprobenumfang n als auch die Differenz $g-n$ hinreichend groß sind.

c) Geben Sie (ausgedrückt durch M_n, σ, n und g) ein um M_n zentriertes zufälliges Intervall $\mathcal{I}_n$ an, welches μ mit Wahrscheinlichkeit $\approx 95\%$ überdeckt.

W3. ***Populationsvarianz geschätzt aus der Stichprobe.*** *In der Situation von W2 bleibt die approximative Überdeckungswahrscheinlichkeit bestehen, wenn im Ausdruck für* $\mathcal{I}_n$ *die Populationsvarianz* σ^2 *durch die* Stichprobenvarianz $s^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - M_n)^2$ *(oder auch durch die sogenannte* modifizierte Stichprobenvarianz $\hat\sigma^2 := \frac{1}{n}\sum_{i=1}^n (X_i - M_n)^2$*) ersetzt wird.*[2]
Überprüfen Sie diese Aussage empirisch anhand des der Situation aus A11b) angepassten R-Programms Mittelwerte.R (mit $g = 100$ und $n = 20$ bzw. $n = 40$).

W4. ***Rayleighverteilung.*** Die Zufallsvariable X sei Exp($\frac{1}{2}$)-verteilt. Berechnen Sie
(i) die Verteilungsfunktion (ii) die Dichte (iii) den Erwartungswert
der Zufallsvariablen $R := \sqrt{X}$.[3]

[1] Weil $\mathbf{E}[M_n] = \mu$ gilt (warum?), nennt man M_n auch einen *erwartungstreuen Schätzer* für μ.

[2] In der Situation von Aufgabe W1 (also für unabhängige $X_1, \ldots, X_n$) ist, wie man zeigen kann, die Stichprobenvarianz s^2 ein erwartungstreuer Schätzer für die Varianz σ^2, d.h. es gilt $\mathbf{E}[s^2] = \sigma^2$. Die modifizierte Stichprobenvarianz $\hat\sigma^2$ ist, kurz gesagt, die "Populationsvarianz der Stichprobe". In der Situation der Aufgabe W1 ist $\hat\sigma^2$ für N(μ, σ^2)-verteilte X_i auch der sogenannte Maximum-Likelihood-Schätzer für σ^2, vg. Buch Seite 124.

[3] Die Verteilung von R heißt *Standard-Rayleighverteilung*, vgl. die Musterlösung zu Blatt 2.

Hinweis zu (i): Schreiben Sie das Ereignis $\{R > b\}$ um zum Ereignis $\{R^2 > b^2\}$ und berechnen Sie so die Wahrscheinlichkeit des Ereignisses $\{R \le b\}$.

Hinweis zu (iii): Hier dürfen Sie verwenden, dass $\int_{-\infty}^{\infty} e^{-u^2/2} du = \sqrt{2\pi}$ gilt, siehe auch W5.

W5. *Die Fläche unter der Gaußschen Glockenkurve.* Zuerst eine Vorbemerkung: Für stetige nichtnegative f_1, f_2 gilt

$$\iint_{\mathbb{R}^2} f_1(x) f_2(y) dx\, dy = \int_{\mathbb{R}} f_1(x) dx \int_{\mathbb{R}} f_2(y) dy,$$

d.h. man kann die Integrationsreihenfolge beliebig vertauschen. Insbesondere gilt also

$$\left(\int_{-\infty}^{\infty} e^{-u^2/2} du \right)^2 = \iint_{\mathbb{R}^2} e^{-(x^2+y^2)/2} dx dy.$$

Verwenden Sie diese Tatsache, um mit der auf F7a1.7 vorgestellten Transformationsformel zu erschließen, dass gilt:

$$\int_{-\infty}^{\infty} e^{-u^2/2} = \sqrt{2\pi}.$$

W6.S *Wie ändert sich der Regressionskoeffizient unter einer linearen Transformation?* G und H seien reellwertige Zufallsvariable; die beste affin lineare Vorhersage von H auf der Basis von G sei $\hat{H} := 3G + 2$. Es sei $X := 2G - 3$ und $Y := 5H + 4$.

a) Drücken Sie den Korrelationskoeffizienten $\kappa_{X,Y}$ durch den Korrelationskoeffizienten $\kappa_{G,H}$ aus.

b) Die beste affin lineare Vorhersage von Y auf der Basis von X sei $\hat{Y} = \beta_1 X + \beta_0$. Bestimmen Sie β_1.

c) Bestimmen Sie in b) auch β_0, wenn bekannt ist, dass $\mu_G = 3$.[4]

W7. *Der Münzwurf lässt grüßen.* a) X_1 sei Binom(n_1, p)-verteilt, X_2 sei Binom(n_2, p)-verteilt und X_1, X_2 seien unabhängig. Wie ist $X_1 + X_2$ verteilt? (Hier ist keine Rechnung nötig!)

b) Y_1 sei Poisson(α)-verteilt, Y_2 sei Poisson(β)-verteilt und Y_1, Y_2 seien unabhängig.

(i) Berechnen Sie für $k = 0, 1, \ldots$ mit dem Binomischen Lehrsatz

$$\mathbf{P}(Y_1 + Y_2 = k) = \sum_{\ell=0}^{k} \mathbf{P}(Y_1 = \ell,\, Y_2 = k - \ell)$$

als Funktion von α, β und k.

(ii) Wie ist $Y_1 + Y_2$ verteilt?

W8. *Bernoulli, Poisson und Gauß geben sich die Hand.* Wir betrachten die drei Verteilungen Binom$(1000, \frac{1}{100})$, Poisson(10) und $N(10, 10)$.

a) Wieso ist damit zu rechnen, dass diese Verteilungen approximativ gleich sind?

b) Plotten Sie die Gewichte $\rho_1(k)$ und $\rho_2(k)$ der ersten beiden Verteilungen sowie die Funktionswerte $\varphi_{10,10}(k)$

(i) für $k = 0, 1, \ldots, 20$,

(ii) für $k = 50, 51, \ldots, 60$.

c) Plotten Sie die relativen Fehler $\frac{\rho_2(k) - \rho_1(k)}{\rho_1(k)}$ und $\frac{\varphi_{10,10}(k) - \rho_1(k)}{\rho_1(k)}$ der Approximation der Binomialgewichte jeweils für die in b) (i) und (ii) angegebenen Bereiche von k.

d) Kommentieren Sie kurz Ihre Ergebnissse aus b) und c).

W9. *Wetten, dass ...* Für lange Winterabende ist hier ein Link zu einer vergnüglichen Lektüre, nicht als Aufgabe, sonderen als Empfehlung und als Einstimmung auf zweistufige Zufallsexperimente und bedingte Wahrscheinlichkeiten, die ein Thema der nächsten Wochen sein werden:
`https://priceonomics.com/the-time-everyone-corrected-the-worlds-smartest/`

[4] Ursprünglich stand hier auch noch "und $\mu_H = -1$". Das verträgt sich aber nicht mit dem Rest der Angabe (warum?). Ein Lob an Frau Tra Mi Tran, die das gemerkt hat!