Statistik für Informatiker **SS 2017**
Prof. Dr. Matthias Birkner
M.Sc. Frederik Klement

Blatt 9

Aufgabe 1: (1 + 1 + 2 + 2)
Seien $X_1, ..., X_n$ unabhängige Bernoulli-verteilte Zufallsvariablen mit Parameter $p \in (0,1)$ ($q := 1-p$). Ein *Run* ist ein maximaler Teilblock von $X := (X_1, ..., X_n)$ der entweder nur aus Nullen oder nur aus Einsen besteht. (0,0,1,1,1,0,1) besteht zum Beispiel aus 4 Runs. Y_n bezeichne die Anzahl der Runs.

a): Argumentieren Sie: $Y_n = 1 + \sum_{i=2}^{n} \mathbf{I}_{\{X_{i-1} \neq X_i\}}$. **Hinweis:** Keine Induktion nötig.

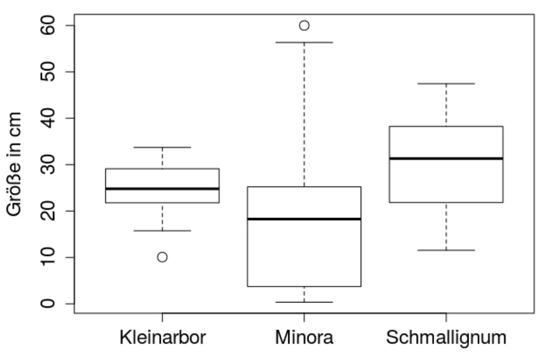
b): Zeigen Sie: $\mathbf{E}[Y_n] = 1 + 2(n-1)pq$.

c): Zeigen Sie:

$$\mathbf{Cov}\big(\mathbf{1}_{\{X_{i-1} \neq X_i\}}, \mathbf{1}_{\{X_{j-1} \neq X_j\}}\big) = \begin{cases} 2pq(1-2pq) & , falls\ i = j \\ pq(1-4pq) & , falls\ i = j \pm 1 \\ 0 & , sonst \end{cases}.$$

d): Zeigen Sie $\mathrm{Var}[Y_n] = (4n-6)pq - (12n-20)p^2q^2$.

Aufgabe 2: (1 + 3 Punkte)
In einer Studie zu Bonsai-Bäumen wurden jeweils **20 Exemplare** der Arten Kleinarbor, Minora und Schmallignum vermessen. Die Ergebnisse dieser Messungen wurden mithilfe von Boxplots illustiert.



a): Bestimmen Sie das 1.Quartil, den Median und das 3.Quartil aller drei Baumarten.

Hinweise: Bei einer Abweichung unter 1.5 cm vom wahren Wert erhalten Sie die volle Punktzahl.

b): Entscheidene Sie, welche der folgenden Aussage wahr oder falsch sind.

i) Mehr als die Hälfte aller Kleinarbor-Bäume ist kleiner als 20 cm.

ii) Es gibt mindestens einen Minora-Baum, der größer als 40 cm ist.

iii) Mehr als die Hälfte aller Schmallignum-Bäume ist größer als 75% aller Minora- und Kleinarbor-Bäume.

iv) Mehr als 5 Schmallignum-Bäume sind kleiner als 30 cm.

v) Es gibt mindestens 5 Schmallignum-Bäume, die größer sind als alle Kleinarbor-Bäume.

vi) Es gibt mindestens ein Schmallignum-Baum, der kleiner ist als alle Kleinarbor-Bäume.

Aufgabe 3: (1 + 1 + 1 Punkte)
Im Rahmen eines Forschungprojekts wurden über das Rhein-Main-Gebiet verstreut sieben Bäume vermessen. Sechs der Messungen haben folgende Ergebnisse geliefert:

Baum	1	2	3	4	5	6
Höhe	5.29m	2.27m	0.01m	6.14m	6.43m	1.67m

Die Messung von Baum Sieben ist leider verloren gegangen, somit ist nichts über seine Höhe bekannt. Was können Sie dennoch über den Median **aller sieben** Bäume sagen?

a):
1. Der Median ist höchstens _____ m groß.
2. Der Median ist mindestens _____ m groß.

Der höchste jemals vermessene Baum besaß eine Höhe von 132,58 Meter. Wenn wir nun annehmen, das Baum Sieben eine Höhe zwischen 0 und 132,58 Meter besitzt, was können Sie dann über den Mittelwert **aller sieben** Bäume sagen?

b):
1. Der Mittelwert ist höchstens _____ m groß.
2. Der Mittelwert ist mindestens _____ m groß.

Eine zweite Testreihe mit sieben Bäumen ergab folgende Tabelle:

Baum	1	2	3	4	5	6	7
Höhe	1.10m	1.26m	3.29m	3.37m	6.96m	15.03m	30.19m

Während sechs der Messungen korrekt sind, wurde bei einem Baum falsch gemessen, sodass dessen wahre Höhe unbekannt ist. Darüber hinaus ist auch nicht bekannt, beim welchem Baum der Fehler passiert ist. Somit kann jeder der obigen Werte falsch sein, während die anderen sechs richtig sind. Was können Sie über den Median der **sieben wahren** Messwerte sagen?

c):
1. Der Median ist höchstens _____ m groß.
2. Der Median ist mindestens _____ m groß.

Aufgabe 4 (2 + 2 Punkte)
Sei $x_1 \leq x_2 \leq x_3 \leq ... \leq x_n$ eine geordnete Stichprobe x einer uns unbekannten Verteilung. Die empirische Verteilungsfunktion der Stichprobe x ist gegeben durch

$$\hat{F}_x(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[x_i,\infty)}(t), \quad t \in \mathbb{R}.$$

Wir vermuten bei der Verteilung von x handelt es sich um die Verteilung P mit Verteilungsfunktion F_P. Ein Maß dafür, wie sehr die Verteilung der Stichprobe der Verteilung P entspricht, ist die Kolmogorov-Smirnov-Statistik, welche gegeben ist durch

$$D^P(x) = \sup_t |\hat{F}_x(t) - F_P(t)|$$

Handelt es sich bei P um die Standardnormalverteilung $\mathcal{N}(0,1)$ mit Verteilungsfunktion Φ, dann lässt sich die $D^{\mathcal{N}}(x)$ berechnen durch

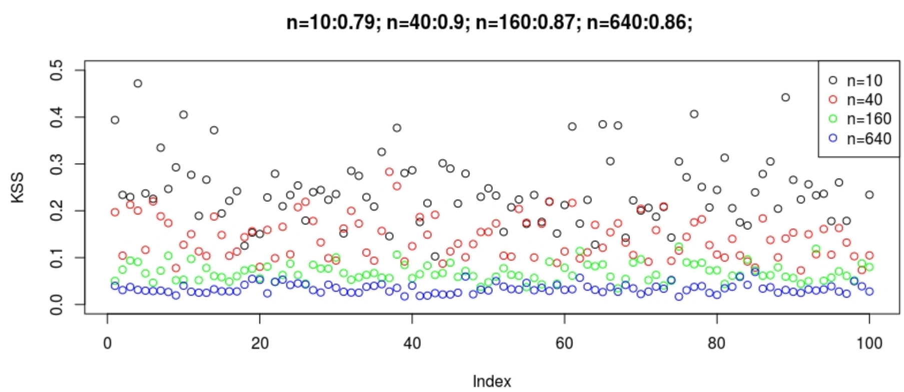
$$D^{\mathcal{N}}(x) = \max\{d_1, d_2, ..., d_n\},$$

wobei $d_i = \max\{|\hat{F}_x(x_i) - \Phi(x_i)|, |\hat{F}_x(x_i) - \Phi(x_{i+1})|\}$ für $i = 1, 2, ..., n-1$ und $d_n = |1 - \Phi(x_n)|$.

a): Schreiben Sie die Funktion `KSS(x)`, welche $D^{\mathcal{N}}$(`x`) berechnet für die möglicherweise ungeordnete Stichprobe `x` (nutzen Sie `sort`).

b): Wir wollen empirisch überprüfen, ob $\hat{F}_x$ mit wachsenden Stichprobenumfang n uniform gegen F_P konvergiert. Simulieren Sie für $n \in \{10, 40, 160, 640\}$ genau 100 mal eine Stichprobe mit dem Umfang n von $X \sim \mathcal{N}(0,1)$ und berechnen Sie für jede der 400 Stichproben die Kolmogorov-Smirnov-Statistik. Plotten Sie Ihre Ergebnisse und verwenden Sie für jedes n eine andere Farbe. Berechnen Sie auch den Mittelwert für jedes n, multiplizieren Sie diesen jeweils mit $\sqrt{n}$ und geben Sie das Ergebnis in der Überschrift Ihres Bildes an.

Hinweis: Ihr Bild sollte in etwa so aussehen:



Achtung: Zusätzlich zur elektronischen Abgabe per Mail, drucken Sie bitte Ihren Code aus und reichen Sie diesen zusammen mit Ihrer restlichen Abgabe ein.

Abgabe: 10 Uhr 30.06.2017