Wissensentdeckung in Datenbanken SoSe 2018

Lösungsvorschlag Übungsblatt 01

Sebastian Buschjäger, Malte Jastrow

Aufgabe 1

a)

Stichprobenraum

$$\Omega = \{\text{Kopf}, \text{Zahl}\}$$

Sigma-Algebren:

$$\mathfrak{A}_1 = \{\emptyset, \Omega\}$$
$$\mathfrak{A}_2 = \{\emptyset, \{\text{Kopf}\}, \{\text{Zahl}\}, \Omega\}$$

Offensichtlich ist nur $\mathfrak{A}_1$ spannend. Weniger spannend ist es hier tatsächlich mal die Bedingungen einer σ-Algebra nachzurechnen. Die Eigenschaften einer Ereignisalgebra gelten offensichtlich immer, die zusätzlichen Eigenschaften ergeben sich direkt aus den Fallunterscheidungen $\emptyset \in A$ und $\Omega \in A$. Falls $\emptyset \in A$ ist, so ist der unendliche Schnitt auch immer $\emptyset$, also in der σ-Algebra. Falls nur Kopf oder nur Zahl und evtl. Ω enthalten sind, ist der Schnitt Kopf bzw. Zahl, also auch enthalten. Wenn sowohl Kopf als auch Zahl enthalten sind, ist der Schnitt wieder leer, was auch enthalten ist. Wenn nur Ω in der Folge enthalten ist, ist der Schnitt auch Ω, damit sind alle möglichen Fälle abgedeckt. Die Eigenschaft über die Vereinigung folgt analog.

b)

Wahrscheinlichkeitsfunktion:

$$P(\emptyset) = 0$$
$$P(\{\text{Kopf}\}) = 0.5$$
$$P(\{\text{Zahl}\}) = 0.5$$
$$P(\Omega) = 1$$

Sei im Folgenden Kopf mit 0 und Zahl mit 1 kodiert.\ Empirische Verteilungsfunktion:

$$F_X(x) = \begin{cases} 0, & \text{falls } x < 0, \\ 0.5, & \text{falls } 0 \leq x < 1 \end{cases}$$

Wahrscheinlichkeitsfunktion:

$$
\begin{aligned}
P(\emptyset) &= 0 \\
P(\{\text{Kopf}\}) &= 0.5 \\
P(\{\text{Zahl}\}) &= 0.5 \\
P(\Omega) &= 1
\end{aligned}
$$

Sei im Folgenden Kopf mit 0 und Zahl mit 1 kodiert.\ Empirische Verteilungsfunktion:

$$
F_X(x) = \begin{cases} 0, & \text{falls } x < 0, \\ 0.5, & \text{falls } 0 \leq x < 1, \\ 1, & \text{falls } x \geq 1. \end{cases}
$$

Empirische Dichtefunktion

$$
f_X(x) = \begin{cases} 0.5, & \text{falls } x \in \{0, 1\}, \\ 0, & \text{sonst.} \end{cases}
$$

c)

Erwartungswert

$$
\mathrm{E}(X) = f_X(0) \cdot 0 + f_x(1) \cdot 1 = 0.5.
$$

Varianz

$$
\mathrm{Var}(X) = (0 - 0.5)^2 \cdot f_X(0) + (1 - 0.5)^2 \cdot f_X(1) = 0.5^2 \cdot 0.5 + 0.5^2 \cdot 0.5 = 0.25
$$

Die Zufallsvariable ist Bernoulli-verteilt mit Parameter $p = 0.5$: $X \sim B(0.5)$. Der Erwartungswert beträgt dementsprechend p, die Varianz $p \cdot (1 - p)$.

d)

d)

Erwartungswert der Summe

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} E(X) = n \cdot E(X_i) = 0.5 \cdot n = \frac{n}{2}$$

Varianz der Summe

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \sum_{i=1}^{n} \mathrm{Var}(X) = n \cdot \mathrm{Var}(X_i) = 0.25 \cdot n = \frac{n}{4}$$

Die erste Gleichheit gilt dabei nur, weil die Zufallsvariablen X_i paarweise unkorreliert sind. Die Summe ist Binomialverteilt. Die Bernoulli-Verteilung ist ein Spezialfall der Binomialverteilung mit $n = 1$.

Aufgabe 2

a)

```
measures <- list(mean, var, median, min, max)
measure.names <- c("mean", "var", "median", "min", "max")
res1 <- sapply(measures, function(measure) apply(iris[, 1:4], 2, measure))
colnames(res1) <- measure.names
res1
```

```
##                  mean       var median min max
## Sepal.Length 5.843333 0.6856935   5.80 4.3 7.9
## Sepal.Width  3.057333 0.1899794   3.00 2.0 4.4
## Petal.Length 3.758000 3.1162779   4.35 1.0 6.9
## Petal.Width  1.199333 0.5810063   1.30 0.1 2.5
```

```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
```

Aufgabe 4

a)

$$
\begin{aligned}
P(\|X\| < 1) =& P(\sqrt{(X_1^2 + X_2^2)} < 1) \\
=& P(X_1^2 + X_2^2 < 1) \\
=& P(X_1^2 < 1 - X_2^2) \\
=& P(-\sqrt{1 - X_2^2} < X_1 < \sqrt{1 - X_2^2}) \\
=& \int_{-1}^{1} \int_{-\sqrt{1-x_2^2}}^{\sqrt{1-x_2^2}} \frac{1}{4} dx_1 dx_2 \\
=& \int_{-1}^{1} \left[\frac{x}{4}\right]_{-\sqrt{1-x_2^2}}^{\sqrt{1-x_2^2}} dx_1 \\
=& \int_{-1}^{1} \frac{2}{4} \sqrt{1 - x_2^2} dx_2 \\
=& \frac{1}{2} \int_{-1}^{1} \sqrt{1 - x_2^2} dx_2 \\
=& \int_{0}^{1} \sqrt{1 - x_2^2} dx_2 \text{ (Tipp)} \\
=& \left[\frac{\arcsin(x)}{2} + \frac{x\sqrt{1-x^2}}{2}\right]_0^1 \\
=& \frac{\pi}{4}
\end{aligned}
$$

b)

Aus a) folgt:

$$\pi = 4 \cdot P(\|X\| < 1)$$

mit dem angegebenen erwartungstreuen Schätzer ergibt sich daraus:

$$\hat{\pi} = 4 \cdot h_n(\mathbf{x})$$

$$
\begin{aligned}
&= \int_0 \sqrt{1 - x_2^2}\, dx_2 \text{ (Tipp)} \\
&= \left[\frac{\arcsin(x)}{2} + \frac{x\sqrt{1-x^2}}{2}\right]_0^1 \\
&= \frac{\pi}{4}
\end{aligned}
$$

b)

Aus a) folgt:

$$\pi = 4 \cdot P(\|X\| < 1)$$

mit dem angegebenen erwartungstreuen Schätzer ergibt sich daraus:

$$\hat{\pi} = 4 \cdot h_n(\mathbf{x})$$

Der Schätzer $\hat{\pi}$ ist erwartungstreu, da:

$$
\begin{aligned}
E(\hat{\pi}) &= E(4 \cdot h_n(\mathbf{x})) \\
&= 4 \cdot \frac{\pi}{4} \text{(weil } h_n(\mathbf{x}) \text{ erwartungstreu ist)} \\
&= \pi
\end{aligned}
$$

c)

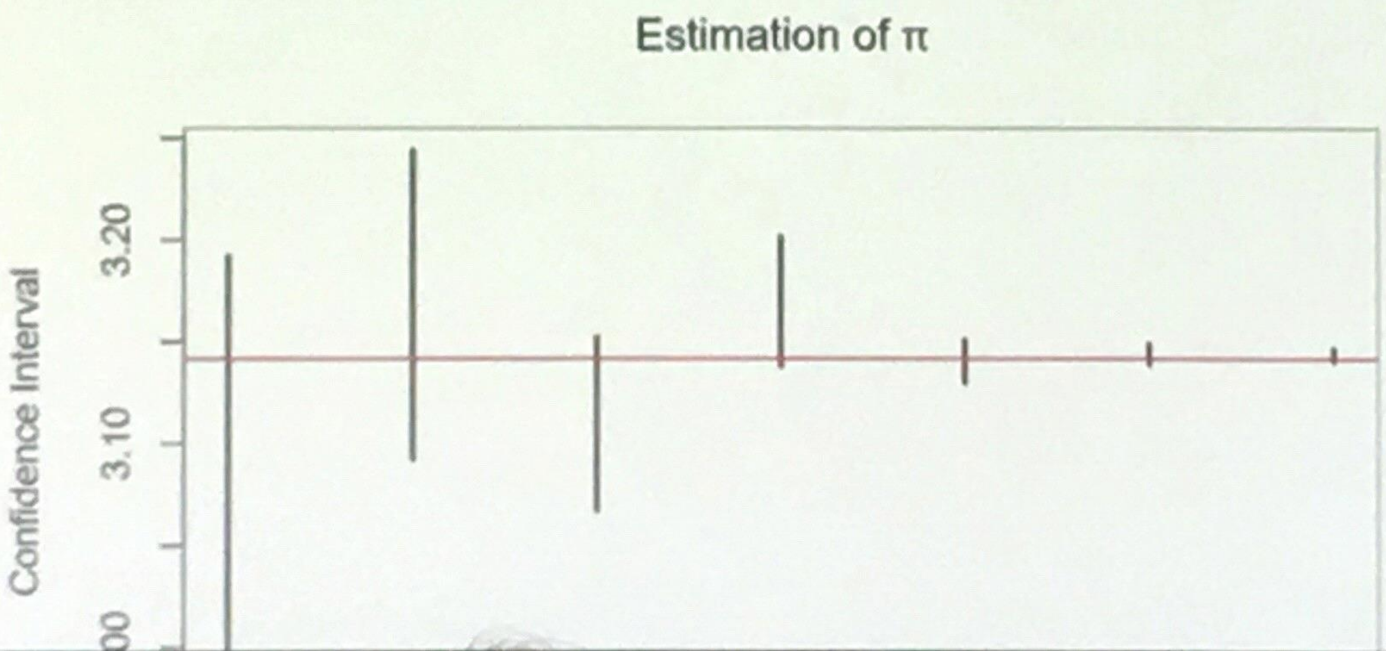
```
piHat <- function(n){
  4 * mean(apply(matrix(runif(2 * n, -1, 1), 2), 2, crossprod) < 1)
}
```

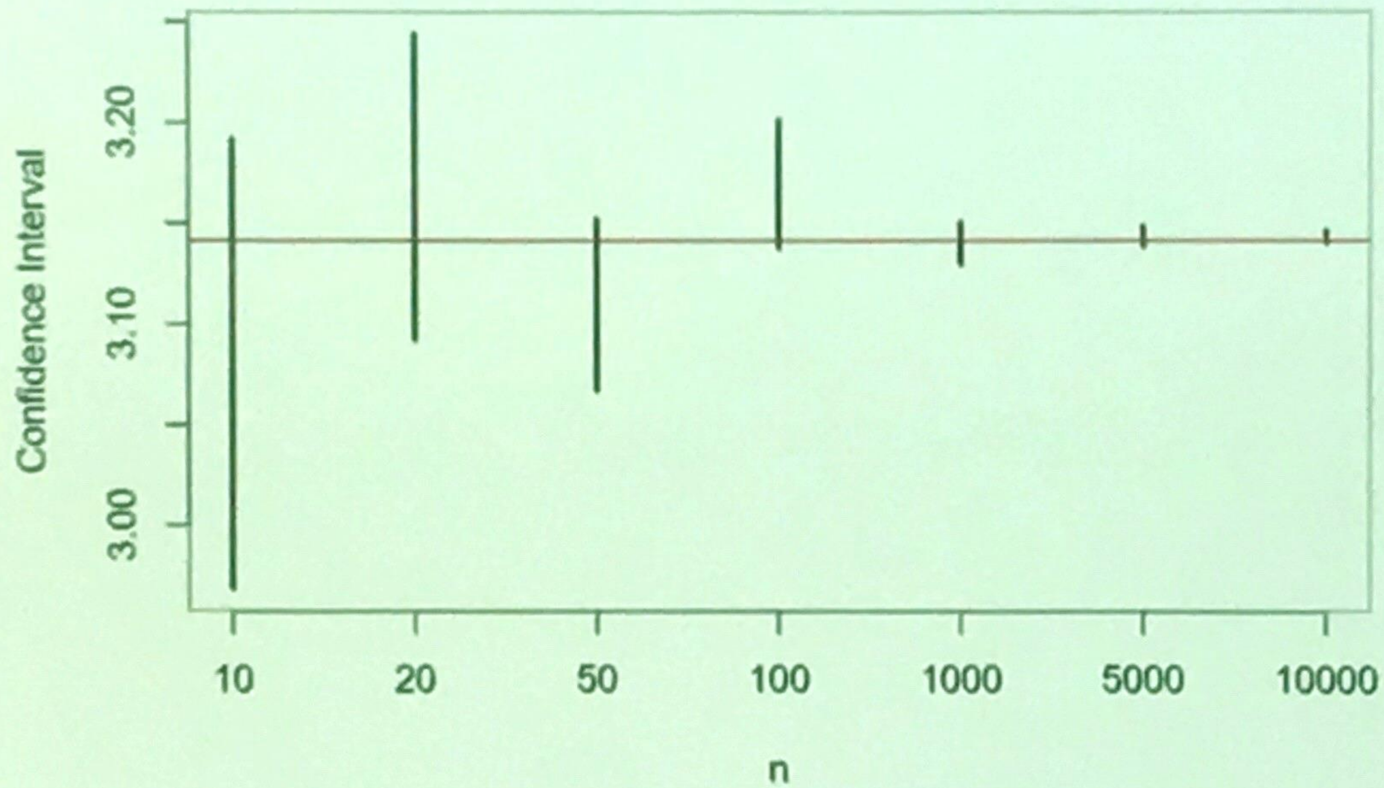
```
set.seed(1273)
```

```
ns <- c(10, 20, 50, 100, 1000, 5000, 10000)
res <- sapply(ns,
  function(n) {
    tmp <- replicate(100, piHat(n))
    return(t.test(tmp)$conf.int)
  }
)
```

Visualisierung:

```
plot(c(1, 1), res[, 1], xlim = c(1, 7), ylim = range(res), type = "l",
  xlab = "n", ylab = "Confidence Interval", xaxt = "n", lwd = 2,
  main = expression(paste("Estimation of ", pi)))
axis(1, at = 1:7, labels = ns)
for (i in 2:7)
  lines(c(i, i), res[, i], type = "l", lwd = 2)
abline(h = pi, col = "red")
```





Die Schätzung scheint Erwartungstreu zu sein (Rote Linie in jedem KI), KIs werden mit steigendem n immer kleiner. Achtung: n ist die Anzahl Zufallszahlen zur Bestimmung von π, nicht die Replikationszahl.

Es zeigt sich wie erwartet: Im Mittelwert wird π ungefähr getroffen, mit steigendem n nimmt die Varianz ab.