Übungen zur Induktiven Statistik

Übung 2 – Punkt- und Intervallschätzer

Wir beobachten den Gewinn pro Stunde von Eiscafés in NRW. Es ist davon auszugehen, dass diese Kennzahl normalverteilt ist.
Es liegt die folgende Stichprobe vor:
Stichprobe 1: $x_j = \{5.22, -5.16, 4.02, -3.25,\ 4.77, -1.46\}$

a.) Erklären Sie in jeweils ein bis zwei Sätzen, welche Vor- und Nachteile die Betrachtung von Stichproben im Vergleich zur Betrachtung der Grundgesamtheit existieren.
b.) Geben Sie den Stichprobenumfang der obigen Stichprobe an.
c.) Geben Sie an, welchen Erwartungswert für die Grundgesamtheit Sie basierend auf der obigen Stichprobe schätzen.
d.) Welches Problem ergibt sich aus Ihrer in c.) verwendeten Methode?
e.) Geben Sie an, welche Varianz und welche Standardabweichung (bzw. Standardfehler) Sie für die Grundgesamtheit basierend auf der obigen Stichprobe schätzen.
Hinweis: Verwenden Sie hierfür die Formeln ohne „minus 1“ im Nenner.
f.) Welches Problem ergibt sich aus Ihrer in e.) verwendeten Methode zusätzlich zu dem in d.) genannten Problem?
g.) Ermitteln Sie basierend auf der obigen Stichprobe ein 95%-Konfidenzintervall für den Erwartungswert der Grundgesamtheit.
h.) Interpretieren Sie Ihr Ergebnis aus g.)
i.) Es ist Ihnen nun bekannt, dass die Standardabweichung in der Grundgesamtheit 5 beträgt. Geben Sie an, welche Länge das auf der obigen Stichprobe basierende 95%-Konfidenzintervall des Erwartungswerts der Grundgesamtheit haben wird.
j.) Gehen Sie davon aus, dass die Standardabweichung in der Grundgesamtheit 5 beträgt. Ermitteln Sie das 95%-Konfidenzintervall des Erwartungswerts der Grundgesamtheit basierend auf der obigen Stichprobe.
k.) Wiederholen Sie den Vorgang aus j.) für die folgenden beiden Stichproben.
Stichprobe 2: $x_j = \{4.46, -6.59, 4.45, -6.49, 1.14, 7.97\}$
Stichprobe 3: $x_j = \{-2.5, 13.4, -6.41, 3.42, -1.12, -1.29$

Vergleichen Sie die Länge Ihrer 4 Intervalle.
Ist dieses Ergebnis so zu erwarten?
l.) Wir gehen wieder von einer Standardabweichung von 5 aus.
Wir nehmen an, Sie ziehen nicht nur 3 Stichproben, sondern *sehr* viele (mehr als 10.000). Sie ermitteln jedes Mal das 95%-Konfidenzintervall für den Erwartungswert der Grundgesamtheit. Welche Beziehung besteht zwischen den Intervallen und dem

tatsächlichen Wert von μ?
Welche Beziehung besteht zwischen dem arithmetischen Mittel der Stichprobenwerte und μ? Welche Varianz hat das arithmetische Mittel der Stichprobenwerte?

Für Interessierte: Die Stichproben wurden mit Excel generiert. Es wurde eine normalverteilte Grundgesamtheit mit $\mu = 2$ und $\sigma = 5$ verwendet.

Die Musterlösungen befinden sich am Ende des Dokuments.

Übung 2 – Punkt- und Intervallschätzer - Musterlösung

Wir beobachten den Gewinn pro Stunde von Eiscafés in NRW. Es ist davon auszugehen, dass diese Kennzahl normalverteilt ist.
Es liegt die folgende Stichprobe vor:
Stichprobe 1: $x_j = \{5.22, -5.16, 4.02, -3.25, \ 4.77, -1.46\}$

a.) Erklären Sie in jeweils ein bis zwei Sätzen, welche Vor- und Nachteile die Betrachtung von Stichproben im Vergleich zur Betrachtung der Grundgesamtheit existieren.
Vorteile: Da Befragungen, Erhebungen oder Auskünfte in der Regel Kosten verursachen, sind Stichproben in der Regel kostengünstiger als die Betrachtung der Grundgesamtheit. In vielen Fällen ist die Betrachtung der Grundgesamtheit nicht möglich.
Nachteile: Da der Stichprobenumfang kleiner ist als der Umfang der Grundgesamtheit müssen wir Hochrechnungen durchführen. Dabei können Schätzfehler auftreten.

b.) Geben Sie den Stichprobenumfang der obigen Stichprobe an.
n=6

c.) Geben Sie an, welchen Erwartungswert für die Grundgesamtheit Sie basierend auf der obigen Stichprobe schätzen.

$$\boldsymbol{\hat{\mu} = Y = \frac{1}{6}(5.22 - 5.16 + 4.02 - 3.25 + 4.77 - 1.46) = 0,69}$$

d.) Welches Problem ergibt sich aus Ihrer in c.) verwendeten Methode?
Der Stichprobenumfang ist gering. Es ist anzunehmen, dass der nächste beobachtete Wert ungleich 0,69 ist. Hätten wir den Stichprobenumfang um eins erhöht, würde sich unser prognostizierter Erwartungswert ändern. Es ist daher zu empfehlen einen Intervallschätzer statt eines Punktschätzers zu verwenden.

e.) Geben Sie an, welche Varianz und welche Standardabweichung (bzw. Standardfehler) Sie für die Grundgesamtheit basierend auf der obigen Stichprobe schätzen.
Hinweis: Verwenden Sie hierfür die Formeln ohne „minus 1" im Nenner.

$$\boldsymbol{\hat{\sigma} = \sqrt{\frac{1}{6} * ((5.22 - 0,69)^2 + (-5.12 - 0,69)^2 + \dots + (-1.46 - 0,69)^2)} \approx 4.14}$$

f.) Welches Problem ergibt sich aus Ihrer in e.) verwendeten Methode zusätzlich zu dem in d.) genannten Problem?
Zusätzlich zu dem grundsätzlichen Problem eines Punktschätzers kommt hier noch hinzu, dass der Schätzer nicht erwartungstreu ist. Es gibt einen relativen Bias, da

$\mathbb{E}[\widehat{\sigma}^2] = \frac{n-1}{n}\sigma^2$ **und somit** $\mathbb{E}[\widehat{\sigma}] = \sqrt{\frac{n-1}{n}}\sigma \neq \sigma$ **gilt.**

Dieses Problem lässt sich umgehen, wenn wir anstelle von $\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i - y]^2}$ **den folgenden unbiased estimator verwenden:**

$$\widehat{\sigma}_{unb} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[x_i - y]^2} = \sqrt{\frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}[x_i - y]^2} = \sqrt{\frac{n}{n-1}\widehat{\sigma}^2} = \sqrt{\frac{n}{n-1}}\widehat{\sigma} \approx 4.54$$

Denn es gilt $\mathbb{E}[\widehat{\sigma}_{unb}] = \mathbb{E}\left[\sqrt{\frac{n}{n-1}}\widehat{\sigma}\right] = \sqrt{\frac{n}{n-1}}\mathbb{E}[\widehat{\sigma}] = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n-1}{n}}\sigma = \sigma.$

g.) Ermitteln Sie basierend auf der obigen Stichprobe ein 95%-Konfidenzintervall für den Erwartungswert der Grundgesamtheit.

$$\left[Y - t_{n-1,1-\frac{\alpha}{2}} * \frac{\widehat{\sigma}}{\sqrt{n-1}};\ Y + t_{n-1,1-\frac{\alpha}{2}} * \frac{\widehat{\sigma}}{\sqrt{n-1}}\right]$$
$$\approx \left[0.69 - 2.5706 * \frac{4.14}{\sqrt{5}};\ 0.69 + 2.5706 * \frac{4.14}{\sqrt{5}}\right]$$
$$\approx [-4.07;\ 5.45]$$

Alternativ:

$$\left[Y - t_{n-1,1-\frac{\alpha}{2}} * \frac{\widehat{\sigma}_{unb}}{\sqrt{n}};\ Y + t_{n-1,1-\frac{\alpha}{2}} * \frac{\widehat{\sigma}_{unb}}{\sqrt{n}}\right]$$
$$\approx \left[0.69 - 2.5706 * \frac{4.54}{\sqrt{6}};\ 0.69 + 2.5706 * \frac{4.54}{\sqrt{6}}\right]$$
$$\approx [-4.07; 5.45]$$

h.) Interpretieren Sie Ihr Ergebnis aus g.)

Würden wir sehr viele Stichproben mit einem jeweiligen Umfang von n=6 generieren und zu jeder Stichprobe ein Konfidenzintervall mithilfe des oben angewandten Algorithmus ermitteln, dann würden 95% dieser Intervalle den wahren Wert von μ beinhalten.
Hierbei ist wichtig: μ ist eine fixe, aber für uns unbekannte Zahl. Sie ist eine deterministische Größe.
Die Stichprobenwerte und folglich auch die Intervallgrenzen sind jedoch nicht exakt prognostizierbar. Daher sprechen hier von stochastischen Größen.

i.) Es ist Ihnen nun bekannt, dass die Standardabweichung in der Grundgesamtheit 5 beträgt. Geben Sie an, welche Länge das auf der obigen Stichprobe basierende 95%-Konfidenzintervall des Erwartungswerts der Grundgesamtheit haben wird.

Bei bekannter Standardabweichung gilt: $L = 2u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} = 2u_{0.975} * \frac{5}{\sqrt{6}}$
$= 2 * 1.96 * \frac{5}{\sqrt{6}} \approx 8$. **Die Länge (bzw. Breite) des Intervalls beträgt etwa 8.**

j.) Gehen Sie davon aus, dass die Standardabweichung in der Grundgesamtheit 5 beträgt. Ermitteln Sie das 95%-Konfidenzintervall des Erwartungswerts der Grundgesamtheit basierend auf der obigen Stichprobe.

$$\left[Y - u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}};\ Y + u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right] = \left[0.69 - 1.96 * \frac{5}{\sqrt{6}};\ 0.69 + 1.96 * \frac{5}{\sqrt{6}}\right]$$
$$[-3.31; 4.69]$$

k.) Wiederholen Sie den Vorgang aus j.) für die folgenden beiden Stichproben.

Stichprobe 2: $x_j = \{4.46, -6.59, 4.45, -6.49, 1.14, 7.97\}$
Stichprobe 3: $x_j = \{-2.5, 13.4, -6.41, 3.42, -1.12, -1.29$

Für Stichprobe 2 gilt:

$$\left[Y - u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}};\ Y + u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right] = \left[0.82\overline{3} - 1.96 * \frac{5}{\sqrt{6}};\ 0.82\overline{3} + 1.96 * \frac{5}{\sqrt{6}}\right]$$
$$\approx [-3.18; 4.82]$$

Für Stichprobe 2 gilt:

$$\left[Y - u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}};\ Y + u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right] = \left[0.91\overline{6} - 1.96 * \frac{5}{\sqrt{6}};\ 0.91\overline{6} + 1.96 * \frac{5}{\sqrt{6}}\right]$$
$$\approx [-3.08; 4.92]$$

Vergleichen Sie die Länge Ihrer 4 Intervalle.

Die Länge der Konfidenzintervalle auf Basis der bekannten Varianz (bzw. Standardabw.) betragen jeweils 8.
Die Länge des Konfidenzintervalls auf Basis der unbekannten Varianz (bzw. Standardabw.) beträgt$5.45 - (-4.07) = 9.52$.

Ist dieses Ergebnis so zu erwarten?

Ja, es ist meist so, dass die Länge des Intervalls auf Basis der unbekannten Varianz größer ist als die Länge des Intervalls auf Basis der bekannten Varianz. Der Grund dafür ist, dass $t_{n-1,1-\frac{\alpha}{2}} > u_{1-\frac{\alpha}{2}}$ **(1) gilt und dass** $\mathbb{E}\left[\frac{\hat{\sigma}}{\sqrt{n-1}}\right] = \mathbb{E}\left[\frac{\hat{\sigma}_{unb}}{\sqrt{n}}\right] = \frac{\sigma}{\sqrt{n}}$ **gilt.**

Es gilt also: $\mathbb{E}\left[2 * u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right] = 2 * u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$ **(2.)**

sowie $\mathbb{E}\left[2 * t_{n-1,1-\frac{\alpha}{2}} * \frac{\hat{\sigma}}{\sqrt{n-1}}\right] = 2 * t_{n-1,1-\frac{\alpha}{2}} * \mathbb{E}\left[\frac{\hat{\sigma}}{\sqrt{n-1}}\right] = 2 * t_{n-1,1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$ **(3.)**

Wegen (1.) ist der Term (3.) größer als Term (2.)

l.) Wir gehen wieder von einer Standardabweichung von 5 aus.
Wir nehmen an, Sie ziehen nicht nur 3 Stichproben, sondern *sehr* viele (mehr als 10.000). Sie ermitteln jedes Mal das 95%-Konfidenzintervall für den Erwartungswert der Grundgesamtheit. Welche Beziehung besteht zwischen den Intervallen und dem tatsächlichen Wert von μ?
Etwa 95% der aufgestellten Intervalle werden den wahren (und fixen, aber für uns unbekannten) Wert von $\boldsymbol{\mu}$ beinhalten.

Welche Beziehung besteht zwischen dem arithmetischen Mittel der Stichprobenwerte und μ?
Es gilt $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}$
Das bedeutet, dass– wenn wir viele Stichproben ziehen und jedes Mal das Stichprobenmittel Y berechnen – der Durchschnitt der ermittelten Werte von Y gegen den wahren (und fixen, aber für uns unbekannten) Wert von $\boldsymbol{\mu}$ konvergiert.

Welche Varianz hat das arithmetische Mittel der Stichprobenwerte?
Es gilt $\mathbb{V}[\boldsymbol{Y}] = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}$
Einsetzen ergibt $\mathbb{V}[\boldsymbol{Y}] = \frac{\mathbf{5^2}}{\mathbf{6}} = \frac{\mathbf{25}}{\mathbf{6}} = \mathbf{4.1\overline{6}}$

Für Interessierte: Die Stichproben wurden mit Excel generiert. Es wurde eine normalverteilte Grundgesamtheit mit $\mu = 2$ und $\sigma = 5$ verwendet.