Bachelorarbeit:

Schätzen der Verteilungsfunktion und Quantilen bei einer PPS-Stichprobe

Autor: Felix Loewe

Matrkelnummer: xxxxxxxxx

Betreuer: Prof. Dr. Göran Kauermann

17. Oktober 2013

Inhaltsverzeichnis

1 Einleitung

Diese Bachelorarbeit vergleicht verschiedene Verfahren für die Schätzung von Verteilungsfunktion und Quantilen mit besonderem Augenmerk auf PPS-Stichproben. Im Rahmen dieser Arbeit wurde eine Simulation erstellt, in der fünf verschiedene Schätzverfahren an zwei Datensätzen verglichen werden. Im ersten Abschnitt werden

die theoretischen Grundlagen der verwendeten Methoden zusammengefasst. Darauf folgt eine Beschreibung des Aufbaus der Simulation und schließlich die Präsentation der Ergebnisse. Ein Fazit fasst nochmal alles zusammen und gibt Anstöße für zukünftige Aufgabenstellungen.

2 Theorie

2.1 Der Horwitz - Tompson Schätzer

Im Folgenden soll π_i die Auswahlwahrscheinlichkeit des i-ten Individuums darstellen in die Stichprobe aufgenommen zu werden. Dabei muss für festen Stichprobenumfang n und Größe der Grundgesamtheit N gelten: $\sum_{i=1}^{N} \pi_i = n$. Des Weiteren sollen die Auswahlwahrscheinlichkeiten zweiter Ordnung durch $\pi_{i,j}$ angegeben werden. Sie bezeichnen die Wahrscheinlichkeit, dass die beiden Elemente i und j gleichzeitig in die Stichprobe aufgenommen werden. Allgemein gilt ein Stichprobendesign als effektiver als ein anderes, wenn die Varianz der Schätzung mit dem ersten Design kleiner ist als die des zweiten, vorausgesetzt der zu betreibende Aufwand ist für beide gleich. Nun kann die Stichprobenvarianz durch geeignete Wahl der Auswahlwahrscheinlichkeiten verkleinert und die Stichprobe dadurch effektiver gemacht werden. Um dies berücksichtigen zu können, schlagen Horvitz und Thompson den inzwischen nach ihnen benannten Horvitz-Thompson-Schätzer vor: (Horvitz and Thompson, 1952)

$$\widehat{\overline{Y}}_{HT} = \frac{1}{N} \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

Dieser Schätzer ist erwartungstreu und für alle Stichprobendesigns anwendbar, bei denen ohne Zurücklegen gezogen wird. Seine Grundidee ist, dass durch die Gewichtung mit der inversen Auswahlwahrscheinlichkeit eine Verzerrung des Schätzers vermieden wird. Wählt man ein Stichprobendesign, in dem alle Elemente die gleiche Auswahlwahrscheinlichkeit $\pi_i = n/N$ besitzen, so ergibt sich als Schätzer das einfache arithmetische Mittel. Um die Varianz zu minimieren, wählt man nun die Auswahlwahrscheinlichkeiten perfekt proportional zu den Werten der interessierenden Größe

Y:

$$\pi_i = \frac{n}{N}\frac{Y_i}{\overline{Y}}$$

Dann ergibt sich:

$$\widehat{\overline{Y}}_{HT} = \frac{1}{N}\sum_{k=1}^{n}\frac{y_k}{\pi_k} = \frac{1}{nN}\sum_{k=1}^{n}\frac{y_k}{y_k}N\overline{Y} = \overline{Y}$$

Man erhält also durch den Schätzer unabhängig von der Stichprobe das arithmetische Mittel. Die Varianz ist folglich 0. Diese Idee versucht sich das sogenannte PPS-Stichprobenverfahren (größenproportinales Stichprobenverfahren) zunutze zu machen. (Kauermann and Küchenhoff, 2011). (siehe Abschnitt 2.2)

Allgemein lässt sich die Varianz des Horvitz-Thompson Schätzers wie folgt beschreiben:

$$Var(\widehat{\overline{Y}}_{HT}) = \frac{1}{N^2}\left[\sum_{i=1}^{N}\frac{\pi_i(1-\pi_i)}{\pi_i^2}Y_i^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}Y_iY_j\right]$$

und ihr Schätzer lautet:

$$\widehat{Var(\widehat{\overline{Y}}_{HT})} = \frac{1}{N^2}\left[\sum_{k=1}^{n}\frac{\pi_i(1-\pi_i)}{\pi_i^2}y_k^2 + \sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\pi_{k,l}-\pi_k\pi_l}{\pi_k\pi_l}y_ky_l\right]$$

Dieser Schätzer ist zwar erwartungstreu, allerdings garantiert er keine positiven Ergebnisse (Kauermann and Küchenhoff, 2011). Deshalb schlagen Yates and Grundy (1953) folgende Umformung vor:

$$Var_{YG}(\widehat{\overline{Y}}_{HT}) = \frac{1}{N^2}\cdot\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}(\pi_i\pi_j-\pi_{i,j})(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j})^2$$

Wie man an dieser Formel gut sehen kann, wird $Var_{YG}(\widehat{\overline{Y}}_{HT}) = 0$, falls π proportional zu Y gewählt wird, da sich dann $\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} = 0$ ergibt.

Der Schätzer für diese Varianzformel lautet:

$$\widehat{Var_{YG}(\widehat{\overline{Y}}_{HT})} = \frac{1}{N^2} \cdot \frac{1}{2} \sum_{k=1}^{n} \sum_{l=1, l \neq k}^{n} \frac{\pi_k \pi_l - \pi_{k,l}}{\pi_{k,l}} (\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l})^2$$

Für Herleitungen und Genaueres verweise ich auf Kauermann and Küchenhoff (2011), die erklären, dass mehrere Simulationsstudien die Überlegenheit der Varianz-Formel von Yates und Grundy gegenüber der von Horvitz und Thompson belegen.

2.2 Die PPS-Stichprobe

Eine spezielle Art Stichproben zu ziehen ist die sogenannte PPS-Stichprobe. Die Abkürzung „PPS“ steht für **P**robability **P**roportional to **S**ize. Anders als bei einer einfachen Stichprobe haben hier nicht alle Elemente der Grundgesamtheit die gleiche Wahrscheinlichkeit in die Stichprobe zu gelangen. Wie der Name vermuten lässt, werden die Auswahlwahrscheinlichkeiten der Elemente so festgelegt, dass sie proportional zur Größe ihres interessierenden Merkmals sind. Je größer also die Ausprägung des interessierenden Merkmals, desto größer ist die Auswahlwahrscheinlichkeit des dazugehörigen Elements der Grundgesamtheit.

Wie in Abschnitt 2.1 gezeigt, lässt sich die Varianz des Horvitz-Thompson Schätzers durch dieses Verfahren auf 0 reduzieren. In der Realität ist aber das interessierende Merkmal natürlich nicht im Voraus bekannt. Dennoch lernen wir, dass die Varianz des Schätzers durch proportionale Wahl der Wahrscheinlichkeiten minimiert wird. In der Regel verwendet man eine bekannte Hilfsgröße, die möglichst proportional zur interessierenden Größe vermutet wird (z.B.: frühere Messungen des gleichen Merkmals), um die Auswahlwahrscheinlichkeiten zu bestimmen. Dieses Hilfsmerkmal nennen wir im Folgenden Z. Nun ergeben sich die zu Z proportionalen Auswahlwahrscheinlichkeiten als

$$\pi_i = n \frac{Z_i}{\sum_{j=1}^{N} Z_j}.$$

Die Mittelwertschätzung erfolgt durch:

$$\widehat{\overline{Y}}_{PPS} = \widehat{\overline{Y}}_{HT} = \frac{1}{N} \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \frac{\sum_{j=1}^{N} Z_j}{N} \frac{1}{n} \sum_{k=1}^{n} \frac{y_k}{z_k} = \overline{Z} \frac{1}{n} \sum_{k=1}^{n} \frac{y_k}{z_k}$$

Der Mittelwertschätzer der einfachen Stichprobe lautet $\widehat{\overline{Y}}_{ES} = \frac{1}{n}\sum_{k=1}^{n} y_k$. Entscheidend für die Streuung des Schätzers ist jeweils der Wert in der Summe. Die Varianz des Schätzers der PPS-Stichprobe ist also kleiner als die der einfachen Stichprobe, falls die Streuung von Y_i größer ist als die von $\frac{Y_i}{Z_i}$. Für diesen Fall ist dann die PPS-Stichprobe der einfachen Stichprobe vorzuziehen. Die Schätzung der Varianz des Mittelwertschätzers $\widehat{\overline{Y}}_{PPS}$ erfolgt durch Yates und Grundy wie in Abschnitt 2.1 beschrieben.

2.3 Schätzung der kommulativen Verteilungsfunktion

Jede Funktion F, die die folgenden drei Eigenschaften erfüllt, ist eine Verteilungsfunktion:

1. F ist monoton wachsend.
2. F ist rechtsseitig stetig
3. $\lim_{x\to-\infty} F(x) = 0$,

Sei F die Verteilungsfunktion von Y. Dann bezeichnet $F(x)$ den Anteil der Elemente in der Grundgesamtheit, deren Wert des interessierenden Merkmals kleiner oder gleich x ist (Fahrmeir et al., 2007). Bei einer einfachen Stichprobe lässt sich die Verteilungsfunktion schätzen durch:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{y_i \leq x\}}$$

$I_{\{A\}}$ ist hierbei die Indikatorfunktion für das Ereignis A.

Dieser Schätzer ist bei Stichproben mit unterschiedlichen Auswahlwahrscheinlichkeiten aber nicht anwendbar. Ein verbreiteter Schätzer für die Verteilungsfunktion bei unterschiedlichen Auswahlwahrscheinlichkeiten lautet:

$$\hat{F}_{HT}(x) = \frac{1}{N}\sum_{i=1}^{n} \frac{I_{\{y_i \leq x\}}}{\pi_i}$$

Allerdings hat dieser Schätzer einen entscheidenden Nachteil: Wenn $\sum_{i=1}^{n} \frac{1}{\pi_i} \neq N$ gilt, ergibt sich keine Wahrscheinlichkeitsfunktion, da $\lim_{x \to \infty} F(x) \neq 1$. Dies ist aber zwingend erforderlich, wenn man mit Hilfe der Verteilungsfunktion später ein Quantil errechnen will (siehe Abschnitt 2.4). Deshalb skalieren wir den Schätzer mit seinem größten Wert und gehen somit sicher, dass der gesamte Wertebereich von 0 bis 1 abgedeckt wird:

$$\hat{F}_S(x) = \frac{1}{N} \sum_{i=1}^{n} \frac{I_{\{y_i \leq x\}}}{\pi_i} / \sum_{i=1}^{n} \pi_i$$

$\hat{F}_S(x)$ ist auf jeden Fall eine Verteilungsfunktion und soll uns deswegen im Folgenden zur Quantil-Schätzung dienen (Kuk, 1988).

2.4 Schätzung der Quantilen

„Jeder Wert x_p mit $0 < p < 1$, für den mindestens ein Anteil p der Daten kleiner/gleich x_p und mindestens ein Anteil $1-p$ größer/gleich x_p ist, heißt p-Quantil.“(Fahrmeir et al., 2007) Formalisiert dargestellt kann man nach Fahrmeir et al. (2007) schreiben:

$$\frac{\text{Anzahl } (x\text{-Werte} \leq x_p)}{n} \geq p \text{ und } \frac{\text{Anzahl } (x\text{-Werte} \geq x_p)}{n} \leq p$$

Das Quantil beatwortet also im Vergleich zur Verteilungsfunktion die umgekehrte Fragestellung. Hier interessiert nicht, welcher Anteil der Daten unter einem bestimmten Wert liegt, sondern unter welchem Wert ein bestimmter Anteil der Daten liegt. Für uns bedeutet das, dass wir die Quantilsfunktion durch die invertierte Verteilungsfunktion schätzen können. $\hat{F}^{-1}(p)$ liefert das gesuchte p-Quantil.

3 Aufbau der Simulation

Ziel der Simulation ist es anhand von fiktiven Daten zu veranschaulichen, welche Schätzmethoden für die Verteilungsfunktion und Quantilen bei einer PPS-Stichprobe am besten geeignet ist. Dafür wird eine Grundgesamtheit zufällig simuliert und an-

schließend durch fünf verschiedene Vorgehensweisen die interessierenden Werte ermittelt. Die entsprechenden Programme sind im **elektronischen Anhang** zu finden.

3.1 Die Grundgesamtheit

Als Grundgesamtheit dienen log-Normalverteilte Zufallswerte Y. Das bedeutet, dass $log(Y)$ einer Normalverteilung folgt. Diese Grundgesamtheit wird als fest angesehen und im Folgenden für alle Szenarien und Simulationsdurchläufe verwendet.

3.2 Einfache Stichprobe

Im ersten Schätzverfahren wird eine einfache Stichprobe gezogen. Verteilung und Quantile der Stichprobe lassen sich hierbei als Schätzungen für die entsprechenden Größen der Grundgesamtheit verwenden. Dieses Verfahren dient als Vergleichsobjekt zu den folgenden Verfahren mit PPS-Stichprobe.

3.3 PPS-Stichprobe ohne Horwitz-Thompson-Schätzer

Aus der Grundgesamheit wird eine PPS-Stichprobe gezogen. Die Quantile der Stichprobe werden als Schätzer für die wahren Werte der Grundgesamtheit verwendet. Es ist zu erwarten, dass dieses Verfahren das wahre Quantil systematisch überschätzt, da große Werte der Grundgesamtheit eine höhere Wahrscheinlichkeit besitzen in die Stichprobe zu gelangen. Dieses Vorgehen kürzen wir im Folgenden mit „PPS_QUANT“ ab.

3.4 PPS-Stichprobe und Horwitz-Thompson-Schätzer

Hier wird aus der gegebenen Grundgesamtheit eine PPS-Stichprobe gezogen. Die Auswahlwahrscheinlichkeiten werden hierbei proportional zu den Y-Werten festgelegt. Verteilungsfunktion und Quantile werden, wie in den Kapiteln 2.3 und 2.4 beschrieben, mit Hilfe von Horwitz-Thompson geschätzt. Dieses Vorgehen kürzen wir im Folgenden mit „PPS_HT“ ab.

3.5 PPS-Stichprobe und Horwitz-Thompson-Schätzer mit unscharfer Grundgesamtheit

Wie bereits erwähnt ist es in der Realität nicht möglich, die Auswahlwahrscheinlichkeiten der PPS-Stichprobe proportional zur unbekannten interessierenden Größe Y festzulegen, da diese vor der Ziehung natürlich nicht bekannt ist. Zudem unterliegen alle Messungen in der Realität einem Messfehler. Um dies nachzustellen, belegen wir die Y Werte mit einem normalverteilten Fehlerterm ϵ. Somit erhalten wir eine Hilfsgröße Z, die gut proportional zur Grundgesamtheit Y ist. Die Auswahlwahrscheinlichkeiten werden dann proportional zur Hilfsgröße Z gewählt (siehe Abschnitt 2.2). Das restliche Vorgehen funktioniert analog zu 3.4. Dieses Vorgehen kürzen wir im Folgendem mit „PPS_HT_EPSILOIN“ ab.

3.6 Alternative Strategie

Man ordnet die Grundgesamtheit der Größe nach und zieht die $n/5$ größten Werte mit einer Wahrscheinlichkeit von 1. Die noch fehlenden $n - n/5$ Einheiten werden durch eine einfache Stichprobe mit gleicher Wahrscheinlichkeit aus der restlichen Grundgesamtheit gezogen. Die Schätzung der interessierenden Werte erfolgt auch hier wie in Kapiteln 2.3 und 2.4. Zur Vereinfachung nennen wir dieses Verfahren im Folgenden „GROßE WERTE“

4 Ergebnisse der Simulation

In diesem Teil der Arbeit werden die Ergebnisse der Simulation vorgestellt. Zuerst werde ich auf die geschätzte Verteilungsfunktion eingehen, um dann auf die Auswirkungen auf die Schätzung der Quantile zu sprechen zu kommen.

Als Grundgesamtheit verwenden wir eine logarithmierte Normalverteilung. Die Grundgesamtheit wird als fest angesehen. Allerdings wurden verschiedene Parameter der Grundgesamtheit getestet.

- Besonders interessant sind die Ergebnisse für den log-Mittelwert -7.8 und der

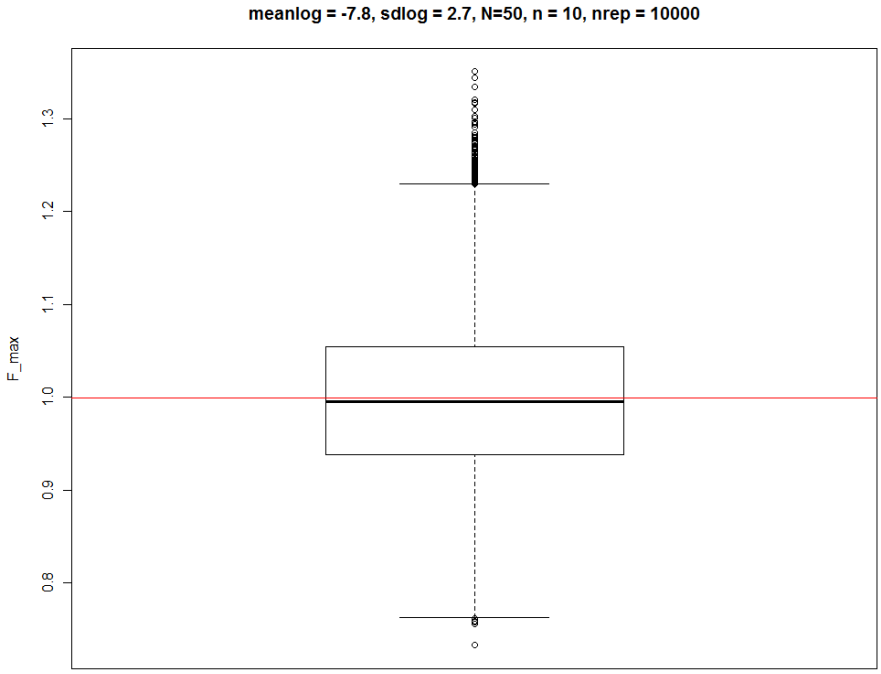

Abbildung 1: Maximum der Verteilungsfunktion, `Grundgesamtheit 1`

log-Standartabweichung 2.7, da diese aus einem realen Datenbeispiel meines Betreuers Herr Prof. Dr. Kauermann stammen. Diese nennen wir im Folgendem `Grundgesamtheit 2`.

- Als Vergleich wird ein log-Mittelwert von 0 und eine log-Standartabweichung von 0.3 betrachtet. Diese nennen wir im Folgendem `Grundgesamtheit 1`.

4.1 Schätzung der Verteilungsfunktion bei einer PPS-Stichprobe

Als ersten Versuch wählen wir `Grundgesamtheit 1` mit der Größe $N = 50$. Wie in Abschnitt 2.3 ausführlich erklärt, wollen wir den Schätzer $\hat{F}_S(x)$ statt $\hat{F}_{HT}(x)$ verwen-

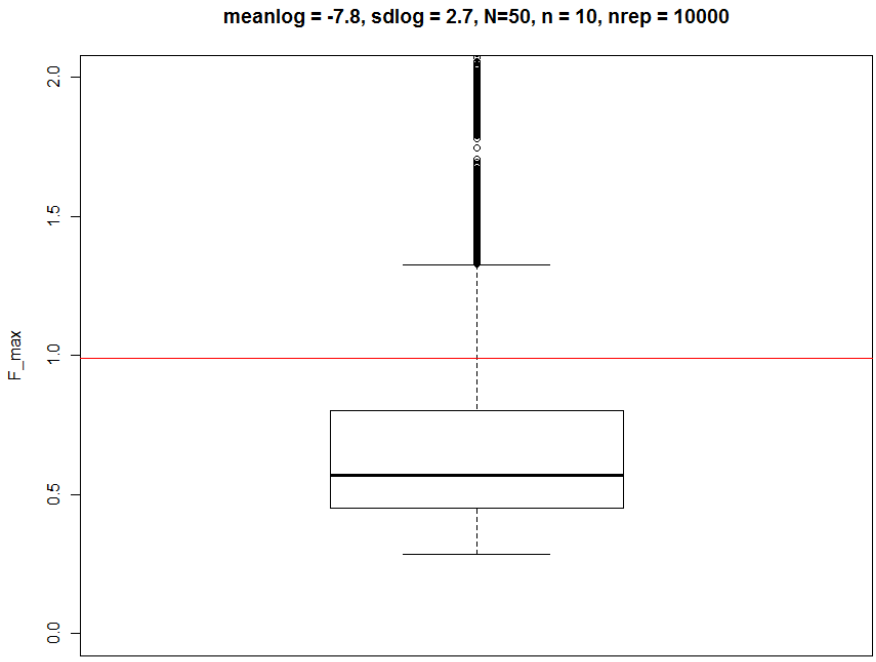

Abbildung 2: Maximum der Verteilungsfunktion, `Grundgesamtheit 2`

den, um sicherzugehen, dass der Wertebereich der geschätzten Verteilungsfunktion das später gesuchtes Quantil überhaupt überdeckt. Um das Problem zu veranschaulichen, sieht man in *Abbildung* 1 den Wert $\frac{1}{N}\sum\frac{1}{\pi}$, der dem Maximum der geschätzten Verteilungsfunktion $\hat{F}_{HT}(x)$ entspricht. Aus der festen `Grundgesamtheit 1` wurde 20000 mal eine PPS-Stichprobe der Größe 10 gezogen und der interessierende Wert errechnet. Die Ergebnisse sind in einem Boxplot dargestellt. Das als roter Strich dargestellte arithmetische Mittel liegt bei 1, das heißt der Schätzer ist erwartungstreu. Auch liegt ein Großteil der Werte nahe um 1. Einige Ausreißer nach oben und unten sind zu sehen, die bei der Schätzung der Quantile Probleme bereiten könnten.

Als nächstes wollen wir versuchen, das Ganze auf ein reales Datenbeispiel anzuwenden. Hierzu wählen wir `Grundgesamtheit 2` und führen den gleichen Vorgang

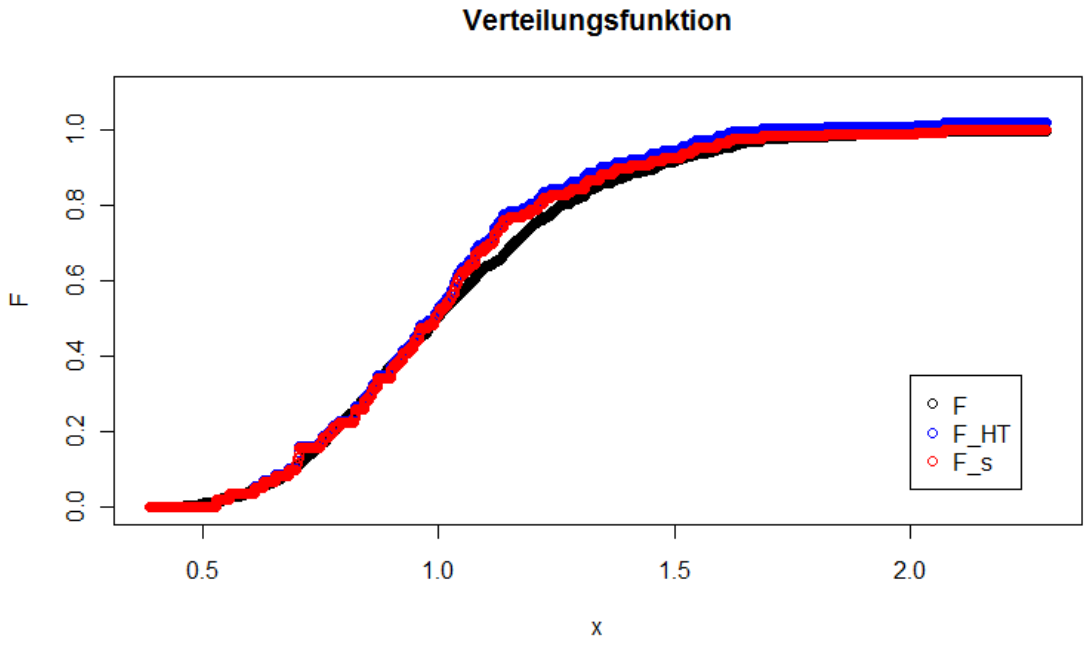

Abbildung 3: Verteilungsfunktion, `Grundgesamtheit 1`

noch einmal durch. Wir erhalten *Abbildung* 2. Auch hier liegt das als roter Strich dargestellte arithmetische Mittel bei 1. Allerdings liegt ein Großteil der Werte zwischen 0.45 und 0.8, womit die spätere Schätzung eines hohen Quantils nicht möglich wäre.

Im Folgenden verwenden wir stets eine Grundgesamtheit der Größe 1000 und eine Stichprobe der Größe 100 was ungefähr dem realen Datenbeispiel entspricht.

In *Abbildung* 3 ist die empirische Verteilungsfunktion $F(x)$ der `Grundgesamtheit 1` zu sehen. Die blaue Kurve bildet die geschätzte Verteilungsfunktion $\hat{F}_{HT}(x)$ ab, während die rote Kurve $\hat{F}_S(x)$ darstellt. Die Funktion verläuft recht gleichmäßig. Die meisten Ausprägungen liegen im mittleren Bereich der deswegen etwas steiler ist. Alle drei abgebildeten Kurven liegen nah beieinander. Die Schätzung würde also mit beiden Schätzmethoden gut gelingen.

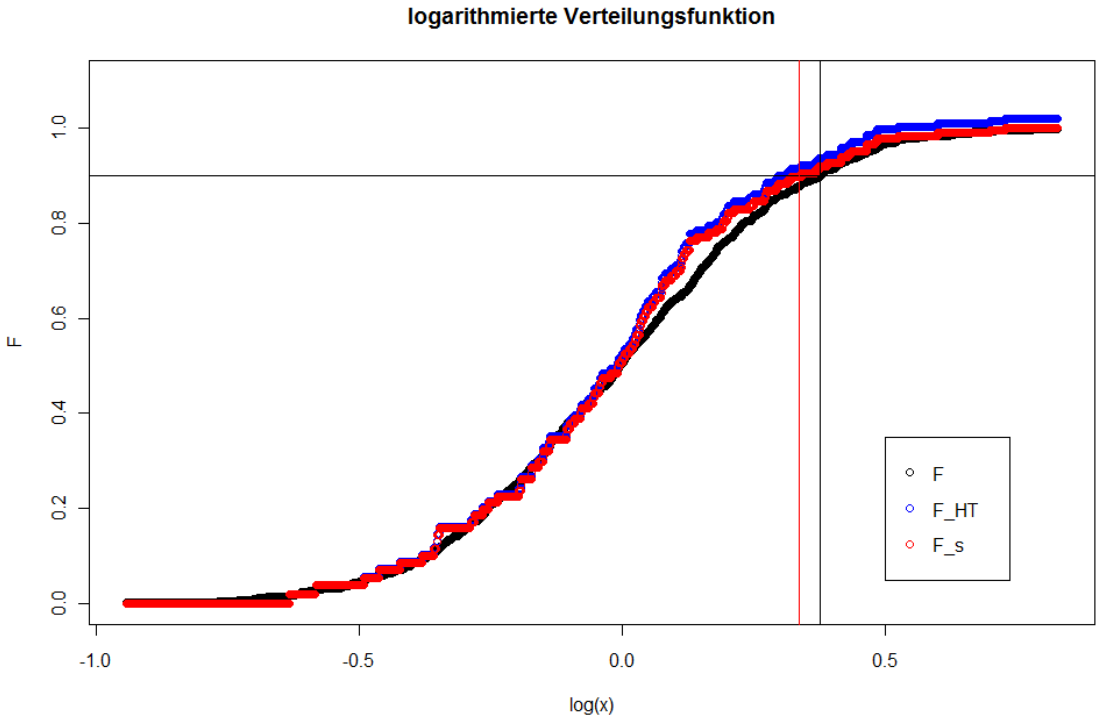

Abbildung 4: logarithmierte Verteilungsfunktion, `Grundgesamtheit 1`

Um die geschätzten Quantile besser erkennen zu können, entzerren wir die x-Achse, indem wir sie logarithmieren. Zieht man eine horizentale Linie auf der Höhe q, so ergibt der x-Wert des Schnittpunkts mit den Kurven das jeweilige q-Quantil der Verteilung. Exemplarisch ist dies hier in *Abbildung* 4 für das 0.9 Quantil geschehen. Die beiden durch die vertikalen Linien dargestellten Quantilswerte liegen trotz der Logarithmierung nahe beieinander. Das deutet darauf hin, dass auch die Schätzung der Quantile gut funktioniert.

Ein anderes Bild ergibt sich für `Grundgesamtheit 2`. Die empirische Verteilungsfunktion in *Abbildung* 5 ist sehr steil. Das bedeutet, dass es sehr viele relativ kleine Werte gibt und wenige mittlere oder große. Wieder bildet die blaue Kurve die geschätzte Verteilungsfunktion $\hat{F}_{HT}(x)$ ab, während die rote Kurve $\hat{F}_S(x)$ darstellt.

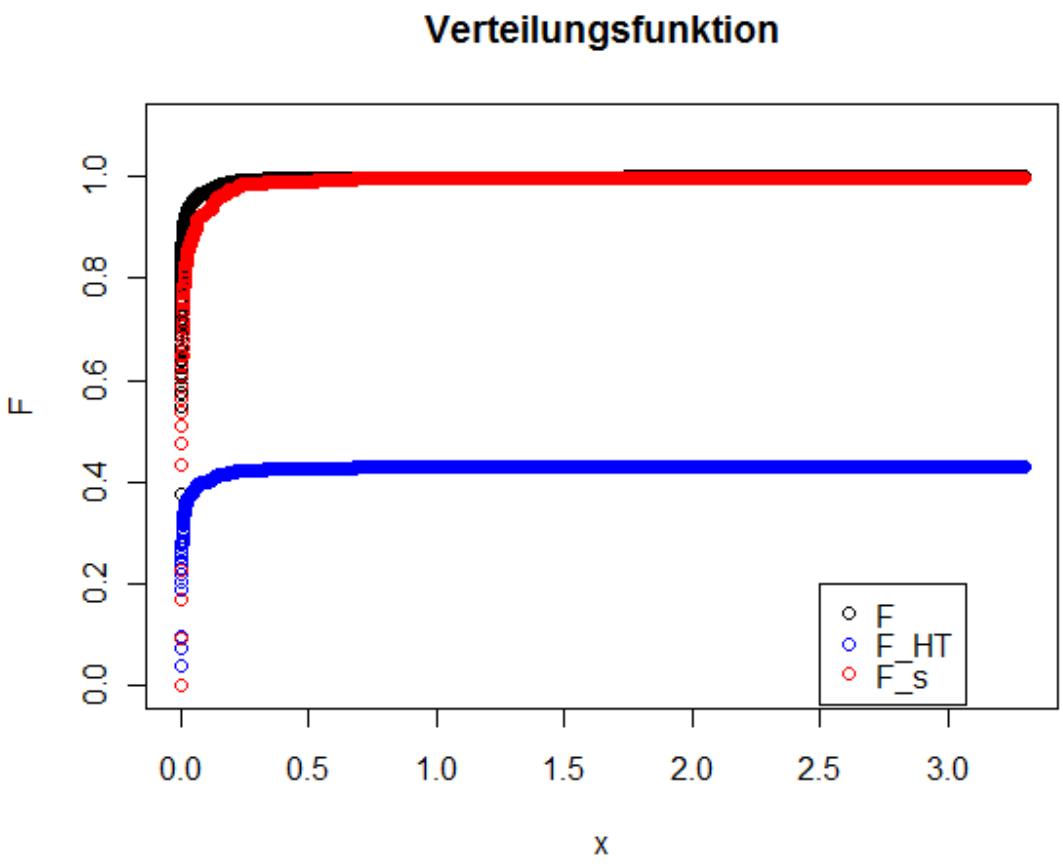

Abbildung 5: Verteilungsfunktion, `Grundgesamtheit 2`

Auf den ersten Blick fällt auf, dass $\hat{F}_{HT}(x)$ den Wert 1 nicht erreicht. Ihr höchster Wert beträgt gerade 0.43. $\hat{F}_S(x)$ hingegen füllt den Wertebereich von $[0, 1]$ aus, ist aber etwas flacher als die wahre Verteilung $F(x)$.

Deutlicher wird das Problem, wenn man wie in *Abbildung* 6 das Ganze etwas entzerrt, indem man die x-Werte logarithmiert. Die gleichen Linien wie bei `Grundgesamtheit 1` werden gezogen. Wegen der flacheren Kurve von $\hat{F}_S(x)$ ergibt sich ein höherer geschätzter Wert des Quantils (-2.8), als er in Wirklichkeit ist (-4.3). Da wie in

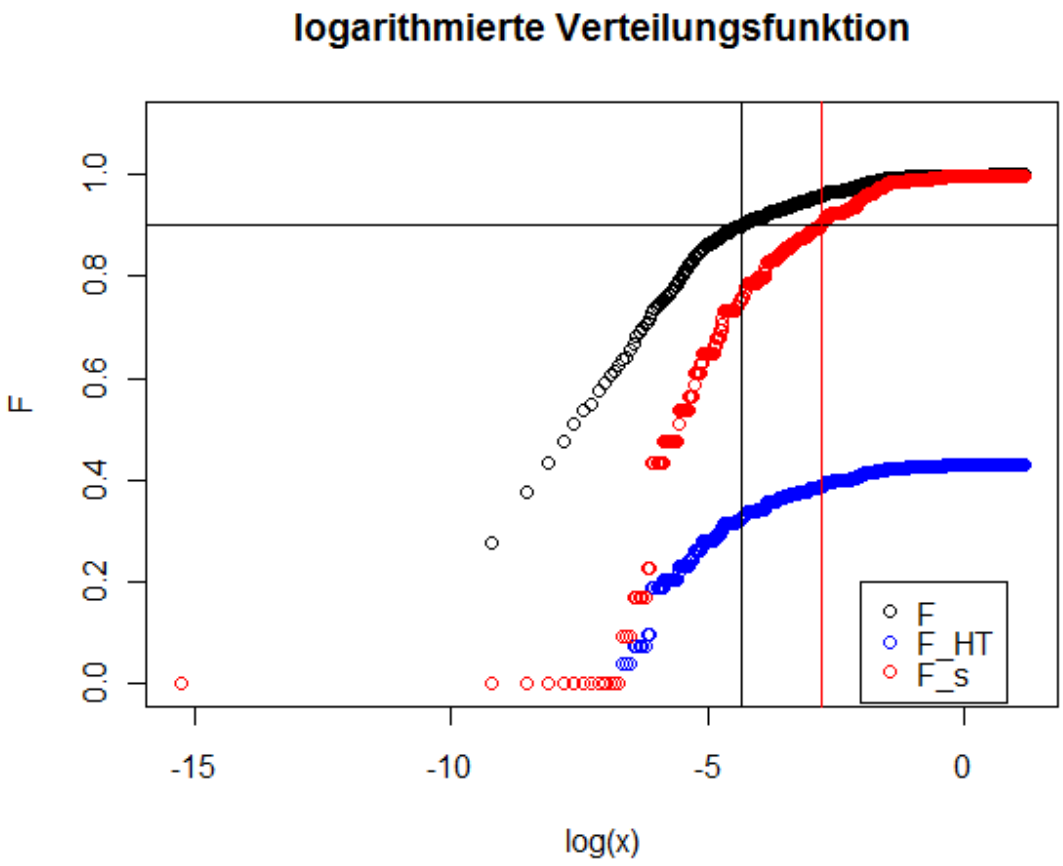

Abbildung 6: logarithmierte Verteilungsfunktion, `Grundgesamtheit 2`

Abbildung 2 gesehen, die meisten $\hat{F}_{HT}(x)$ deutlich unter eins bleiben, kann man die Vermutung aufstellen, dass dies zu einer systematischen Überschätzung der Quantile führen kann.

Die Schätzung des Quantils mit Hilfe von $\hat{F}_{HT}(x)$ ist nicht möglich, da hier kein Schnittpunkt existiert.

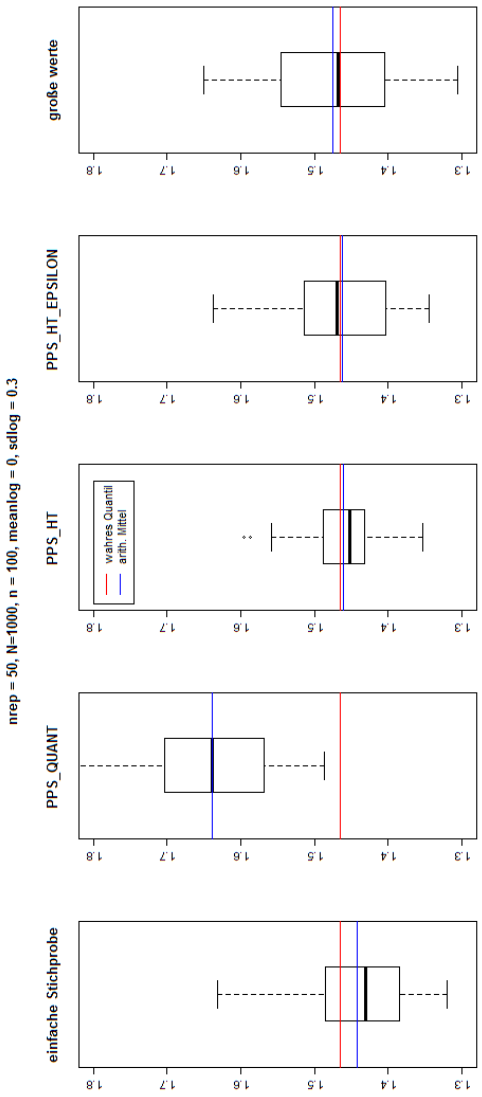

Abbildung 7: 0.9-Quantilsschätzer, `Grundgesamtheit 1`

4.2 Schätzung der Quantilen

Nach den Erfahrungen in Abschnitt 4.1 verwenden wir nun nur noch die geschätzte Verteilungsfunktion $\hat{F}_S(x)$ zur Schätzung der Quantile.

Für einen ersten Eindruck starten wir wieder mit `Grundgesamtheit 1` und dem

0.9-Quantil. In *Abbildung* 7 sind fünf Boxplots. Jeder repräsentiert 50 Schätzungen mit den in Abschnitt 3 beschriebenen Schätzverfahren. Die roten horizontalen Linien symbolisieren das wahre Quantil, während die blauen horizontalen Linien das arithmetische Mittel der 50 Schätzungen darstellen. Im Idealfall sollte das arithmetische Mittel der Schätzungen und das wahre Quantil möglichst nahe beieinander (Erwartungstreue) und die Box möglichst eng darum herum liegen(Streuung). Wie erwartet wird bei PPS_QUANT das Quantil stark überschätzt. Dieses Verfahren wird im Folgendem nicht weiter betrachten. Die anderen Schätzungen sind wesentlich besser. Das arithmetische Mittel liegt bei der EINFACHEN STICHPROBE am weitesten vom wahren Wert entfernt, während die Box bei PPS_HT am kleinsten ist. Bei den GROßEN WERTEN liegen arithmetisches Mittel und wahres Quantil nahe beinander, jedoch ist die Box recht groß. PPS_HT_EPSILON hat eine größere Streuung als PPS_HT. Das Ergebnis entspricht den vorangegangenen theoretischen Überlegungen.

Nun wieder der Blick auf die `Grundgesamtheit 2` in *Abbildung* 8: Diesmal ist PPS_QUANT weggelassen. Wie nach Abbildung 6 vermutet, liegt das arithmetische Mittel der Schätzer PPS_HT und PPS_HT_Epsilon deutlich über dem wahren Quantil. Der Eindruck scheint sich zu bestätigen, dass die Verwendung von $\hat{F}_S(x)$ zumindest bei einer steilen Verteilung zur Überschätzung der Quantile führt. Hinzu kommt, dass die Boxen der PPS-Methoden deutlich größer sind als die der EINFACHEN STICHPROBE. Anders als bei `Grundgesamtheit 1` haben die PPS Verfahren also keine geringere Streuung als die Schätzung durch eine EINFACHE STICHPROBE, wie man es aus der Theorie erwarten würde (Abschnitt 2.2). Die Methode der GROßEN WERTE hat ein ähnliches Ergebnis wie das der EINFACHEN STICHPROBE. Auch hier liegen wahres Quantil und arithmetisches Mittel der Schätzungen nahe beieinander, und die Box hat eine ähnliche Größe. In den *Abbildungen* 9 und 10 wurde die Simulation für die Quantile 0.8 und 0.9 wiederholt. Das Ergebnis unterscheidet sich kaum von dem der 0.9 Quantilschätzer.

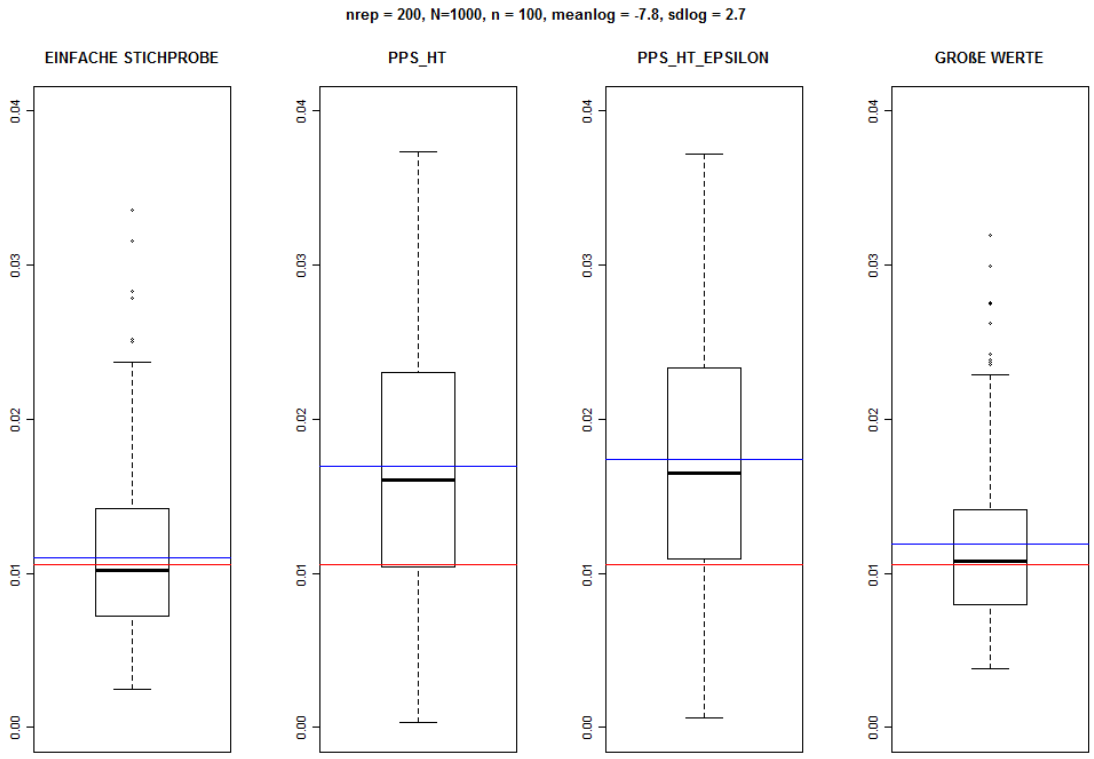

Abbildung 8: 0.9-Quantilsschätzer, `Grundgesamtheit 2`

5 Fazit

Im theoretischem Teil der Arbeit wurde erarbeitet, dass die PPS-Stichprobe im Vergleich zur einfachen Stichprobe effizientere Schätzer mit geringerer Varianz ermöglicht. Im praktischem Teil hat sich dies aber nur zum Teil bestätigt. Bei einer aus der Realität nachempfundenen Grundgesamtheit, deren Verteilung sehr steil war, wurden die Quantile durch eine einfache Stichprobe genauer geschätzt. Hinzu kam eine deutliche systematische Überschätzung der Quantile bei PPS-Stichproben. Der Grund dafür ist in der zu flachen Schätzung der Verteilungsfunktion durch $\hat{F}_S(x)$ zu finden. Die alternative Schätzmethode, die gewissermaßen einfache und PPS Stichprobe mischt, zeigte sich als ähnlich effizient wie die Schätzung mit der einfachen Stichprobe. Al-

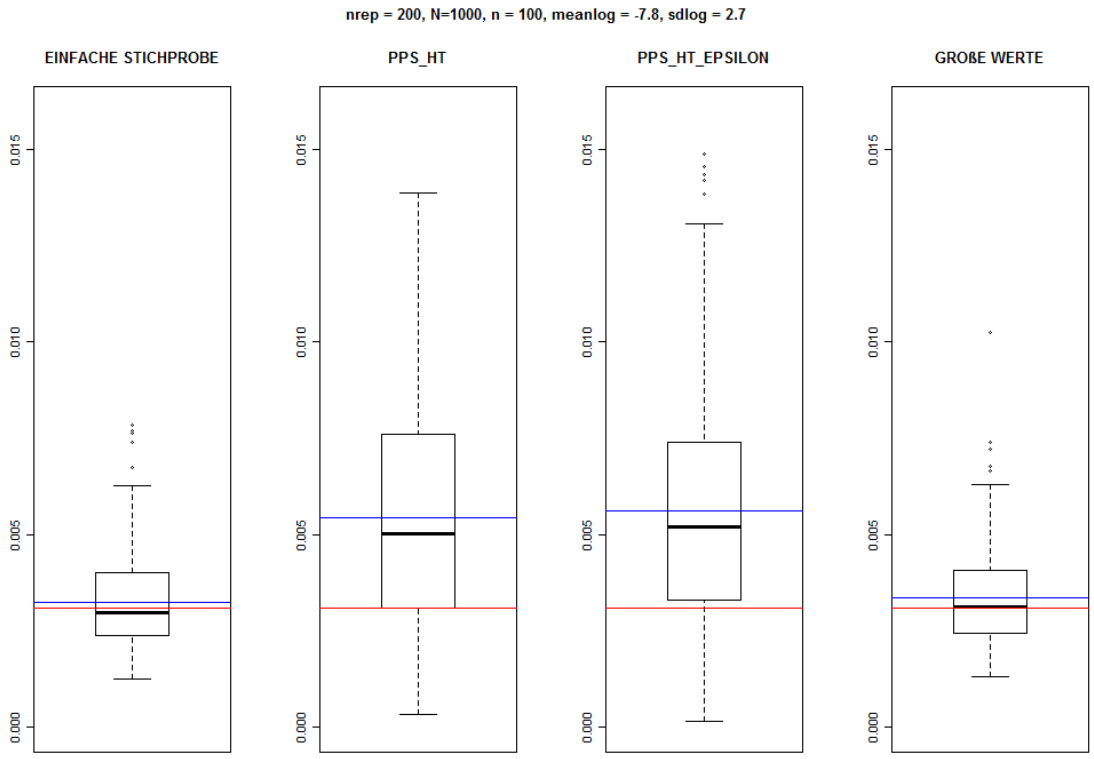

Abbildung 9: 0.8-Quantilsschätzer, `Grundgesamtheit 2`

lerdings ist sie in der Praxis wohl schwieriger zu realisieren.

Bei einer „einfacheren“ fiktiven Grundgesamtheit hingegen konnten die theoretischen Überlegungen bestätigt werden. Zukünftige Arbeiten könnten sich damit befassen, wie man diese theoretischen Vorteile der PPS-Stichprobe auch auf alle Grundgesamtheiten aus der Praxis übertragen kann.

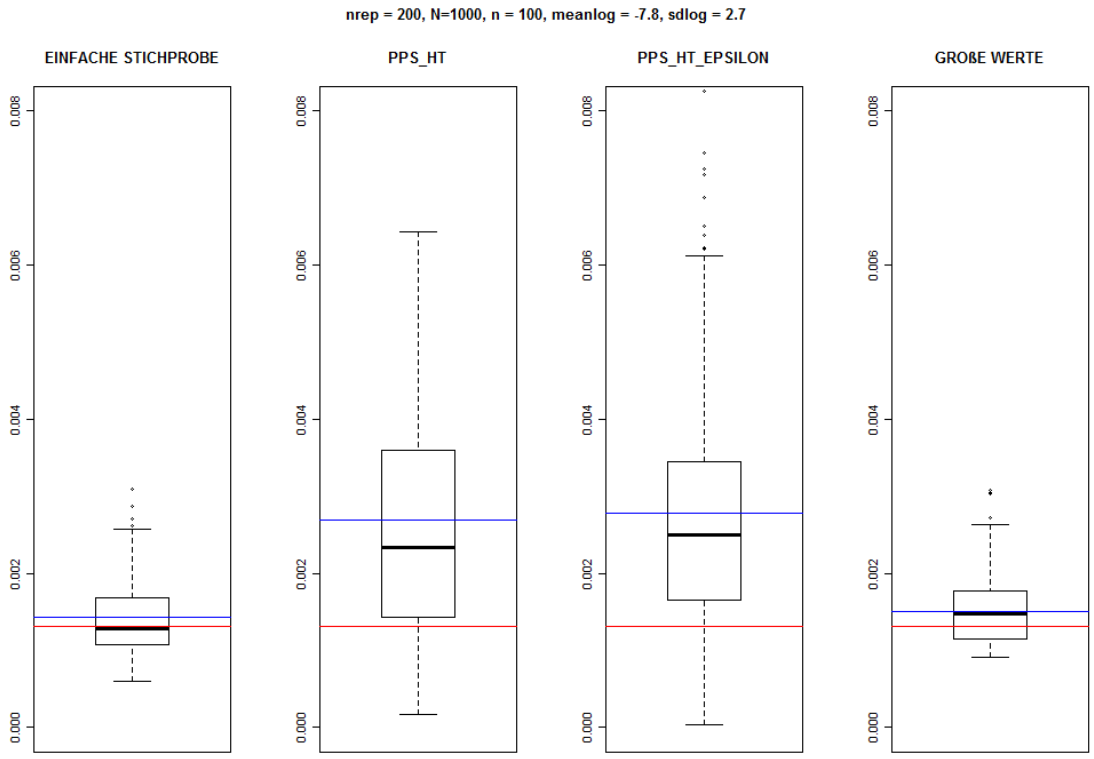

Abbildung 10: 0.7-Quantilsschätzer, `Grundgesamtheit 2`

Literatur

Fahrmeir, L., Künstler, R., Pigeot, I. and Tutz, G. (2007). *Statistik*, Springer DE, pp. 49–67.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**(260): 663–685.

Kauermann, G. and Küchenhoff, H. (2011). *Stichproben: Methoden und praktische Umsetzung mit R*, Springer DE.

Kuk, A. Y. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities, *Biometrika* **75**(1): 97–103.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 253–261.

Erklärung zur Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 17. Oktober 2013

(Felix Loewe)