Wieviele Kraftfahrzeuge gibt es im Kanton Schaffhausen? – Anzahlschätzung anhand von Seriennummern

Klaus Pommerening

Juni 2020 – letzte Änderung 23. September 2020

Zusammenfassung

Aus einem ganzzahligen Intervall $\{1, \ldots, N\}$ – etwa von Seriennummern aus einer Fertigung – werden zufällig n Zahlen beobachtet. Daraus soll auf die Gesamtanzahl N geschlossen werden. Es werden verschiedene Schätzer für N diskutiert.

1 Fragestellung

In der Schweiz werden Zulassungsnummern für Kraftfahrzeuge pro Kanton sequenziell vergeben – das erste zugelassene Fahrzeug bekommt die 1 an das Kürzel des Kantons (im Beispiel SH für Schaffhausen) angehängt, das zweite die 2, usw. Die größte vergebene Nummer[1] gibt also die Gesamtzahl der aktuell zugelassenen Kraftfahrzeuge im Kanton wieder.[2]

Beobachtet man einige Fahrzeuge zufällig, so wird mit ziemlicher Sicherheit das Fahrzeug mit der höchsten Nummer N nicht dabei sein; aber man erhält eine Stichprobe aus dem ganzzahligen Intervall $\{1, \ldots, N\}$. Was kann man daraus über N schließen?

Trivialerweise ist $N \geq M$, wenn M die maximale beobachtete Nummer ist, und $M \geq n$, der Anzahl der verschiedenen beobachteten Nummern.

Eine Fahrt über Stein am Rhein führte zu den zwanzig Nummern

SH 64920	SH 18224	SH 60075	SH 53810	SH 57292
SH 34205	SH 43945	SH 54804	SH 43996	SH 64279
SH 59972	SH 32275	SH 40779	SH 36524	SH 51811
SH 41200	SH 43609	SH 51998	SH 17771	SH 35554

Wieviele Kraftfahrzeuge dürfen wir also im Kanton SH vermuten?

Diese Fragestellung ist für das konkrete Beispiel natürlich Spielerei – eine Anfrage beim Strassenverkehrsamt des Kantons würde vermutlich ohne weitere Rechnerei zu

[1] Das mobile Radarmessgerät des Kantons SH hat allerdings eine Nummer jenseits von 75 000, die anscheinend nicht in dieses Schema passt.

[2] Frei werdende Nummern bleiben ihrem Besitzer erhalten, der sie in der Regel sofort wieder verwendet oder an jemand anders weitergibt.

einer exakten Zahl führen.[3] Sie hat aber auch ernste Anwendungen: So schätzten die Alliierten im zweiten Weltkrieg z. B. die Anzahl der gebauten deutschen Panzer[4] anhand der Seriennummern von Bauteilen der erbeuteten Exemplare [5] und, wichtiger noch, die Produktionsraten. Für weitere Anwendungen siehe [1].

2 Verschiedene Schätzwerte

Für die Qualität einer statistischen Schätzung gibt es verschiedene Kriterien, die nicht notwendig alle gleichzeitig erfüllbar sind:

Erwartungstreue bedeutet, dass bei vielen Wiederholungen der Schätzwert um den wahren Wert streut, genauer, dass sein Erwartungswert mit dem wahren Wert übereinstimmt. Diese Eigenschaft nennt man auch Unverzerrtheit.

Fehlerminimierung bedeutet, dass der Erwartungwert für die Abweichung des Schätzwerts vom wahren Wert möglichst klein ist.

Plausibilität bedeutet, dass die Schätzung keine Werte liefern kann, die von vornherein ausgeschlossen sind, d. h., dass der Schätzwert offensichtliche Nebenbedingungen respektiert. In unserem konkreten Fall bedeutet das etwa, dass er nicht kleiner als das beobachtete Maximum sein sollte.

Die wichtigste unter diesen Eigenschaften ist sicherlich die Erwartungstreue. Sie bedeutet in diesem Fall: Beobachtet man *alle* n-elementigen Teilmengen, so ist der Mittelwert aller daraus berechneten Schätzwerte *genau* N. Oder praktisch gedacht: Wiederholt man das Experiment „Ziehung einer n-elementigen Stichprobe“ mehrfach, so werden die ermittelten Schätzwerte um N herum streuen. Im Falle der Erwartungstreue bedeutet die Fehlerminimierung, dass diese Streuung möglichst eng, die Varianz (oder die Standardabweichung) des Schätzwerts also möglichst klein ist. Ein direkteres Maß ist der mittlere absolute Fehler, aber der ist meist schwieriger zu bestimmen.

Von einem formalen Gesichtspunkt aus betrachten wir den Wahrscheinlichkeitsraum $\Omega = \mathfrak{P}_n(\{1, \ldots, N\})$, der genau aus den $\binom{N}{n}$ verschiedenen n-elementigen Teilmengen $A \subseteq \{1, \ldots, N\}$ besteht, mit der Gleichverteilung – jede solche Teilmenge A hat also die Chance $1/\binom{N}{n}$, zufällig beobachtet zu werden. Ein Schätzer ist, formal betrachtet, eine Zufallsvariable

$$Z \colon \Omega \longrightarrow \mathbb{R},$$

und wir wollen statistische Kennzahlen dieser Zufallsvariablen wie Erwartungswert und Varianz bestimmen. (Die Werte $Z(\omega)$ für $\omega \in \Omega$ nennen wir dann Schätzwerte.)

[3] Oder man konsultiert den Schweizer Autoindex `https://autokennzeichen.halterauskunft.ch/`. Natürlich ändert sich die Zahl täglich.

[4] Daher wird die Aufgabe im Englischen auch “German Tank Problem” genannt, auf Deutsch dagegen ganz zivil „Taxiproblem“.

Schätzer 1: das Maximum

Wir stellen uns vor, wir haben eine zufällige Menge $A \in \Omega$ beobachtet; ihr maximales Element sei M. Sehr naiv ist die Schätzung der Anzahl N durch M, also durch den Wert

$$N^{(1)} := M \quad \text{bzw. korrekter } N^{(1)}(A) = \max A.$$

Im Zahlenbeispiel der Schaffhauser Autonummern wäre das $N^{(1)} = 64920$, was uns nicht als wahrscheinliche Anzahl zu überzeugen vermag. Diese Schätzung wird sich in der Tat als verzerrt herausstellen, und das ist auch plausibel, da sie stets $\leq N$ und in den meisten Fällen sicherlich $< N$ ist. Aber immerhin erfüllt $N^{(1)}$ das Plausibilitätskriterium, immer $\geq M$ zu sein. Der minimale Wert, den $N^{(1)}$ annehmen kann, ist n, der maximale N. Die Breite seiner Verteilung ist also $N - n + 1$.

Trotz der ins Auge stechenden Fehlerhaftigkeit dieser Schätzung ist es instruktiv, $N^{(1)}$ zu diskutieren, denn ein üblicher Ansatz für statistische Schätzprobleme ist die Maximum-Likelihood-Methode. Der Maximum-Likelihood-Schätzer für N wäre aber, wie sich zeigen wird, tatsächlich gerade gleich $N^{(1)}$ – sie ist in diesem Fall also verzerrt.

Schätzer 2: Maximum plus Minimum

Wir sollten für eine bessere Schätzung das beobachtete Maximum M etwas nach oben korrigieren. Aufgrund einer Symmetrie-Erwägung erwarten wir, dass M von der tatsächlichen Maximalzahl N ungefähr so weit entfernt ist wie das beobachtete Minimum m von der tatsächlichen Minimalzahl 1. Das motiviert uns, ausgehend von der hypothetischen Gleichheit $N - M = m - 1$, den Schätzer (hinfort „Max+Min-Schätzer“ genannt)

$$N^{(2)} := M + m - 1 \quad \text{bzw. korrekter } N^{(2)}(A) = \max A + \min A - 1$$

ins Auge zu fassen. Im Zahlenbeispiel wäre das $N^{(2)} = 64920 + 17771 - 1 = 82680$.

Der minimal mögliche Wert von $N^{(2)}$ wird für $m = 1$, $M = n$, angenommen, ist also n, der maximal mögliche Wert, für $M = N$, $m = N - n + 1$, ist $2N - n$. Daher streut $N^{(2)}$ fast doppelt so breit wie $N^{(1)}$.

Satz 1 *Die Verteilung von $N^{(2)}$ ist symmetrisch um den Wert N, insbesondere ist N der Mittelwert und somit $N^{(2)}$ erwartungstreu.*

Beweis. Dazu betrachten wir die Bijektion

$$\varphi\colon \{1, \ldots, N\} \longrightarrow \{1, \ldots, N\}, \quad a \mapsto N + 1 - a,$$

die die Reihenfolge der Zahlen $1, \ldots, N$ genau umkehrt. Sie induziert auf den n-elementigen Teilmengen eine Bijektion

$$\Phi\colon \Omega \longrightarrow \Omega, \quad \Phi(\{a_1, \ldots, a_n\}) = \{\varphi a_1, \ldots, \varphi a_n\},$$

mit den Eigenschaften:

- Ist $m = \min A$, so $N + 1 - m = \max \Phi(A)$,
- Ist $M = \max A$, so $N + 1 - M = \min \Phi(A)$.

Ist also $x = N^{(2)}(A) = M + m - 1$, so

$$N^{(2)}(\Phi(A)) = N + 1 - m + N + 1 - M - 1 = 2N - M - m + 1 = 2N - x.$$

Der Wert $2N - x$ kommt also genau so oft vor wie der Wert x. $\diamond$

Schätzer 3: der doppelte Mittelwert

Eine Erweiterung dieser Überlegung führt dazu, den Mittelwert nicht nur über die beiden extremen Beobachtungen M und m, sondern über alle beobachteten Werte $a_1, \ldots, a_n$ zu bilden, wieder mit 2 multipliziert und 1 subtrahiert:

$$N^{(3)} := \frac{2}{n} \cdot \sum_{i=1}^{n} a_i - 1 \quad \text{oder } N^{(3)}(A) = \frac{2}{n} \cdot \Sigma(A) \; - 1, \quad \text{wobei } \Sigma(A) = \sum_{a \in A} a\,.$$

Schließlich wird auf diese Weise die gesamte beobachtete Information ausgenützt! Eine andere Sicht auf diesen Schätzer ist die näherungsweise Übereinstimmung zwischen dem Mittelwert $\frac{1}{n}\Sigma(A)$ der Beobachtungen mit dem Mittelwert $\frac{N+1}{2}$ der Grundgesamtheit $\{1, \ldots, N\}$. Im Zahlenbeispiel ist $N^{(3)} = 2 \cdot 45352 - 1 = 90703$. Auf jeden Fall ist dieser Schätzer $N^{(3)}$ ebenfalls erwartungstreu, da alle Einzelbeobachtungen a_i den Erwartungswert $\frac{N+1}{2}$ haben:

$$E(N^{(3)}) = \frac{2}{n} \cdot \sum_{i=1}^{n} E(a_i) - 1 = \frac{2}{n} \cdot n \cdot \frac{N+1}{2} - 1 = N.$$

Satz 2 *Die Verteilung von $N^{(3)}$ ist symmetrisch um den Erwartungswert N.*

Beweis. Wie oben betrachten wir die Bijektion Φ von Ω. Es ist

$$\begin{aligned} N^{(3)}(\Phi A) &= \frac{2}{n} \cdot \sum_{a \in A} \varphi a \; - 1 = \frac{2}{n} \cdot \sum_{a \in A} (N + 1 - a) \; - 1 \\ &= 2N + 2 - \sum_{a \in A} a \; - 1 = 2N - N^{(3)}(A), \end{aligned}$$

analog zu Satz 1. $\diamond$

Aber o weh! – schon einfache Beispiele zeigen, dass $N^{(3)} < M$ sein kann, dieser Schätzer also gegen unser Plausibilitätskriterium verstößt.

Beispiel: $N = 10$, $n = 3$, $a_1 = 1$, $a_2 = 2$, $M = a_3 = 9$, mit Mittelwert 4, also $N^{(3)} = 2 \cdot 4 - 1 = 7$.

Die Schätzung $N^{(3)}(A)$ wird minimal, wenn $A = \{a_1, \ldots, a_n\} = \{1, \ldots, n\}$ und hat dann den Wert

$$\frac{2}{n} \cdot \sum_{i=1}^{n} i \; - 1 = \frac{2}{n} \cdot \frac{n \cdot (n+1)}{2} - 1 = n.$$

Das Maximum wird im Fall $A = \{N - n + 1, \ldots, N\}$ angenommen und ist

$$\frac{2}{n} \cdot \sum_{i=1}^{n} (N + 1 - i) - 1 = \frac{2}{n} \cdot \left[nN + n - \frac{n \cdot (n+1)}{2} \right] - 1 = 2N + 2 - (n+1) - 1 = 2N - n.$$

Die Breite der Streuung stimmt also genau mit der von $N^{(2)}$ überein.

Ohne weitere Begründungung sei hier erwähnt, dass die Mittelwerte der n-elementigen Teilmengen $A = \{a_1, \ldots, a_n\}$ annähernd die Varianz $(N^2 - 1)/12n$ haben – exakt wäre dieser Wert, wenn die Elemente a_i unabhängig voneinander gewählt würden, was im Urnenmodell dem Ziehen mit Zurücklegen entspräche [4]. Daher hat $N^{(3)}$ annähernd die Varianz $4 \cdot (N^2 - 1)/12n = (N^2 - 1)/3n$. Der exakte Wert ist, wie sich in Abschnitt 8 herausstellen wird, $(N + 1)(N - n)/3n$.

Schätzer 4: das Maximum mit Korrekturfaktor

Ein weiterer Ansatz ist, das beobachtete Maximum M mit einem Korrekturfaktor $\gamma > 1$ zu multiplizieren:

$$N^{(4)} = M \cdot \gamma,$$

wobei sich $\gamma = 1 + 1/n$ (im wesentlichen) als geeignet herausstellen wird. An diesem Schätzer ist bemerkenswert, dass von der durch Beobachtung gewonnenen Information nur das Maximum und die Zahl der Beobachtungen verwendet werden. Zu einem plausiblen Wert für den Faktor γ kommt man auf folgende Weise, siehe [3, 3.3.2]: Die Menge A der Einzelbeobachtungen sei der Größe nach angeordnet, $A = \{a_1, \ldots, a_n\}$ mit $1 \leq a_1 < \ldots < a_n \leq N$, also $m = a_1$, $M = a_n$. Dann ist die Lücke zwischen zwei größenmäßig benachbarten Beobachtungen $a_{i+1} - a_i$ für $1 \leq i < n$. Die Lücke am Anfang hat die Größe $a_1 - 1$. Mit dem Mittelwert dieser n Lücken schätzen wir die unbekannte Lücke $N - a_n$ am Ende (hoffentlich genauer als nur durch die Anfangslücke $a_1 - 1$ wie bei $N^{(2)}$) durch

$$x = \frac{1}{n} \cdot [(a_1 - 1) + (a_2 - a_1) + \cdots + (a_n - a_{n-1})] = \frac{1}{n} \cdot (a_n - 1) = \frac{1}{n} \cdot M - 1,$$

also N durch $a_n + x = M + \frac{1}{n} \cdot M - 1 = M \cdot (n+1)/n - 1$, und das motiviert die Definition

$$N^{(4)} = M \cdot (1 + \frac{1}{n}) - 1, \quad \text{bzw. korrekter } N^{(4)}(A) = \max A \cdot \left(1 + \frac{1}{\#A}\right) - 1,$$

also den Korrekturfaktor $\gamma = 1 + 1/n$ (mit der kleinen additiven Adjustierung -1)[5]. Mit den zwanzig beobachteten Schaffhauser Autokennzeichen gäbe das die Schätzung

[5] Würden die Seriennummern gemäß der Zählweise des Informatikers mit der natürlichsten aller Zahlen, der 0, beginnen, entfiele die Adjustierung um -1 bei $N^{(2)}$, $N^{(3)}$ und $N^{(4)}$. Bei großem N ist sie ohnehin praktisch ohne Bedeutung.

$N^{(4)} = 64920 \cdot 21/20 - 1 = 68165$ für die Anzahl der zugelassenen Kraftfahrzeuge[6], die zumindest in diesem Zahlenbeispiel wesentlich plausibler erscheint als die bisherigen Schätzungen $N^{(1)} = 64920$, $N^{(2)} = 82690$ und $N^{(3)} = 90703$.

Die Verteilung von $N^{(4)}$ ist eine leicht gedehnte Variante der von $N^{(1)}$, die wir im Folgenden näher untersuchen wollen. Jedenfalls streut $N^{(4)}$ zwischen den Werten $n \cdot \frac{n+1}{n} - 1 = n$ und $N \cdot \frac{n+1}{n} - 1$, wobei im Falle der (hier durchaus sinnvollen) Rundung auf ganze Zahlen einzelne Lücken im Wertebereich auftreten werden.

3 Die Verteilung des Maximums

Die Verteilung des beobachteten Maximums M von n zufällig beobachteten Werten zwischen 1 und N wird im Folgenden mithilfe einer kombinatorischen Beschreibung im Urnenmodell genauer analysiert.

Wir stellen uns – wie bei der Ziehung der Lottozahlen – eine Urne mit N Kugeln vor, die mit den ganzen Zahlen von 1 bis N beschriftet sind. Aus dieser Urne werden n Kugeln zufällig gezogen, wir wählen also eine zufällige Teilmenge

$$A = \{a_1, \ldots, a_n\} \subseteq \{1, \ldots, N\}$$

aus und fragen zunächst nach dem Erwartungswert für das Maximum $\max A$. Dazu zählen wir, wieviele der insgesamt $\binom{N}{n}$ Teilmengen A der Größe n ein bestimmtes Maximum haben.

- Das Maximum ist N, wenn $N \in A$. Für die Restmenge $A - \{N\}$ gibt es dann noch $\binom{N-1}{n-1}$ Möglichkeiten – dies ist also die Anzahl der Teilmengen mit Maximum N.
- Das Maximum ist genau dann $N-1$, wenn $N \notin A$, $N-1 \in A$. Dafür gibt es $\binom{N-2}{n-1}$ Möglichkeiten, denn auf genau so viele Weisen können wir die übrigen $n-1$ Elemente in $\{1, \ldots, N-2\}$ noch wählen.
- Allgemein ist das Maximum genau dann M, wenn $M+1, \ldots, N \notin A$, aber $M \in A$. Die Zahl der Möglichkeiten dafür ist $\binom{M-1}{n-1}$.

Das geht, solange $M \geq n$, denn für eine n-elementige Teilmenge $A \subseteq \{1, \ldots, N\}$ ist $\max A \geq n$. Wir haben gezeigt:

Hilfssatz 1 *Sei $1 \leq n \leq M \leq N$ für natürliche Zahlen $n, M, N \in \mathbb{N}$. Dann ist die Anzahl aller n-elementigen Teilmengen $A \subseteq \{1, \ldots, N\}$ mit $\max A = M$ gleich*

$$H^{(1)}_{N,n}(M) := \binom{M-1}{n-1}.$$

[6] also etwas vergröbert: Maximum plus 5% – bei 100 beobachteten Werten wäre die Faustregel: „Maximum plus 1%“.

Also ist die Wahrscheinlichkeit, das Maximum M zu beobachten,

$$p^{(1)}_{N,n}(M) = \binom{M-1}{n-1} \Big/ \binom{N}{n} .$$

Durch diese Formel wird bei fester Anzahl N und festem Beobachtungsumfang n die Verteilung des beobachteten Maximums M vollständig beschrieben. Aus dem Bildungsgesetz des Pascalschen Dreiecks,

$$\binom{M}{n-1} = \binom{M-1}{n-1} + \binom{M-1}{n-2},$$

folgt, dass im Fall $n \geq 2$

$$p^{(1)}_{N,n}(M+1) \; > \; p^{(1)}_{N,n}(M) \quad \text{für } n \leq M \leq N-1,$$

ergänzt durch $p^{(1)}_{N,1}(M+1) = 1/N = p^{(1)}_{N,1}(M)$.

Die Verteilung von $p^{(1)}_{N,n}(M)$ für M zwischen n und N ist also monoton wachsend vom Minimum $p^{(1)}_{N,n}(n) = 1/\binom{N}{n}$ für den Anfangspunkt $M = n$ bis zum Maximum $p^{(1)}_{N,n}(N) = n/N$ für den Endpunkt $M = n$ dieses Intervalls. Insbesondere ist die Verteilung rechtssteil (oder linksschief). Abbildung 1 illustriert dies für die Parameter $N = 50$, $n = 10$. Dazwischen, mit wachsendem $n = 1, \ldots, N$, nimmt die Schiefe zu, d. h., die Verteilung verschiebt sich immer mehr nach rechts. Im Fall $n = 1$ haben wir die Gleichverteilung $p^{(1)}_{N,1}(M) = 1/N$ für $M = 1, \ldots, N$, im Fall $n = N$ konzentriert sich die ganze Verteilung in $p^{(1)}_{N,N}(N) = 1$.

Zusammengefasst und in Abbildung 1 veranschaulicht:

Satz 3 *Die Verteilung des Maximums einer n-elementigen Teilmenge von $\{1, \ldots, N\}$ ist*

$$p^{(1)}_{N,n}(x) = \begin{cases} \binom{x-1}{n-1} \Big/ \binom{N}{n} & \textit{für } x \in \mathbb{N}, n \leq x \leq N, \\ 0 & \textit{sonst.} \end{cases}$$

Sie ist im Falle $n = 1$ konstant $= 1/N$. Im Falle $N > n \geq 2$ wächst sie im ganzzahligen Intervall $\{n, \ldots, N\}$ streng monoton von $p^{(1)}_{N,n}(n) = 1/\binom{N}{n}$ bis $p^{(1)}_{N,n}(N) = n/N$.

Insbesondere ist die Chance, dass der Schätzer $N^{(1)}$ zufällig den richtigen Wert N trifft, gleich $p^{(1)}_{N,n}(N) = n/N$.

Bemerkung 1. Bei festen n und M ist der Zähler von $p^{(1)}_{N,n}(M)$ konstant und der Nenner als Funktion von N monoton wachsend. Also ist $p^{(1)}_{N,n}(M)$ mit wachsendem N monoton fallend. Es nimmt sein Maximum also am linken Rand, d. h. für $N = M$ an. Dieses, nämlich der Schätzwert $N^{(1)}$, wäre also die Maximum-Likelihood-Schätzung für N, wie im Abschnitt 2 behauptet.

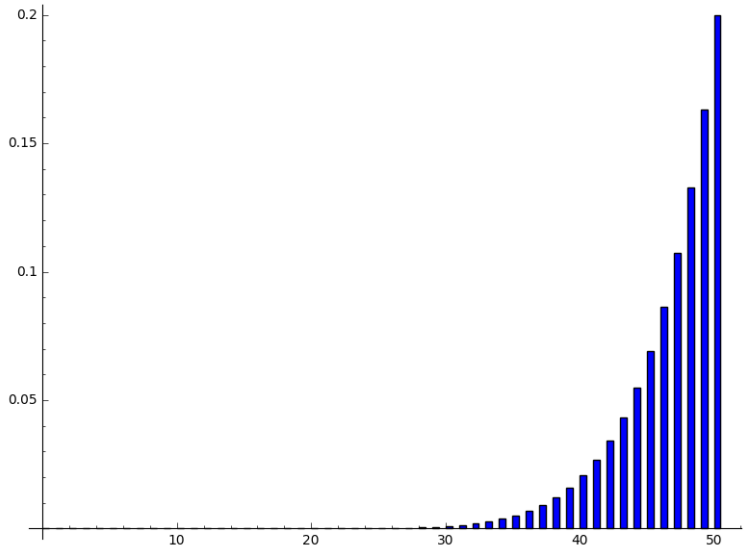

Abbildung 1: Die Verteilung des Maximums M für $N = 50$ und $n = 10$ (SageMath-Code in Anhang C.1)

Bemerkung 2. Nach Anhang B, Korollar 4 (i) gilt (mit $i = N - x + 1$)

$$\sum_{x=n}^{N} p_{N,n}^{(1)}(x) = \frac{1}{\binom{N}{n}} \cdot \sum_{x=n}^{N} \binom{x-1}{n-1} = \frac{1}{\binom{N}{n}} \cdot \sum_{i=1}^{N-n+1} \binom{N-i}{n-1} = \frac{\binom{N}{n}}{\binom{N}{n}} = 1,$$

d. h., die Wahrscheinlichkeiten addieren sich zu 1, wie es sich gehört.

Zur späteren Verwendung bei der Diskussion des Max+Min-Schätzwerts $N^{(2)}$, siehe Abschnitt 7, halten wir noch fest:

Hilfssatz 2 *Sei $1 \le n \le M \le N$ für natürliche Zahlen $n, M, N \in \mathbb{N}$ und $m \in \mathbb{N}$ mit $1 \le m \le M-n+1$. Dann ist die Anzahl aller n-elementigen Teilmengen $A \subseteq \{1, \ldots, N\}$ mit $\max A = M$ und $\min A = m$ gleich*

$$H_{N,n}^{(1a)}(M, m) := \binom{M-m-1}{n-2}.$$

Beweis. Für eine solche Menge A ist $A - \{m, M\}$ eine beliebige Teilmenge von $\{m+1, \ldots, M-1\}$, und diese letztere Menge hat $M - m - 1$ Elemente. $\diamond$

4 Der Erwartungswert des Maximums

In die Definition von Erwartungswert und Varianz werden die in Satz 3 bestimmten Wahrscheinlichkeiten eingesetzt. Das ergibt:

Hilfssatz 3 *Der Erwartungswert für das Maximum einer n-elementigen Teilmenge von* $\{1, \ldots, N\}$ *ist*

$$E(N^{(1)}) = \frac{\sum_{x=n}^{N} x \cdot \binom{x-1}{n-1}}{\binom{N}{n}} := \mu(N, n)\,.$$

Die Varianz ist

$$V(N^{(1)}) = \frac{\sum_{x=n}^{N} x^2 \cdot \binom{x-1}{n-1}}{\binom{N}{n}} - \mu(N, n)^2 := \tau(N, n)\,.$$

Die Summe

$$S(N, n) = \sum_{x=n}^{N} x \cdot \binom{x-1}{n-1}$$

im Zähler der ersten Formel von Hilfssatz 3 für den Erwartungswert wird mithilfe von Korollar 4 in Anhang B ausgewertet (für $i = N - x + 1$):

$$\begin{aligned} S(N, n) &= \sum_{x=n}^{N} x \cdot \binom{x-1}{n-1} = \sum_{i=1}^{N-n+1} (N - i + 1) \cdot \binom{N-i}{n-1} \\ &= (N+1) \cdot \sum_{i=1}^{N-n+1} \binom{N-i}{n-1} - \sum_{i=1}^{N-n+1} i \cdot \binom{N-i}{n-1} \\ &= (N+1) \cdot \binom{N}{n} - \binom{N+1}{n+1} = \frac{(N+1)!}{n!(N-n)!} - \frac{(N+1)!}{(n+1)!(N-n)!} \end{aligned}$$

Da $1 - 1/(n+1) = n/(n+1)$, folgt:

Hilfssatz 4

$$S(N, n) = n \cdot \binom{N+1}{n+1}\,.$$

Das wird in die Formel für den Erwartungswert aus Hilfssatz 3 eingesetzt:

$$\begin{aligned} \mu(N, n) &= \frac{S(N, n)}{\binom{N}{n}} = \frac{n \cdot \binom{N+1}{n+1}}{\binom{N}{n}} = \frac{n \cdot (N+1)! \cdot n! \cdot (N-n)!}{(n+1)! \cdot (N-n)! \cdot N!} \\ &= \frac{n}{n+1} \cdot (N+1). \end{aligned}$$

Damit ist gezeigt:

Satz 4 *Der Erwartungswert für das Maximum einer n-elementigen Teilmenge von* $\{1, \ldots, N\}$ *ist*

$$\mu(N, n) = \frac{n}{n+1} \cdot (N+1).$$

Damit ist auch gezeigt, dass der Schätzer $N^{(1)}$ mit seinem Erwartungswert $E(N^{(1)}) = (N+1) \cdot n/(n+1)$ nicht erwartungstreu, also verzerrt ist.

Im Fall einer einelementigen Stichprobe, $n = 1$, geht die Formel aus Satz 4 über in den Erwartungswert $(N+1)/2$ für ein zufälliges Element aus $\{1, \ldots, N\}$. Im Zahlenbeispiel von Abbildung 1 errechnen wir den Erwartungswert $\mu(50, 10) = \frac{10}{11} \cdot 51 \approx 46.4$.

Korollar 1 *Der Erwartungswert für das Minimum einer n-elementigen Teilmenge von $\{1, \ldots, N\}$ ist*

$$\nu(N, n) = \frac{1}{n+1} \cdot (N+1).$$

Beweis. Wir wenden den Satz auf die Menge der zu $\{a_1, \ldots, a_n\}$ „komplementären" Zahlen $\{N+1-a_1, \ldots, N+1-a_n\}$ an.[7] Der Erwartungswert für ihr Maximum ist $\mu(N, n) = (N+1) \cdot n/(n+1)$. Der Erwartungswert für das Minimum von $\{a_1, \ldots, a_n\}$ ist also $N + 1 - \mu(N, n) = (N+1) \cdot [1 - n/(n+1)] = (N+1) \cdot 1/(n+1)$. $\diamond$

Hieraus folgt noch einmal etwas, was wir schon aus Satz 1 wissen;

Korollar 2 *Der Erwartungswert für den Max+Min-Schätzer $N^{(2)}$ einer n-elementigen Teilmenge von $\{1, \ldots, N\}$ ist*

$$E(N^{(2)}) = N.$$

Beweis. Dieser Erwartungswert ist die Summe $\mu(N, n) + \nu(N, n) - 1$. $\diamond$

5 Die Varianz des Maximums

Um die Varianz $\tau(N, n)$ zu bestimmen, haben wir analog die Summe

$$T(N, n) = \sum_{x=n}^{N} x^2 \cdot \binom{x-1}{n-1}$$

aus Hilfssatz 3 auszuwerten. Das ist natürlich etwas mühsamer:

[7] D. h., wir wenden die Bijektion $\varphi : x \mapsto N + 1 - x$ von $\{1, \ldots, N\}$ an, die die Ordnung umkehrt.

$$
\begin{aligned}
T(N,n) &= \sum_{i=1}^{N-n+1} (N+1-i)^2 \cdot \binom{N-i}{n-1} \\
&= (N+1)^2 \sum_{i=1}^{N-n+1} \binom{N-i}{n-1} - (2N+3) \sum_{i=1}^{N-n+1} i \cdot \binom{N-i}{n-1} + 2 \cdot \sum_{i=1}^{N-n+1} \frac{i^2+i}{2} \binom{N-i}{n-1} \\
&= (N+1)^2 \cdot \binom{N}{n} - (2N+3) \cdot \binom{N+1}{n+1} + 2 \cdot \binom{N+2}{n+2} \\
&= \binom{N}{n} \left[(N+1)^2 - (2N+3) \cdot \frac{N+1}{n+1} + 2 \cdot \frac{(N+1)(N+2)}{(n+1)(n+2)} \right] \\
&= \binom{N}{n} \cdot \frac{N+1}{(n+1)(n+2)} \cdot [(N+1)(n+1)(n+2) - (2N+3)(n+2) + 2N + 4] \\
&= \binom{N}{n} \cdot \frac{N+1}{(n+1)(n+2)} \cdot \left[n^2 N + nN + n^2 \right] \\
&= \binom{N}{n} \cdot \frac{n\,(N+1)\,(nN+N+n)}{(n+1)\,(n+2)}\,.
\end{aligned}
$$

Das wird in die Formel für die Varianz eingesetzt:

$$
\begin{aligned}
\tau(N,n) &= \frac{1}{\binom{N}{n}} \cdot T(N,n) - \frac{n^2\,(N+1)^2}{(n+1)^2} = \frac{n\,(N+1)}{(n+1)} \cdot \left[\frac{nN+N+n}{n+2} + \frac{n\,(N+1)}{n+1} \right] \\
&= \frac{n\,(N+1)}{(n+1)^2\,(n+2)} \cdot \left[(n+1)\,(nN+N+n) - (n^2+2n)\,(N+1) \right] \\
&= \frac{n\,(N+1)}{(n+1)^2\,(n+2)} \cdot [N-n]
\end{aligned}
$$

Zusammengefasst:

Satz 5 *Die Varianz des Maximums einer n-elementigen Teilmenge von $\{1,\ldots,N\}$ ist*

$$\tau(N,n) = \frac{n\,(N+1)\,(N-n)}{(n+1)^2\,(n+2)}.$$

Im Zahlenbeispiel von Abbildung 1 mit $N = 50$, $n = 10$, ist die Varianz also $\tau(50,10) = (10 \cdot 51 \cdot 40)/(11^2 \cdot 12) \approx 14.05$, die Standardabweichung die Wurzel daraus, also $\sigma(50,10) \approx 3.75$.

6 Schätzung der Anzahl durch das Maximum mit Korrekturfaktor

Kommen wir zum Schätzer $N^{(4)}$ zurück. Satz 4 legt nahe, die Gleichung nach N aufzulösen und als Schätzwert für N den schon bekannten Wert

$$N^{(4)}(A) = M \cdot \frac{n+1}{n} - 1$$

zu nehmen. Erfreulich ist, dass $N^{(4)}$ das wichtigste Kriterium erfüllt:

Satz 6 *Der Schätzer*

$$N^{(4)} = N^{(1)} \cdot \frac{n+1}{n} - 1$$

für N ist erwartungstreu. Seine Varianz ist

$$V(N^{(4)}) = \frac{(N+1)\,(N-n)}{n\,(n+2)}\,.$$

Beweis. Der Erwartungswert für $N^{(4)}$ ist

$$E(N^{(4)}) = \mu(M,n) \cdot \frac{n+1}{n} - 1 = \frac{n}{n+1} \cdot (N+1) \cdot \frac{n+1}{n} - 1 = N + 1 - 1 = N.$$

Die Aussage über die Varianz folgt aus

$$V(N^{(4)}) = V(N^{(1)}) \cdot \frac{(n+1)^2}{n^2}\,.$$

◇

Die Verteilung von $N^{(4)}$ ist, bis auf die Verschiebung um -1, eine mit dem Faktor $1 + 1/n$ skalierte Version davon, sieht also auch so aus wie in Abbildung 1. In der Praxis wird man den Schätzwert auf die nächste ganze Zahl runden, was natürlich kleinere Störungen im Säulendiagramm bewirkt, siehe Abbildung 2. [8]

7 Die Verteilung des Max+Min-Schätzers

Wann nimmt die Zufallsvariable $N^{(2)} = M + m - 1$ in $A \in \Omega$ den Wert x an? Die Nebenbedingung $M = x - m + 1 \geq m + n - 1$ erzwingt $x - n + 2 \geq 2m$, also

$$1 \leq m \leq \frac{x-n}{2} + 1.$$

Daher bleiben genau die Fälle

$$\begin{aligned} m &= 1 & M &= x \\ &\vdots & &\vdots \\ m &= \left\lfloor \frac{x-n}{2} \right\rfloor + 1 & M &= \left\lceil \frac{x+n}{2} \right\rceil \end{aligned}$$

Da $M \leq N$, beginnt diese Aufzählung allerdings nur so lange bei $m = 1$, wie $x \leq N$.

[8] in der die Werte 37 und 48 nicht vorkommen, weil etwa $44 \cdot 1.1 - 1 = 47.4$ von SageMath zu 47 gerundet wird und $45 \cdot 1.1 - 1 = 48.5$ bereits zu 49.

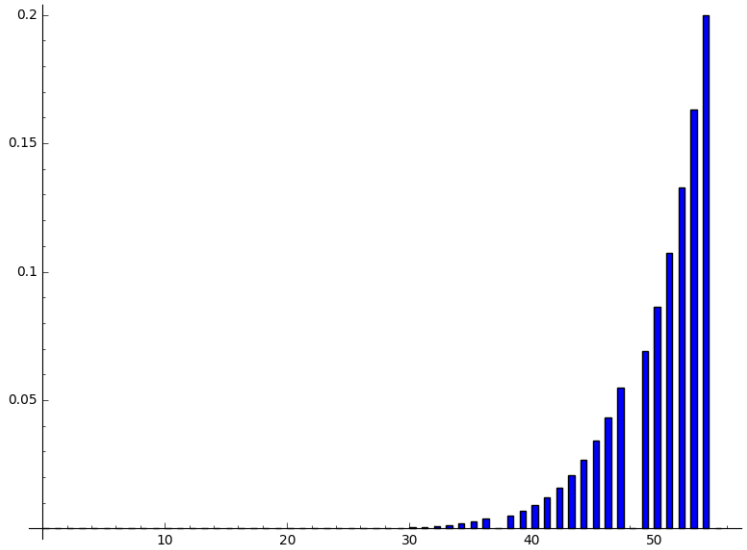

Abbildung 2: Die Verteilung des Schätzers $N^{(4)}$ für $N = 50$ und $n = 10$ (SageMath-Code in Anhang C.1)

Nehmen wir $n \geq 2$ an, so tritt jeder dieser Fälle genau so oft auf, wie es $(n-2)$-elementige Teilmengen der $(M-m-1)$-elementigen Menge $\{m+1, \ldots, M-1\}$ gibt, also

$$\binom{M-m-1}{n-2} = \binom{x-m+1-m-1}{n-2} = \binom{x-2m}{n-2}$$

mal.

Im trivialen Fall $n = 1$ ist A eine einelementige Menge $A = \{a\}$ mit $1 \leq a \leq N$, Maximum und Minimum sind $= a$, und somit $N^{(2)}(A) = 2a - 1$.

Damit ist der folgende Satz bewiesen:

Satz 7 *Seien n und N ganze Zahlen mit $1 \leq n \leq N$.*

(i) *Für $n = 1$ und $1 \leq x \leq 2N-1$ ist die Anzahl der einelementigen Mengen $A \in \Omega$ mit $N^{(2)}(A) = x$ gleich*

$$H^{(2)}_{N,1}(x) = \begin{cases} 1, & \text{falls } x \text{ ungerade,} \\ 0, & \text{falls } x \text{ gerade.} \end{cases}$$

(ii) *Falls $n \geq 2$, ist für jede ganze Zahl x mit $n \leq x \leq N$ die Anzahl der n-elementigen Mengen $A \in \Omega$ mit $N^{(2)}(A) = x$ gleich*

$$H^{(2)}_{N,n}(x) = \sum_{m=1}^{\lfloor \frac{x-n}{2} \rfloor + 1} \binom{x-2m}{n-2} =: F(x,n)\,.$$

(iii) *Für jede ganze Zahl x mit $N \le x \le 2N - n$ ist $n \le 2N - x \le N$ und die Anzahl der n-elementigen Mengen $A \in \Omega$ mit $N^{(2)}(A) = x$ genau gleich*

$$H^{(2)}_{N,n}(x) = H^{(2)}_{N,n}(2N - x)\,.$$

Dabei folgt (iii) aus der Symmetrie der Verteilung um N, siehe Satz 1: Der Wert $2N - x$ kommt genau so oft vor wie der Wert x.

Abbildung 3 zeigt wieder exemplarisch die Verteilung $p^{(2)}_{N,n}(x) = H^{(2)}_{N,n}(x)/\binom{N}{n}$ für $N = 50$ und $n = 10$.

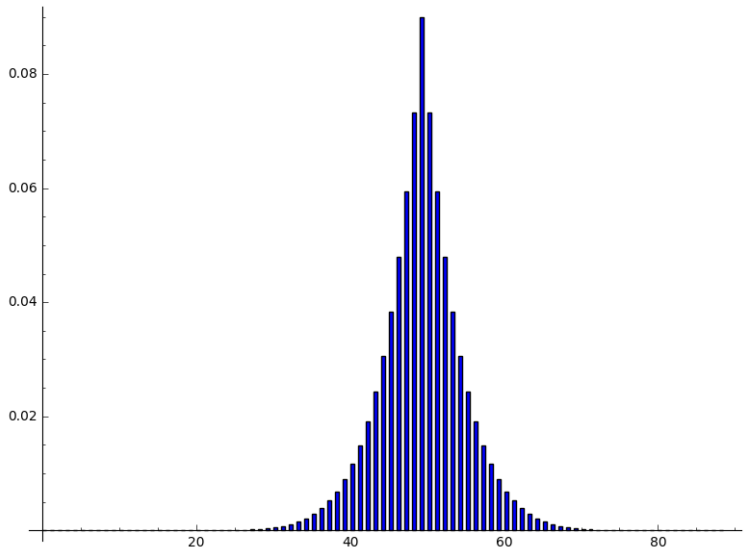

Abbildung 3: Die Verteilung des Max+Min-Schätzers $N^{(2)}$ für $N = 50$ und $n = 10$ (SageMath-Code in Anhang C.1)

Bemerkenswerterweise ist der Wert $H^{(2)}_{N,n}(x) = F(x, n)$ (für $n \le x \le N$) von N unabhängig, und es gilt:

Korollar 1 *Für beliebige $x \in \mathbb{N}$ ist $F(x, n)$ die Anzahl aller n-elementigen Mengen A von natürlichen Zahlen ≥ 1 mit $\max A + \min A = x$.*

Aus (ii), also der Definitionsformel der Häufigkeitsfunktion F, läßt sich eine Rekursionsformel herleiten (die im Fall $n = 1$ trivialerweise gilt). Dazu nehmen wir erst $x - n$ als gerade an. Dann ist nach Korollar 4 in Anhang B

$$\begin{aligned} F(x,n) &= \binom{x-2}{n-2} + \cdots + \binom{n}{n-2} + \binom{n-2}{n-2}, \\ F(x+1,n) &= \binom{x-1}{n-2} + \cdots + \binom{n+1}{n-2} + \binom{n-1}{n-2}, \\ F(x,n) + F(x+1,n) &= \binom{x}{n-1}. \end{aligned}$$

Falls $x-n$ ungerade ist, schließt man genauso:

$$\begin{aligned} F(x,n) &= \binom{x-2}{n-2} + \cdots + \binom{n+1}{n-2} + \binom{n-1}{n-2}, \\ F(x+1,n) &= \binom{x-1}{n-2} + \cdots + \binom{n}{n-2} + \binom{n-2}{n-2}, \\ F(x,n)+F(x+1,n) &= \binom{x}{n-1}. \end{aligned}$$

Damit folgt

Korollar 2 *Die Häufigkeitsfunktion $F(x,n)$ erfüllt für alle n die Rekursionsformel*

$$\begin{aligned} F(n,n) &= 1, \\ F(x,n) &= \binom{x-1}{n-1} - F(x-1,n) \quad \textit{für } x \geq n+1. \end{aligned}$$

Die Definition von $F(x,n)$ wird für $n > x$ durch beliebige Werte (am besten durch Nullen) ergänzt und für $n = 0$ durch

$$F(x,0) = (-1)^x = \begin{cases} -1, & \text{falls } x \text{ ungerade,} \\ 1, & \text{falls } x \text{ gerade.} \end{cases}$$

Dann gilt (mit $a = 1$ in der Definition aus Abschnitt B):

Satz 8 *Die Häufigkeitsfunktion F ist ein Pascal-Tableau in $\mathbb{Z}$, d. h.,*

$$F(x,n) = F(x-1,n-1) + F(x-1,n) \quad \textit{für } x > n \geq 1.$$

Dieses Pascal-Tableau ist in Abbildung 4 illustriert.
Beweis. Für $n = 1$ ist

$$F(x,1) = \begin{cases} 1 = 1+0 = F(x-1,0) + F(x-1,1), & \text{falls } x \text{ ungerade,} \\ 0 = -1+1 = F(x-1,0) + F(x-1,1), & \text{falls } x \text{ gerade.} \end{cases}$$

Von jetzt an sei $n \geq 2$.

Fall 1, $x-n$ gerade: Dann ist $(x-1)-n$ ungerade und $(x-1)-(n-1)$ gerade, also

$$\begin{aligned} F(x-1,n) &= \sum_{i=1}^{\frac{x-n}{2}} \binom{x-1-2i}{n-2}, \\ F(x-1,n-1) &= \sum_{i=1}^{\frac{x-n}{2}+1} \binom{x-1-2i}{n-3} = \underbrace{\binom{n-3}{n-3}}_{1} + \sum_{i=1}^{\frac{x-n}{2}} \binom{x-1-2i}{n-3}, \end{aligned}$$

$$\begin{aligned} F(x-1,n)+F(x-1,n-1) &= \underbrace{\binom{n-2}{n-2}}_{1} + \sum_{i=1}^{\frac{x-n}{2}} \underbrace{\left[\binom{x-1-2i}{n-2}+\binom{x-1-2i}{n-3}\right]}_{\binom{x-2i}{n-2}} \\ &= \sum_{i=1}^{\frac{x-n}{2}+1} \binom{x-2i}{n-2} = F(x,n). \end{aligned}$$

Fall 2, $x-n$ ungerade: Dann ist $(x-1)-n$ gerade und $(x-1)-(n-1)$ ungerade, also

$$\begin{aligned} F(x-1,n) &= \sum_{i=1}^{\frac{x-n+1}{2}} \binom{x-1-2i}{n-2}, \\ F(x-1,n-1) &= \sum_{i=1}^{\frac{x-n+1}{2}} \binom{x-1-2i}{n-3}, \\ F(x-1,n)+F(x-1,n-1) &= \sum_{i=1}^{\frac{x-n+1}{2}} \underbrace{\left[\binom{x-1-2i}{n-2}+\binom{x-1-2i}{n-3}\right]}_{\binom{x-2i}{n-2}} \\ &= \sum_{i=1}^{\frac{x-n+1}{2}+1} \binom{x-2i}{n-2} = F(x,n). \end{aligned}$$

$\diamond$

Daher können wir für F die Summationsformeln aus Anhang B anwenden. Als erstes rechnen wir zur Probe nach:

$$\begin{aligned} \sum_{x=n}^{2N-n} H_{N,n}^{(2)}(x) &= \sum_{x=n}^{N-1} H_{N,n}^{(2)}(x) + H_{N,n}^{(2)}(N) + \sum_{x=n}^{N-1} H_{N,n}^{(2)}(2N-x) \\ &= \sum_{x=n}^{N-1} F(x,n) + F(N,n) + \sum_{x=n}^{N-1} F(x,n) \\ &= 2\,F(N,n+1) + F(N,n) = F(N,n+1) + F(N+1,n+1) \\ &= \binom{N}{n} \end{aligned}$$

nach Korollar 2. Die Wahrscheinlichkeiten addieren sich also zu 1, wie es sein muss.

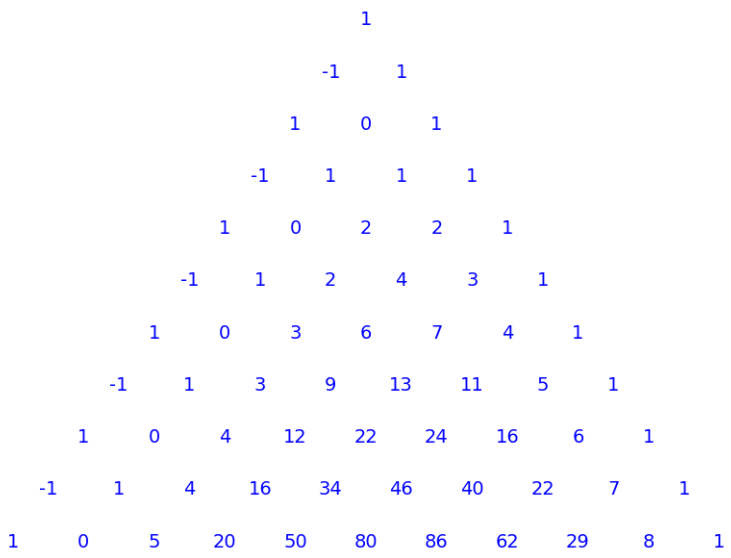

Abbildung 4: Das Pascal-Tableau der Funktion F (SageMath-Code in Anhang C.6)

Als weitere Probe überprüfen wir den Erwartungswert:

$$\begin{aligned}
\binom{N}{n} \cdot E(N^{(2)}) &= \sum_{x=n}^{2N-n} x \cdot H^{(2)}_{N,n}(x) \\
&= \sum_{x=n}^{N-1} x \cdot F(x,n) + N \cdot F(N,n) + \sum_{x=n}^{N-1} (2N-x) \cdot F(x,n) \\
&= \sum_{x=n}^{N-1} (x + 2N - x) \cdot F(x,n) + N \cdot F(N,n) \\
&= 2N \cdot F(N,n+1) + N \cdot F(N,n) = N \cdot \binom{N}{n},
\end{aligned}$$

erhalten also für den Erwartungswert den schon bekannten Wert N.

Auf die gleiche Weise wird jetzt die Varianz des Schätzers $N^{(2)}$ bestimmt:

$$\begin{aligned}
\binom{N}{n} \cdot \left[V(N^{(2)}) + N^2\right] &= \sum_{x=n}^{2N-n} x^2 \cdot H^{(2)}_{N,n}(x)) \\
&= \sum_{x=n}^{N-1} x^2 \cdot F(x,n) + N^2 \cdot F(N,n) + \sum_{x=n}^{N-1} (2N-x)^2 \cdot F(x,n)
\end{aligned}$$

$$\begin{aligned}\binom{N}{n}\cdot\left[V(N^{(2)})+N^2\right] &= N^2\cdot F(N,n)+\sum_{x=n}^{N-1}\left[4N^2-4Nx+2x^2\right]\cdot F(x,n)\\ &= N^2\cdot F(N,n)+2\cdot\sum_{x=n}^{N-1}\left[N^2+(N-x)^2\right]\cdot F(x,n)\\ &= N^2\cdot\left[F(N,n)+2\cdot F(N,n+1)\right]+2\cdot\sum_{x=n}^{N-1}(N-x)^2\cdot F(x,n).\end{aligned}$$

Nach Korollar 2 ist

$$F(N,n)+2\cdot F(N,n+1)=F(N+1,n+1)+F(N,n+1)=\binom{N}{n}.$$

Die übrige Summe wird mit Korollar 6 aus Anhang B für $m=N-1$ und $k=x$ ausgewertet:

$$\begin{aligned}\sum_{x=n}^{N-1}(N-x)^2\cdot F(x,n) &= 2\cdot F(N+2,n+3)-F(N+1,n+2)\\ &= 2\cdot F(N+1,n+3)+F(N+1,n+2)\\ &= F(N+1,n+3)+F(N+2,n+3)=\binom{N+1}{n+2},\end{aligned}$$

letzteres wieder nach Korollar 2. Das zusammengefasst ergibt nun

$$\begin{aligned}\binom{N}{n}\cdot\left[V(N^{(2)})+N^2\right] &= N^2\cdot\binom{N}{n}+2\cdot\binom{N+1}{n+2}\\ V(N^{(2)})+N^2 &= N^2+2\cdot\frac{n!\,(N-n)!}{N!}\cdot\frac{(N+1)!}{(n+2)!\,(N-n-1)!}\\ &= N^2+2\cdot\frac{(N+1)(N-n)}{(n+1)(n+2)}.\end{aligned}$$

Damit ist gezeigt:

Satz 9 *Die Varianz des Schätzers $N^{(2)}$ ist*

$$V(N^{(2)})=2\cdot\frac{(N+1)(N-n)}{(n+1)(n+2)}.$$

8 Die Verteilung des Mittelwerts

Um die Varianz von $N^{(3)}$ zu analysieren, starten wir mit der Häufigkeitsfunktion

$$\begin{aligned}G_{N,n}\colon\mathbb{Z} &\longrightarrow \mathbb{Z},\\ G_{N,n}(x) &= \#\{A\subseteq\{1,\ldots,N\}\mid \#A=n,\,\Sigma(A)=x\},\end{aligned}$$

Sie beschreibt die Verteilung der Mittelwerte von Teilmengen $A \subseteq \{1, \dots, N\}$ indirekt durch die Formel

$$q_{N,n}(t) = \frac{1}{\binom{N}{n}} \cdot G_{N,n}(nt) \quad \text{for } t \in \mathbb{R}.$$

Das ist der Anteil der insgesamt $\binom{N}{n}$ Teilmengen $A \subseteq \{1, \dots, N\}$ aus n Elementen, die die Summe nt bzw. den Mittelwert $t = \Sigma(A)/n$ haben.

Beispiel Für $N = 4$ und $n = 2$ sind die Teilmengen A:

$$\begin{aligned}
A &= \{1,2\} \quad \text{mit} \quad \Sigma(A) = 3,\\
A &= \{1,3\} \quad \text{mit} \quad \Sigma(A) = 4,\\
A &= \{1,4\} \quad \text{mit} \quad \Sigma(A) = 5,\\
A &= \{2,3\} \quad \text{mit} \quad \Sigma(A) = 5,\\
A &= \{2,4\} \quad \text{mit} \quad \Sigma(A) = 6,\\
A &= \{3,4\} \quad \text{mit} \quad \Sigma(A) = 7.
\end{aligned}$$

Die Häufigkeitsfunktion ist also gegeben durch

$$G_{4,2}(x) = \begin{cases} 2 & \text{für } x = 5,\\ 1 & \text{für } x = 3,4,6,7,\\ 0 & \text{sonst.} \end{cases}$$

Hier sind ein paar triviale Formeln

Hilfssatz 5 *Sei $N \in \mathbb{N}$. Dann gilt:*

(i) $G_{N,n}(x) = 0$ *konstant für* $n > N$.

(ii) $G_{N,0}(x) = 1$ *für* $x = 0$, *und* $G_{N,0}(x) = 0$ *für* $x \neq 0$.

(iii) $G_{N,1}(x) = 1$ *für* $1 \leq x \leq N$, *und* $G_{N,1}(x) = 0$ *für* $x \leq 0$ *oder* $x \geq N + 1$.

(iv) (*Symmetrie*) $G_{N,n}(x) = G_{N,n}(n\,(N+1) - x)$ *für alle* $x \in \mathbb{Z}$.

(v) (*Rekursion*) $G_{N,n}(x) = G_{N-1,n}(x) + G_{N-1,n-1}(x - N)$ *für* $1 \leq n < N$.

Beweis. (i) Es gibt keine Teilmengen mit n Elementen.
(ii) Die leere Menge hat die Summe 0.
(iii) Für eine einelementige Teilmenge A gilt $\Sigma(A) = x$ genau dann, wenn $A = \{x\}$.
(iv) Die bijektive Abbildung

$$\varphi\colon \{1, \dots, N\} \longrightarrow \{1, \dots, N\}, \quad a \mapsto N + 1 - a,$$

kehrt die Ordnung von $\{1, \dots, N\}$ um und induziert eine Bijektion

$$\Phi\colon \mathcal{P}(\{1, \dots, N\}) \longrightarrow \mathcal{P}(\{1, \dots, N\})$$

auf der Potenzmenge durch die Zuordnung $\Phi(\{a_1,\ldots,a_n\}) = \{\varphi a_1,\ldots,\varphi a_n\}$. Damit ist

$$\Sigma(\Phi(A)) = \sum_{a\in A} \varphi a = \sum_{a\in A} N+1-a = n\cdot(N+1) - \Sigma(A).$$

Also nimmt Σ den Wert $n\,(N+1)-x$ genau so oft an wie den Wert x.

(v) Sei $A \subseteq \{1,\ldots,N\}$ eine n-elementige Menge. Wir unterscheiden zwei Fälle, je nachdem, ob $N \in A$:

Fall 1. Die Teilmengen $A \subseteq \{1,\ldots,N-1\}$ tragen $G_{N-1,n}(x)$-mal die Summe x bei.

Fall 2. Die Teilmengen mit $N \in A$ haben die Summe $\Sigma(A) = \Sigma(A-\{N\}) + N$, tragen also $G_{N-1,n-1}(x-N)$-mal die Summe x bei.

$\diamond$

Als Anwendung der Rekursionsformel beweisen wir:

Hilfssatz 6 *Für $1 \le n \le N$ gilt*

$$\sum_{x\in\mathbb{Z}} x^2 \cdot G_{N,n}(x) = \frac{n\,(N+1)}{12}\cdot(3nN+N+2n)\cdot\binom{N}{n}.$$

Beweis. Wir starten den Induktionsbeweis bei $N=1$, also auch $n=1$. Dann hat die linke Seite genau einen nichtverschwindenden Summanden, nämlich: $1^2\cdot G_{1,1}(1) = 1$. Die rechte Seite ist $\frac{1\cdot 2}{12}\cdot(3+1+2)\cdot\binom{1}{1} = 1$.

Sei nun $N \ge 2$. Verwendet wird die Identität

$$x^2 = [(x-N)+N]^2 = (x-N)^2 + 2N\,(x-N) + N^2.$$

Mithilfe der Rekursionsformel folgt durch Induktion:

$$\begin{aligned}
\sum_{x\in\mathbb{Z}} x^2\cdot G_{N,n}(x) &= \sum_{x\in\mathbb{Z}} x^2\cdot G_{N-1,n}(x) + \sum_{x\in\mathbb{Z}} (x-N)^2\cdot G_{N-1,n-1}(x-N) \\
&\quad + 2N\cdot\sum_{x\in\mathbb{Z}} (x-N)\cdot G_{N-1,n-1}(x-N) + N^2\cdot\sum_{x\in\mathbb{Z}} G_{N-1,n-1}(x-N) \\
&= \frac{n\,N}{12}\cdot\binom{N-1}{n}\cdot[3n\,(N-1)+(N-1)+2n] \\
&\quad + \frac{(n-1)\,N}{12}\cdot\binom{N-1}{n-1}\cdot[3\,(n-1)(N-1)+(N-1)+2\,(n-1)] \\
&\quad + 2N\cdot\binom{N-1}{n-1}\cdot\frac{(n-1)\,N}{2} + N^2\cdot\binom{N-1}{n-1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{n\,N}{12}\cdot\frac{N-n}{N}\cdot\binom{N}{n}\cdot\underbrace{[3nN-3n+N-1+2n]}_{3nN+N-n-1}\\
&\quad+\frac{(n-1)\,N}{12}\cdot\frac{n}{N}\cdot\binom{N}{n}\cdot\underbrace{[3nN-3n-3N+3+N-1+2n-2]}_{3nN-2N-n}\\
&\quad+N\cdot\frac{n}{N}\cdot\binom{N}{n}\cdot[nN-N]+N^2\cdot\frac{n}{N}\cdot\binom{N}{n}\\
&= \frac{n}{12}\cdot\binom{N}{n}\cdot\left[3nN^2+N^2-nN-N-3n^2N-nN+n^2+n\right]\\
&\quad+\frac{n}{12}\cdot\binom{N}{n}\cdot\left[3n^2N-2nN-n^2-3nN+2N+n\right]\\
&\quad+\frac{n}{12}\cdot\binom{N}{n}\cdot[12nN-12N+12N]\\
&= \frac{n}{12}\cdot\binom{N}{n}\cdot\left[3nN^2+N^2+5nN+N+2n\right]\\
&= \frac{n}{12}\cdot\binom{N}{n}\cdot[3nN\,(N+1)+N\,(N+1)+2n\,(N+1)]\,.
\end{aligned}
$$

$\diamond$

Daraus kann man einen gechlossenen Ausdruck für die Varianz des Stichprobenmittels herleiten:

$$
\begin{aligned}
\sum_{t\in\mathbb{R}} t^2\,q_{N,n}(t)-\left(\frac{N+1}{2}\right)^2 &= \frac{1}{\binom{N}{n}}\sum_{t\in\mathbb{R}} t^2\,G_{N,n}(nt)-\left(\frac{N+1}{2}\right)^2\\
&= \frac{1}{\binom{N}{n}}\,\frac{1}{n^2}\underbrace{\sum_{x\in\mathbb{R}} x^2\,G_{N,n}(x)}_{\frac{n\,(N+1)}{12}\binom{N}{n}(3nN+N+2n)}-\left(\frac{N+1}{2}\right)^2\\
&= \frac{N+1}{12\,n}\cdot(3nN+N+2n)-\frac{N+1}{12\,n}\cdot(3nN+3n)\\
&= \frac{(N+1)(N-n)}{12\,n}.
\end{aligned}
$$

Damit ist als Ergebnis bewiesen:

Satz 10 *Let $1\le n\le N$. Mit $q_{N,n}$ sei die Verteilung der Mittelwerte der Stichproben vom Umfang n aus der Menge $\{1,\dots,N\}$ bezeichnet. Deren Varianz ist*

$$V_{N,n}=\frac{(N+1)(N-n)}{12\,n}.$$

Daraus gewinnen wir die Formel für die Varianz von $N^{(3)}$, da $N^{(3)}(A) = 2\times$ der Mittelwert von A, bis auf die Verschiebung um -1 . Diese spielt für die Varianz keine Rolle, und der Faktor 2 ergibt einen Faktor 4 im Vergleich zur Formel in Satz 10.

Korollar 1 *Die Varianz des Schätzers $N^{(3)}$ ist*

$$V(N^{(3)}) = \frac{(N+1)(N-n)}{3n}\,.$$

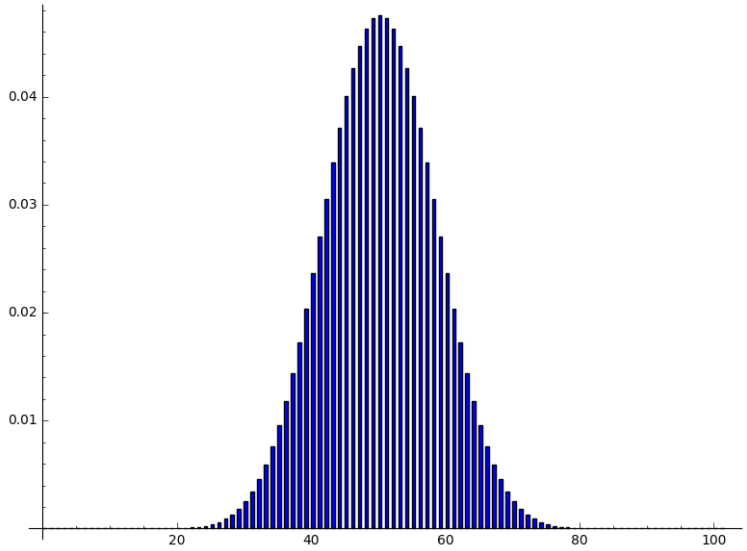

Abbildung 5: Die Verteilung des Schätzers $N^{(3)}$ für $N = 50$ und $n = 10$ (SageMath-Code in Anhang C.2)

Mithilfe der Rekursionsformel für G lässt sich auch das Histogramm der Verteilung $p^{(3)}_{N,n}$ von $N^{(3)}$ für konkrete, nicht allzu große Werte von N und n bestimmen, siehe Abbildung 5. Die Umrechnungsformel ist

$$p^{(3)}_{N,n}(t) = \frac{1}{\binom{N}{n}} \cdot G_{N,n}\left(\frac{n}{2}\,(t+1)\right) \quad \text{für } t \in \mathbb{R}.$$

Die angenommenen Werte $t = N^{(3)}(A)$ liegen diskret auf der reellen Achse, sind aber im Allgemeinen nicht ganzzahlig. Rundet man sie auch hier auf ganze Zahlen, so ist die Formel statt dessen

$$p^{(3a)}_{N,n}(s) = \frac{1}{\binom{N}{n}} \sum_{x\in\mathbb{N},\ s=\lfloor \frac{2}{n}\,x-1\rceil} G_{N,n}\left(\frac{n}{2}\,(t+1)\right) \quad \text{für } s \in \mathbb{N},$$

wobei $\lfloor\bullet\rceil$ die Rundung zur nächsten ganzen Zahl bedeutet. Mit diesem modifizierten Wert wurde Abbildung 5 erstellt. Leichte Unregelmäßigkeiten durch Rundungseffekte sind im Histogramm optisch nicht zu erkennen.

Schätzer	Erw.-wert	Varianz	Eigenschaften
$N^{(1)} = \text{Maximum}$	$(N+1)/\lambda$ (Satz 4)	$\frac{n\,(N+1)(N-n)}{(n+1)^2(n+2)}$ (Satz 5)	↯ erwartungstreu ↯ sehr linksschief
$N^{(2)} = \text{Max} + \text{Min} - 1$	N (Satz 1)	$\frac{2\,(N+1)(N-n)}{(n+1)(n+2)}$ (Satz 9)	symmetrisch erwartungstreu
$N^{(3)} = 2\times\text{Mittelwert} - 1$	N (Satz 2)	$\frac{(N+1)(N-n)}{3n}$ (Korollar zu Satz 10)	↯ Randbed. ↯ symmetrisch erwartungstreu breite Streuung
$N^{(4)} = \lambda\times\text{Max} - 1$	N (Satz 6)	$\frac{(N+1)(N-n)}{n\,(n+2)}$ (Satz 6)	erwartungstreu sehr linksschief

Tabelle 1: Eigenschaften der Schätzer ($\lambda = 1 + 1/n$)

9 Vergleich der Schätzer

Die bisherigen Erkenntnisse über die Schätzer $N^{(1)}$ bis $N^{(4)}$ sind in Tabelle 1 zusammengefasst. Bemerkenswert ist, dass die Varianz von $N^{(1)}$ und $N^{(4)}$ jeweils die Größenordnung $\approx N^2/n^2$ hat, die von $N^{(2)}$ die ungefähr doppelt so große $\approx 2\,N^2/n^2$, die von $N^{(3)}$ dagegen einen um den Faktor n größeren Wert $\approx N^2/n$.

Die Verteilung von $N^{(1)}$ ist explizit in Abschnitt 3 beschrieben und in Abbildung 1 durch ein Säulendiagramm veranschaulicht, die Verteilung von $N^{(4)}$ in Abschnitt 6 und Abbildung 2, die von $N^{(2)}$ in Abschnitt 8, die von $N^{(3)}$ in Abschnitt 7.

Die vergleichsweise geringe Varianz von $N^{(1)}$ ist für die Qualitätsbeurteilung allerdings wenig aussagekräftig, da dieser Schätzer ja verzerrt ist. Statt dessen sollte man zum Vergleich die mittlere quadratische Abweichung vom wahren Wert N heranziehen, also den Wert

$$\Delta(N,n) = \frac{1}{\binom{N}{n}} \cdot \underbrace{\sum_{\substack{A\subseteq\{1,\ldots,N\}\\ \#A=n}} \Big(N^{(1)}(A) - N\Big)^2}_{=:D(N,n)}.$$

Nach der Betrachtung in Abschnitt 3 nimmt $N^{(1)}(A) - N$ Werte von 0 bis $n - N$ an, und zwar den Wert $M - N$ genau $\binom{M-1}{n-1}$-fach. Damit wird die Summe $D(N,n)$ zu

$$\begin{aligned} D(N,n) &= \sum_{M=n}^{N} (N-m)^2 \cdot \binom{M-1}{n-1} \overset{i=N-M}{=} \sum_{i=0}^{N-n} i^2 \cdot \binom{N-i-1}{n-1} \\ &\overset{N'=N-1}{=} \sum_{i=1}^{N'-n+1} i^2 \cdot \binom{N'-i}{n-1} = \binom{N'+2}{n+2} + \binom{N'+1}{n+2} \\ &= \binom{N+1}{n+2} + \binom{N}{n+2}. \end{aligned}$$

Die mittlere quadratische Abweichung ist also

$$\begin{aligned}\Delta(N,n) &= \frac{\binom{N+1}{n+2}+\binom{N}{n+2}}{\binom{N}{n}} = \frac{(N+1)!\,n!\,(N-n)!}{(n+2)!\,(N-n-1)!\,N!} + \frac{N!\,n!\,(N-n)!}{(n+2)!\,(N-n-2)!\,N!} \\ &= \frac{N-n}{(n+2)(n+1)} \cdot \underbrace{\big[(N+1)+(N-n-1)\big]}_{2N-n}.\end{aligned}$$

Noch leichter intuitiv verständlich ist der mittlere absolute Fehler, definiert als der Erwartungswert des Absolutbetrags[9] der Abweichung der Schätzung vom wahren Wert, also in diesem Fall

$$E(|N^{(1)}-N|) = E(N-N^{(1)}) = N - \frac{n}{n+1}\cdot(N+1) = \frac{N-n}{n+1}\,.$$

Als Ergebnis zusammengefasst:

Satz 11 *Für gegebenes N und n hat der Schätzer $N^{(1)}$*

(i) *die mittlere quadratische Abweichung*

$$\Delta(N,n) = \frac{(N-n)(2N-n)}{(n+1)(n+2)}\,,$$

(ii) *den mittleren absoluten Fehler*

$$\delta^{(1)}(N,n) = \frac{N-n}{n+1}\,.$$

Die mittlere quadratische Abweichung von $N^{(1)}$ hat also die Größenordnung $\approx 2\,N^2/n^2$, ist also knapp doppelt so groß wie die Varianz, und $N^{(1)}$ in dieser Hinsicht nur unwesentlich genauer als $N^{(2)}$ und deutlich schlechter als $N^{(4)}$.

Für einen direkteren Vergleich ist natürlich der mittlere absolute Fehler am besten geeignet, wenn er sich denn in geeigneter Form explizit bestimmen lässt. Für $N^{(1)}$ war das trivial. Versuchen wir es mit $N^{(2)}$, wo wir die Symmetrie ausnützen können:

$$\begin{aligned}\binom{N}{n}\cdot E(|N^{(2)}-N|) &= \sum_{x=n}^{N-1}(N-x)\,F(x,n) + 0\cdot F(N,n) + \sum_{x=N+1}^{2N-n}(x-N)\,F(2N-x,n) \\ &= \sum_{x=n}^{N-1}(N-x)\,F(x,n) + \sum_{y=n}^{N-1}(N-y)\,F(y,n) \\ &= 2\cdot\sum_{x=n}^{N-1}(N-x)\,F(x,n) = 2\cdot F(N+1,n+2)\end{aligned}$$

mit der Funktion F aus Abschnitt 7, die nach Satz 8 ein Pascal-Tableau darstellt, wo bei der Umsummierung $y = 2N-x$ gesetzt und das Korollar 1 aus Anhang B verwendet wurde. Bewiesen ist damit:

[9] und daher oft unhandlich zu berechnen

	$N = 100$, $n = 10$, $r = 20$		
Schätzer	Mittelwert	Standardabw.	mittl. abs. Fehler
$N^{(1)}$	92.5 (91.8)	6.4 (7.9)	7.5 (8.2)
$N^{(2)}$	100.0 (100)	11.3 (11.7)	9.0 (8.7)
$N^{(3)}$	96.2 (100)	14.2 (17.4)	11.8 (?)
$N^{(4)}$	100.8 (100)	7.1 (8.7)	6.0 (?)
	$N = 10000$, $n = 50$, $r = 500$		
Schätzer	Mittelwert	Standardabw.	mittl. abs. Fehler
$N^{(1)}$	9791 (9805)	205 (191)	209 (195)
$N^{(2)}$	9984 (10000)	293 (274)	206 (196)
$N^{(3)}$	9950 (10000)	793 (814)	633 (?)
$N^{(4)}$	9986 (10000)	209 (195)	156 (?)

Tabelle 2: Empirische Kennzahlen der Schätzer (in Klammern der theoretische Wert) (SageMath-Code in Anhang C.3)

Satz 12 *Für gegebenes N und n hat der Schätzer $N^{(2)}$ den mittleren absoluten Fehler*

$$\delta^{(2)}(N,n) = 2 \cdot F(N+1, n+2) \Big/ \binom{N}{n}.$$

So elegant dieses Ergebnis auch aussieht – für den direkten Vergleich mit den anderen Schätzern ist es erst geeignet, wenn man daraus einen expliziten rationalen Ausdruck in N und n herleiten kann, ähnlich wie in Satz 11. Immerhin ist es zur Berechnung konkreter Werte bei nicht allzugroßem N geeignet.

10 Empirische Verteilung der Schätzwerte

Da die mittleren absoluten Fehler von $N^{(3)}$ und $N^{(4)}$ nicht theoretisch abgesichert sind, sollen ein paar numerische Experimente[10] einen ungefähren Eindruck von diesen Größen vermitteln. Dazu werden die Parameter $N = 100$ und $n = 10$ bei $r = 20$ Wiederholungen bzw. $N = 10000$, $n = 50$, $r = 500$ verwendet. Tabelle 2 enthält die Ergebnisse dieses Experiments. Auch sie weist den Schätzer $N^{(4)}$ als klaren Sieger aus. Dass $N^{(4)}$ eine geringfügig größere Standardabweichung hat als $N^{(1)}$, ist von vorneherein klar: auch diese wird ja mit $\lambda = 1 + 1/n$ skaliert, also mit 1.1 bzw. 1.02.

Die geringere Standardabweichung für $N^{(1)}$ ist aber ohnehin irrelevant, da sie wegen der Verzerrtheit von $N^{(1)}$ den Fehler unterschätzt. Nimmt man für $N^{(1)}$ statt dessen die Quadratwurzel der mittleren quadratischen Abweichung, so ist $N^{(1)}$ deutlich schlechter als $N^{(4)}$. Im Zahlenbeispiel mit $N = 100$, $n = 10$ wie in Tabelle 2 ist der theoretische Wert $\Delta(100, 10) \approx 129.5$ und seine Quadratwurzel ≈ 11.4. Im Vergleich mit der Standardabweichung von $N^{(2)}$ schneidet $N^{(1)}$ also geringfügig besser ab, und immer noch

[10] mit SageMath

Schätzer	Min	2.5%	25%	Median	75%	97.5%	Max
$N^{(1)}$	23	36	45	47	49	50	50
$N^{(2)}$	25	38	47	50	53	61	74
$N^{(3)}$	21	34	44	50	56	66	78
$N^{(4)}$	24	39	49	51	53	54	54

Tabelle 3: Empirische Quantile der Schätzer ($N = 50$, $n = 10$, $r = 1000$) (SageMath-Code in Anhang C.4)

deutlich besser als $N^{(3)}$. Diese Bewertung wird auch durch die Betrachtung der mittleren absoluten Fehler untermauert. Die Formel aus Satz 12 gibt im zweiten Fall den Wert

$$2 \cdot F(10001, 52) / \binom{10000}{50} \approx 195.6,$$

zu dessen Berechnung[11] die Definitionsformel von F verwendet wurde.

Für die graphische Darstellung der empirischen Verteilungen werden zur besseren Vergleichbarkeit mit den theoretischen Verteilungen in den Abbildungen 1 und 3 die Parameter $N = 50$ und $n = 10$ bei $r = 1000$ Wiederholungen gewählt. Die entsprechenden empirischen Verteilungen sind in den Abbildungen 6 bis 9 wiedergegeben. Die dabei beobachteten Quantile stehen in Tabelle 3 und werden in den Abbildungen durch Boxplots repräsentiert.

A Nachbetrachtung

Eine weitere Beobachtungsreihe im Kanton Schaffhausen führte zu 20 weiteren Fahrzeug-Kennzeichen:

SH 9039	SH 18973	SH 7577	SH 64079	SH 42871
SH 19244	SH 17132	SH 18222	SH 42457	SH 53074
SH 3965	SH 12101	SH 52115	SH 1104	SH 29826
SH 67732	SH 11455	SH 12571	SH 63162	SH 4520

Daraus ergeben sich (mit $n = 40$) verbesserte Schätzwerte:

$$\begin{aligned} N^{(1)} &= 67732 \\ N^{(2)} &= 68835 \\ N^{(3)} &= 72912 \\ N^{(4)} &= 69424 \end{aligned}$$

Diese liegen schon recht nah beieinander und lassen die Aussage wagen:

Im Kanton Schaffhausen gibt es knapp 69000 Kraftfahrzeuge.

[11] SageMath-Code im Anhang C.6

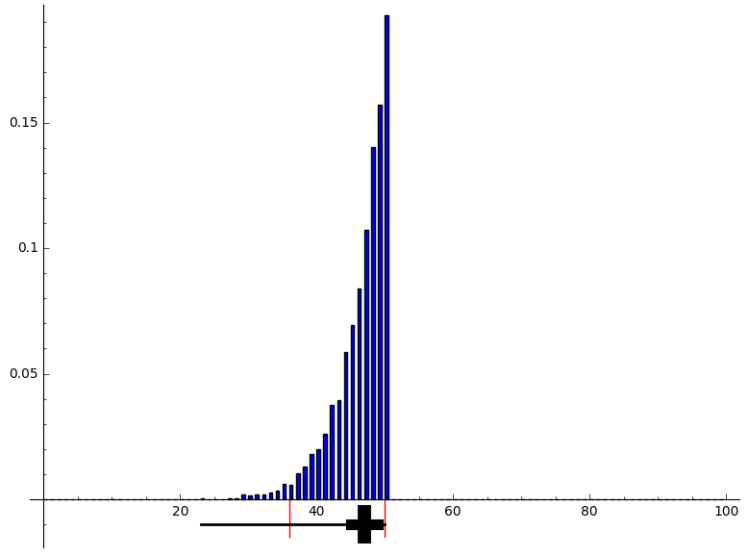

Abbildung 6: Empirische Verteilung von $N^{(1)}$; rot: 95%-Konfidenzintervall, schwarz: Boxplot (SageMath-Code in Anhang C.4)

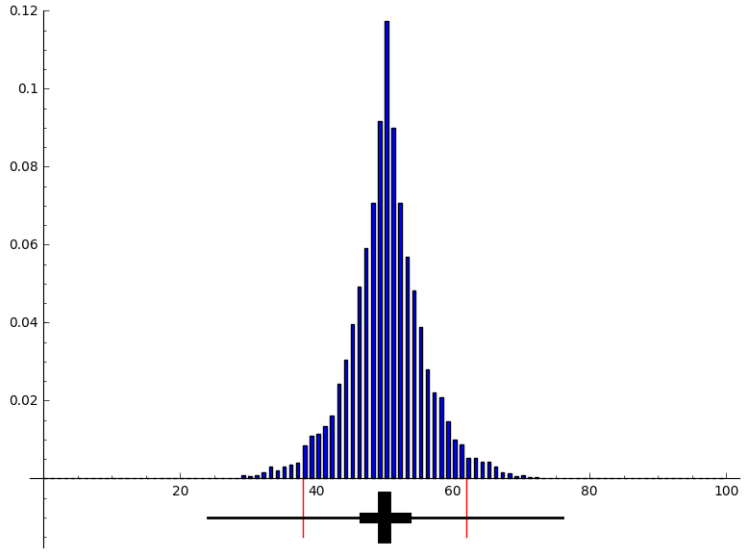

Abbildung 7: Empirische Verteilung von $N^{(2)}$ (SageMath-Code in Anhang C.4)

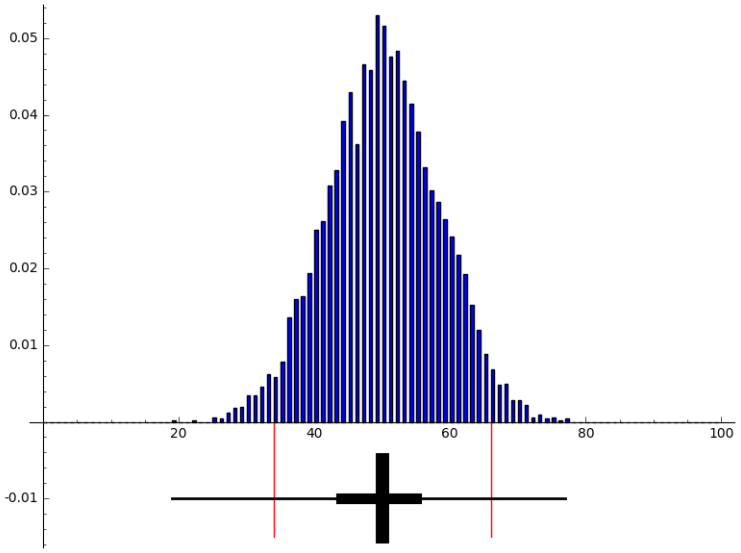

Abbildung 8: Empirische Verteilung von $N^{(3)}$ (SageMath-Code in Anhang C.4)

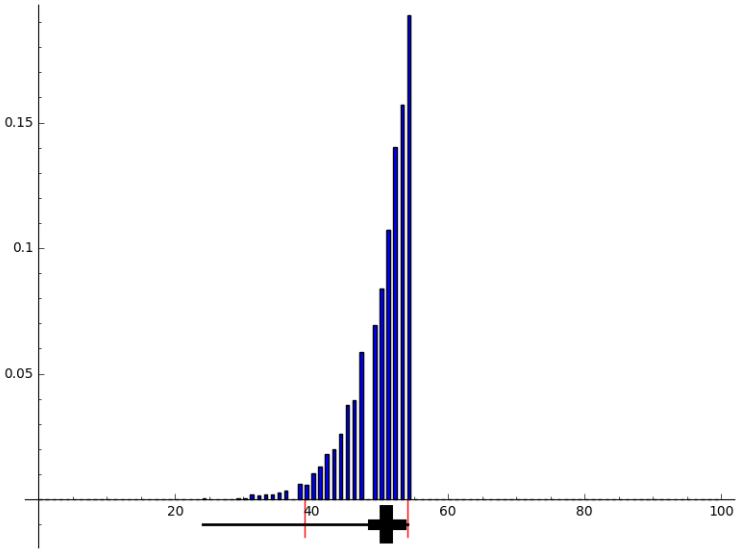

Abbildung 9: Empirische Verteilung von $N^{(4)}$ (SageMath-Code in Anhang C.4)

B Identitäten für Binomialkoeffizienten und Pascal-Tableaus

Die Abbildungen 10 und 11 illustrieren zwei Summationsformeln für Binomialkoeffizienten, die im folgenden in ziemlich allgemeiner Form bewiesen werden.

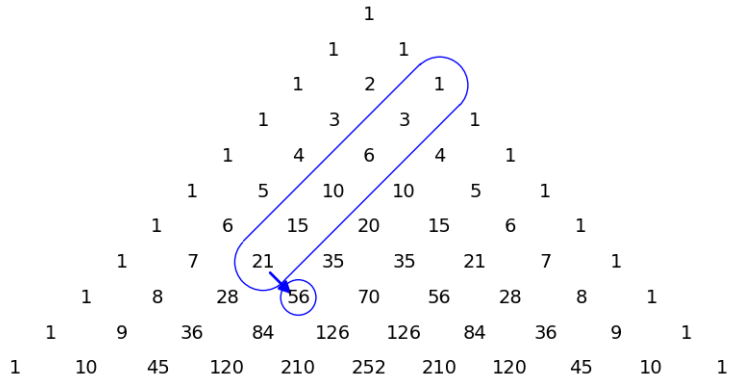

Abbildung 10: Pascalsches Dreieck mit der Beziehung $\binom{2}{2}+\binom{3}{2}+\binom{4}{2}+\binom{5}{2}+\binom{6}{2}+\binom{7}{2}=\binom{8}{3}$ (SageMath-Code in Anhang C.5)

Definition. Sei M ein $\mathbb{Z}$-Modul (bei Anwendungen ist fast immer $M=\mathbb{Z}$). Ein **Pascal-Tableau** in M ist eine Abbildung

$$T\colon \mathbb{N}\times\mathbb{N}\longrightarrow M$$

mit folgenden Eigenschaften, siehe Abbildung 12:

(i) $T(m,0)\in M$ beliebig für alle $m\in\mathbb{N}$.

(ii) $T(m,n)\in M$ beliebig für alle $n>m$.

(iii) $T(n,n)=a$ für alle $n\geq 1$ mit einem festen $a\in M$.

(iv) Für $m>n\geq 1$ gilt

$$T(m,n)=T(m-1,n-1)+T(m-1,n).$$

Beispiel. $M=\mathbb{Z}$, $T(m,n)=\binom{m}{n}$, $a=1$.

Die von Abbildung 10 illustrierte Regel sieht in diesem allgemeinen Rahmen nun so aus:

Hilfssatz 1 *Sei T ein Pascal-Tableau in M. Dann gilt für ganze Zahlen $m\geq n\geq 1$:*

$$\sum_{k=n}^{m} T(k,n)=T(m+1,n+1)\,.$$

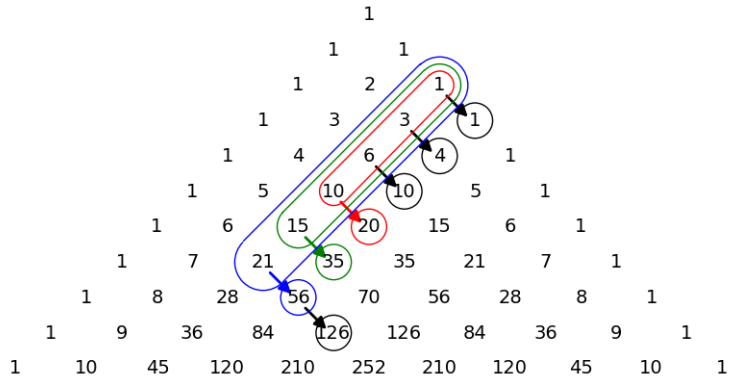

Abbildung 11: Pascalsches Dreieck mit der Beziehung $6 \cdot \binom{2}{2} + 5 \cdot \binom{3}{2} + 4 \cdot \binom{4}{2} + 3 \cdot \binom{5}{2} + 2 \cdot \binom{6}{2} + \binom{7}{2} = \binom{9}{4}$ (SageMath-Code in Anhang C.5)

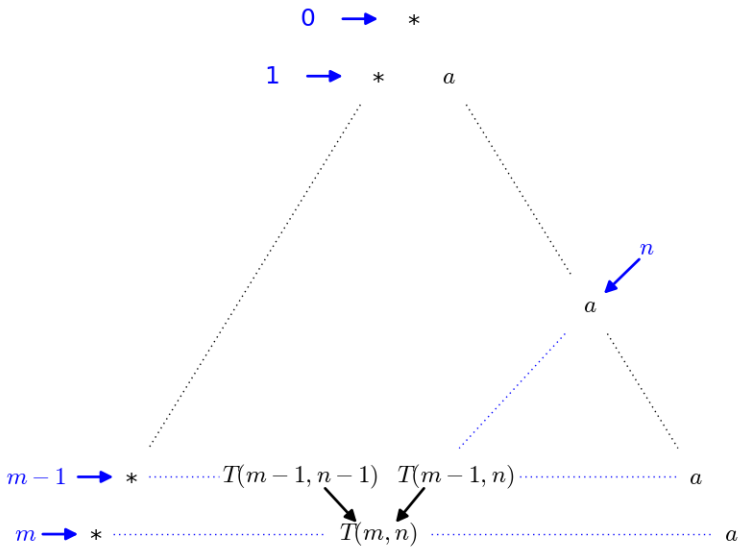

Abbildung 12: Ein Pascal-Tableau (SageMath-Code in Anhang C.5)

Beweis. Induktion über $m = n, n+1, \ldots$ bei festem n. Der Induktionsanfang $m = n$ ist trivial, denn die linke Seite ist $T(n,n) = a$, die rechte $T(n+1, n+1) = a$.

Sei nun $m \geq n+1$. Dann folgt mit Induktion

$$\sum_{k=n}^{m} T(k,n) = T(m,n) + \underbrace{\sum_{k=n}^{m-1} T(k,n)}_{=T(m,n+1)} = T(m+1, n+1).$$

$\diamond$

Dieses Beweismuster trägt aber, illustriert durch Abbildung 11, viel weiter:

Satz 1 *Sei T ein Pascal-Tableau in M. Dann gilt für ganze Zahlen $m \geq n \geq 1$ und $r \geq 0$:*

$$\sum_{k=n}^{m} \binom{m-k+r}{r} T(k,n) = T(m+r+1, n+r+1)\,.$$

Beweis. Durch doppelte Induktion über r und m. Für $r = 0$ ist die Aussage in Hilfssatz 1 bewiesen. Für $m = n$ (bei beliebigem r) ist sie trivial: denn auf der linken Seite steht $\binom{r}{r} T(n,n) = a$, auf der rechten $T(n+r+1, n+r+1) = a$.

Sei nun also $m \geq n+1$ und $r \geq 1$. Dann zerlegen wir die Summe

$$\sum_{k=n}^{m} \binom{m-k+r}{r} T(k,n) = \sum_{k=n}^{m} \left[\binom{m-k+r-1}{r-1} + \binom{m-k+r-1}{r} \right] T(k,n)$$

und werten die beiden Summanden einzeln aus:

$$\sum_{k=n}^{m} \binom{m-k+r-1}{r-1} T(k,n) = T(m+r, n+r)$$

nach Induktion über r,

$$\begin{aligned} \sum_{k=n}^{m} \binom{m-k+r-1}{r} T(k,n) &= \sum_{k=n}^{m-1} \binom{m-k+r-1}{r} T(k,n) \\ &= \sum_{k=n}^{q} \binom{q-k+r}{r} T(k,n) = T(q+r+1, n+r+1) \end{aligned}$$

nach Induktion über m (da $q = m-1$). Zusammengefasst ergibt die Summe

$$T(m+r, n+r) + T(m+r, n+r+1) = T(m+r+1, n+r+1)$$

nach dem Bildungsgesetz des Pascal-Tableaus. $\diamond$

Der Spezialfall $r = 0$ steht schon im Hilfssatz 1, die Spezialfälle $r = 1$ und $r = 2$ sehen explizit so aus:

Korollar 1 *In einem Pascal-Tableau T gilt für $m \geq n \geq 1$:*

$$\text{(i)} \qquad T(m+2, n+2) = \sum_{k=n}^{m} (m-k+1)\, T(k,n),$$

$$\text{(ii)} \qquad T(m+3, n+3) = \sum_{k=n}^{m} \frac{(m-k+1)(m-k+2)}{2}\, T(k,n).$$

Speziell für $n = 0$ ergibt der Satz eine Formel, die beliebige Einträge eines Pascal-Tableaus durch die „erste Spalte“ ausdrückt:

$$\sum_{k=0}^{m} \binom{m-k+r}{r} T(k,0) = T(m+r+1, r+1)\,.$$

Setzt man hierin $q = m+r+1$ und $n = r+1$ (mit geänderter Bedeutung des Buchstabens n), also $m = q-n$, wird die Formel zu:

Korollar 2 *In einem Pascal-Tableau T gilt für $q \geq n \geq 1$:*

$$T(q,n) = \sum_{k=0}^{q-n} \binom{q-k-1}{n-1} T(k,0)\,.$$

Auf die Binomialkoeffizienten $T(m,n) = \binom{m}{n}$ angewendet, ergibt Satz 1 die Formel

$$\sum_{k=n}^{m} \binom{m-k+r}{r} \binom{k}{n} = \binom{m+r+1}{n+r+1}.$$

Setzt man $N = m+1$ und $q = n+1$, so entsteht die Variante

$$\binom{N+r}{q+r} = \sum_{k=q-1}^{N-1} \binom{N-1-k+r}{r} \binom{k}{q-1} = \sum_{i=1}^{N-q+1} \binom{i-1+r}{r} \binom{N-i}{q-1},$$

also mit erneuter Umbenennung der Größen:

Korollar 3 *Für ganze Zahlen $N \geq n \geq 1$ und $r \geq 0$ gilt:*

$$\binom{N+r}{n+r} = \sum_{i=1}^{N-n+1} \binom{i+r-1}{r} \binom{N-i}{n-1}.$$

Die wichtigsten Spezialfälle $r = 0, 1, 2$ werden nochmal einzeln explizit formuliert:

Korollar 4 *Für ganze Zahlen $N \geq n \geq 1$ gilt:*

(i) $$\binom{N}{n} = \sum_{i=1}^{N-n+1} \binom{N-i}{n-1},$$

(ii) $$\binom{N+1}{n+1} = \sum_{i=1}^{N-n+1} i \cdot \binom{N-i}{n-1},$$

(iii) $$\binom{N+2}{n+2} = \sum_{i=1}^{N-n+1} \frac{i\,(i+1)}{2} \cdot \binom{N-i}{n-1}.$$

Da $i\,(i+1) = i^2 + i$, also $i^2 = 2 \cdot \frac{i\,(i+1)}{2} - i$, kann man die Formel (iii) noch etwas umformulieren:

$$\begin{aligned}
\sum_{i=1}^{N-n+1} i^2 \cdot \binom{N-i}{n-1} &= 2 \cdot \binom{N+2}{n+2} - \binom{N+1}{n+1} \\
&= 2 \cdot \binom{N+1}{n+1} + 2 \cdot \binom{N+1}{n+2} - \binom{N+1}{n+1} \\
&= \binom{N+2}{n+2} + \binom{N+1}{n+2},
\end{aligned}$$

erhält also als Ergebnis:

Korollar 5 *Für ganze Zahlen $N \geq n \geq 1$ gilt:*

$$\sum_{i=1}^{N-n+1} i^2 \cdot \binom{N-i}{n-1} = \binom{N+2}{n+2} + \binom{N+1}{n+2}.$$

Analog folgt aus Korollar 1 auch allgemein:

Korollar 6 *In einem Pascal-Tableau T gilt für $m \geq n \geq 1$:*

$$\begin{aligned}
\sum_{k=n}^{m} (m-k+1)^2\, T(k,n) &= 2\,T(m+3,n+3) - T(m+2,n+2) \\
&= T(m+3,n+3) + T(m+2,n+3)
\end{aligned}$$

C SageMath-Code

C.1 Verteilung der Schätzer 1, 2 und 4

```
NN = 50
n = 10
N4 = (NN*(1 + 1/n)).round()

p1list = [0]*(NN+1)
p4list = [0]*(N4+1)
q = binomial(NN,n)
for i in range(1,NN+1):
    r = binomial(i-1,n-1)
    p1list[i] = N(r/q)
    j = (i*(1 + 1/n)).round() - 1
    p4list[j] = N(r/q)
bar_chart(p1list)
bar_chart(p4list)

Hlist = [0] * n
for x in range(n,NN):
    sum = 0
    XX = floor((x-n)/2) + 1
    for m in range(1,XX+1):
        sum += binomial(x-2*m,n-2)
    Hlist.append(sum)
for x in range(NN+2, 2*NN-n+1):
    Hlist.append(Hlist[2*NN-x])
p2list = []
for i in range(len(Hlist)):
   p2list.append(Hlist[i]/binomial(NN,n))
bar_chart(p2list)
```

C.2 Verteilung des Schätzers 3

```
NN = 50
NMax = NN+1
nsize = 10
GNlist = [[1]]
for NN in range(1,NMax):
    NSq = int(NN*(NN+1)/2)
    NSr = int(NN*(NN-1)/2)
    Nulllist = [0]*(NSq + 1)
    Oldlist = copy(GNlist)
    GNlist = []
    for i in range(NN+1):
        GNlist.append(copy(Nulllist))
    GNlist[0][0] = 1
    GNlist[NN][NSq] = 1
    for n in range(1,NN):
        for x in range(NN):
            GNlist[n][x] = Oldlist[n][x]
        for x in range(NN,NSr+1):
            GNlist[n][x] = Oldlist[n][x] + Oldlist[n-1][x-NN]
        for x in range(NSr+1,NSq+1):
            GNlist[n][x] = Oldlist[n-1][x-NN]

lp = round(NMax*(NMax-1)/nsize)
p3list = [0]*(lp)
slist = []
for x in range(ll):
    s = round(2*x/nsize - 1)
    slist.append(s)
    if s >= 0:
        p3list[s] += GNlist[nsize][x]
Novern = binomial(NMax-1,nsize)
for s in range(lp):
    p3list[s] = N(p3list[s]/Novern)
bar_chart(p3list[:2*NMax])
```

C.3 Empirische Kennzahlen der Schätzer

```
NN = 10000
n = 50
r = 500
N1list = []
N2list = []
N3list = []
N4list = []
F1list = []
F2list = []
F3list = []
F4list = []

for i in range(r):
    Ch = sample(range(1,NN+1),n)
    M = max(Ch)
    mm = min(Ch)
    N1 = M
    N2 = M + mm - 1
    N3 = round(N(2*mean(Ch)-1),0)
    N4 = round(N(M*(n+1)/n - 1),0)
    F1 = abs(NN-N1)
    F2 = abs(NN-N2)
    F3 = abs(NN-N3)
    F4 = abs(NN-N4)
    N1list.append(N1)
    N2list.append(N2)
    N3list.append(N3)
    N4list.append(N4)
    F1list.append(F1)
    F2list.append(F2)
    F3list.append(F3)
    F4list.append(F4)

[N(mean(N1list)),N(std(N1list)),N(mean(F1list))]
[N(mean(N2list)),N(std(N2list)),N(mean(F2list))]
[N(mean(N3list)),N(std(N3list)),N(mean(F3list))]
[N(mean(N4list)),N(std(N4list)),N(mean(F4list))]
```

```
NN = 100
n = 10
r = 20
N1list = []
N2list = []
N3list = []
N4list = []
F1list = []
F2list = []
F3list = []
F4list = []

for i in range(r):
    Ch = sample(range(1,NN+1),n)
    M = max(Ch)
    mm = min(Ch)
    N1 = M
    N2 = M + mm - 1
    N3 = round(N(2*mean(Ch)-1),0)
    N4 = round(N(M*(n+1)/n - 1),0)
    F1 = abs(NN-N1)
    F2 = abs(NN-N2)
    F3 = abs(NN-N3)
    F4 = abs(NN-N4)
    N1list.append(N1)
    N2list.append(N2)
    N3list.append(N3)
    N4list.append(N4)
    F1list.append(F1)
    F2list.append(F2)
    F3list.append(F3)
    F4list.append(F4)

[N(mean(N1list)),N(std(N1list)),N(mean(F1list))]
[N(mean(N2list)),N(std(N2list)),N(mean(F2list))]
[N(mean(N3list)),N(std(N3list)),N(mean(F3list))]
[N(mean(N4list)),N(std(N4list)),N(mean(F4list))]
```

C.4 Empirische Verteilung der Schätzer

```
NN = 50
n = 10
r = 5000
N1list = []
N2list = []
N3list = []
N4list = []
N1ctr = [0]*(2*NN)
N2ctr = [0]*(2*NN)
N3ctr = [0]*(2*NN)
N4ctr = [0]*(2*NN)
N1err = 0
N2err = 0
N3err = 0
N4err = 0

for i in range(r):
    Ch = sample(range(1,NN+1),n)
    M = max(Ch)
    mm = min(Ch)
    N1 = M
    N2 = M + mm - 1
    N3 = round(N(2*mean(Ch)-1),0)
    N4 = round(N(M*(n+1)/n - 1),0)
    N1list.append(N1)
    N2list.append(N2)
    N3list.append(N3)
    N4list.append(N4)
    N1ctr[N1] += 1
    N2ctr[N2] += 1
    N3ctr[N3] += 1
    N4ctr[N4] += 1
    if N1 > NN:
        N1err += N1-NN
    else:
        N1err += NN-N1
    if N2 > NN:
        N2err += N2-NN
    else:
        N2err += NN-N2
```

```
    if N3 > NN:
        N3err += N3-NN
    else:
        N3err += NN-N3
    if N4 > NN:
        N4err += N4-NN
    else:
        N4err += NN-N4

def quantile(l,p):
    ll = sorted(l)
    nl = len(ll)
    np = floor(nl*p)
    if np == floor(nl*p):
        quant = N((ll[np-1] + ll[np])/2)
    else:
        quant = ll[np]
    return quant

N1quant = [min(N1list),quantile(N1list,0.025),quantile(N1list,0.25),
quantile(N1list,0.5),quantile(N1list,0.75),quantile(N1list,0.975),max(N1list)]
N2quant = [min(N2list),quantile(N2list,0.025),quantile(N2list,0.25),
quantile(N2list,0.5),quantile(N2list,0.75),quantile(N2list,0.975),max(N2list)]
N3quant = [min(N3list),quantile(N3list,0.025),quantile(N3list,0.25),
quantile(N3list,0.5),quantile(N3list,0.75),quantile(N3list,0.975),max(N3list)]
N4quant = [min(N4list),quantile(N4list,0.025),quantile(N4list,0.25),
quantile(N4list,0.5),quantile(N4list,0.75),quantile(N4list,0.975),max(N4list)]

N1quant
N2quant
N3quant
N4quant

N1distr = [0]*(2*NN)
N2distr = [0]*(2*NN)
N3distr = [0]*(2*NN)
N4distr = [0]*(2*NN)
```

```
for i in range(2*NN):
    N1distr[i] = N1ctr[i]/r
    N2distr[i] = N2ctr[i]/r
    N3distr[i] = N3ctr[i]/r
    N4distr[i] = N4ctr[i]/r

p1 = bar_chart(N1distr)
p1 += line([(N1quant[1],-0.015),(N1quant[1],0)],color="red")
p1 += line([(N1quant[5],-0.015),(N1quant[5],0)],color="red")
p1 += line([(N1quant[0],-0.01),(N1quant[6],-0.01)],color="black",thickness=2)
p1 += line([(N1quant[2],-0.01),(N1quant[4],-0.01)],color="black",thickness=8)
p1 += line([(N1quant[3],-0.015),(N1quant[3],-0.005)],color="black",thickness=10)
p1.show()

p2 = bar_chart(N2distr)
p2 += line([(N2quant[1],-0.015),(N2quant[1],0)],color="red")
p2 += line([(N2quant[5],-0.015),(N2quant[5],0)],color="red")
p2 += line([(N2quant[0],-0.01),(N2quant[6],-0.01)],color="black",thickness=2)
p2 += line([(N2quant[2],-0.01),(N2quant[4],-0.01)],color="black",thickness=8)
p2 += line([(N2quant[3],-0.015),(N2quant[3],-0.005)],color="black",thickness=10)
p2.show()

p3 = bar_chart(N3distr)
p3 += line([(N3quant[1],-0.015),(N3quant[1],0)],color="red")
p3 += line([(N3quant[5],-0.015),(N3quant[5],0)],color="red")
p3 += line([(N3quant[0],-0.01),(N3quant[6],-0.01)],color="black",thickness=2)
p3 += line([(N3quant[2],-0.01),(N3quant[4],-0.01)],color="black",thickness=8)
p3 += line([(N3quant[3],-0.015),(N3quant[3],-0.005)],color="black",thickness=10)
p3.show()

p4 = bar_chart(N4distr)
p4 += line([(N4quant[1],-0.015),(N4quant[1],0)],color="red")
p4 += line([(N4quant[5],-0.015),(N4quant[5],0)],color="red")
p4 += line([(N4quant[0],-0.01),(N4quant[6],-0.01)],color="black",thickness=2)
p4 += line([(N4quant[2],-0.01),(N4quant[4],-0.01)],color="black",thickness=8)
p4 += line([(N4quant[3],-0.015),(N4quant[3],-0.005)],color="black",thickness=10)
p4.show()
```

C.5 Pascal-Tableaus

```
nmax = 11

p3 = point((0,0), pointsize=0)
for i in range(nmax):
    for j in range(i+1):
        nn = binomial(i,j)
        p3 += text(str(nn),(-i+2*j,-i), fontsize=14, color = "black")
p3+= arc((2,-2),0.8,sector=(-pi/4,3*pi/4), color= "blue")
p3+= arc((-3,-7),0.8,sector=(3*pi/4,7*pi/4), color= "blue")
p3+= line([(1.45,-1.4),(-3.55,-6.4)], color= "blue")
p3+= line([(2.57,-2.55),(-2.43,-7.55)], color= "blue")
p3+= circle((-2,-8), 0.5, color= "blue")
p3+= arrow((-2.8,-7.3),(-2.2,-7.9), color= "blue")
p3.axes(False)
p3.show()

p4 = point((0,0), pointsize=0)
for i in range(nmax):
    for j in range(i+1):
        nn = binomial(i,j)
        p4 += text(str(nn),(-i+2*j,-i), fontsize=14, color = "black")
p4+= arc((2,-2),0.8,sector=(-pi/4,3*pi/4), color= "blue")
p4+= arc((-3,-7),0.8,sector=(3*pi/4,7*pi/4), color= "blue")
p4+= line([(1.45,-1.4),(-3.55,-6.4)], color= "blue")
p4+= line([(2.57,-2.55),(-2.43,-7.55)], color= "blue")
p4+= circle((-2,-8), 0.5, color= "blue")
p4+= arc((2,-2),0.6,sector=(-pi/4,3*pi/4), color= "green")
p4+= arc((-2,-6),0.6,sector=(3*pi/4,7*pi/4), color= "green")
p4+= line([(1.57,-1.57),(-2.45,-5.6)], color= "green")
p4+= line([(2.45,-2.4),(-1.6,-6.45)], color= "green")
p4+= circle((-1,-7), 0.5, color= "green")
p4+= arc((2,-2),0.4,sector=(-pi/4,3*pi/4), color= "red")
p4+= arc((-1,-5),0.4,sector=(3*pi/4,7*pi/4), color= "red")
p4+= line([(1.71,-1.71),(-1.27,-4.7)], color= "red")
p4+= line([(2.3,-2.25),(-0.7,-5.3)], color= "red")
p4+= circle((0,-6), 0.5, color= "red")
p4+= circle((1,-5), 0.5, color= "black")
p4+= circle((2,-4), 0.5, color= "black")
p4+= circle((3,-3), 0.5, color= "black")
p4+= arrow((-2.8,-7.3),(-2.2,-7.9), color= "blue")
p4+= arrow((-1.8,-6.3),(-1.2,-6.9), color= "green")
p4+= arrow((-0.8,-5.3),(-0.2,-5.9), color= "red")
```

```
p4+= arrow((0.2,-4.3),(0.8,-4.9), color= "black")
p4+= arrow((1.2,-3.3),(1.8,-3.9), color= "black")
p4+= arrow((2.2,-2.3),(2.8,-2.9), color= "black")
p4+= arrow((-1.8,-8.3),(-1.2,-8.9), color= "black")
p4+= circle((-1,-9), 0.5, color= "black")
p4.axes(False)
p4.show()

p5 = point((0,0), pointsize=0)
p5 += text("*",(0,-0.1), fontsize=18, color = "black")
p5 += arrow((-2,0),(-1,0), color = "blue")
p5 += text("0",(-3,0), fontsize=18, color = "blue")
p5 += text("*",(-1,-1.1), fontsize=18, color = "black")
p5 += arrow((-3,-1),(-2,-1), color = "blue")
p5 += text("1",(-4,-1), fontsize=18, color = "blue")
p5 += text("$a$",(1,-1), fontsize=18, color = "black")
p5 += line([(-1.5,-1.5),(-7.5,-7.5)], color="black", linestyle="dotted")
p5 += line([(1.5,-1.5),(4.5,-4.5)], color="black", linestyle="dotted")
p5 += line([(5.5,-5.5),(7.5,-7.5)], color="black", linestyle="dotted")
p5 += text("$a$",(5,-5), fontsize=18, color = "black")
p5 += line([(4.3,-5.5),(1.3,-7.5)], color="blue", linestyle="dotted")
p5 += text("$n$",(6.6,-4), fontsize=18, color = "blue")
p5 += arrow((6.4,-4.2),(5.4,-4.8), color = "blue")
p5 += text("*",(-8,-8.1), fontsize=18, color = "black")
p5 += text("$T(m-1,n-1)$",(-3.2,-8), fontsize=18, color = "black")
p5 += text("$T(m-1,n)$",(1.2,-8), fontsize=18, color = "black")
p5 += arrow((-2.5,-8.2),(-1.6,-8.8), color = "black")
p5 += arrow((0.3,-8.2),(-0.5,-8.8), color = "black")
p5 += text("$a$",(8,-8), fontsize=18, color = "black")
p5 += text("*",(-9,-9.1), fontsize=18, color = "black")
p5 += text("$T(m,n)$",(-1,-9), fontsize=18, color = "black")
p5 += text("$a$",(9,-9), fontsize=18, color = "black")
p5 += text("$m$",(-11,-9), fontsize=18, color = "blue")
p5 += arrow((-10.5,-9),(-9.5,-9), color = "blue")
p5 += text("$m-1$",(-10.7,-8), fontsize=18, color = "blue")
p5 += arrow((-9.5,-8),(-8.5,-8), color = "blue")
p5 += line([(-8.5,-9),(-2.5,-9)], color="blue", linestyle="dotted")
p5 += line([(0.5,-9),(8.5,-9)], color="blue", linestyle="dotted")
p5 += line([(-7.5,-8),(-5.5,-8)], color="blue", linestyle="dotted")
p5 += line([(3.0,-8),(7.5,-8)], color="blue", linestyle="dotted")
p5.axes(False)
p5.show()
```

C.6 Die Häufigkeitsfunktion F

```
def HF(k,n):
    if n < 0:
        return 0
    if n == 0:
        return 1 - 2*(k%2)
    if n == 1:
        return k%2
    q = 1 + floor((k-n)/2)
    hf = 0
    for i in range(1,q+1):
        hf += binomial(k-2*i,n-2)
    return hf

p3 = point((0,0), pointsize=0)
for i in range(11):
    for j in range(i+1):
        nn = HF(i,j)
        p3 += text(str(nn),(-i+2*j,-i), fontsize=14)
p3.axes(False)
p3.show()
```

Literatur

[1] R. Johnson: Estimating the size of a population. Teaching Statistics 16 (1994), 50–52

[2] D. Knuth: *The Art of Computer Programming*, Vol. 1. Addison-Wesley, Reading 1973 (2nd Ed.)

[3] G. E. Noether: *Introduction to Statistics – The Nonparametric Way*. Springer, New York 1990

[4] Wikipedia: Diskrete Gleichverteilung. `https://de.wikipedia.org/wiki/Diskrete_Gleichverteilung` [aufgerufen am 14. Juli 2020]

[5] Wikipedia: German tank problem. `https://de.wikipedia.org/wiki/German_tank_problem` [aufgerufen am 14. Juli 2020]