Die Suche nach emergenten deterministischen empirischen Gesetzen als Alternative zur wahrscheinlichkeitsbasierten Statistik

Vorläufige Version Dezember 2015

André Kuck und Hans Frischhut

1. Einleitung

"Another thing I must point out is that you cannot prove a vague theory wrong." Richard Feynman (1964) [1]

In der nachfolgenden Arbeit soll dem in den Wirtschaftswissenschaften üblichen, wahrscheinlichkeitsbasierten Gesetzesbegriff ein deterministischer empirischer Gesetzesbegriff gegenübergestellt werden.

Dabei wird gezeigt, dass die Verwendung des traditionellen stochastischen Gesetzesbegriffes dazu führt, dass Gesetzeshypothesen nicht objektiv und nicht endgültig falsifiziert werden können. Dies ist die Ursache für die Unschärfe vieler Aussagen aus dem Bereich der Wirtschaftswissenschaften und verhindert einen kumulativen Aufbau immer besser gesicherten Wissens. Auch wird so kein geeignetes Fundament zur definitiven und fehlerfreien Anwendung der zweiwertigen Logik auf empirische Fragestellungen konstruiert.

Als Vorschlag zur Lösung dieses fundamentalen methodischen Problems wird das Konzept des emergenten deterministischen empirischen Gesetzes vorgeschlagen, welches diese Mängel nicht aufweist. Die Verwendung emergenter deterministischer empirischer Gesetze erlaubt es einerseits, viele bisher wahre Gesetze zu finden und andrerseits, Hypothesen objektiv und endgültig zu falsifizieren. Dies hat zur Konsequenz, dass auch die empirische Wahrheit logischer Schlüsse mit zweiwertiger Logik zwar nicht für zukünftige Beobachtungen, zumindest aber in der Menge der bisherigen Beobachtungen einschließlich der bisher gemachten und überprüften Prognosen eindeutig wird. Außerdem wird der Aufbau einer Menge von bisher immer wahren Allaussagen über die Menge der bisher gemachten Beobachtungen – von Wissen – möglich.

Weiterhin wird gezeigt, dass der grundsätzliche Ansatz, bisher wahre empirische Gesetze durch einen Induktionsschluss zur Prognose zukünftiger Beobachtungen zu verwenden nicht nur logisch, sondern auch empirisch begründbar ist. Die messbare Güte von Prognosen mit bisher wahren Gesetzen steigt mit zunehmender Anzahl der Bestätigungen (DiV – Degree of Inductive Verifikation) dieser Gesetze an. Deshalb wird hier der DiV als zentrale Statistik zur Beurteilung der Güte von Gesetzen vorgeschlagen.

Ob die aus wahrscheinlichkeitsbasierten Gesetzesannahme abgeleitete Schätz- und Entscheidungsregeln (die im folgenden Heuristiken genannt werden) hingegen „gute" Ergebnisse liefern, ist immer eine empirische Frage, die nicht deduktiv aus Annahmen über die stochastische Struktur der Welt abgeleitet werden kann. Es wird die Auffassung vertreten, dass die Interpretation von aus wahrscheinlichkeitsbasierten Gesetzeshypothesen abgeleiteten Schätz- und Entscheidungsheuristiken als Gesetze nicht notwendig ist. Stattdessen sollte man auch die mit solchen Heuristiken gemachten Schätzungen auf empirische Gesetze über die relative Vorteilhaftigkeit von Sequenzen von Ergebnissen gegenüber alternativen Schätzheuristiken hin untersuchen. Dies ähnelt der z. B. im Maschinenlernen verwendeten Praxis, die so empirisch begründet und systematisch ausgearbeitet werden kann.

2. Konzeptuelle Grundlage der Suche nach emergenten Gesetzen

Der im Folgenden entwickelte empirische Gesetzesbegriff basiert auf der Generalisierung einfacher Methoden der beschreibenden Statistik. Dabei wird die Tatsache genutzt, dass bei geeigneter Setzung der Menge der zur Berechnung einer einzelnen Funktion verwendeten Beobachtungen auch für die Funktionen von Mengen von Beobachtungen der beschreibenden Statistik zahlreiche Beobachtungen generiert werden können.

Dadurch ist es auch in der Statistik wieder möglich, das grundlegende Konzept eines empirischen Gesetzes – einer bisher immer wahren und auch für Prognosen exakt überprüfbaren Allaussage – sinnvoll zu definieren und zu verwenden.

Um zu zeigen, dass bisher immer wahre Allaussagen über Eigenschaften von Mengen von Beobachtungen tatsächlich existieren, werden einige Beispiele in der nachfolgenden Arbeit gezeigt. Da solche immer wahren Allaussagen in der Realität häufig auftreten, ist es sinnvoll, sie als empirische deterministische Gesetze zu interpretieren und durch einen „Induktionsschluss" zur Prognose der noch nicht beobachteten Zukunft zu generalisieren.

Die so gewonnenen Hypothesen über zukünftige Eigenschaften von Mengen von Beobachtungen können dann nach jeder neuen Beobachtung zweifelsfrei überprüft und entweder bestätigt oder aber endgültig als falsch verworfen

[1] Dieses Zitat wurde David H. Baily, Jonathan M. Borwein, Marcos Lopez de Prado und Oiji Jim Zhu: Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance in: Notices of the AMS, Volume 61, No.5 (2014) entnommen.

werden.

Außerdem lassen sich solche Gesetze mittels einfacher Metriken zu Mengen „gleichartiger“ Gesetze zusammenfassen, deren bisherige Prognoseperformance dann als empirische Begründung für die in Zukunft zu erwartende Prognoseleistung von Gesetzen dieser Art verwendet werden kann.

2.1. Sequenzen von Kennzahlen der beschreibenden Statistik

In der deskriptiven Statistik werden Kennzahlen als Funktionen auf Basis einer Menge oder Sequenz von Beobachtungen (mit M_t als gemessener Merkmalsausprägung) mit ex-ante definierter Mächtigkeit T formuliert.

$$f(\{M_{t-T} \dots M_t\}) = f_{t,T}$$

Die Anzahl der Beobachtungen in der Stichprobe (T) stehen also ex-ante fest. Dabei wird T meist so gewählt (ohne dies explizit zu erwähnen), dass die Sequenz aller Beobachtungen $G = \{M_1 \dots M_t\}$ zur Berechnung des Funktionswertes zum aktuellen Zeitpunkt t verwendet wird ($t = T_G = T$). In diesem Fall hat man jeweils nur eine einzige Beobachtung des Funktionswertes der Mengenfunktion $f_{t,T}$.

Wählt man hingegen $T < T_G$, so erhält man eine Sequenz von Eigenschaften von Teilsequenzen von Beobachtungen. Man erhält für jeden Zeitpunkt $t > T$ jeweils einen Wert für $f_{t,T}$.

Beispiele solcher Funktionen sind u.a. gleitende Mittelwerte ($f_{1,t,T}$).

$$f_{1,t,T} = \bar{M}_{t,T} = \frac{1}{T}\sum_{\tau=t-T}^{t} M_\tau$$

Nimmt man z. B. als Merkmal den täglichen Temperaturanstieg von 11.00 Uhr bis 12.00 Uhr in Washington DC und berechnet die drei-Tages Mittelwert der Temperaturanstiege, so ergibt sich folgende Sequenz von Mittelwerten:

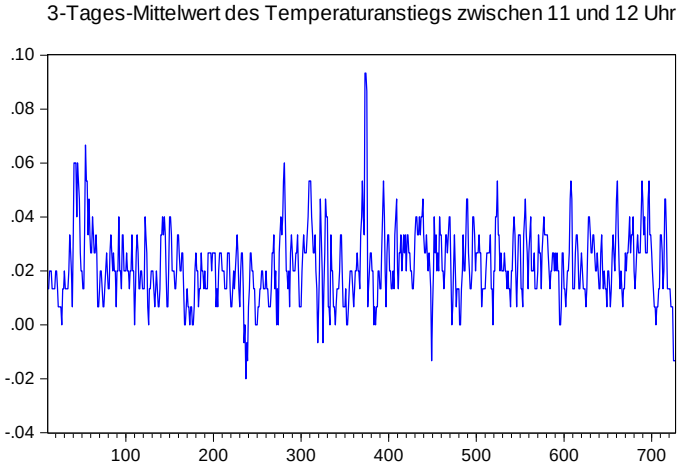

Abb. 1 3-Tages Gleitende Durchschnitte des Temperaturanstiegs zwischen 11 und 12 Uhr in Washington DC[2]

Aber auch Sequenzen von relativen Häufigkeiten ($f_{2,S,t,T}$) in einem bestimmten Intervall S (mit der Indikatorfunktion $I()$ sind sinnvoll berechenbar.

$$f_{2,S,t,T} = RelH_{S,t,T} = \frac{1}{T}\sum_{\tau=t-T}^{t} I(M_\tau \in S, 1, 0)$$

Nimmt man wieder die Temperaturanstiege in Washington DC, setzt $S = [0..\infty]$ und $T = 10$ (Berechnet also den Anteil der Tage in jedem 10 Tagesintervall mit einer Veränderung der Temperatur >=0) so erhält man das folgende Bild:

[2] Die Daten stammen aus Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3. (Verfügbar unter: https://archive.ics.uci.edu/)

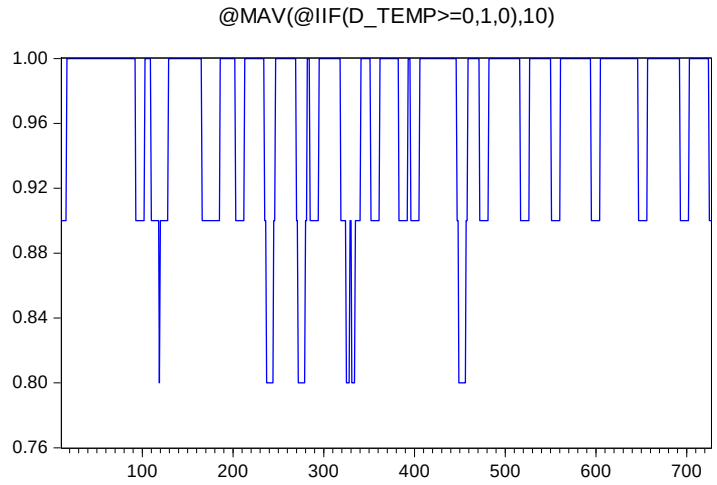

Abb. 2 10-Tages Gleitende Relative Häufigkeit von Temperaturanstiegen zwischen 11 und 12 Uhr in Washington DC

Die Menge der Funktionen, die hier verwendet werden können, ist beliebig erweiterbar. Wir werden später noch andere Beispiele vorstellen.

2.2. Wahre Allaussagen über Sequenzen statistischer Kennzahlen als emergente deterministische empirische Gesetze

Da für die Berechnung der jeweiligen Funktion f die Menge der verwendeten Beobachtungen für einen Zeitpunkt t auf $\{M_{t-T} \dots M_t\}$ beschränkt ist, können für $T_G > T$ mehrere Funktionswerte berechnet werden. Dies erlaubt es dann, sinnvolle Allaussagen über diese Funktion von Teilsequenzen von Messungen zu formulieren.

So kann zum Beispiel eine Aussage – eine Gesetzeshypothese H_T - über alle bisherigen Mittelwerte $f_{1,t,T}$ für alle $t > T$ (die ersten T Beobachtungen braucht man, um den ersten Wert von $f_{1,t,T}$ zu berechnen) der Indexmenge B (also für alle berechenbaren Beobachtungen) bei gegebenem T folgendermaßen formuliert werden:

$$H_T: \forall t > T \in B\big(I\big(f_{1,t,T} \geq 0{,}1{,}0\big) = 1\big)$$

„Für alle $t > T$ in der beobachteten Indexmenge B und bei gegebenem T gilt, dass der jeweilige Mittelwert $f_{1,t,T}$ immer größer 0 ist.“

Man kann dann im folgenden Schritt eine ganze Menge solcher Gesetzeshypothesen für unterschiedliche T bilden und dann überprüfen, ob tatsächlich ein T existiert, für das eine solche Allaussage wahr ist.

$$\{H_{T \leq \frac{T_G}{2}}: (\forall t > T \in B\big(I\big(f_{1,t,T} \geq 0{,}1{,}0\big) = 1\big))\}$$

Man „sucht“ empirische Gesetze aus einer Menge von Hypothesen über exakt berechenbare Eigenschaften (Funktionen) von (unterschiedlichen) Mengen von Beobachtungen – nicht indem man eine einzelne Hypothese überprüft.

Wenn man durch „Suchen“ in einer Menge von Hypothesen über beobachtbare Eigenschaften von Mengen von Beobachtungen wahre Allaussagen findet, dann wollen wir sagen, dass ein empirisches Gesetz **emergiert**. Man findet es nur dadurch, dass Funktionen über unterschiedliche Mengen von Beobachtungen gebildet und daraufhin untersucht werden, ob sie bisher immer wahr waren.

Im Beispiel findet man für $T = 9$ kein $t > T$ mehr, für das der mittlere Temperaturanstieg nicht positiv gewesen ist. Der Wert der Indikatorfunktion ist an jedem Tag 1.

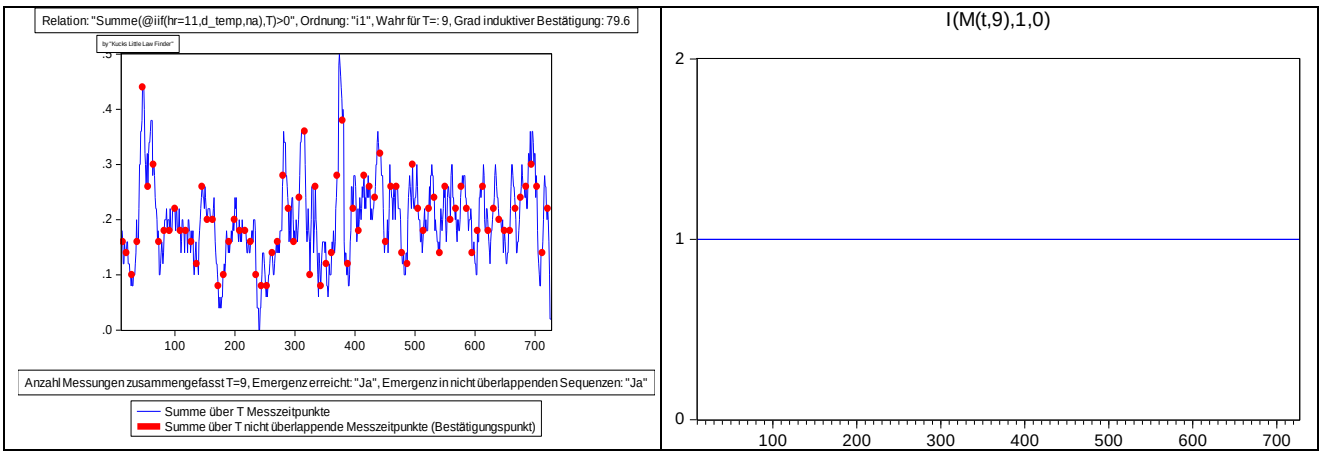

Abb. 3 Empirisches Gesetz über Summen von Temperaturanstiegen zwischen 11 und 12 Uhr in Washington DC

Genauso ist der Anteil der Tage mit Temperaturanstieg in jeder Zehn-Tages-Sequenz größer oder gleich 0.8 gewesen. (Vgl. Abb. 2)

Da die Gesetzeseigenschaft hier nicht für jede einzelne gemessene Merkmalsausprägung M_t gilt, sondern aktiv in einer Menge von Funktionen über Mengen von Beobachtungen gesucht werden muss, sprechen wir von einem **emergenten** empirischen Gesetz.[3]

Da diese Gesetze auch für jede einzelne Beobachtung der Mengenfunktion $f_{1,t,T}$ bisher wahr sind, sind sie dennoch **deterministisch**.

Dabei sollte $T \leq \frac{T_G}{2}$ sein, damit eine Aussage einmal als Hypothese empirisch gefunden und mindestens einmal in nicht überlappenden Stichproben bestätigt (bzw. falsifiziert) werden kann. Nur dann wollen wir davon sprechen, dass eine Hypothese „immer" wahr war.

Die Kennzahl

$$DiV = \frac{T_G}{T} - 1$$

die zählt, in wie vielen nicht überlappenden Stichproben der Gesamtmenge an Beobachtungen eine Aussage bisher bestätigt wurde, wird **Grad der induktiven Bestätigung** (Degree of inductive Verification DiV) genannt.

Im Beispiel ist die Hypothese, dass der Neuner-Mittelwert der Temperaturanstiege immer größer oder gleich Null ist genau $DiV = \frac{727}{9} - 1 = 79{,}6$ mal in nicht überlappenden Sequenzen bestätigt worden. Die Hypothese, dass der Anteil der Temperaturanstiege größer oder gleich 0,8 ist, ist $DiV = \frac{727}{10} - 1 = 71{,}7$ mal bestätigt.

Auch auf Finanzmärkten, bei denen üblicherweise zufälliges Verhalten der Beobachtungen unterstellt wird, gibt es eine große Zahl solcher bisher immer wahren Allaussagen über Funktionen von Mengen von Messungen.

Betrachtet man zum Beispiel die stetige Rendite $r_t = \ln(\frac{P_t}{P_{t-1)}})$ des S&P500[4] und berechnet die kumulierte Rendite über T Handelstage $f_{3,t,T} = \sum_{\tau=t-T}^{t} r_\tau$, und sucht in der Menge der Hypothesen:

$$\{H_{T<\frac{T_G}{2}}: \forall t > T \in B\big(I\big(f_{3,t,T} \geq 0{,}1{,}0\big) = 1\big)\}$$

nach wahren Gesetzen, so stellt man fest, dass für $T \geq 11231$ die untersuchten Relationen immer wahr waren.

---

[3] Für eine exakte Definition und eine weitergehende Diskussion vgl. Kuck A., Harries und Kuck E.(2015) „Der Weg zu wahren Gesetzen und rationalem Handeln; Durch Emergenz statt der Illusion metaphysischer Wahrheit zu empirischer Erkenntnis, S.20ff.

[4] Dieser Aktienindex kombiniert die Marktpreise von 500 großen US-amerikanischer Unternehmen.

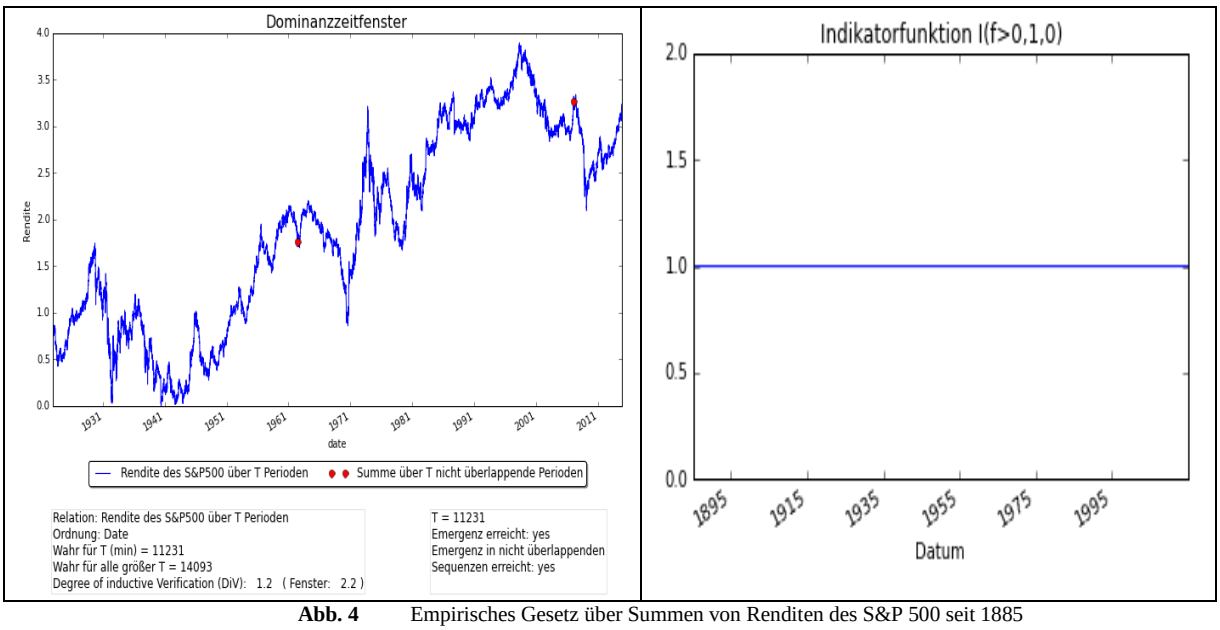

Abb. 4 Empirisches Gesetz über Summen von Renditen des S&P 500 seit 1885

Da man aus der Ordnung der Summe über T stetige Renditen auf die Ordnung der Kontostände nach T Perioden bei zum Zeitpunkt $t-T$ gleichem Kapitaleinsatz schließen kann, ist es bisher wahr gewesen, dass man nach T=11231 Tagen immer einen höheren Kontostand durch Anlage in den S&P500 erreicht hätte als durch Bargeldhaltung. Allerdings ist dieses Gesetz bisher nur $DiV = \frac{35651}{11231} - 1 = 1{,}17$ mal in nicht überlappenden Sequenzen bestätigt.

Viel häufiger bestätigte Gesetze für Finanzmärkte findet man für Bereiche, deren Entscheidungskonsequenzen nicht sofort klar sind.

Betrachtet man z. B. eine der empirischen Kovarianz sehr ähnliche Kennzahl (nur für die Berechnung des Mittelwertes wird jeweils eine etwas andere Stichprobe verwendet) zwischen dem S&P500 und dem Dow Jones Index.

$$f_{4,t,T,S\&P,DJ} = Cov_{t,T,S\&P,DJ} = \frac{1}{T}\sum_{\tau=t-T}^{T}(r_{\tau,S\&P} - \bar{r}_{\tau,T,S\&P}) \cdot (r_{\tau,Dj} - \bar{r}_{\tau,T,DJ})$$

und berücksichtigt, dass in der Definition des Korrelationskoeffizienten

$$\rho_{t,T,S\&P,DJ} = \frac{Cov_{t,T,S\&P,DJ}}{s_{t,T,S\&P} \cdot s_{t,T,DJ}}$$

der Nenner nicht negativ sein kann, dann sieht man sofort, dass die Korrelation positiv ist, wenn $Cov_{t,T,S\&P,DJ}$ positiv ist. Durchsucht man die Hypothesenmenge

$$\{H_{T<\frac{T_G}{2}}: \forall t > T \in B\big(I\big(Cov_{t,T,S\&P,DJ} \geq 0{,}1{,}0\big) = 1\big)\}$$

nach T‘s, für die die Beziehung bisher immer wahr war, so findet man schon für $T \geq 17$ nur noch Bestätigungen.

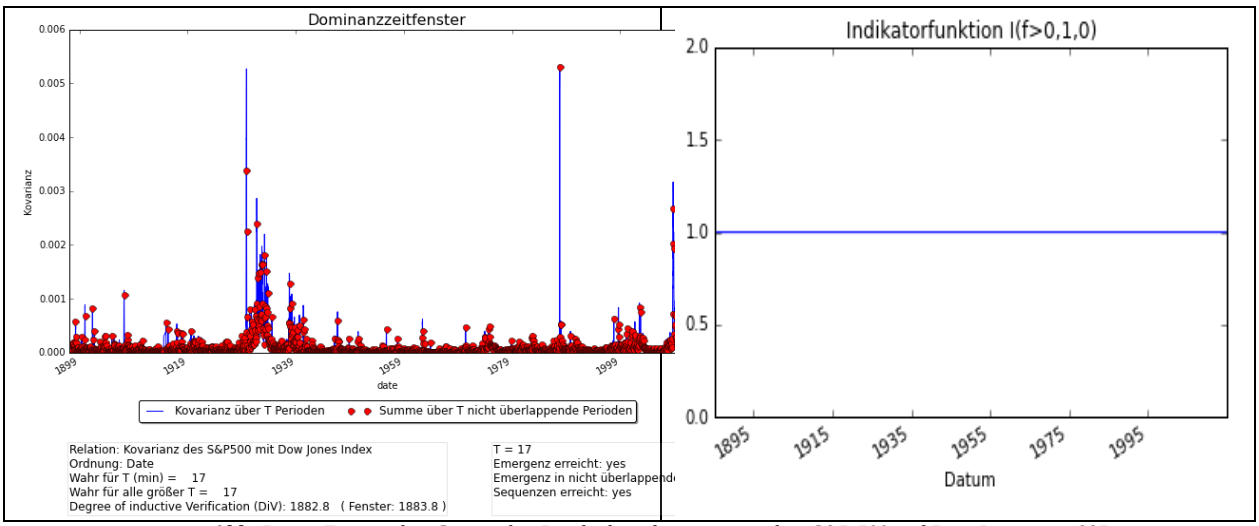

Abb. 5 Empirisches Gesetz über Renditekorrelationen zwischen S&P 500 und Dow Jones seit 1885

Es ist also tatsächlich so, dass es seit 1885 noch kein Fenster größer als 16 Tage gibt, für das die Korrelation zwischen den Renditen des S&P500 und des Dow Jones nicht größer oder zumindest gleich 0 gewesen wäre. Diese Beziehung ist $DiV = \frac{32042-17}{17} = 1882{,}82$ mal bestätigt worden.

Da beide Aussagen tatsächlich für alle t gelten, können sie mit Fug und Recht als deterministische empirische Gesetze bezeichnet werden.

In Übereinstimmung mit den wissenschaftstheoretischen Anforderungen an empirische Gesetze wird nicht die Gültigkeit des Gesetzes bis in alle Ewigkeit verlangt oder zwingend erwartet. Es reicht zunächst aus, dass die Aussage bisher immer wahr war.

2.3. Induktionsschluss und Prognose auf Basis empirisch wahrer Gesetze

Es scheint naheliegend, solche bisher immer wahren Gesetze zu verwenden, um auch zukünftige Eigenschaften von Mengen von Beobachtungen (die also nicht in der Menge der Beobachtungen G liegen) zu prognostizieren. Dazu bedarf es allerdings einer Generalisierung der Gesetze auf noch nicht gemachte Beobachtungen – einen Induktionsschluss.

Dazu wird der Anwendungsbereich des Gesetzes von der begrenzten Indexmenge der Beobachtungen $B = \{1 \dots T_G\}$ auf eine unendliche Indexmenge vergrößert, die auch noch nicht gemachte Beobachtungen enthält - $B^{\infty} = \{1 \dots T_G \dots \infty\}$.

So wird aus dem bisher wahren Gesetz (auch) eine **Gesetzeshypothese für zukünftige Beobachtungen**:

$$\forall t \in \{1 \dots T_G \dots \infty\}(I(f_{1,t,T} \geq 0{,}1{,}0) = 1)$$

Dabei bleibt die Gesetzeshypothese trotz dieses Induktionsschrittes und der Verwendung des metaphysischen Konzeptes „Unendlich“ nach wie vor auch zukünftig eindeutig überprüfbar. So liefert das Gesetz z. B. auch für das konkrete Intervall $T_G \dots T_G + T$ eine eindeutige, überprüfbare Prognose: $I(f_{1,T_G+T,T} \geq 0{,}1{,}0) = 1$.

Das Symbol „∞“ in der Beschreibung des Geltungsbereichs zeigt also lediglich die „Hoffnung“, dass dieses bisher gültige Gesetz auch für eine unbestimmte Zahl zukünftiger Beobachtungen gelten wird. Es bedeutet gerade nicht, dass die Wahrheit des Gesetzes erst im Unendlichen überprüft werden kann.

Für alle Indizes $t > T_G$ handelt es sich dementsprechend nicht um ein wahres Gesetz, sondern nur um eine erst in Zukunft überprüfbare Gesetzeshypothese.

Man glaubt oder hofft also, dass sich von der bisherigen universellen Wahrheit des Gesetzes auch auf seine zukünftige Wahrheit schließen lässt. Natürlich gibt es für die Richtigkeit dieses Schlusses keinen deduktiven „Beweis“. Es wird auch gar nicht behauptet, dass mit bisher wahren Gesetzen in Zukunft immer wahre Prognosen gemacht werden können. Es gibt keine Annahme metaphysischer Wahrheit empirischer Gesetze und daraus abgeleiteter Prognosen. Ganz im Gegenteil besteht bei jeder Prognose mit einem solchen Gesetz immer die Möglichkeit, dass es falsifiziert wird.

2.4. Begründung des auf empirischen Gesetzen basierenden Induktionsschlusses und des DiV als zentraler Gütestatistik

Dass es trotz des Fehlens eines deduktiven „Beweises“ Sinn macht, bisher wahre Allaussagen über Funktionen von Mengen von Beobachtungen zur Prognose zu verwenden, hat einen logischen und einen empirischen Grund.

Logisch führt eine Verwendung bisher wahrer empirischer Gesetze dazu, dass man zumindest in der Menge der bisherigen Beobachtungen auch richtige Schlüsse in die jeweilige Zukunft hätte ziehen können. Jede Prognose mit einem bisher wahren empirischen Gesetzes ist zumindest bisher immer wahr gewesen. Definitive Wahrheit von Aussagen und logischen Schlüssen bleibt zwar auf die Menge der Aussagen über Eigenschaften bisheriger Beobachtungen beschränkt, ist aber auch für bisherige Prognosen eindeutig und objektiv.

So kann nun empirisch sinnvoll auch die Frage gestellt werden, wie gut bisherige Prognosen mit DiV-mal bestätigten Gesetzen generell funktioniert haben. Dies ist möglich, da erstens die bisherige Anzahl Bestätigungen eine objektiv messbare Eigenschaft der Gesetze zu jedem Zeitpunkt $t-T$ ist. Zweitens ist die Prognose, die mit einem solchen Gesetz gemacht wird, objektiv nach weiteren T Beobachtungen überprüfbar. Man ist also in der Lage, neue statistische Merkmalsträger (Gesetze die zum Zeitpunkt $(t-T)$ DiV-mal bestätigt sind) und ein zu untersuchendes Merkmal (wird die Prognose nach weiteren T Beobachtungen zum Zeitpunkt t bestätigt, $P_{i,DiV(t-T),t,T}$) zu definieren.

So ist es möglich zu zählen, wie viele irgendwann DiV-mal bestätigten Gesetze man bisher gefunden hat und wie oft diese dann zutreffende Prognosen gemacht haben. Dabei wird unterstellt, dass zum Zeitpunkt t nur die letzten $(DiV+2)\cdot T$ Beobachtungen vorlagen. Das Gesetz war in dem eingeschränkten Zeitraum $t-T(DiV+2)\ldots t-T$ also immer wahr. Zusammenfassend wird eine statistische, die Güte bisheriger Prognosen mit DiV-mal bestätigten Gesetzen beschreibende Funktion konstruiert:

$$f_{H,t,T,DiV}=\frac{1}{N}\sum_{\tau=T+1}^{t}\sum_{i=1}^{N_\tau} I(P_{i,DiV(\tau-T),\tau,T}=wahr,1,0)$$

mit $I()$: Indikatorfunktion
N: Gesamtzahl aller DiV-mal bestätigten Gesetze die, zu allen Zeitpunkten $t-T$ gefunden wurden
N_τ: Anzahl der mit DiV-mal bestätigten Gesetzen gemachte Prognosen, die zum Zeitpunkt τ überprüft wurden

Dabei ist es egal, für welche Daten und welche Zeitpunkte die Gesetze gefunden wurden. Betrachtet man also Gesetze, die zu irgendeinem Zeitpunkt in der Vergangenheit einem bestimmten Grad an Bestätigung hatten, so kann man die bisherige Übertragbarkeit von Prognosen in die jeweilige Zukunft messen. Um dies anschaulich zu machen, soll nun eine Menge unterschiedlicher Zeitreihen auf Gesetze hin untersucht und die Ergebnisse insgesamt zusammengefasst werden.

Es sollen z. B. die Variablen $M_{i,hr}$: stündliche Veränderung zur Tageszeit hr der Temperaturen (temp: in Klammern die Variablennamen in der Datenbank) der gefühlten Temperatur (atemp) der Windgeschwindigkeit (windspeed) der Luftfeuchtigkeit (hum) der Anzahl verliehener Fahrräder (cnt) der an Vertragspartner verliehenen Fahrräder (registered) und der an nicht Vertragspartner verliehenen Fahrräder (casual) aus der schon weiter oben verwendeten Datenbank[5] untersucht werden. Dabei wird nach emergenten Gesetzen über das Vorzeichen der durchschnittlichen Veränderungen

$\forall t>T(I(\bar{M}_{i,hr,t,T}>0,1,0)=1)$ oder $\forall t>T(I(\bar{M}_{i,hr,t,T}<0,1,0)=1)$

in Sequenzen der Länge T=4 mit Bestätigung $DiV\in[1..50]$ gesucht.

Für die beschriebene Menge von Zeitreihen kann man sehr viele Hypothesen finden, die zu irgendeinem Zeitpunkt ein DiV-mal bestätigtes Gesetz waren. In Abhängigkeit von DiV ergibt sich folgende Anzahl gefundener Gesetze:

[5] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3. (Verfügbar unter: https://archive.ics.uci.edu/)

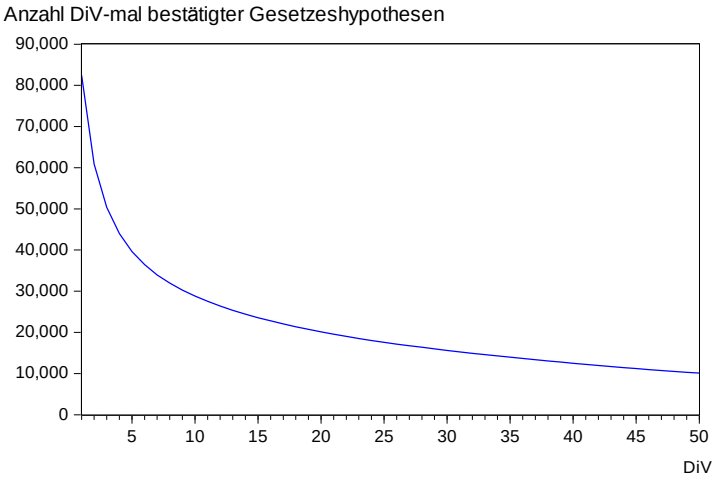

Abb. 6 Anzahl gefundener Gesetze nach Grad Induktiver Bestätigung der Gesetze

Überprüft man dann, wie viele der mit diesen Gesetzen erstellten Prognosen richtig waren, so erkennt man, dass die relative Häufigkeit richtiger Prognosen mit den Grad der Bestätigung ansteigt.

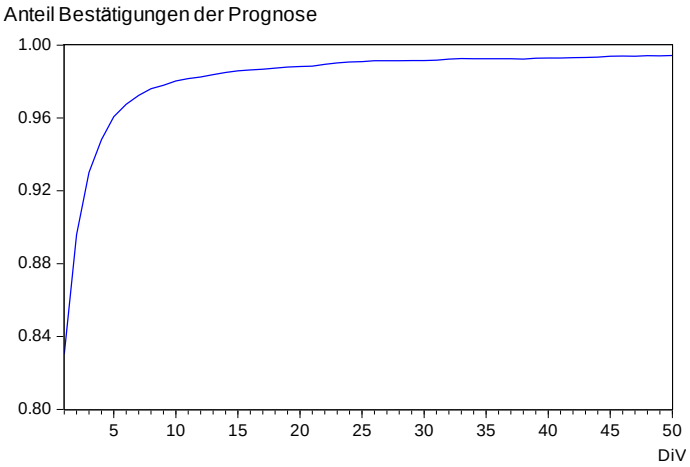

Abb. 7 Anteil bestätigter Prognosen nach Grad Induktiver Bestätigung der Gesetze

Es ist zwar nicht immer wahr, dass eine Prognose mit einem bisher wahren Gesetz auch selber wahr ist, aber der Anteil wahrer Prognosen steigt mit der bisherigen empirischen Bestätigung des zur Prognose verwendeten Gesetzes.
Es wäre also erstens sinnvoll gewesen, zur Prognose Regeln über Eigenschaften von Mengen von Beobachtungen zu verwenden, die bisher immer wahr waren. Zweitens ist es sinnvoll, möglichst gut bestätigte Gesetze zu verwenden.

Man kann sogar empirische Gesetze über die Prognosequalität empirischer Gesetze finden: Betrachtet man nur Prognosen mit mindesten 40-mal bestätigten Gesetzen, so gab es noch nie ein Zeitfenster von 200 Stunden, in dem nicht mindestens 90% der Prognosen bestätigt wurden.

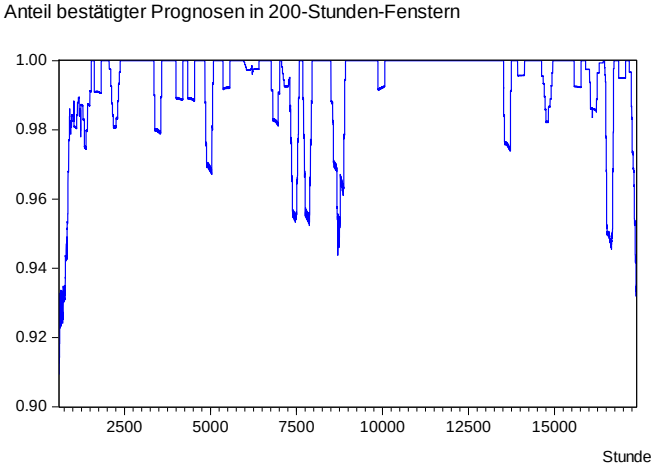

Abb. 8 Anteil bestätigter Prognosen von Gesetzen mit $DiV \geq 40$ in 200 Stunden-Fenstern

Zusammenfassend kann man also sagen, dass der hier vorgestellte Begriff des emergenten deterministischen empirischen Gesetzes eine Vorgehensweise erlaubt, die die folgenden Eigenschaften hat:

- Es existiert ein eindeutiges endgültiges Ausschlusskriterium für falsche Gesetzeshypothesen.
- Der Induktionsschluss von bisher immer wahren Gesetzen auf zukünftige Beobachtungen führt mit zunehmender Anzahl bisheriger Bestätigungen des Gesetzes empirisch zu einer größeren relativen Häufigkeit zutreffender Prognosen.
- Mit zunehmender Erfahrung „überleben" tendenziell Gesetze mit einer größeren Anzahl von Bestätigungen. Damit sind die überlebenden Gesetze immer besser bestätigt und man kann erwarten, dass Prognosen mit noch nicht falsifizierten Gesetzen immer besser werden.
- Es gibt eine klare Antwort auf die Frage nach der Verwendbarkeit von Logik auf empirische Fragestellungen. Richtiges logisches Schließen im Rahmen der zweiwertigen Logik ist mit bisher wahren Gesetzen in der Menge der bisherigen Beobachtungen auch für Prognosen möglich.

Die Formulierung von Gesetzeshypothesen über Eigenschaften von Mengen von Beobachtungen in einer Form, die ihre exakte empirische Überprüfbarkeit sicherstellt und der Verzicht auf den Anschein metaphysischer Wahrheit von Gesetzen erlauben also (unserer Einschätzung nach) die Lösung der methodischen Kernprobleme der Wirtschaftswissenschaften.
In Form von emergenten deterministischen empirischen Gesetzen kann Wissen akkumuliert werden. Nach unserer Definition bezieht sich Wissen nur auf die Vergangenheit und besteht aus Regeln, mit denen bisher immer wahre Prognosen gemacht werden konnten. Außerdem kann man Metawissen darüber generieren, wie gut Prognosen auf Basis solchen Wissens bisher funktioniert haben.
Natürlich ist es ein erstrebenswertes Ziel, dass man Gesetze mit einem Grad an Bestätigung findet, für den noch nie ein Gesetz nachfolgend falsifiziert wurde. Solche Gesetze sind (unserer Einschätzung nach) die größte Annäherung an metaphysische Wahrheit, die in empirischen Wissenschaften überhaupt möglich ist.

2.5. Heuristiken und Gesetze über ihre relative Vorteilhaftigkeit (T-Dominanz)

Im Folgenden soll deshalb von einer **Prognose** verlangt werden, dass sie mit einer Regel gemacht wird, die bisher immer richtige Prognosen generiert hat. Der Begriff Prognose wird also auf **Vorhersagen auf Basis bisher immer wahrer Gesetze** beschränkt. Es wird unterstellt, dass man nur in diesem Fall einen Grund hat zu glauben, dass auch die nächste Vorhersage wieder zutreffend sein wird. Diese Sichtweise ist wie oben gezeigt objektiv und erlaubt eine zukünftige exakte empirische Überprüfung einer Prognose.

Außerdem waren bisher auch alle logischen Folgerungen aus der Anwendung dieser Gesetze immer richtig – Prognosen mit einem wahren Gesetz hätten bisher das zutreffende logische Schließen in die jeweilige Zukunft erlaubt. Die Menge der uns bekannten Gesetze hätte bisher immer eine zutreffende Prognose der bisher gefundenen vorhersehbaren Zustände (als Eigenschaften von Mengen von Beobachtungen interpretiert) in unseren Daten erlaubt.

Hingegen soll eine **Regel**, deren Verwendung bisher **nicht immer richtige Vorhersagen** mit sich brachte, im Folgenden **Heuristik** genannt werden. Vorhersagen, die aus der Anwendung von Heuristiken stammen, sollen **Schätzungen** heißen. Da die Schätzungen noch nicht einmal in der Menge der bisherigen Beobachtungen richtig waren, soll auch ihre zukünftige Wahrheit nicht unterstellt werden. Logische Schlussfolgerungen aus Schätzungen wären auch in der Vergangenheit - wenn überhaupt - dann nur zufällig ab und an empirisch richtig gewesen.

Eine Schätzheuristik ist also z. B. eine bereits falsifizierte Regel der Art : $\forall t > T \in B\left(y_{t+i} = f_{t,T}\right)$.[6]

Die derzeitige Statistik verwendet (fast) nur solche bereits falsifizierten Vorhersageregeln. Es bedarf daher einer Methode, mit der man aus unterschiedlichen schon falsifizierten Heuristiken ggf. eine „beste" auswählen kann.

Emergente deterministische empirische Gesetze bieten in vielen Fällen eine Möglichkeit, objektiv und eindeutig die bisherige relative Vorteilhaftigkeit unterschiedlicher Schätzheuristiken - gegeben eine Bewertungsmetrik - zu beurteilen.

[6] Auf die Frage, warum es trotzdem sinnvoll und notwendig ist, solche Heuristiken zu verwenden, kann man nur dann eine Antwort finden, wenn man als ultimatives Ziel von Lernen das Lernen von Handlungsregeln unterstellt. „Die Notwendigkeit zu entscheiden reicht weiter, als die Möglichkeit zu erkennen." (Immanuel Kant) Die Anwendung von Entscheidungs- und Prognoseheuristiken ist genau deshalb notwendig, weil Menschen häufiger entscheiden müssen, als sie ihre Einzelfallentscheidungen mit bisher immer wahren Gesetzen über deren Ergebnisse begründen können.

Zwei häufig verwendete Regeln zur Schätzung sind z. B.:

- Morgen wird es genau so sein wie gestern: $\hat{y}_{1,t+1} = y_t$
- Der (rekursiv zu verwendende) KQ-Schätzer: $\hat{y}_{2,t+1} = \beta_{KQ,0,t} + \beta_{KQ,1,t} \cdot y_t$

mit

$$\beta_{KQ,1,t} = \frac{\sum_{\tau=1}^{t}(y_\tau - \bar{y}_\tau) \cdot (y_{\tau-1} - \bar{y}_{\tau-1})}{\sum_{\tau=1}^{t}(y_{\tau-1} - \bar{y}_{\tau-1})^2}$$

$$\bar{y}_t = \frac{1}{t}\sum_{\tau=1}^{t} y_\tau$$

$$\beta_{KQ,0,t} = \bar{y}_t - \beta_{KQ,1,t} \cdot \bar{x}_t$$

Beides sind (ohne stochastischen oder interpretatorischen Überbau) einfach Regeln zur Berechnung einer Zahl aus Beobachtungen, die bis zum Zeitpunkt t vorlagen. Dabei wird die berechnete Zahl jeweils dem nächsten Zeitpunkt t+1 zugewiesen.

Im schon oben verwendeten Beispiel der Temperaturänderungen in Washington DC ergeben sich exemplarisch folgende Werte für die Zeitpunkte t=211…215.

t	y_t	$\hat{y}_{1,t}$	$\beta_{KQ,0,t-1}$	$\beta_{KQ,1,t-1}$	$\hat{y}_{2,t}$
211	0.02	0.02	0.01941436	0.07428559	0.02090007
212	0.02	0.02	0.01940991	0.07429326	0.02089578
213	0	0.02	0.01940551	0.07430085	0.02089153
214	0.02	0	0.01930331	0.07447706	0.01930331
215	0.02	0.02	0.01930988	0.07431699	0.02079622

Nun stellt sich die Frage nach der Bedeutung der so berechneten Zahlen. In der üblichen Interpretation sind $\hat{y}_{1,t}$ und $\hat{y}_{2,t}$ zwei Prognosen, die zum Zeitpunkt $t-1$ gemacht wurden, um y_t zu prognostizieren. Man kann sofort erkennen, dass die Regel $\hat{y}_{2,t}$ zum Zeitpunkt t=211 und die Regel $\hat{y}_{1,t}$ zum Zeitpunkt t=213 aus den uns vorliegenden Beobachtungen falsifiziert werden.

Da auf diesen Regeln basierenden Prognosen für die Temperaturänderung in Washington DC zwischen 11 und 12 Uhr ab diesen Zeitpunkten mindestens einmal falsifiziert sind, können die Allaussagen, dass Vorhersagen mit diesen Regeln immer zu richtigen Prognosen führen, nie wieder wahr werden. Beide Regeln sind also endgültig falsifiziert.

Trotzdem ist die Frage sinnvoll, ob eine Regel immer bessere Ergebnisse liefert als die Alternative. Um dies zu beurteilen, benötigt man eine Metrik zur Bewertung der Güte der Schätzungen. In der Weiterführung des obigen Beispiels wird die Summe über die absoluten Schätzfehler in Fenstern der Größe T (für $i = 1{,}2$) verwendet.

$$S_{i,t,T} = \sum_{\tau-T}^{t} |y_\tau - \hat{y}_{i,\tau}|$$

Berechnet man diese Bewertungsmetrik für alle Fenster der Größe $T < TG/2$, dann stellt man fest, dass für alle $T \geq 27$ das Gesetz

$$\forall t(S_{1,t,T} > S_{2,t,T})$$

gilt. Man kann also sagen, dass in allen Fenstern der Größe $T \geq 27$ die Summe der absoluten Schätzfehler von Schätzungen mit der KQ-Regel geringer ist als die der Schätzung mit der einfachen „Morgen gleich Heute"-Regel. Dieses Ergebnis ist bisher $DiV = \frac{727}{27} - 1 = 25{,}9$ mal bestätigt. Diese Eigenschaft nennen wir **T-Dominanz**. In Bezug auf das Kriterium „Absoluter Schätzfehler" dominiert die KQ-Heuristik die einfache Schätzung mit dem Vortageswert. Sie war zwar nicht für jede einzelne Schätzung, aber bisher „immer" nach höchstens 27 Schätzungen insgesamt besser.

- Man kann also prognostizieren, dass auch in Zukunft immer nach 27 Schätzungen mit den beiden Verfahren das Gesamtergebnis der KQ-Heuristik nach dem Kriterium absoluter Schätzfehler insgesamt besser sein wird.
- Diese Prognose ist exakt überprüfbar und ggf. ist die Hypothese für immer falsifiziert, wenn das Ergebnis der KQ-Heuristik nach 27 weiteren Beobachtungen nicht besser ist.
- Bisher sind alle logischen Schlüsse über Prognosen mit diesem Gesetz richtig gewesen.
- Die zu erwartende Güte dieser Prognose kann auf Basis der Einordnung des Gesetzes in die Menge aller bisher

irgendwann einmal DiV-mal bestätigten Gesetze abgeschätzt werden.

- Und schließlich gibt es Gesetze über die Eigenschaften solcher DiV-mal bestätigten Gesetze.

Die Verwendung von empirischen Gesetzen über die relative Vorteilhaftigkeit von Schätzheuristiken bietet ein stimmiges, rein empirisches System ganz ohne Annahmen über die ewigen stochastischen Eigenschaften der Welt.

Man kann aus diesen Ergebnissen nicht schließen, dass die KQ-Heuristik auch bei anderen Schätzproblemen zu besseren Ergebnissen führt, weil es für sie „eine theoretische Begründung (z.B. BLUE-Eigenschaft von KQ-Schätzern)“ gibt. Empirisch kann man feststellen, dass bei Schätzung der Luftfeuchtigkeit (hum) $\hat{y}_{1,t+1}$ (die Schätzung mit dem Vortageswert) immer nach $T \geq 320$ Perioden die kleinere Summe absoluter Schätzfehler liefert. Hier T-dominiert also die Schätzung auf Basis des Vortageswerts.

Auch die Vorstellung, dass damit eine allgemeine Aussage über die Güte des Schätzverfahrens gegeben den Datensatz und das Schätzproblem für beliebige Bewertungsmetriken gemacht werden kann, ist falsch. Trotz der Tatsache, dass nach dem Kriterium Absoluter Prognosefehler die Schätzung auf Basis des Vortageswertes die KQ-Heuristik bei der Schätzung der Luftfeuchtigkeit T-dominiert, wird sie nach dem Kriterium „Quadrierter Absoluter Prognosefehler“ vom KQ-Verfahren T-dominiert.

Dies alles widerspricht der Vorstellung, dass es (in den meisten Fällen) ein „wahres“ stochastisches Modell zur Beschreibung der Welt gibt und die Aufgabe der Statistik darin besteht, dieses wahre Modell aus einer Menge von Modellen durch optimale Auswahlregeln „aufzudecken“.

Dieser Ansicht widerspricht auch die im Maschinenlernen gängige Praxis, Modelle nach Ihrer Prognosegüte (und nicht rein auf Basis deduzierter Optimalitätseigenschaften) auszuwählen.[7]

3. Empirische Interpretation frequentistischer, wahrscheinlichkeitsbasierter Statistik

„In everyday language we call random those phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking, there is no ground to believe that random phenomena should possess any definite probability. Therefore, we should distinguish between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory). There emerges the problem of finding reasons for the applicability of the mathematical theory of probability to the real world.”[8]

Das Konzept Wahrscheinlichkeit ist auch in der heutigen Anwendung zuerst einmal ein rein mathematisches Konzept. Die allgemein akzeptierte Grundlage dieses Konzepts sind die Kolmgorov‘schen Axiome.[9] Es liegt eine riesige Zahl von Theoremen vor, die aus diesen Axiomen deduziert wurden, und von denen in allen wissenschaftlichen Disziplinen reger Gebrauch gemacht wird.

Wahrscheinlichkeitsbegriffe - also Interpretationen dieses mathematischen Konzepts - sollten unserer Einschätzung nach eine definitive Antwort auf die Frage geben, unter welchen Bedingungen es richtig ist, den mathematischen Apparat der Stochastik auf konkrete empirische Fragestellungen anzuwenden. Wir wollen zeigen, dass die derzeit verwendeten Wahrscheinlichkeitskonzepte keine befriedigende Antwort auf diese Frage geben können.

Unsere Argumentation basiert jeweils auf der Frage, welche Beobachtungen eine Wahrscheinlichkeitsaussage empirisch wahr machen würden. Wir verlangen also von **sinnvollen Aussagen**, dass die **Regeln zu ihrer definitiven empirischen Überprüfung a-priori** formulierbar sind.

Ersetzt man in den Kolmogorov’schen Axiomen[10] Elementarereignisse durch Merkmalsausprägungen, den

[7] Vgl. z.B. Rich Caruana, Alexandru Niculescu-Mizil (2006): An Empirical Comparision of Supervised Lerning Algorthms oder Leo Breiman (2001): Statistical Modeling: The Two Cultures, in Statistical Science, Vol. 16, No. 3, S.199-231.

[8] Kolmogorov, zitiert nach Ming Li, Paul Vitányi: An Introduction to Kolmogorov Complexity and Its Applications, Second Edition, S. 52.

[9] Vgl. die Darstellung der Axiome z.B. in Ming Li, Paul Vitányi: An Introduction to Kolmogorov Complexity and Its Applications, Second Edition, S. 18. Vgl. auch Hájek, Alan, "Interpretations of Probability", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2012/entries/probability-interpret/, Kapitel 1.

[10] Vgl. die Darstellung der Axiome in Ming Li, Paul Vitányi: An Introduction to Kolmogorov Complexity and Its Applications, Second Edition, S. 18.

Stichprobenraum durch die Menge der beobachteten Merkmalsausprägungen, den Ereignisraum durch die Potenzmenge über der Menge der beobachteten Merkmalsausprägungen und jede Wahrscheinlichkeit durch eine relative Häufigkeit, so erkennt man, dass hier (auch) die Axiome zur Ableitung der Rechenregeln für relative Häufigkeiten zu finden sind. Wir unterstellen deshalb, dass die beobachtete relative Häufigkeit immer eine richtige empirische Übersetzung des Wahrscheinlichkeitsbegriffs ist. Inhaltlich sind die Unterschiede zwischen Wahrscheinlichkeiten und relativen Häufigkeiten dann die folgenden:

- Erstens werden nicht bisher beobachtete Merkmalsausprägungen, sondern in irgendeinem nicht explizit definierten Sinne mögliche Merkmalsausprägungen (Elementarereignisse) betrachtet.
- Zweitens wird keine explizite Regel zur Bestimmung der konkreten relativen Häufigkeiten angegeben.

Für eine empirische Interpretation von Wahrscheinlichkeiten muss man also unterstellen, dass man erstens alle Merkmalsausprägungen, die möglich sind, kennt (objektivistisch) bzw. man legt offen, welche Merkmalsausprägungen man subjektiv für möglich hält (subjektivistisch). Zweitens muss man a-priori definieren, welche beobachteten relativen Häufigkeiten die Wahrscheinlichkeitsaussagen wahr machen würden.[11]

Unser zentrales Argument gegen die Eignung von Wahrscheinlichkeiten als fundamentalem Konzept empirischer Wissenschaft ist es dann, dass es mit den derzeit verwendeten Interpretationen und Methoden nicht gelingen wird, eine objektive und endgültige Falsifikation wahrscheinlichkeitsbasierter, stochastischer Gesetzeshypothesen zu gewährleisten. Diese ist aber notwendig, um objektives definitives „Wissen“ zu generieren.

Dazu untersuchen wir zunächst die Sichtweise der Frequentisten. Wir zeigen zunächst das alt bekannte Problem, dass frequentistische stochastische Gesetzeshypothesen nicht suis generis exakt überprüfbar sind. Daher wird in statistischen Tests immer noch eine Idee von Jacob Bernoulli (die heute als Cournot‘s Prinzip bekannt ist) aus dem 18. Jahrhundert zur empirischen Überprüfung stochastischer Gesetzeshypothesen verwendet.

“An event with very small probability is morally impossible; it will not happen.”[12]

Wir werden argumentieren, dass die notwendige (immer aus nicht überprüfbaren Annahmen generierte) a-priori Auswahl der falsifizierenden - der moralisch unmöglichen - Beobachtung im statistischen Testen erstens nicht objektiv erfolgen kann und zweitens eine einmal erfolgte Falsifikation nicht zwingend endgültig ist. Außerdem zeigen wir, dass auch die von den Frequentisten verwendete Form des Induktionsschlusses zu nicht objektiv und endgültig überprüfbaren Gesetzeshypothesen führt.

3.1. Der frequentistische Wahrscheinlichkeits- und Gesetzesbegriff

Die frequentistische Wahrscheinlichkeitsinterpretation definiert Wahrscheinlichkeit als relative Häufigkeit von als Mengen von Elementarereignissen definierten Ereignissen. In ihrer empirischen Übersetzung sind Elementarereignisse (mögliche) Merkmalsausprägungen und Ereignisse Mengen von möglichen Merkmalsausprägungen. Es geht also um die relative Häufigkeit des Eintretens von Mengen von möglichen Merkmalsausprägungen. Dabei werden als Beispiele zur Begründung der Vorgehensweise meist konstruierte Zufallsexperimente wie der Münzwurf, der Würfelwurf oder das Roulett verwendet.

Die Aussage „Die Wahrscheinlichkeit für Ereignis $\{„Kopf“\}$ beim einfachen Münzwurf beträgt 50%“ wird in der klassischen oder Laplace’schen Interpretation durch die relative Häufigkeit der Elementarereignisse im Ereignis $\{„Kopf“\}$ an den möglichen Elementarereignissen im Stichprobenraum $S = \{„Kopf“, „Zahl“\}$ begründet.

$$Prob_K(A) = \frac{|A|}{|S|}$$

Natürlich ist allein durch die Definition der klassischen Wahrscheinlichkeit als relative Häufigkeit der Elemente von Mengen von Elementarereignissen am Stichprobenraum sichergestellt, dass das Rechnen mit so definierten Wahrscheinlichkeiten auf Basis von aus den kolmogorov’schen Axiomen abgeleiteten Theoremen konsistente Ergebnisse hat. An dieser Stelle können wir aber auch erkennen, dass sich auch in dieser Definition der

[11] In der frequentistischen Interpretation sind dies die relativen Häufigkeiten im Unendlichen. In subjektivistischer Interpretation gibt es gar keine empirische Bedeutung von Wahrscheinlichkeitsaussagen.

[12] Jacob Bernoulli: Ars Conjectandi, zitiert nach Glenn Shafer, Vladimir Vovk: The origins and legacy of Kolmogorov’s Grundbegriffe, S.7.

Wahrscheinlichkeit schon eine subjektive Komponente versteckt. Die Definition dessen, was man für möglich hält: $S = \{„Kopf“, „Zahl“\}$ ist keinesfalls definitiv. Es gibt keinen objektiven Grund das Elementarereignis „die Münze bleibt auf dem Rand liegen“ für unmöglich zu halten. In diesem Fall wäre $S_R = \{„Kopf“, „Zahl“, "Rand"\}$ und es würden sich ganz andere klassische Wahrscheinlichkeiten ergeben.[13]

Klassische Wahrscheinlichkeiten werden in der Praxis zur Prognose beobachtbarer relativer Häufigkeiten verwendet. Es geht also letztendlich um die Vorhersage der relativen Häufigkeit, mit der in Sequenzen von Münzwürfen (mit einer bestimmten Münze? Mit einer bestimmten Art von Münzen?) die Beobachtung a_t {„Kopf bei Wurf t“} beobachtet werden wird $A_T = \{a_1 \dots a_T\}$. Die klassische Wahrscheinlichkeit wird also dazu verwendet, die relative Häufigkeit des Eintritts von A zu schätzen.

$|A_T|$: Anzahl Würfe bei denen das Ereignis A: „Kopf“ beobachtet wurde.
T : Anzahl der beobachteten Würfe.
s : Schätzoperator, es wird keine exakte Gleichheit unterstellt.

$$Prob_K(A)\ s\ \frac{|A_T|}{T}$$

Eine Anweisung, die genau sagt, wann man eine solche Prognose als bestätigt und wann man sie als falsifiziert betrachten soll, ist hier nicht zu erkennen. Wenn man statt zu sagen, dass die klassische Wahrscheinlichkeit die relative Häufigkeit nur „schätzt“, behaupten würde, dass die beobachteten relativen Häufigkeiten (für alle T) genau gleich der Wahrscheinlichkeit sein sollen, so wäre die dann exakt überprüfbare Hypothese immer schnell falsifiziert. Das Konzept der klassischen Wahrscheinlichkeit ist in seiner Verwendung als Schätzregel für beobachtete relative Häufigkeiten von Ereignissen nicht überprüfbar und muss dies auch sein, um nicht immer falsifiziert zu werden. [14]

Deshalb ist es üblich, als „Bedeutung“ der Wahrscheinlichkeit den Grenzwert der relativen Häufigkeit im Unendlichen zu betrachten.

$$Prob_K(A)\ s\ Prob(A) = \lim_{T\to\infty} \frac{|A_T|}{T}$$

Die Wahrscheinlichkeit ist zwar die relative Häufigkeit des Ereignisses, aber die Gleichheit zwischen beobachteter relativer Häufigkeit und Wahrscheinlichkeit ist erst nach unendlich vielen Beobachtungen exakt.

Wahrscheinlichkeiten werden also erst als im Unendlichen beobachtbar definiert. Im Unendlichen sind dann alle möglichen Elementarereignisse wirklich eingetreten und somit zu Merkmalsausprägungen geworden und die relative Häufigkeit des Eintritts von Mengen von Merkmalsausprägungen ist genau gleich der Wahrscheinlichkeit des Ereignisses.

Im Umkehrschluss wäre die obige frequentistische Wahrscheinlichkeitsaussage auf Basis des klassischen Schätzprinzips daher nur dann wahr, wenn die sich nach unendlich vielen Beobachtungen ergebende empirisch beobachtete relative Häufigkeit $\lim_{T\to\infty} RelH_T(\{„Kopf“\})$ genau 50% betrüge. Dies ist aber prinzipiell nicht überprüfbar.

Da das einfache Abzählen der für möglich gehaltenen Elementarereignisse oft keine „guten“ Schätzungen für die relative Häufigkeit, mit der diese Elementarereignisse zukünftig auftreten werden, liefert, ist heute die gängige Wahrscheinlichkeitsinterpretation vom Zwang zur Schätzung durch Anteile an den möglichen Elementarereignissen befreit. Damit ist sie unabhängig von der zu ihrer Schätzung verwendeten Methode definiert als Grenzwert einer Mengenfunktion im Unendlichen:

---

[13] Aber auch der oft beschrittene Weg, den Stichprobenraum möglichst umfassend (alles was logisch möglich ist) zu definieren, führt oft zu leicht erkennbaren Fehlern. Nimmt man z. B. die der Normalverteilung zugrundeliegende Annahme, dass alle reellen Zahlen mögliche Elementarereignisse eines Zufallsexperiments sind, so macht man in ihrer Anwendung z. B. auf diskrete Renditen von Aktien einige logische Fehler. So sind zuerst einmal diskrete Renditen kleiner als -1 aus institutionellen Gründen ausgeschlossen. Wenn man außerdem nur ganze Zahlen (z. B. Preise in Ct.) als Aktienpreise verwendet, so sind schon aus logischen Gründen nur rationale Zahlen als Ergebnis einer konkreten Renditeberechnung möglich. Was im Unendlichen alles passiert sein wird, ist unserer Einschätzung nach prinzipiell nicht objektiv erkennbar.

[14] Da wir die Möglichkeit der a-priori Angabe der die Aussage verifizierenden (falsifizierenden) Beobachtungen verlangen, ist die Interpretation als “physikalische” Wahrheit oder als “Propensity” (eine Neigung hin zu einer relativen Häufigkeit) nicht wirklich von Bedeutung. (Für eine Diskussion dieser Begriffe z. B. Vgl. Hájek, Alan, "Interpretations of Probability", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2012/entries/probability-interpret/).

$$Prob(A) = \lim_{T\to\infty} \frac{|A_T|}{T} = \lim_{T\to\infty} f_{t,T} = E(f)$$

Diese Interpretation frequentistischer Aussagen als Hypothesen über Grenzwerte von Funktionen von Mengen von Beobachtungen im Unendlichen lässt sich auf alle stochastischen Hypothesen generalisieren.

Die exakte Wahrheit einer stochastischen Hypothese kann erst überprüft werden, wenn sich die Zahl der verwendeten Beobachtungen T dem Unendlichen annähert:

$$\lim_{T\to\infty} f_{t,T} = E(f)$$

Soll die stochastische Aussage als Allaussage, also als Gesetz gelten, so muss der Zusammenhang immer (also für alle Zeitpunkte t) gelten, wenn unendlich viele Beobachtungen T zur Berechnung verwendet wurden:

$$\forall t(\lim_{T\to\infty} f_{t,T} = E(f))$$

Erst wenn man für unendlich viele Beobachtungen den Mittelwert gebildet hat, behauptet man exakt auf den Mittelwert der nächsten unendlich vielen Beobachtungen schließen zu können.

Natürlich lässt sich dies in keiner Weise empirisch überprüfen, da nicht unendlich viele Beobachtungen (bzw. für verschiedene Zeitpunkte t jeweils unendliche Sequenzen von Beobachtungen) vorliegen.

„The frequency concept based on the notion of limiting frequency as the number of trials increases to infinity does not contribute anything to substantiate the application of the results of probability theory to real practical problems where we always have to deal with a finite number of trials."[15]

Stochastische Gesetzeshypothesen basieren also erstens immer auf einer subjektiven Einschätzung, was im Unendlichen alles passiert sein wird und entziehen sich zweitens (bisher immer) jeglicher empirischen Überprüfung. Sie sind in der empirischen Anwendung nicht im Rahmen der zweiwertigen Logik behandelbar. Man muss sie deshalb in diesem Sinne als subjektive, metaphysische Aussagen betrachten.

3.2. Cournot‘s Prinzip und statistisches Testen

Unserer Kenntnis nach wird die prinzipielle Subjektivität der Konstruktion des Stichprobenraums in der frequentistischen (objektiven) Statistik nicht weiter beachtet. Hingegen ist die fehlende Möglichkeit zur empirischen Überprüfung stochastischer Hypothesen ein altbekanntes Problem. Die auch derzeit praktizierte Vorgehensweise lässt sich mit folgendem Zitat kompakt beschreiben:

"Popper taught that in general scientific theories make contact with reality by providing opportunities for falsification. Cournot's principle tells us how to find such opportunities with a probabilistic theory: single out an event of very small probability and see if it happens."[16]

Man kann statistische Tests als die moderne Version der Anwendung des Cournot'schen Prinzips interpretieren. Sie sollen überprüfen, ob eine stochastische Gesetzeshypothese H_0 mit den gemachten Beobachtungen in dem Sinne vereinbar ist, das die beobachteten Daten nicht zu „extrem" und damit zu unwahrscheinlich (bezogen auf ein bestimmtes „Signifikanzniveau" α) sind, wenn die stochastische Gesetzeshypothese wahr wäre.

Es soll also die Frage beantwortet werden, ob die vorliegenden Beobachtungen bei Geltung des stochastischen Gesetzes „moralisch unmöglich" sind.

So wird auf Basis der Hypothese H_0 (und J immer zusätzlich zur Berechnung der Wahrscheinlichkeiten notwendigen Annahmen HZ_j) ein hypothetisches Ereignis A mit

$$A = \{A_i = \{M_{i,1} \dots . M_{i,T=t}\} | f_{H_0,HZ_1,..HZ_J}(M_{i,1} \dots . M_{i,T=t}) < \alpha\}$$

[15] Kolmogorov, zitiert nach: Ming Li, Paul Vitányi: An Introduction to Kolmogorov Complexity and Its Applications, Second Edition, S.55.

[16] Glenn Shafer, Vladimir Vovk (2003) The origins and legacy of Kolmogorov's Grundbegriffe, S. 63.

konstruiert, dessen Beobachtung uns dazu veranlasst, die Hypothese als falsch einzustufen.

Ein Beispiel sei der Test auf den Mittelwert für die Veränderung der Temperatur in Washington DC von 14.00-15.00 Uhr. Die Hypothese sei, dass der Mittelwert im Unendlichen gleich 0 ist. ($H_0: E(M) = 0$). Zusätzlich wird davon ausgegangen, dass M normalverteilt sei $HZ_1: M \approx N(0, \sigma_M)$ mit der korrigierten empirischen Standardabweichung als Schätzwert für die Wurzel aus der mittleren quadrierten Abweichung vom Mittelwert im Unendlichen ($HZ_2: \sigma_M = \sqrt{\frac{1}{t-1}\sum_{\tau=1}^{t}(M_t - \bar{M}_t)^2}$). Weiterhin soll davon ausgegangen werden, dass M iid ist und daraus folge, dass die Standardabweichung des Stichprobenmittelwerts im Unendlichen $HZ_3: \sigma_{\bar{M}_t} = \frac{1}{\sqrt{t}}\sigma_M$ ist. Weiterhin soll die Grenze für moralische Unmöglichkeit bei $\alpha = 0.01$ gezogen werden. Zusammenfassend machen wir also folgende a-priori Setzungen:

$$H_0: E(M) = 0,\ HZ_1: M \approx N(0, \sigma_M),\ HZ_2: \sigma_M = \sqrt{\frac{1}{t-1}\sum_{\tau=1}^{t}(M_t - \bar{M}_t)^2},\ HZ_3: \sigma_{\bar{M}_t} = \frac{1}{\sqrt{t}}\sigma_M,\ \alpha = 0.01$$

Dann können wir die Menge der a-priori als moralisch unmöglich oder falsifizierend zu betrachtenden Beobachtungen mit der folgenden, empirisch entscheidbaren, Relation beschreiben:

$$f_{H_0,HZ_1,..HZ_J}(M_1 \ldots . M_{T=t}) = F_{H_0,HZ_1,..HZ_J}(T_t) = 1 - F_N\left(\frac{\bar{M}_t}{\frac{1}{\sqrt{t}}\sqrt{\frac{1}{t-1}\sum_{\tau=1}^{t}(M_t - \bar{M}_t)^2}}\right) = P_t < 0.01$$

Dabei lässt sich die Funktion $f_{H_0,HZ_1,..HZ_J}(M_1 \ldots . M_{T=t})$ als die Verkettung der Berechnung einer Teststatistik

$$T_t = \frac{\bar{M}_t}{\frac{1}{\sqrt{t}}\sqrt{\frac{1}{t-1}\sum_{\tau=1}^{t}(M_t - \bar{M}_t)^2}}$$

mit einer Verteilungsfunktion $F_{H_0,HZ_1,..HZ_J}(T_t)$ darstellen.

Mit allen $t = TG = 729$ Beobachtungen berechnet ergibt sich ein P_t-Wert[17] von 0.0000185 und wir sagen, die Daten sind tatsächlich ein Elementarereignis, welches H_0 falsifiziert. Die Beobachtungen hätten wir a-priori für „moralisch unmöglich“ gehalten.

3.3. Die zentralen Schwächen statistischer Testverfahren

Subjektivität

Man muss erkennen, dass die Konstruktion des falsifizierenden Ereignisses bei der Anwendung von Testverfahren in dem Sinne subjektiv ist, dass sie auf einer Reihe von Annahmen beruht, deren Wahrheit nicht überprüft ist. Mit derselben Berechtigung hätten wir Gleichverteilung oder eine andere Verteilung oder irgendeine Form von Autokorrelation und vieles andere mehr unterstellen können. Natürlich könnten wir die Annahmen wieder testen – wir müssten aber dafür wieder Annahmen über die Verteilung der zu testenden Größen machen. Die fundmentale – nicht durch Wahrheit von Beobachtungen begründbare – Subjektivität der Annahmen kann man aber letztendlich nicht verhindern.

Vorläufigkeit der Falsifikation

Ein weiterer schwerwiegender Konstruktionsfehler dieses Falsifikationsprinzips zeigt sich in der Weiterführung des obigen Beispiels:

[17] Für eine aktuelle Diskussion der Vertrauenswürdigkeit von P-Werten vgl. z.B. Regina Nuzzo: (Scientific method:) P-values, the „gold standard“ of statistical validity, are not as reliable as many scientists assume, in Nature Volume 506, 12 February 2014. (http://www.nature.com/news/scientific-method-statistical-errors-1.14700)

Hätte man nur die Beobachtungen bis t=192 gemacht und auf dieser Basis den Test durchgeführt, so wäre man zu dem Ergebnis gekommen, dass H_0 moralisch unmöglich ist.

t	P_t	Entscheidung
192	0.0011	Moralisch unmöglich
193	0.0262	Nicht moralisch unmöglich

Hingegen wird mit der Hinzunahme der nächsten (der 193.) Beobachtung zur Beurteilung der Wahrheit der Hypothese dieses Ergebnis hinfällig. Die dann gemachte Beobachtung der statistischen Funktion liegt nicht in der Menge der moralisch unmöglichen. Falsifikationen mit Hilfe statistischer Tests sind also nicht endgültig. Letztendlich muss man eingestehen, dass die Verwendung dieses Falsifikationsprinzips dazu führt, dass keine Hypothese endgültig verworfen werden kann.

Sich gegenseitig ausschließende stochastische Gesetze können gleichzeitig wahr sein

Betrachtet man die stündliche Veränderung der Windgeschwindigkeit in Washington DC und führt einen Test der Hypothese $H_0: E(M) = 0$ aus so errechnet man einen P-Wert von 0.434. Man kann also die Hypothese nicht verwerfen. Es wurde keine Beobachtung gemacht, die a-priori als moralisch unmöglich definiert wurde. Ändert man die Hypothese auf $H_{0,A}: E(M) = 0.0001$, fragt man also, ob es im Unendlichen im Durchschnitt eine leichte Erhöhung der Windgeschwindigkeit gibt, so erhält man einen P-Wert von 0.49. Auch diese Hypothese darf weiterhin als wahr gelten.

Beide Hypothesen sind aber kontradiktorisch – Sie können aus logischen Gründen nicht gleichzeitig wahr sein.

Es gibt nicht falsifizierbare Hypothesen

Abschließend soll die Hypothese $H_{0,A}: E(M) = \bar{M}_t$ betrachtet werden. Hier handelt es sich um eine prinzipiell (wenn man unendliche viele Beobachtungen machen kann) empirisch falsifizierbare Hypothese. Aus der Tatsache, dass der bisherige Mittelwert $\bar{M}_t$ ist, kann man nicht logisch schließen, dass der Mittelwert im Unendlichen $\bar{M}_t$ sein muss. Wir können aber feststellen, dass bei obigem Test diese Hypothese immer einen P-Wert von 0.5 hat. Sie kann also – obwohl nicht aus logischen Gründen wahr - prinzipiell nicht falsifiziert werden.

Zusammenfassung

Die weit verbreitete Verwendung statistischer Tests zur Falsifikation von stochastischen Gesetzeshypothesen weist also (zumindest) vier schwerwiegende Konstruktionsmängel auf:

1. Statistisches Testen ist inhärent subjektiv.
2. Falsifikation durch statistisches Testen ist nicht endgültig.
3. Kontradiktorische Hypothesen können gleichzeitig als wahr gelten.
4. Es gibt nicht logisch wahre Hypothesen, die trotzdem nicht empirisch falsifizierbar sind.

Diese Mängel statistischer Tests sind unserer Einschätzung nach die wichtigste Ursache für den fehlenden Fortschritt in den Wirtschaftswissenschaften. Wir können nicht objektiv, endgültig und widerspruchsfrei entscheiden, ob eine Geldmengenerhöhung Inflation auslöst, ob Asset-Märkte effizient sind oder ob ein Mindestlohn zu weniger Beschäftigung führt. Letztendlich kann jeder mit statistischen Tests das Ergebnis begründen, das ihm (oder seinem Auftraggeber) gefällt. Wir sitzen auf einem immer größer werdenden Berg von aus irgendwelchen Annahmen deduzierten Gesetzeshypothesen, ohne die falschen endgültig aussortieren zu können.

3.4. Hypothesendefinition und -überprüfung bei der Suche nach emergenten empirischen Gesetzen

Aus der Perspektive der Suche nach emergenten empirischen Gesetzen stellt sich die Frage nach der metaphysischen Wahrheit von H_0 im Unendlichen gar nicht. Stattdessen wird eine Menge von eindeutig überprüfbaren Hypothesen definiert.

$$\{H_T: \forall t(\bar{M}_{t,T} = 0)\}$$

mit

$$\bar{M}_{t,T} = \frac{1}{T}\sum_{\tau=t-T}^{t} M_\tau$$

Und es wird die Frage gestellt, ob es ein $T \leq \frac{T_G}{2}$ gibt, für das der Mittelwert immer Null ist

$$\exists T \leq \frac{T_G}{2}(\forall t(\bar{M}_{t,T} = 0)).$$

Diese Frage ist schnell beantwortet.

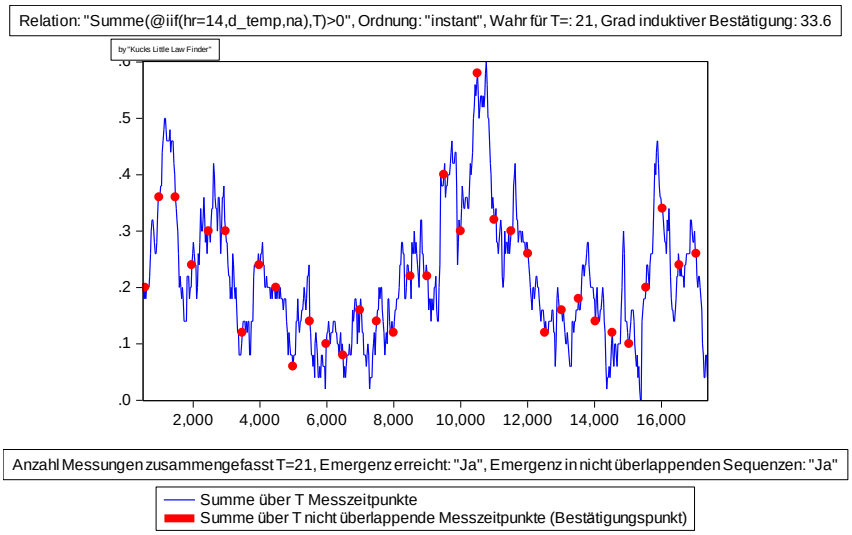

Abb. 9 Gesetz über Summen von Temperaturanstiegen zwischen 14 und 15 Uhr in Washington DC

Da für T>=21 alle Mittelwerte größer Null sind, gibt es kein solches T. Damit sind alle Gesetzeshypothesen H_T: $\forall t(\bar{M}_{t,T} = 0)$ für $T \leq T_G$ für immer falsifiziert. Das falsifizierende Ereignis ist jeweils einfach und klar definiert.

Man kann weiterhin fragen, ob es ein T gibt, für das die Mittelwerte bisher immer größer (oder kleiner) als 0 gewesen sind.

$\exists T \leq \frac{T_G}{2}(\forall t(\bar{M}_{t,T} > 0))$ oder $\exists T \leq \frac{T_G}{2}(\forall t(\bar{M}_{t,T} < 0))$

Man kommt zu dem Ergebnis, dass der Mittelwert des Temperaturanstiegs bisher in jedem Zeitfenster der Länge $T \geq 21$ größer als 0 war. Es ist also für immer falsifiziert, dass der Mittelwert in Fenstern bis zur Länge $TG/2$ immer kleiner oder gleich 0 ist. Außerdem haben wir ein 33,6-mal bestätigtes Gesetz gefunden. Natürlich kann dieses Gesetz in Zukunft durch eindeutig definierte Beobachtungen falsifiziert werden. Außerdem wird es nie wieder wahr werden, dass der Mittelwert für $T < 21$ immer Größer 0 ist.

Die Hypothesen sind bei diesem Verfahren so eindeutig formuliert, dass keine nicht durch Beobachtungen begründbaren Zusatzannahmen nötig sind. Die Entscheidung ist daher objektiv.

Einmal falsifizierte Hypothesen der Form $\forall t(f_{t,T} = 0)$ können nie wieder wahr werden. Es ist einfach unmöglich dass, wenn ein $f_{t,T} \neq 0$ beobachtet wurde, irgendwann die Aussage, dass alle $f_{t,T} = 0$ sind, wieder wahr wird.

Sich gegenseitig ausschließende Aussagen wie $\forall t(f_{t,T} = 0)$ und $\exists t(f_{t,T} \neq 0)$ können nie gleichzeitig wahr sein. Möglich ist nur, dass unterschiedliche Aussagen - die aber logisch miteinander verträglich sind (wie z.B. $\forall t(f_{t,T} > 0)$ und $\forall t(f_{t,T} > 0.1)$) - gleichzeitig wahr sind.

Nicht falsifizierbar sind nur Hypothesen, die schon aus logischen Gründen immer wahr sind (z. B. $\forall t(RelH_{t,T} \leq 1)$).

3.5. Der frequentistische Induktionsschluss und die frequentistische Begründung der Auswahl von Schätzheuristiken

Wenn wie im obigen Beispiel die Nullhypothese abgelehnt wird, geht man davon aus, dass der Mittelwert im Unendlichen größer als 0 ist. Dies wird als Begründung dafür genommen, den Wert der Mengenfunktion im Unendlichen aus den Daten zu schätzen.

Empirische Grundlage des frequentistischen Induktionsschlusses ist dabei eine einzelne Beobachtung einer Kennzahl der deskriptiven Statistik – im Beispiel der bisherige arithmetische Mittelwert über die gesamte Zahl der Beobachtungen:

$$M_{t=T=T_G,T} = \frac{1}{T}\sum_{t=1}^{T} M_t$$

Im nächsten Schritt wird dann versucht, von einer Beobachtung des Mittelwerts auf den metaphysischen Mittelwert im Unendlichen – den Erwartungswert – zu schließen (frequentistischer Induktionsschluss).

Nach dieser Interpretation kann der Erwartungswert dann als Mittelwert für unendlich viele Beobachtungen gesehen werden

$$M_{t=T=T_G,T} \; \boldsymbol{s} \; \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} M_t = E(M)$$

Dabei soll die Verwendung des Operators „$\boldsymbol{s}$“ (schätzt) anstatt von „=“ deutlich machen, dass keine Gleichheit zwischen dem bisherigen Mittelwert einer endlichen Stichprobe und dem Mittelwert im Unendlichen unterstellt wird. Es handelt sich nur um eine Schätzung, deren Wahrheit in der Stochastik gar nicht behauptet wird.

Die Induktion erfolgt also nicht von einem beobachteten Mittelwert auf zukünftig beobachtbare Mittelwerte (empirischer Induktionsschluss), sondern auf einen niemals beobachtbaren Mittelwert im Unendlichen. Dabei wird noch nicht einmal behauptet, dass diese Schätzung tatsächlich richtig ist. Dies macht aber auch nichts, da der Erwartungswert als Grenzwert im Unendlichen niemals beobachtet werden kann. Auch wenn man Gleichheit von bisherigem Stichprobenmittelwert und Erwartungswert fordern würde, so ließe sich die Wahrheit dieser Relation niemals überprüfen.

Begründet wird diese Art von Induktion in der Stochastik deduktiv. Dazu werden Annahmen über die Art, wie Sequenzen von Beobachtungen generiert werden, benötigt.

Die am häufigsten verwendete Annahme ist die sog. iid-Annahme (independent and identically distributed).

Im Unendlichen sollen alle Arten von Stichproben (Sequenzen) der Länge T mit der relativen Häufigkeit auftreten, mit der sie kombinatorisch möglich sind.

Allerdings ist selbst unter Berücksichtigung der iid-Annahme kein deduktiver Schluss vom Stichprobenmittelwert auf den Erwartungswert möglich, aber man kann unter Anwendung dieser Annahme beispielsweise auf die sog. Erwartungstreue des Mittelwerts einer Stichprobe (Realisation) als Schätzwert für den Erwartungswert (der Zufallsvariable) deduktiv schließen.

Kurz gesagt behauptet man, dass es eine gute Idee ist, aus dem Mittelwert einer Stichprobe auf den Erwartungswert der Zufallsvariable zu schließen, da der Mittelwert eine erwartungstreue Schätzfunktion für den Erwartungswert ist:

$$\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} \bar{M}_{\tau,T} = \lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} M_\tau = E(M)$$

Wenn man unendlich viele Mittelwerte aus Stichproben der Länge T beobachten würde, dann könnte man bei Geltung der iid-Annahme sicher sein, dass der Durchschnitt dieser Mittelwerte dem Durchschnitt der Beobachtungen im Unendlichen (dem Erwartungswert) entspricht.

Natürlich kann nicht definitiv überprüft werden, ob die iid-Annahme in einem konkreten Anwendungsfall gerechtfertigt ist und es können (zumindest bisher) nicht unendlich viele Beobachtungen berücksichtigt werden, um den Erwartungswert tatsächlich zu bestimmen.

Das Argument der Erwartungstreue des Stichprobenmittelwertes wird nur als Begründung dafür verwendet, dass es sinnvoll ist, von einem beobachteten Mittelwert auf den Erwartungswert induktiv zu schließen.

Dabei wird aber weder unterstellt, dass der Erwartungswert tatsächlich dem bisherigen Mittelwert entspricht, noch, dass die aufeinanderfolgenden Stichprobenmittelwerte identisch sein müssen.

Wenn also der nächste Mittelwert über T Beobachtungen nicht mit dem vorherigen Mittelwert über T Beobachtungen übereinstimmt, so falsifiziert dies die induzierte Hypothese einer Übertragbarkeit des Stichprobenmittelwertes auf den Erwartungswert nicht.

Ohne dies explizit zu machen, ist die zentrale Verwendung des nicht falsifizierbaren stochastischen Induktionsschlusses in der frequentistischen Statistik aber die Erstellung von Prognosen. In der statistischen Praxis begründet man z. B. mit der (unter geeigneten Annahmen deduzierbaren) Erwartungstreue des Mittelwerts seine Geeignetheit als Schätzheuristik auch für einzelne Beobachtungen.

Im Beispiel aus dem letzten Kapitel würde man aus der Ablehnung der stochastischen Hypothese (dass der Erwartungswert der Temperaturänderungen gleich 0 ist) dazu übergehen, den konkreten Mittelwert $\bar{M}_t$ als Prognose für den Erwartungswert zu verwenden. Praktisch werden dann auch (unter Verwendung von $\bar{M}_t$ als Heuristik) Schätzungen für jede einzelne zukünftige Beobachtung gemacht. Natürlich behauptet man nicht, dass man glaubt, dass die so erstellte Prognose für den nächsten Wert tatsächlich eintreten wird.

3.6. Empirische Gesetze über den relativen Vorteil statt metaphysischer Wahrheit als Prinzip zur Auswahl von Schätzheuristiken

Die hier vorgeschlagene Begründung für die Wahl einer Schätzheuristik ist immer ihre gesetzmäßige relative Vorteilhaftigkeit gegenüber zumindest einer konkreten Alternative bezüglich einer Bewertungsmetrik.

Die beiden oben nur implizit verglichenen Werte der Schätzheuristiken

$$f_{1,t} = 0, und\ f_{2,t} = \bar{M}_{t-1,t-1} = \frac{1}{t-1}\sum_{\tau=1}^{t-1} M_\tau$$

müssen rekursiv bestimmt und ihre Prognoseleistung mit einer Bewertungsmetrik, z. B. der Summe des absoluten Prognosefehler bewertet werden.

$S_{1,t,T} = \sum_{t=t-T}^{t} |M_t - f_{1,t}|$ und $S_{2,t,T} = \sum_{t=t-T}^{t} |M_t - f_{2,t}|$

Dann werden die Hypothesen, dass eines der beiden Verfahren nach T Schätzungen immer ein besseres Gesamtergebnis erzeugt hätte, untersucht:

$$\{H_{T,>}: \forall t\left(S_{1,t,T} > S_{2,t,T}\right), H_{T,<}: \forall t\left(S_{1,t,T} < S_{2,t,T}\right)\}$$

Im Beispiel käme man zu dem Ergebnis, dass keine der Hypothesen für irgendein $T < \frac{T_G}{2}$ immer zu einem besseren Gesamtergebnis – einer kleineren Summe absoluter Prognosefehler - geführt hätte. Und dies obwohl es in Fenstern $T \geq 21$ immer einen positiven mittleren Temperaturanstieg gegeben hat. Es ist also endgültig falsifiziert, dass die Schätzung mit dem bisherigen Mittelwert in Fenstern $T < \frac{T_G}{2}$ nach dem Kriterium Summe der absoluten Prognosefehler immer zu besseren Ergebnissen führt als die Schätzung mit der Heuristik Temperaturanstieg gleich 0. Genauso ist es endgültig falsifiziert, dass die Heuristik Temperaturanstieg=0 in Fenstern $T < \frac{T_G}{2}$ immer zu besseren Ergebnissen führt. Keines der beiden Verfahren ist T-dominant. [18]

[18] Natürlich erzeugt das Kriterium der T-Dominanz nicht immer eine vollständige Ordnung alternativer Heuristiken. Da eine Diskussion möglicher Vorgehensweise in diesem Fall den Rahmen dieses Aufsatzes deutlich sprengen würde, sei hier nur auf einige grundlegende Überlegungen in A. Kuck, J. Harries, E. Kuck: Der Weg zu wahren Gesetzen und rationalen Entscheidungen, S.49ff. verwiesen.

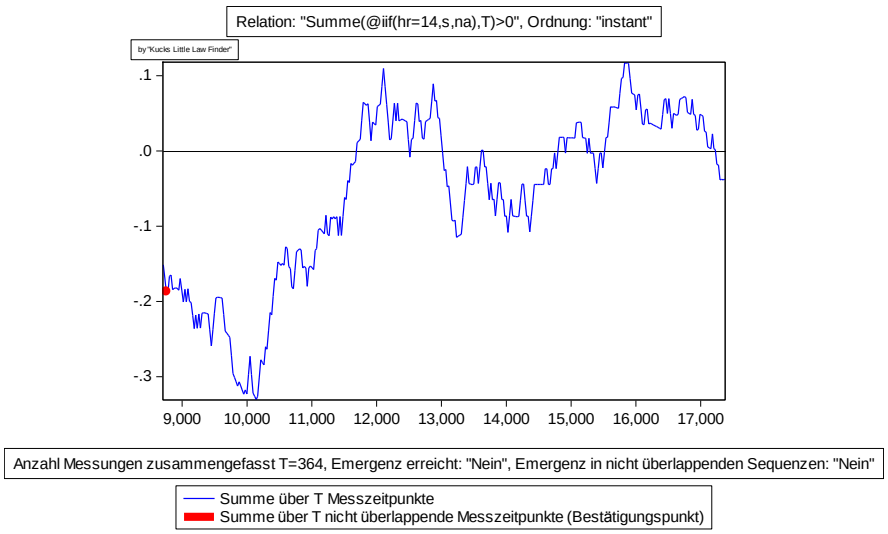

Abb. 10 Falsifizierte Gesetzeshypothesen über die relative Vorteilhaftigkeit zweier Schätzheuristiken

Man sieht hier deutlich, dass die Schlüsse, die man aus Testergebnissen glaubt ziehen zu können, nicht immer gerechtfertigt sind. Aus der Ablehnung der Hypothese, dass der Erwartungswert gleich Null ist und der Eigenschaft der Erwartungstreue des Mittelwerts, folgt nur logisch (im Unendlichen und unter Zusatzannahmen), nicht aber empirisch die Überlegenheit der Schätzung mit dem Mittelwert als Schätzheuristik. Die als logische Implikation von Testergebnissen und stochastischen Annahmen erwarteten Eigenschaften der Anwendung der Verfahren sind oft gar nicht zu beobachten.

Man mag einwenden, dass man natürlich vor der Anwendung der Heuristik die Annahmen hätte testen müssen – wie aber oben gezeigt, ist dies niemals vollständig möglich.

Nehmen wir wieder die Daten aus Fanaee-T, Hadi, and Gama, Joao, und überprüfen mit welcher relativen Häufigkeit bei der Auswahl zwischen

$$f_{1,t} = 0, und\ f_{2,t} = \bar{M}_{i,hr,t-1,t-1} = \frac{1}{t-1}\sum_{\tau=1}^{t-1} M_\tau$$

nach dem obigen Kriterium (Summe des absoluten Prognosefehlers in Fenstern der Länge 4) in Abhängigkeit von DiV eine richtige Vorhersage der Relation der Summe der Schätzfehler im jeweils folgenden 4-Stundenzeitraum gemacht worden wäre.

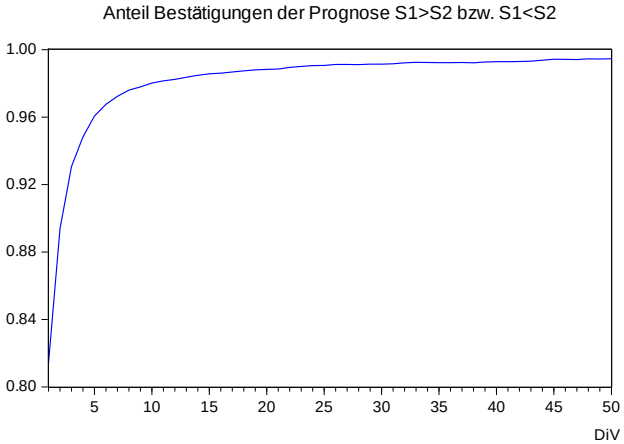

Abb. 11 Anteil bestätigter Prognosen über die relative Vorteilhaftigkeit zweier Schätzheuristiken

In Abb. 11 können wir erkennen, dass auch hier mit dem Grad der Bestätigung der Überlegenheit der Heuristik auch die Überlegenheit der Prognose nach dem Kriterium Summe der absoluten Prognosefehler zunimmt. Es macht also Sinn Heuristiken zu verwenden, die bisher immer besser waren als betrachtete Alternativen. Natürlich wird hier kein Anspruch auf Optimalität oder das Vorliegen wünschenswerter Eigenschaften im Unendlichen begründet. Um die Verwendung der Heuristiken zu rechtfertigen reicht es, dass sie bisher immer besser waren und dass man weiß, dass das, was bisher immer besser war, stark (und quantifizierbar) dazu neigte, auch zukünftig besser zu sein.

Lernen wird als die Suche nach immer besseren Heuristiken in immer endlichen Mengen ausprobierter Heuristiken gesehen und es wird keine Anspruch auf Optimalität in irgendeinem unendlichen Funktionenraum erhoben.

Unserer Einschätzung nach ist es ein fundamentaler Fehler, Schätzheuristiken auf Basis stochastischer Induktionsschlüsse mit angenommenen wünschenswerten Eigenschaften im Unendlichen auszuwählen. (Und de-facto vertraut man diesen Schlüssen auch nicht, sondern überprüft die Prognoseleistung.) Man muss sich bewusst machen, dass die wünschenswerten Eigenschaften im Unendlichen immer aus nicht auf Basis von Beobachtungen objektiv begründbaren und nicht endgültig falsifizierbaren Annahmen über die Gültigkeit stochastische Gesetze im Unendlichen deduziert werden.

3.7. Zusammenfassung der konzeptuellen Unterschiede zwischen frequentistischer Statistik und der Suche nach empirischen Gesetzen

Die wesentlichen Unterschiede zwischen der frequentistischen Statistik und dem auf der Suche nach emergenten Gesetzen basierenden Verfahren sind noch einmal in der folgenden Tabelle zusammengefasst:

Tabelle 1 Unterschiede zwischen frequentistischen Statistik und dem auf der Suche nach emergenten Gesetzen basierenden Verfahren

Konzept	Frequentistische Statistik	Emergente Gesetze
Gesetzeshypothesen	$\forall t(\lim_{T\to\infty} f_{t,T} = E(f))$ Grenzwert einer Mengenfunktion im Unendlichen	$\{H_T: \forall t \in B(f_{t,T} = c)\}$ Es gibt in einer wohldefinierten Menge von Hypothesen zumindest eine konkrete Mengenfunktion $f_{t,T}$, die bisher immer $\forall t \in B$ die Eigenschaft c hatte
Falsifizierendes Ereignis	$A = \{A_i = \{M_{i,1} \dots . M_{i,T=t}\} \mid f_{H_0,HZ_1,..HZ_J}(M_{i,1} \dots . M_{i,T=t}) = P_{i,t} < \alpha\}$ Moralisch unmögliches Ereignis. Ereignis mit kleiner Eintrittswahrscheinlichkeit gegeben die Wahrheit der Hypothese und einiger Zusatz-annahmen	$\exists t \in B(f_{t,T} \neq c)$ Zu irgendeinem Zeitpunkt war der Wert der Mengenfunktion für ein gegebenes T ungleich c
Induktionsschluss	Wenn $f_{H_0,HZ_1,..HZ_J}(M_1 \dots . M_{T=t}) = P_t \geq \alpha$ dann $\lim_{T\to\infty} f_{t,T} = E(f) = 0$ sonst $f_{t=T=T_G,T}\ s\ \lim_{T\to\infty} f_{t,T} = E(f)$ Wenn die gemachte Beobachtung nicht moralisch unmöglich ist, dann ist im Unendlichen die Hypothese wahr ansonsten schätze mit Hilfe der Beobachtungen den Wert der Mengenfunktion im Unendlichen.	Wenn $\forall t \in B(f_{t,T} = c)$ dann $\forall t \in \{1 \dots T_G \dots \infty\}(f_{t,T} = c)$ Wenn eine Mengeneigenschaft bisher immer beobachtet wurde dann wird sie auch für die nächsten Beobachtungen wahr sein.
Prognose	Es ist nicht klar, was alles prognostiziert wird. Ist die gesamte Testverteilung eine Prognose für die Verteilung der Teststatistik im Unendlichen? Oder gilt nur obiger Induktionsschluss?	$f_{t+T,T} = c$ Nach weiteren T Beobachtungen wird der Wert der Mengenfunktion wieder c sein.
Maß für die Güte des Gesetzes	$P_t = f_{H_0,HZ_1,..HZ_J}(M_{i,1} \dots . M_{i,T=t}) = \lim_{t\to\infty}(RelH(T_{t,T} > T_{T,T}))$	DiV_t

	Im Unendlichen erwartete relative Häufigkeit von Werten der Teststatistik, die größer sind als die beobachtete. HHh	Anzahl Bestätigungen des betrachteten Gesetzes
Maß für die zu erwartende Prognosegüte	Unterschiedliche Verfahren (P-Wert, Standardabweichung / empirische Prognose-güte)	$f_{H,t,T,DiV} = \frac{1}{N} \sum_{\tau=T+1}^{t} \sum_{i=1}^{N_\tau} I(P_{i,DiV(\tau-T),\tau,T} = wahr, 1{,}0)$ Relative Häufigkeit der Wahrheit von bisher mit DiV-mal bestätigten Gesetzen gemachten Prognosen
Schätzheuristik für nächste Beobachtung	$f_{t=T=T_G,T}\ s\ f_{t+1,1} = E(f)$ Üblicherweise wird der Wert für die nächste Beobachtung mit dem im Unendlichen als wahr unterstellten Gesetz geschätzt.	Wähle eine nach einer Bewertungs-metrik nicht T-Dominierte Heurisik (oder Kombinationen aus nicht T-dominierten Heuristiken)
Maß für die Güte der Schätzheuristik	Unterschiedliche Verfahren (P-Wert, Standard-abweichung / empirische Prognosegüte)	DiV_t Anzahl Bestätigungen des betrachteten Gesetzes

4. Empirischer Gehalt der subjektivistischen Wahrscheinlichkeitsinterpretation

Im folgenden Kapitel werden wir zeigen, dass Subjektivisten die Beobachtungen, auf die sich die subjektiv interpretierte Wahrscheinlichkeitsaussage beziehen, gar nicht definieren. Damit sind die Überprüfung subjektivistischer Wahrscheinlichkeitsaussagen und ihre Reformulierung als empirisches Gesetz ausgeschlossen. Eine empirische Überprüfbarkeit würde sich erst durch die Zusammenfassung mehrerer gleichartiger, von einem Subjekt als Merkmalsträger gemachten, subjektiven Wahrscheinlichkeitsaussagen ergeben. Die Interpretation der subjektiven Wahrscheinlichkeitsaussagen als Heuristik und ihre Überprüfung auf relative T-Dominanz würde dann eine objektive Definition eines rational prognostizierenden Individuums erlauben.

Auch die von den Subjektivisten als Brücke in die Empirie verwendete Bayes‘sche Lernregel kann nur als eine (von den Subjektivisten als universell gültig betrachtete) Schätzheuristik interpretiert werden. Ihre Anwendung muss aus unserer Perspektive durch T-Dominanz von Sequenzen von Schätzergebnissen im Einzelfall gerechtfertigt werden.

4.1. Die subjektivistische Wahrscheinlichkeitsinterpretation

In der einfachsten Form der subjektivistischen Sichtweise handelt es sich bei einer Wahrscheinlichkeit um die entscheidungsrelevante subjektive Einschätzung oder den subjektiven Glauben eines Individuums.

Yes or no: was there once life on Mars? […] Suppose your odds were 1:9 for life […] corresponding to [a] probability of 1/10 […]. Here is another way of saying the same thing: You would think a price of one dollar just right for a ticket worth ten if there was life on Mars […].[19]

Für Subjektivisten ist eine Wahrscheinlichkeit nur ein handlungsrelevanter Glaubenszustand eines Individuums. Der Glaube an eine Eintrittswahrscheinlichkeit von 0.1 hat zur Konsequenz, dass ein (risikoneutrales) Individuum gerade bei einem Preis von 1 US$ indifferent wäre zwischen erstens dem sicheren Besitz von einem US$ oder zweitens einer Wette, die 10 US$ zahlt wenn man herausfindet, dass es Leben auf dem Mars gab. Dies ist die Einschätzung eines Einzelfalls, von dem wir gar nicht erwarten können, dass er im frequentistischen Sinne unendlich oft (oder auch nur mehrmals) auftritt. Es gab Leben auf dem Mars oder eben nicht.

Im Kern versucht die Theorie subjektiver Wahrscheinlichkeiten nur abzuleiten, welche Eigenschaften solche

[19] Richard Jeffrey: Subjective Probability – The Real Thing, 2002, S. 8.

subjektiven Wahrscheinlichkeiten für rationale Menschen haben sollten.[20]

Wenn man die Wahrscheinlichkeitsaussage im Beispiel empirisch ernst nehmen würde, so ist von vorn herein klar, dass sie falsch ist. Die relative Häufigkeit mit der wir herausfinden, dass es Leben auf dem Mars gibt, wird sicher nicht 0.1 sein. Es kann also entweder gar nicht gemeint sein, dass die Aussagen in irgendeiner Menge von Beobachtungen in 10% der Fälle wahr ist oder aber es ist von vornherein klar, dass die Aussage falsch ist.

So sind auch subjektivistisch interpretierte Wahrscheinlichkeitsaussagen prinzipiell nicht empirisch überprüfbar.

4.2. Interpretation von Sequenzen subjektivistischer Wahrscheinlichkeitsaussagen eines Individuums als Schätzheuristik und die empirische Überprüfung ihrer Rationalität

In unserer Terminologie kann es sich also bestenfalls um die Beschreibung des Ergebnisses einer Schätzheuristik für relative Häufigkeit handeln, von dem man a-priori schon weiß, dass es nicht im Einzelfall zutrifft.

Will man trotzdem versuchen, subjektive Wahrscheinlichkeitsaussagen empirisch zu interpretieren, so könnte man subjektive Wahrscheinlichkeitsaussagen zu Sequenzen gleichartiger Aussagen zusammenzufassen:

Ein Individuum schätzt (auf Basis von ihm verwendeter Heuristiken) dass die Aussage „Es gab Leben auf dem Mars“ zu einer Gruppe von (zeitlich geordneten) Aussagen $\{A_{P,t}\}$ gehört, von denen es prognostiziert, dass Prob% der Aussagen wahr sind. Eine Heuristik in dem Sinne, dass eine Schätzung auf Basis eines Beobachtungsmusters erfolgt, ergibt sich z. B. dann, wenn das Individuum seine Wahrscheinlichkeitsschätzungen zum Zeitpunkt t-1 veröffentlicht und ein Beobachter die nachfolgenden tatsächlich gemachten Beobachtungen, ob die $A_{P,t}$ eingetreten sind, überprüft.

$$RelH\big(A_{P,t} = wahr\big) = Prob(A)$$

Um die Hypothese, dass ein Individuum im Durchschnitt richtig liegt, empirisch überprüfen zu können, muss auch hier wieder die Gesamtheit definiert werden, auf die sich die Wahrscheinlichkeitsaussage bezieht. Natürlich könnte man die frequentistische Sichtweise einnehmen, ein stochastisches Gesetz für unendlich viele Wahrscheinlichkeitsaussagen des Individuums formulieren:

$$\forall t(\lim_{T\to\infty}\Big(RelH_{t,T}\big(A_{P,\tau} = wahr\big)\Big) = Prob(A))$$

Allerdings ist diese Sichtweise wieder mit den schon oben ausführlich beschriebenen Problemen behaftet.

In unserer Terminologie verwendet ein Individuum „rationale“ Heuristiken zur Schätzung der subjektiven Wahrscheinlichkeiten und somit ist die Verwendung der von ihm geäußerten Wahrscheinlichkeit als Schätzheuristik empirisch rational, wenn es bisher nicht immer (systematisch) geirrt hat. Wenn es also keine Sequenzlänge T$\leq \frac{T_G}{2}$ gibt, für die es die relative Häufigkeit der Wahrheit der Aussagen $A_{P,t}$ immer über- bzw. unterschätzt hat:

$$\neg\exists T \leq \frac{T_G}{2}(\forall t\Big(RelH_{t,T}\big(A_{P,\tau} = wahr\big) > Prob(A)\Big)) \wedge \neg\exists T \leq \frac{T_G}{2}(\forall t\Big(RelH_{t,T}\big(A_{P,\tau} = wahr\big) < Prob(A)\Big))$$

Allerdings ist es gar nicht notwendig zu unterstellen, dass Individuen bei ihren Entscheidungen Ereignisse in Gruppen mit gleichen subjektiven Eintrittswahrscheinlichkeiten einordnen und ihre Entscheidungen (über Wetten mit der Auszahlung Y_t und dem Einsatz E_t) dann aus den geschätzten relativen Erfolgshäufigkeiten deduzieren.

$$Wenn(E_t < Prob(A_t) \cdot Y_t, Y_t - E_t, 0)$$

Um entscheiden zu können würde es reichen, wenn die Grenze $E(Y_t)$, ab der das Individuum die Wette eingeht direkt geschätzt wird.

$$Wenn(E_t < E(Y_t), Y_t - E_t, 0)$$

Damit ein Individuum rational genannt werden kann, sollte bisher nur beobachtet worden sein, dass es seine Wetten bisher nicht systematisch verliert.

[20] Still, we know that any probabilities anyone might think acceptable for those two hypotheses out to satisfy certain rules[…].Richard Jeffrey: Subjective Probability – The Real Thing, 2002, S. 9.

$$Z_t = Wenn(E_t < E(Y_t), Y_t - E_t, 0), S_{t,T} = \sum_{\tau=t-T}^{t} Z_t$$

$$\neg\exists T < \frac{T_G}{2}, \forall t(S_{t,T} \leq 0)$$

Unsere empirische Interpretation subjektiver Wahrscheinlichkeit ist also die einer Beobachtung, mit der „gleichartige" Handlungsergebnisse zusammengefasst werden können. Die mögliche empirische Bedeutung subjektiver Wahrscheinlichkeitsaussagen liegt in der Funktion des Zusammenfassen von „gleichartigen" Schätzungen eines Subjekts, die dann wieder auf die Mengeneigenschaften der Validität der Schätzungen hin überprüft werden können. Die subjektive Wahrscheinlichkeit ist eher analog zur DiV-Kennzahl sinnvoll empirisch interpretierbar. Durch die geäußerte Wahrscheinlichkeit lassen sich Sequenzen von gleichartigen Beobachtungen bilden, die dann mit Hilfe unseres Ansatzes auf Gesetzmäßigkeiten hin überprüft werden können.

Allerdings besteht kein Grund sich auf subjektive Aussagen über relative Häufigkeiten zu beschränken, um Schätzungen und deren Ergebnisse zu Mengen zusammenzufassen. Allgemein können gleichartige Aussagen über die Begründung von Entscheidungen oder aber auch Entscheidungsergebnisse allein diese Funktion erfüllen.

4.3. Die Interpretation des Bayes-Lernens als Schätzheuristik

Bei der Definition subjektiver Wahrscheinlichkeiten spielt die Heuristik, mit der diese Wahrscheinlichkeiten vom Subjekt geschätzt werden gar keine Rolle. Die von den Bayesianern vorgeschlagene Kombination aus dem Satz von der totalen Wahrscheinlichkeit und der bayes'schen Lernregel zur Bestimmung „rationaler" Wahrscheinlichkeiten der Wahrheit von Hypothesen und zukünftigen Ereignissen kann aus unserer Perspektive nur als Vorschlag für eine Schätzheuristik interpretierbar werden.

Grundlage ihrer Vorgehensweise sind zwei Theoreme der Wahrscheinlichkeitsrechnung, deren Wahrheit für empirische relative Häufigkeiten logisch aus der Definition relativer Häufigkeiten für disjunkte Mengen von Merkmalsausprägungen (Ereignisse) H_i folgt. (mit N_X als beobachtete absolute Häufigkeit des Ereignisses X und N als gesamte Anzahl von Beobachtungen)

Erstens der Satz von der totalen Wahrscheinlichkeit.

$$\frac{N_A}{N} = RelH(A) = \sum_i RelH(A|H_i) \cdot RelH(H_i) = \sum_i \frac{N_{A \wedge H_i}}{N_{H_i}} \cdot \frac{N_{H_i}}{N}.$$

und zweitens der Satz von Bayes.

$$\frac{N_{H_i \wedge A}}{N_A} = RelH(H_i|A) = \frac{RelH(A|H_i) \cdot RelH(H_i)}{RelH(A)} = \frac{\frac{N_{A \wedge H_i}}{N_{H_i}} \cdot \frac{N_{H_i}}{N}}{\frac{N_A}{N}}$$

Sie sind logisch wahr, besagen aber nur dass, wenn man alle Häufigkeiten N_A, $N_{A \wedge H_i}$, N_{H_i} und N gezählt hat und die relativen Häufigkeiten $RelH(A|H_i) = \frac{N_{A \wedge H_i}}{N_{H_i}}$ und $RelH(H_i) = \frac{N_{H_i}}{N}$ ausgerechnet hat, dann kann man ohne N_A zählen zu müssen $RelH(A) = \frac{N_A}{N}$ und $RelH(H_i|A) = \frac{N_{H_i \wedge A}}{N_A}$ einfach ausrechnen oder logisch deduzieren.

Diese beiden Theoreme bilden die Grundlage der von den Bayesianern vorgeschlagenen Schätzheuristik.

Um ihre Vorgehensweise transparent zu machen, soll sie in drei Schritten dargestellt werden:

In einem ersten Schritt wird ein „rationaler" Induktionsschluss vorgeschlagen. Man soll von den bisher (bis zum Zeitpunkt t) beobachteten relativen Häufigkeiten $RelH_{t,t=T}(A|H_i)$ und $RelH_{t,t=T}(H_i)$ auf die subjektiven Wahrscheinlichkeiten $Prob_{t+1}(A)$ und $Prob_{t+1}(H_i|A)$ schließen. Also letztendlich wird eine Schätzheuristik für Relative Häufigkeiten konstruiert.

$Prob_{t+1}(A) = \sum_i RelH_{t,t=T}(A|H_i) \cdot RelH_{t,t=T}(H_i)$

$$Prob_{t+1}(H_i|A) = \frac{RelH_{t,t=T}(A|H_i) \cdot RelH_{t,t=T}(H_i)}{Prob_{t+1}(A)}$$

Im zweiten Schritt wird die Forderung aufgegeben, dass die relativen Häufigkeiten $RelH_{t,t=T}(H_i)$ und $RelH_{t,t=T}(A|H_i)$ beobachtete relative Häufigkeiten sein müssen. Sie werden durch subjektive Wahrscheinlichkeiten $Prob_{a-priori}(H_i)$ und $Prob_{a-priori}(A|H_i)$ ersetzt. In unserer Interpretation durch das Ergebnis einer von einem Subjekt verwendeten, unbekannten Heuristik zur Schätzung dieser Wahrscheinlichkeiten.

Nun kann man in einem dritten Schritt die beobachtbaren Ereignisse H_i durch stochastische Gesetzeshypothesen aus dem Handwerkskasten der frequentistischen Statistik ersetzt. Zum Beispiel durch die Hypothesen, dass das Ereignis A bernoulliverteilt (B) mit dem Parameter p_i (der Eintrittswahrscheinlichkeit) ist.

$$H_i: A \approx B(p_i)$$

Wenn $RelH_{t,t=T}(A) = \frac{N_{A,1..t}}{N_{1..t}}$ bisher beobachtet wurde, dann folgt logisch (im Unendlichen) dass $N_{A,1..t}|H_i$ binomialverteilt mit $Prob(N_{A,1..t}|H_i) = fb(N_{A,1..t}, N_{1..t}, p_i)$ ist - die Wahrscheinlichkeit die Daten (Anzahl eingetretene A) zu beobachten, wenn A tatsächlich bernoulliverteilt mit Eintrittswahrscheinlichkeit p_i ist.

Nun braucht man nur noch eine (subjektive) a-priori-Schätzung für die Wahrscheinlichkeit, dass die Hypothese H_i wahr ist, und man erhält eine Schätzheuristik für die Wahrscheinlichkeit von A und die Wahrscheinlichkeit der Wahrheit der subjektiven A-priori-Hypothesen H_i im Lichte von Beobachtungen von A:

$Prob_{t+1}(A) = \sum_i fb(N_{A,1..t}, N_{1..t}, p_i) \cdot Prob_{a-priori}(H_i)$

$$Prob_{t+1}(H_i|A) = \frac{fb(N_{A,1..t}, N_{1..t}, p_1) \cdot Prob_{a-priori}(H_i)}{Prob_{t+1}(A)}$$

Und man kann dann bei jeder neuen Beobachtung ($\tau \geq 1$) von A seine subjektiven Wahrscheinlichkeiten rekursiv updaten:

$Prob_{t+\tau+1}(A) = \sum_i fb(N_{A,t+\tau}, 1, p_i) \cdot Prob_{t+\tau}(H_i|A)$

$$Prob_{t+\tau+1}(H_i|A) = \frac{fb(N_{A,t+\tau}, 1, p_i) \cdot Prob_{t+\tau}(H_i|A)}{Prob_{t+\tau}(A)}$$

Insgesamt kann man also sagen, dass es sich beim Bayes-Lernen um eine aus Rechenregeln für relative Häufigkeiten abgeleitete Schätzheuristik handelt. Es wird kein Anspruch auf irgendeine messbare Form von empirischer Wahrheit erhoben. Nur ein „rationaler“ Prozess zur Berücksichtigung von Daten in subjektiven Wahrscheinlichkeitsschätzungen wird hier vorgeschlagen.

Es wird (außer in Grenzfällen mit Hypothesen wie $fb(x, n, 0)$ und $fb(x, n, 1)$) nie eine Hypothese falsifiziert. Anders als in der Testtheorie ist es im Bayes-Lernen meist unmöglich Hypothesen H_i überhaupt zu falsifizieren. Jede am Anfang verwendete Hypothese bleibt bis in alle Ewigkeit möglich (meist sogar wahrscheinlich) und muss in Prognosen auf ewig weiter berücksichtigt werden.

Ob es Sinn macht, Bayes-Heuristiken überhaupt und auf Basis welcher subjektiv gewählter Hypothesen und a-Priori-Wahrscheinlichkeiten zu verwenden, ist unserer Einschätzung nach nicht theoretisch, sondern nur empirisch zu beantworten. Das von uns vorgeschlagene Verfahren dazu haben wir bereits ausführlich diskutiert.

5. Zusammenfassende Würdigung

1. Sind Assetmärkte informationseffizient?
2. Löst eine Erhöhung der Geldmenge Inflation aus?
3. Sind Menschen risikoavers?
4. Gibt es einen positiven Grenznutzen des Geldes und ist die Von-Neumann-Morgenstern Nutzenfunktion zur Beschreibung menschlichen Verhaltens richtig?
5. Ist die frequentistische oder die subjektivistische Wahrscheinlichkeitsinterpretation richtig?

Auf all diese Fragen gibt die Wirtschaftswissenschaft keine definitive Antwort. Dies liegt unserer Einschätzung daran, dass

- die Hypothesen zu unscharf formuliert sind und
- Operationalisierungen dieser Hypothesen meist stochastisch formuliert sind und somit nicht objektiv, endgültig und widerspruchsfrei falsifiziert werden können.

Aus unserer Perspektive ist die Antwort auf etwas genauer formulierte Hypothesen mit einer nicht technischen Anwendung des Grundgedankens emergenter Gesetze in den meisten Fällen definitiv und einfach möglich.

Ad 1.:
Fragt man sich, ob es Subjekte oder Institutionen gibt, die über eine Handelsheuristik verfügen, die bisher immer höher Renditen generiert hat als der S&P500, so hätte man im Jahr 2005 zumindest James „Jim“ Simons mit seinem Renaissance-Medaillon-Fonds gefunden. Renaissance Medaillon hatte zwar nicht in jedem einzelnen Jahr eine höhere Rendite als der S&P500, aber die Summe der Renditen über zwei Jahre war immer höher als die des S&P500.

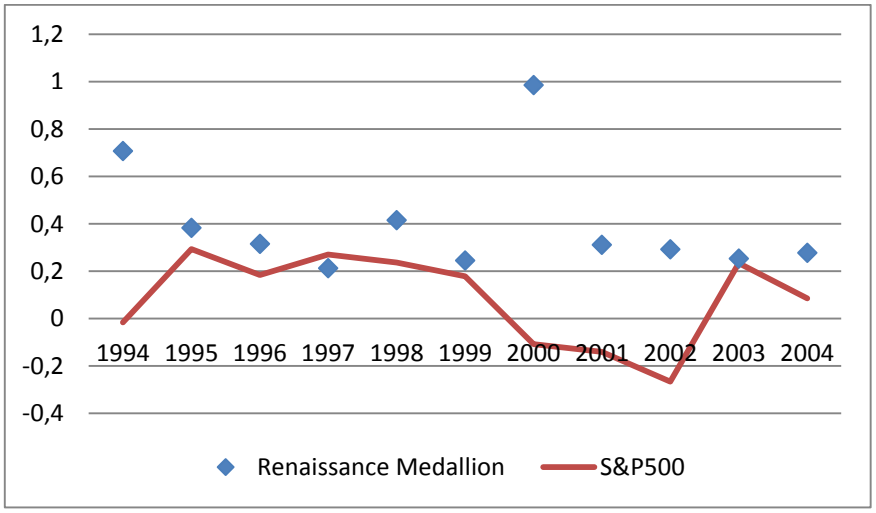

Abb. 12 Ein-Jahres-Renditen: Renaissance-Medaillon versus S&P500[21]

Dieses empirische Gesetz wäre also $DiV = 11/2 - 1 = 4.5$ mal bestätigt gewesen. Wir hätten vorhergesagt, dass er es auch in den folgenden zwei Jahren wieder schaffen wird. Da eine solche Strategie existiert ist die Effizienzmarkthypothese falsifiziert. Es ist für immer falsifiziert, dass es noch nie dauerhaft (hier 11 Jahre) Überrenditen gegeben hat. Damit ist nicht bestätigt, dass es sie für immer geben wird. Dass es Überrenditen auch weiterhin geben wird, ist nur eine Hypothese.

Wahrscheinlichkeitstheoretiker hätten dagegen wie folgt argumentiert: Wenn wir unterstellen, dass die beiden Elementarereignisse

$$Rendite\ Renaissance >= S\&P500\ und\ Rendite\ Renaissance < S\&P500$$

gleichwahrscheinlich sind, dann ist die Wahrscheinlichkeit, dass ein Investor 5-mal nacheinander in zwei Jahreszeiträumen den Index schlägt genau

[21] Daten aus: Rachel E S Ziemba,William T Ziemba(2013) Investing in the Modern Age, S.58.

$$f_{Binomial}(x = 5, n = 5, p = 0.5) = 0.03125$$

Wenn wir eine Millionen Investoren haben ist die Wahrscheinlichkeit, dass mindestens einer zufällig in fünf zwei-Jahreszeiträumen immer den Index schlägt

$$1 - f_{Binomial}(x = 0, n = 1000000, p = 0.03125) \cong 1$$

Man sagt also, dass es fast sicher ist, dass ein solches Ergebnis zufällig zu beobachten sein wird. Man würde eine solche Beobachtung nicht als Evidenz gegen die Effizienzmarkthypothese anerkennen. Die Beobachtung ist nicht moralisch unmöglich.

Da der Fonds seit 2005 keine Publikumsgelder mehr verwaltet, sondern nur noch die Gelder der Eigentümer, gibt es ab dem Jahr 2005 nur noch anekdotische Evidenz über seine Performance. Es wurde aber bekannt, dass der Fonds seit 2005 kein Quartal mit Verlust abgeschlossen hat und die Rendite im Jahr 2014 bei ca. 20% gelegen haben soll. Aber auch das würde sicher nicht als Evidenz gegen die Gültigkeit der Effizienzmarkthypothese ausreichen, die eher als „Glaubensdogma“ an vielen Betriebswirtschaftlichen Fakultäten Millionen von Studenten „gelehrt“ wird.

Man wird nie wissen ob die Strategien von Renaissance Medaillon im Unendlichen auch besser abgeschnitten haben als der S&P500 - aber bisher war es immer so. Bisher sind die Assetmärkte eindeutig in diesem Sinne nicht effizient. Aber mit Wahrscheinlichkeitsüberlegungen wird eine Hypothese eben nie so exakt formuliert, dass man sie eindeutig falsifizieren kann.

Ad 2.:
Fragen wir uns, ob es ein Messverfahren für Inflation und ein Messverfahren für die Geldmenge gibt, sodass für jedes Land, immer wenn die Geldmenge in einem Jahr höher war als die bisher durchschnittliche Geldmenge auch die Inflationsrate des Folgejahres höher war als die bisher durchschnittliche, so würden wir zum Ergebnis kommen, dass es solche Messverfahren nicht gibt.

Ad 3.:
Wenn wir die Frage so formulieren, ob Menschen in jeder Situation, in der sie zwischen einem sicheren Geldbetrag und einer Lotterie mit einer erwarteten Auszahlung kleiner als dem sicheren Geldbetrag den sicheren Geldbetrage wählen, so können wir sofort erkennen, dass die Tatsache, dass weltweit Milliarden Menschen wöchentlich Lotto mit einer erwarteten Auszahlung, die kleiner als der Spieleinsatz ist, spielen, diese Hypothese empirisch falsifiziert.

Ad 4.: Aus der Tatsache das Menschen in jedem Jahr viele Milliarden Euro für wohltätige Zwecke spenden, können wir sofort erkennen, dass nicht jeder in jeder Situation die Handlungsalternative wählt, die ihm eine höhere erwartete Auszahlung verspricht. (Auch wenn viele glauben, Spenden durch die - immer nur einen Bruchteil des Spendenbetrags ausmachende - Steuerersparnis erklären zu können.)

Ad. 5.: Da beide Konzepte nicht durch Beobachtungen überprüft werden können, ist diese Frage unserer Einschätzung nach einfach nicht sinnvoll.

Was man dann aus Theoremen, die aus der Deduktion aus obigen nicht mit den derzeitigen Methoden falsifizierbaren Annahmen folgen, lernen soll und warum so viele Wissenschaftler hochkomplizierte Modelle aus Annahmen dieser Art konstruieren ist uns nicht wirklich klar.

Problematisch wird es, wenn auf Basis dieser „Glaubenssätze“ über politische Maßnahmen entschieden wird. Ob eine Geldmengenerhöhung tatsächlich Inflation verursacht, kann durch einen findigen Wirtschaftswissenschaftler statistisch problemlos bestätigt oder widerlegt werden. Die aktuellen Staatsanleihekäufe der EZB auf Basis einer drohenden Inflation zu kritisieren, ist daher völlig irrelevant, da auch umgekehrt argumentiert werden könnte. Tatsächlich haben die Käufe der EZB bisher keine Inflation im Euroraum ausgelöst. Als alternatives Argument für die Übertragung von Geldmengenerhöhung auf Inflation dann kurzerhand eine Asset-Inflation zu unterstellen, entspricht eher einer Ausrede.

Wir wollen damit nicht behaupten, dass eine Ausweitung der Geldmenge keine negativen Konsequenzen mit sich bringen wird - wir geben nur zu, dass wir es nicht wissen. Ob eine (genau bestimmte) längerfristige expansive Geldpolitik immer irgendwann in den nächsten T Jahren zu einer (genau bestimmten) größeren Rezessionen oder anderen Verwerfungen geführt hat oder ob das durchschnittliche Wachstum der nachfolgenden T Jahren in der Gruppe der 20 Länder mit der höchsten Geldmengenerhöhung immer kleiner war als das in der Gruppe mit dem geringsten Geldmengenwachstum - haben wir nicht untersucht. Vielleicht gibt es ja sogar eine nach einem Wachstumskriterium T-Dominate Geldmengenregel.

Bei einem Überblick über die wirtschaftswissenschaftliche Literatur erkennt man rasch, dass sich eine beliebig lange Liste unterschiedlicher Hypothesen, Meinungen und Aussagen zu verschiedenen Themenbereichen finden lässt, die sich ergänzen, gegenseitig überdecken oder widersprechen, die aber niemals wieder verschwinden.

Insgesamt ist nicht absehbar, inwieweit mit der derzeit verwendeten Methodik eine Ordnung in diese überbordende Fülle an Hypothesen gebracht, geschweige denn eine Konsolidierung erzielt werden kann, da es bisher gar nicht möglich ist, Aussagen eindeutig (aber vorläufig) zu akzeptieren oder endgültig zu verwerfen.

Die Wirtschaftswissenschaften generieren im Moment lediglich eine unfassbar große Anzahl an Hypothesen, die gar nicht falsifizierbar sind. Es kann dementsprechend keine Konsolidierung von Wissen und damit keine Weiterentwicklung des Faches selbst stattfinden.

Es können also die bisherigen wissenschaftlichen Arbeiten auf dem Gebiet nicht als „Riesen“ genutzt werden, auf deren Schultern dann die „Zwerge“ neue Erkenntnisse erblicken können. Stattdessen bildet jede Generation einen eigenen Berg an Hypothesen, ohne die Falschen aussortieren zu können.

Unserer Meinung nach wäre es durchaus sinnvoll, wenn den Wirtschaftswissenschaften eine Grundlagenkrise widerfahren würde, dank derer das Fach als solches auf neue stabile Füße gehoben werden kann, die echten Fortschritt und Weiterentwicklung ermöglichen.

Da wir nicht nur die derzeitige Methodik kritisieren sondern auch eine (unserer Meinung nach) funktionsfähige Alternative anbieten, hoffen wir eine solche Grundlagendiskussion auslösen zu können.

6. Literaturverzeichnis

Baily D H, Borwein J M, Lopez de Prado M, Zhu O J (2014) Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. Notices of the AMS, Volume 61, No.5, 458-471.

Bernoulli J (1713), Ars Conjectandi. Basel. Übersetzung ins Deutsche von R. Haussner, Oswalds Klassiker der exakten Wissenschaften, Band 107, Leipzig 1899

Breiman, L (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statist. Sci. 16 (2001), no. 3, 199-231. doi:10.1214/ss/1009213726.

Caruana R, Niculescu-Mizil A (2006) An Empirical Comparison of Supervised Learning Algorithms, ICML '06 Proceedings of the 23rd international conference on Machine learning, pp 161-168, ACM New York, NY, USA. doi:10.1145/1143844.1143865

Fanaee-T H, Gama J (2013), "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence : pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

Hájek, A (2012) "Interpretations of Probability", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.). http://plato.stanford.edu/archives/win2012/entries/probability-interpret/

Jeffrey, R (2002) Subjective Probability – The Real Thing, https://www.princeton.edu/~bayesway/Book*.pdf

Kant I (1986), Kritik der reinen Vernunft, Reclam.

Kuck A, Harries J P, Kuck E (2015) „Der Weg zu wahren Gesetzen und rationalem Handeln; Durch Emergenz statt der Illusion metaphysischer Wahrheit zu empirischer Erkenntnis. http://udpl.de/files/Kuck_Harries_WahreGesetze.pdf

Li M, Vitányi P (1997): An Introduction to Kolmogorov Complexity and Its Applications, Second Edition, Springer New York

Nuzzo, R (2014) (Scientific method:) P-values, the „gold standard“ of statistical validity, are not as reliable as many scientists assume, Nature Volume 506, 12 February 2014, http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Shafer G, Vovk V (2003): The origins and legacy of Kolmogorov's Grundbegriffe, The Game-Theoretic Probability and Finance Project ,Working Paper #4, http://www.probabilityand_nance.com

Ziemba R E S, Ziemba W T (2013) Investing in the Modern Age, World Scientific Publishing Company