Entscheidungen unter Unsicherheit

Grundlagen, neuere Entwicklungen und Anwendungen —
Skript zur Vorlesungseinheit am Beispiel der Personalplanung im Krankenhaus

Dr. Celso Gustavo Stall Sikora

Oktober 2026

Zweck dieses Skripts. Dieses Skript enthält den vollständigen Stoff der Vorlesungseinheit in ausformulierter Form. Wer die Vorlesung besucht hat, findet hier alle Herleitungen und Zahlen zum Nachlesen; wer nicht dabei sein konnte, kann sich den Stoff vollständig selbst erarbeiten. Das begleitende Jupyter-Notebook enthält alle Berechnungen zum Nachrechnen und Verändern; die Übungsaufgaben stehen am Ende des Notebooks. Beides unter:
https://sikora.one/events/interview-talk-bielefeld-2026/

1 Das Problem: eine Samstagnacht auf Station 7

Anfang Oktober erstellt die Stationsleitung einer Krankenhausstation den Dienstplan für den November — darunter die Nachtdienste für vier Samstagnächte. Zum Zeitpunkt der Planung ist der *Personalbedarf* dieser Nächte unbekannt: Er hängt von der Belegung der Station, von Akutfällen und von kurzfristigen Krankmeldungen ab. Über gesetzliche Personalschlüssel (in der Pflege: PPR 2.0) übersetzt sich eine unsichere Belegung mechanisch in einen unsicheren Personalbedarf.

Was die Planung aber kennt, ist die *Verteilung* des Bedarfs: Die Situation wiederholt sich Monat für Monat, und aus der Vergangenheit lässt sich schätzen, wie wahrscheinlich welcher Bedarf ist. Für unsere Beispielstation nehmen wir an:

Bedarf D (Pflegekräfte)	2	3	4	5	6	7
Wahrscheinlichkeit	5 %	20 %	30 %	25 %	15 %	5 %

Der Erwartungswert beträgt $\mathbb{E}[D] = 4{,}4$ Pflegekräfte. Die zentrale Frage der Einheit lautet: **Wie viele Kräfte plant man fest ein — und in welcher Form?**

Was steht auf dem Spiel? Beide Fehlplanungen kosten — aber auf sehr unterschiedliche Weise. Plant man *zu wenige* Kräfte ein, muss die Station in der Nacht reagieren: zuerst wird — falls eingeplant — der Bereitschaftsdienst aktiviert; reicht auch das nicht, muss jemand kurzfristig *einspringen* — aus dem Frei geholt, mit Prämie und Mehrarbeitszuschlägen, im Zweifel über teure kurzfristige Leiharbeit. Häufige Einspringer belasten zudem das Team: ungeplante Dienste zerstören die Erholungsphasen, erhöhen Unzufriedenheit und Krankenstand — und verschärfen damit genau das Problem, das sie lösen sollen. Plant man dagegen *zu viele* Kräfte ein, bezahlt die Station Nacht für Nacht Schichten, die in den meisten Nächten nicht gebraucht werden — gebundenes Personal und gebundenes Budget in einem Umfeld, in dem beides knapp ist. Zu wenig und zu viel sind also beide teuer, aber *nicht gleich* teuer — diese Asymmetrie wird sich in Abschnitt 4 als der Kern des Problems erweisen.



Eine ruhige Nacht: eingeplante Kapazität, die (diesmal) nicht gebraucht wurde.



Eine volle Nacht: jetzt hilft nur noch Aktivieren — oder teures Einspringen.

Symbolbilder, KI-generiert (ChatGPT).

2 Die Handlungsoptionen und ihre Kosten

Das deutsche Tarifrecht stellt der Station eine ganze *Kostenleiter* von Instrumenten bereit (TV-Ärzte/VKA, sinngemäß auch TVöD-K für die Pflege):

Instrument	**festgelegt**	**Kosten (stilisiert)**	**Grundlage**
Vollarbeit (Nachtdienst)	im Voraus	1,0 pro Schicht	Tarifentgelt
Bereitschaftsdienst (im Haus)	im Voraus	0,75 + 0,3 je Einsatz	§ 12, Stufe II
Rufbereitschaft (zu Hause)	im Voraus	$\approx$ 0,3 Pauschale + Einsatz	§ 11 Abs. 3
Einspringen / kurzfristiger Ersatz	*in der Nacht*	3,0	Prämie, Zuschläge, Leiharbeit

Alle Kosten sind in *Schichtäquivalenten* ausgedrückt (1,0 = Kosten einer regulären Nachtschicht inklusive Zuschlägen). Die Werte sind stilisiert, aber tariflich verankert: Ein Bereitschaftsdienst wird nach § 12 TV-Ärzte/VKA je nach erfahrungsgemäßer Arbeitsleistung mit 60 % (Stufe I, bis 25 % Arbeitsleistung), 75 % (Stufe II, bis 40 %) oder 90 % (Stufe III, bis 49 %) als Arbeitszeit bewertet — wir rechnen mit Stufe II. Die +0,3 je Einsatz bilden ab, dass regelmäßig eingesetzte Bereitschaften in höhere Stufen rutschen (*Stufendrift*): Die Zuordnung zur Stufe erfolgt als Nebenabrede zum Arbeitsvertrag auf Basis der erfahrungsgemäßen Inanspruchnahme. Ein aktivierter Bereitschaftsdienst kostet damit insgesamt $1{,}05 > 1{,}0$ — *teurer* als eine von vornherein eingeplante Vollkraft. Das Einspringen aus dem Frei schließlich ist immer möglich, aber am teuersten: dokumentierte Einspringprämien (z. B. LMU Klinikum: 75–125 € pro Dienst), Mehrarbeitszuschläge, im Zweifel kurzfristige Leiharbeit zum Zwei- bis Dreifachen des Tarifentgelts.

Zwei rechtliche Leitplanken begrenzen die Optionen: Bereitschaftsdienst darf **keine Regelarbeit ersetzen** (§ 10 Abs. 1: die Zeit ohne Arbeitsleistung muss erfahrungsgemäß überwiegen) — wir modellieren das als Obergrenze von zwei Bereitschaftsdiensten pro Nacht. Und Rufbereitschaft ist nur zulässig, wenn Einsätze die *Ausnahme* bleiben (§ 10 Abs. 8); für eine Station mit regelmäßigem Nachtbedarf ist sie daher nicht das passende Instrument, und wir lassen sie im Modell weg (eine lohnende Übungsaufgabe: siehe Notebook).

Eine ehrliche Einordnung: Kosten *und* Modell sind bewusste Vereinfachungen. Die Kostenwerte (1,0 / 0,75 / +0,3 / 3,0) sind stilisierte Größenordnungen, tariflich verankert, aber kein Abbild der Lohnabrechnung eines konkreten Hauses — dort hängen sie von Entgeltgruppe, Stufen, Zuschlägen und lokalen Regelungen ab. Ebenso betrachtet das Modell *eine* Nacht isoliert, mit homogenen, beliebig teilbaren Kräften, ohne Qualifikationen, Teilzeitquoten oder Übergänge zwischen

Diensten. Das ist Absicht: Die Vereinfachung macht die *Struktur* der Entscheidung sichtbar — jetzt binden oder später teuer reagieren —, und genau diese Struktur bleibt erhalten, wenn man die Zahlen durch die des eigenen Hauses ersetzt. Im begleitenden Notebook sind alle Parameter veränderbar: Probieren Sie es aus.

3 Die Struktur: Entscheiden — Beobachten — Reagieren

Die Entscheidungssituation hat eine Struktur, die weit über Krankenhäuser hinausreicht:

1. **Entscheiden** (Anfang Oktober): Der Dienstplan wird festgelegt — x_V Vollkräfte und $x_B \le 2$ Bereitschaftsdienste pro Samstagnacht. Diese Entscheidung fällt, *ohne* den Bedarf zu kennen.
2. **Beobachten** (in der Nacht): Der Bedarf D realisiert sich — der Zufall entscheidet, nicht wir.
3. **Reagieren**: Reicht x_V nicht aus, wird die Bereitschaft aktiviert; reicht auch das nicht, muss jemand einspringen. Die Reaktion ist die *beste Antwort auf das, was eingetreten ist.*

Formal suchen wir den Plan (x_V, x_B) mit minimalen erwarteten Gesamtkosten:

$$\min_{x_V,\, x_B \le 2} \underbrace{1{,}0\, x_V + 0{,}75\, x_B}_{\text{jetzt, sicher}} + \underbrace{\mathbb{E}_D\big[\,0{,}3 \cdot B(D) + 3{,}0 \cdot S(D)\,\big]}_{\text{später, im Erwartungswert}}$$

mit $B(D) = \min\big(\max(D - x_V,\, 0),\, x_B\big)$ (aktivierte Bereitschaften) und $S(D) = \max(D - x_V - x_B,\, 0)$ (Einspringer). In der Sprache der Optimierung ist das ein *zweistufiges stochastisches Programm*: Die erste Stufe sind die Hier-und-Jetzt-Entscheidungen, die zweite Stufe ist der *Recourse* — die Reaktion nach Auflösung der Unsicherheit. Weil die letzte Stufe (Einspringen) immer verfügbar ist, ist jeder Plan zulässig, nur unterschiedlich teuer (*vollständiger Recourse*). Der Erwartungswert ist hier das richtige Kriterium, weil sich die Situation monatlich wiederholt: Über viele Nächte hinweg zählen die durchschnittlichen Kosten.

Das vollständige Optimierungsmodell

Da der Bedarf nur endlich viele Werte annehmen kann, lässt sich das zweistufige Problem als gewöhnliches (gemischt-)ganzzahliges Programm ausschreiben — das sogenannte *deterministische Äquivalent* oder die *extensive Form*. Jede mögliche Bedarfsrealisierung $d \in \{2, \dots, 7\}$ wird ein *Szenario* mit Wahrscheinlichkeit p_d und eigenen Reaktionsvariablen b_d (aktivierte Bereitschaften) und s_d (Einspringer):

$$\begin{aligned}
\min_{x_V, x_B, b_d, s_d} \quad & 1{,}0\, x_V \;+\; 0{,}75\, x_B \;+\; \sum_{d=2}^{7} p_d \left(0{,}3\, b_d + 3{,}0\, s_d\right) && \\
\text{s. t.} \quad & x_V + b_d + s_d \;\ge\; d && \text{für alle } d \text{ (Bedarf decken)} \\
& b_d \;\le\; x_B && \text{für alle } d \text{ (nur Eingeplante aktivierbar)} \\
& x_B \;\le\; 2 && (\S\,10\text{: keine Regelarbeit}) \\
& x_V,\, x_B,\, b_d,\, s_d \;\in\; \mathbb{Z}_{\ge 0}. &&
\end{aligned}$$

Die erste Zeile der Nebenbedingungen koppelt Erststufe und Szenario: In jeder denkbaren Nacht muss der Bedarf aus eingeplanter Arbeit, aktivierter Bereitschaft und Einspringern gedeckt werden. Genau dieses Modell löst das begleitende Notebook.

Warum die zweite Stufe hier leicht ist — und wann sie es nicht ist. Bei festem Plan (x_V, x_B) und beobachtetem Bedarf d ist die beste Reaktion offensichtlich: Fehlende Kapazität wird zuerst mit Bereitschaften gedeckt (0,3 pro Einsatz), erst dann mit Einspringern (3,0) — billig vor

teuer. Deshalb konnten wir die Reaktion oben in geschlossener Form schreiben ($B(D)$, $S(D)$) und das Gesamtproblem durch einfaches Durchrechnen aller Pläne lösen. Das ist *nicht* selbstverständlich: In vielen realen Problemen ist die zweite Stufe selbst ein Optimierungsproblem — etwa wenn die Reaktion auf einen Maschinenausfall eine Umplanung der ganzen Produktionsreihenfolge ist oder die Reaktion auf eine Auftragsspitze neue Fahrzeugtouren verlangt. Dann muss für *jedes Szenario* ein eigenes Planungsproblem gelöst werden, und man greift zu Dekompositionsverfahren (z. B. Benders- bzw. L-shaped-Verfahren), die zwischen der Erststufenentscheidung und den Szenario-Teilproblemen iterieren, bis beide zusammenpassen.

4 Warum "für den Durchschnitt planen" scheitert

Der naheliegende Plan lautet: $\mathbb{E}[D] = 4{,}4$, also planen wir 4 Kräfte. Diese Logik hat einen systematischen Fehler: Sie ignoriert, dass die *Fehlerkosten asymmetrisch* sind. Eine Kraft zu viel kostet eine bezahlte ruhige Nacht (1,0); eine Kraft zu wenig kostet einen Einspringer (3,0) — ein Verhältnis von $3:1$. Bei asymmetrischen Kosten ist der Erwartungswert des Bedarfs *nicht* die optimale Kapazität: Man plant so weit über dem Mittel, bis zusätzliche Sicherheit teurer wird als das verbleibende Risiko. Dieses Prinzip — im Spezialfall ohne Bereitschaft als *kritisches Verhältnis* des Newsvendor-Modells bekannt ($c_E/(c_E + c_V)$) — ist die Grundlage jeder Kapazitätsplanung unter Unsicherheit. Die verbreitete Fehlplanung mit dem Durchschnitt trägt einen eigenen Namen: *Flaw of Averages*.

5 Die Lösung: ein Portfolio schlägt beide Extreme

Zwei Pläne von Hand durchgerechnet

Wie entsteht eine Zahl wie "erwartete Kosten 6,10"? Rechnen wir zwei Pläne vollständig durch.

Plan $(x_V, x_B) = (4, 0)$ — der Durchschnittsplan. Die sicheren Grundkosten betragen $4 \times 1{,}0 = 4{,}00$. Bereitschaft gibt es nicht; fehlt Kapazität ($D > 4$), muss eingesprungen werden, $S(D) = \max(D - 4,\, 0)$:

Bedarf D	$p(D)$	Einspringer $S(D)$	Kosten $3{,}0 \cdot S$	$p\cdot$ Kosten
2–4	0,55	0	0	0
5	0,25	1	3,0	0,75
6	0,15	2	6,0	0,90
7	0,05	3	9,0	0,45
			erwartete Reaktionskosten	2,10

Gesamt: $4{,}00 + 2{,}10 = \mathbf{6{,}10}$.

Plan $(x_V, x_B) = (3, 2)$ — das Portfolio. Die sicheren Grundkosten betragen $3 \times 1{,}0 + 2 \times 0{,}75 = 4{,}50$. Fehlt Kapazität, werden erst bis zu zwei Bereitschaften aktiviert (+0,3 je Einsatz), dann wird eingesprungen:

Bedarf D	$p(D)$	aktiviert $B(D)$	Einspringer $S(D)$	Kosten $0{,}3B + 3{,}0S$	$p\cdot$ Kosten
2–3	0,25	0	0	0	0
4	0,30	1	0	0,3	0,09
5	0,25	2	0	0,6	0,15
6	0,15	2	1	3,6	0,54
7	0,05	2	2	6,6	0,33
				erwartete Reaktionskosten	1,11

Gesamt: 4,50 + 1,11 = **5,61**. Das Portfolio hat höhere *sichere* Kosten als der Durchschnittsplan (4,50 > 4,00) — und gewinnt trotzdem, weil es die teuren Nächte drastisch entschärft. Genau dieselbe Rechnung führt das Notebook für alle Pläne aus; ändern Sie dort Kosten oder Verteilung und beobachten Sie, wie die Tabelle unten kippt.

Da $x_V \leq 7$ und $x_B \leq 2$, lässt sich das Problem vollständig durchrechnen (alle Werte im Notebook). Die erwarteten Gesamtkosten pro Samstagnacht:

	$x_B = 0$	$x_B = 1$	$x_B = 2$
$x_V = 2$	9,20	7,38	6,11
$x_V = 3$	7,35	6,08	**5,61**
$x_V = 4$	6,10	5,63	5,85
$x_V = 5$	5,75	5,96	6,58
$x_V = 6$	6,15	6,76	7,51

Drei Beobachtungen:

1. Beide Extreme sind (fast) gleich teuer. Der "Durchschnittsplan" (4 Vollkräfte: 6,10) und der "Auf-Nummer-sicher-Plan" (6 Vollkräfte: 6,15) kosten praktisch dasselbe — der eine wegen häufiger teurer Einspringer (in 45 % der Nächte!), der andere wegen dauerhaft bezahlter Überkapazität.

2. Das Optimum ist ein Portfolio. Der beste Plan kombiniert einen festen Kern mit einem flexiblen Puffer: **3 Vollkräfte + 2 Bereitschaftsdienste** (5,61) — rund 8 % bzw. eine halbe Schicht pro Nacht günstiger als beide Extreme. Einspringen wird nur noch in 20 % der Nächte nötig. Über Stationen und Jahre summiert sich das zu sechsstelligen Beträgen.

3. Der Wert der stochastischen Lösung (VSS). Der VSS beantwortet die Frage: *Was bringt es überhaupt, die Unsicherheit ins Modell zu holen?* Dazu vergleicht man zwei sauber definierte Größen. Erstens das stochastische Optimum RP (*recourse problem*): die beste erwartete Gesamtkostenzahl, hier RP = 5,61. Zweitens löst man das *Erwartungswertproblem*: Man ersetzt den zufälligen Bedarf durch seinen Mittelwert $\mathbb{E}[D] = 4{,}4$ und plant, als wäre dieser Bedarf sicher — die Lösung sind 5 Vollkräfte und, wohlgemerkt, *keine* Bereitschaft (gegen sicheren Bedarf lohnt sie nie). Setzt man diese Erwartungswert-Lösung dann der echten, unsicheren Welt aus, kostet sie im Erwartungswert EEV = 5,75 (*expected result of using the expected-value solution*). Die Differenz

$$\text{VSS} \;=\; \text{EEV} - \text{RP} \;=\; 5{,}75 - 5{,}61 \;=\; 0{,}14$$

ist der messbare Gewinn der stochastischen Planung — hier gut 2 % der Nachtkosten, Monat für Monat, Station für Station. Wichtiger als die Zahl ist die *Struktur*: Bei bekanntem Bedarf enthielte kein optimaler Plan Bereitschaft — aktiviert kostet sie 1,05 > 1,0, und rechtlich wäre sichere Arbeit ohnehin Vollarbeit (die Bereitschaftsstufen enden bei 49 % Arbeitsleistung).

> **Der Bereitschaftsdienst existiert nur, weil es Unsicherheit gibt — und die Optimierung sagt uns, wie viel davon wir brauchen.**

6 Neuere Entwicklungen

Daten statt Verteilungen. Klassisch wird die Bedarfsverteilung aus historischen Häufigkeiten geschätzt — wie in unserem Beispiel. Moderne Prognosemodelle gehen weiter: Sie nutzen *Merkmale* (Wochentag, Saison, Belegung, Grippewelle, Ferien) und liefern für jede einzelne Nacht eine eigene bedingte Verteilung. Dabei gilt eine zentrale Warnung: **Die beste Prognose ist nicht die beste Entscheidung.** Ein Modell, das auf kleinsten Prognosefehler trainiert ist, liefert den bedingten *Mittelwert* — und wer den in die Planung steckt, begeht den Flaw of Averages mit teurerem Vorbau. Zwei

Nächte mit identischer Punktprognose "4", aber unterschiedlicher Streuung, verlangen unterschiedliche Pläne. Die Konsequenz heißt *kontextuelle* bzw. *entscheidungsfokussierte Optimierung*: entweder ganze Verteilungen prognostizieren und dagegen optimieren — oder das Lernverfahren direkt auf die *Entscheidungskosten* trainieren statt auf den Prognosefehler (Stichworte: Smart Predict-then-Optimize, prescriptive analytics).

Maschinelles Lernen als Löser-Beschleuniger. Der echte Dienstplan umfasst einen Monat, mehrere Stationen, Qualifikationen, Ruhezeiten und Fairness — Optimierungsmodelle mit Millionen Variablen. Ein aktueller Forschungsansatz lässt maschinelles Lernen vorhersagen, *welche Teile des Modells in guten Lösungen überhaupt vorkommen*, reduziert das Modell entsprechend und löst dann exakt (*learning to reduce*). Für die Dienstplanung wurde dieser Ansatz an der Universität Bielefeld entwickelt: Porrmann & Römer (2021), *Learning to Reduce State-Expanded Networks for Multi-Activity Shift Scheduling*, CPAIOR 2021, S. 383–391 — die reduzierten Modelle bleiben dort in fast allen Fällen optimal lösbar.

7 Ausblick: dieselbe Struktur, überall

Was hier an einer Nachtschicht entwickelt wurde, trägt weit: Der echte Personaleinsatzplan bringt Ruhezeiten (seit dem EuGH-Urteil *Jaeger* 2003 zählt Bereitschaft arbeitszeitrechtlich voll als Arbeitszeit), Präferenzen und Qualifikationen hinzu — das Forschungsfeld der Personaleinsatzplanung (Überblick: Erhard et al., *State of the art in physician scheduling*, EJOR 265(1), 2018). Und die Struktur *Entscheiden–Beobachten–Reagieren* findet sich überall dort, wo Kapazität vor der Nachfrage festgelegt werden muss: Produktionskapazitäten unter Nachfrageunsicherheit, Lager- und Ersatzteilbestände, Energiebeschaffung, Infrastrukturplanung. Die nächste Vorlesung formalisiert das kritische Verhältnis — und zeigt, was passiert, wenn die Verteilung aus Daten gelernt wird.

Zum Weiterarbeiten — ausdrücklich erwünscht: Das begleitende Notebook enthält das vollständige Modell und alle Rechnungen dieses Skripts. Alles darin ist veränderbar, und genau dazu laden wir ein: Senken Sie den Einspringpreis c_E und beobachten Sie, wann der Durchschnittsplan doch gewinnt; setzen Sie die Aktivierungskosten c_{akt} auf null und erklären Sie, was mit der Bereitschaft passiert; ersetzen Sie die Verteilung durch eine symmetrische oder eine mit mehr Streuung; lockern Sie die Grenze $x_B \leq 2$. Vier Übungsaufgaben mit genau diesen Experimenten stehen am Ende des Notebooks: `https://sikora.one/events/interview-talk-bielefeld-2026/`.

Kostenwerte stilisiert nach TV-Ärzte/VKA (§§ 10–12); Einspringprämien nach öffentlich dokumentierten Regelungen (z. B. LMU Klinikum München, Klinikum Ernst von Bergmann Potsdam). Literatur: Birge & Louveaux, *Introduction to Stochastic Programming*, Springer; Ban & Rudin (2019), *The Big Data Newsvendor*, Management Science 65(9); Bertsimas & Kallus (2020), *From Predictive to Prescriptive Analytics*, Management Science 66(3).